

---

# JMIR Medical Informatics

---

Impact Factor (2023): 3.1

Volume 9 (2021), Issue 7 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

---

## Contents

### Reviews

- Effect of Interventions With a Clinical Decision Support System for Hospitalized Older Patients: Systematic Review Mapping Implementation and Design Factors ([e28023](#))  
Birgit Damoiseaux-Volman, Nathalie van der Velde, Sil Ruige, Johannes Romijn, Ameen Abu-Hanna, Stephanie Medlock. . . . . 4
- The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities ([e21929](#))  
Muhammad Ayaz, Muhammad Pasha, Mohammed Alzahrani, Rahmat Budiarto, Deris Stiawan. . . . . 368

### Viewpoint

- Contact Tracing Apps: Lessons Learned on Privacy, Autonomy, and the Need for Detailed and Thoughtful Implementation ([e27449](#))  
Katie Hogan, Briana Macedo, Venkata Macha, Arko Barman, Xiaoqian Jiang. . . . . 15

### Original Papers

- Evaluation of Three Feasibility Tools for Identifying Patient Data and Biospecimen Availability: Comparative Usability Study ([e25531](#))  
Christina Schüttler, Hans-Ulrich Prokosch, Martin Sedlmayr, Brita Sedlmayr. . . . . 35
- Measuring the Interactions Between Health Demand, Informatics Supply, and Technological Applications in Digital Medical Innovation for China: Content Mapping and Analysis ([e26393](#))  
Jian Du, Ting Chen, Luxia Zhang. . . . . 53
- Assessing the Performance of Clinical Natural Language Processing Systems: Development of an Evaluation Methodology ([e20492](#))  
Lea Canales, Sebastian Menke, Stephanie Marchesseau, Ariel D'Agostino, Carlos del Rio-Bermudez, Miren Taberna, Jorge Tello. . . . . 71
- Social Media Insights During the COVID-19 Pandemic: Infodemiology Study Using Big Data ([e27116](#))  
Huyen Tran, Shih-Hao Lu, Ha Tran, Bien Nguyen. . . . . 84
- A Tool for Evaluating Medication Alerting Systems: Development and Initial Assessment ([e24022](#))  
Wu Zheng, Bethany Van Dort, Romaric Marcilly, Richard Day, Rosemary Burke, Sepehr Shakib, Young Ku, Hannah Reid-Anderson, Melissa Baysari. . . . . 103

<b>Predicting Unscheduled Emergency Department Return Visits Among Older Adults: Population-Based Retrospective Study (e22491)</b>	
Rai-Fu Chen, Kuei-Chen Cheng, Yu-Yin Lin, I-Chiu Chang, Cheng-Han Tsai. . . . .	113
<b>The Association Between Using a Mobile Version of an Electronic Health Record and the Well-Being of Nurses: Cross-sectional Survey Study (e28729)</b>	
Tarja Heponiemi, Anu-Marja Kaihlanen, Kia Gluschkoff, Kaija Saranto, Sari Nissinen, Elina Laukka, Tuulikki Vehko. . . . .	123
<b>A Biomedical Knowledge Graph System to Propose Mechanistic Hypotheses for Real-World Environmental Health Observations: Cohort Study and Informatics Application (e26714)</b>	
Karamarie Fecho, Chris Bizon, Frederick Miller, Shepherd Schurman, Charles Schmitt, William Xue, Kenneth Morton, Patrick Wang, Alexander Tropsha. . . . .	133
<b>Frequency of Participation in External Quality Assessment Programs Focused on Rare Diseases: Belgian Guidelines for Human Genetics Centers (e27980)</b>	
Joséphine Lantoine, Anne Brysse, Vinciane Dideberg, Kathleen Claes, Sofie Symoens, Wim Coucke, Valérie Benoit, Sonia Rombout, Martine De Rycke, Sara Seneca, Lut Van Laer, Wim Wuyts, Anniek Corveleyn, Kris Van Den Bogaert, Catherine Rydlewski, Françoise Wilkin, Marie Ravoet, Elodie Fastré, Arnaud Capron, Nathalie Vandevelde. . . . .	144
<b>Preferences of the Public for Sharing Health Data: Discrete Choice Experiment (e29614)</b>	
Jennifer Viberg Johansson, Heidi Bentzen, Nisha Shah, Eik Haraldsdóttir, Guðbjörg Jónsdóttir, Jane Kaye, Deborah Mascalzoni, Jorien Veldwijk. . . . .	154
<b>Predicting Biologic Therapy Outcome of Patients With Spondyloarthritis: Joint Models for Longitudinal and Survival Analysis (e26823)</b>	
Carolina Barata, Ana Rodrigues, Helena Canhão, Susana Vinga, Alexandra Carvalho. . . . .	169
<b>Relation Classification for Bleeding Events From Electronic Health Records Using Deep Learning Systems: An Empirical Study (e27527)</b>	
Avijit Mitra, Bhanu Rawat, David McManus, Hong Yu. . . . .	186
<b>Machine Learning Methods for the Diagnosis of Chronic Obstructive Pulmonary Disease in Healthy Subjects: Retrospective Observational Cohort Study (e24796)</b>	
Shigeo Muro, Masato Ishida, Yoshiharu Horie, Wataru Takeuchi, Shunki Nakagawa, Hideyuki Ban, Tohru Nakagawa, Tetsuhisa Kitamura. . . . .	203
<b>Ambulatory Risk Models for the Long-Term Prevention of Sepsis: Retrospective Study (e29986)</b>	
Jewel Lee, Sevda Molani, Chen Fang, Kathleen Jade, D O'Mahony, Sergey Kornilov, Lindsay Mico, Jennifer Hadlock. . . . .	215
<b>Digital Medical Device Companion (MylUS) for New Users of Intrauterine Systems: App Development Study (e24633)</b>	
Toeresin Karakoyun, Hans-Peter Podhaisky, Ann-Kathrin Frenz, Gabriele Schuhmann-Giampieri, Thais Ushikusa, Daniel Schröder, Michal Zvolanek, Agnaldo Lopes Da Silva Filho. . . . .	229
<b>Predicting Antituberculosis Drug-Induced Liver Injury Using an Interpretable Machine Learning Method: Model Development and Validation Study (e29226)</b>	
Tao Zhong, Zian Zhuang, Xiaoli Dong, Ka Wong, Wing Wong, Jian Wang, Daihai He, Shengyuan Liu. . . . .	243
<b>Predicting Writing Styles of Web-Based Materials for Children's Health Education Using the Selection of Semantic Features: Machine Learning Approach (e30115)</b>	
Wenxiu Xie, Meng Ji, Yanmeng Liu, Tianyong Hao, Chi-Yin Chow. . . . .	255
<b>Patient Representation From Structured Electronic Medical Records Based on Embedding Technique: Development and Validation Study (e19905)</b>	
Yanqun Huang, Ni Wang, Zhiqiang Zhang, Honglei Liu, Xiaolu Fei, Lan Wei, Hui Chen. . . . .	271

<b>A Machine Learning–Based Algorithm for the Prediction of Intensive Care Unit Delirium (PRIDE): Retrospective Study (e23401)</b>	
Sujeong Hur, Ryoung-Eun Ko, Junsang Yoo, Juhyung Ha, Won Cha, Chi Chung. ....	283
<b>Candidemia Risk Prediction (CanDETEC) Model for Patients With Malignancy: Model Development and Validation in a Single-Center Retrospective Study (e24651)</b>	
Junsang Yoo, Si-Ho Kim, Sujeong Hur, Juhyung Ha, Kyungmin Huh, Won Cha. ....	296
<b>Prediction Model of Anastomotic Leakage Among Esophageal Cancer Patients After Receiving an Esophagectomy: Machine Learning Approach (e27110)</b>	
Ziran Zhao, Xi Cheng, Xiao Sun, Shanrui Ma, Hao Feng, Liang Zhao. ....	310
<b>Effects of Background Colors, Flashes, and Exposure Values on the Accuracy of a Smartphone-Based Pill Recognition System Using a Deep Convolutional Neural Network: Deep Learning and Experimental Approach (e26000)</b>	
KyeongMin Cha, Hyun-Ki Woo, Dohyun Park, Dong Chang, Mira Kang. ....	323
<b>Experiences With Internet Triaging of 9498 Outpatients Daily at the Largest Public Hospital in Taiwan During the COVID-19 Pandemic: Observational Study (e20994)</b>	
Ding-Heng Lu, Chia-An Hsu, Eunice Yuan, Jun-Jeng Fen, Chung-Yuan Lee, Jin-Lain Ming, Tzeng-Ji Chen, Wui-Chiang Lee, Shih-Ann Chen. ....	3
<b>Social Media Opinions on Working From Home in the United States During the COVID-19 Pandemic: Observational Study (e29195)</b>	
Ziyu Xiong, Pin Li, Hanjia Lyu, Jiebo Luo. ....	342
<b>Development, Acceptance, and Concerns Surrounding App-Based Services to Overcome the COVID-19 Outbreak in South Korea: Web-Based Survey Study (e29315)</b>	
Jihwan Park, Jinhyun Han, Yerin Kim, Mi Rho. ....	353
<b>Multifeature Fusion Attention Network for Suicide Risk Assessment Based on Social Media: Algorithm Development and Validation (e28227)</b>	
Jiacheng Li, Shaowu Zhang, Yijia Zhang, Hongfei Lin, Jian Wang. ....	389
<b>Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation (e28754)</b>	
Lu Ren, Hongfei Lin, Bo Xu, Shaowu Zhang, Liang Yang, Shichang Sun. ....	399
<b>Automatic Extraction of Lung Cancer Staging Information From Computed Tomography Reports: Deep Learning Approach (e27955)</b>	
Danqing Hu, Huanyao Zhang, Shaolei Li, Yuhong Wang, Nan Wu, Xudong Lu. ....	412
<b>Head and Tail Entity Fusion Model in Medical Knowledge Graph Construction: Case Study for Pituitary Adenoma (e28218)</b>	
An Fang, Pei Lou, Jiahui Hu, Wanqing Zhao, Ming Feng, Huiling Ren, Xianlai Chen. ....	429

Review

# Effect of Interventions With a Clinical Decision Support System for Hospitalized Older Patients: Systematic Review Mapping Implementation and Design Factors

Birgit A Damoiseaux-Volman<sup>1</sup>, MSc; Nathalie van der Velde<sup>2</sup>, PhD; Sil G Ruige<sup>1</sup>, BSc; Johannes A Romijn<sup>3</sup>, PhD; Ameen Abu-Hanna<sup>1</sup>, PhD; Stephanie Medlock<sup>1</sup>, PhD

<sup>1</sup>Department of Medical Informatics, Amsterdam Public Health Research Institute, Amsterdam UMC, University of Amsterdam, Amsterdam, Netherlands

<sup>2</sup>Section of Geriatric Medicine, Amsterdam Public Health Research Institute, Amsterdam UMC, University of Amsterdam, Amsterdam, Netherlands

<sup>3</sup>Department of Medicine, Amsterdam Public Health Research Institute, Amsterdam UMC, University of Amsterdam, Amsterdam, Netherlands

**Corresponding Author:**

Birgit A Damoiseaux-Volman, MSc  
Department of Medical Informatics  
Amsterdam Public Health Research Institute  
Amsterdam UMC, University of Amsterdam  
Meibergdreef 15  
Amsterdam, 1105AZ  
Netherlands  
Phone: 31 20 5666204  
Email: [b.a.damoiseaux@amsterdamumc.nl](mailto:b.a.damoiseaux@amsterdamumc.nl)

## Abstract

**Background:** Clinical decision support systems (CDSSs) form an implementation strategy that can facilitate and support health care professionals in the care of older hospitalized patients.

**Objective:** Our study aims to systematically review the effects of CDSS interventions in older hospitalized patients. As a secondary aim, we aim to summarize the implementation and design factors described in effective and ineffective interventions and identify gaps in the current literature.

**Methods:** We conducted a systematic review with a search strategy combining the categories *older patients*, *geriatric topic*, *hospital*, *CDSS*, and *intervention* in the databases MEDLINE, Embase, and SCOPUS. We included controlled studies, extracted data of all reported outcomes, and potentially beneficial design and implementation factors. We structured these factors using the Grol and Wensing Implementation of Change model, the GUIDES (Guideline Implementation with Decision Support) checklist, and the two-stream model. The risk of bias of the included studies was assessed using the Cochrane Collaboration's Effective Practice and Organisation of Care risk of bias approach.

**Results:** Our systematic review included 18 interventions, of which 13 (72%) were effective in improving care. Among these interventions, 8 (6 effective) focused on medication review, 8 (6 effective) on delirium, 7 (4 effective) on falls, 5 (4 effective) on functional decline, 4 (3 effective) on discharge or aftercare, and 2 (0 effective) on pressure ulcers. In 77% (10/13) effective interventions, the effect was based on process-related outcomes, in 15% (2/13) interventions on both process- and patient-related outcomes, and in 8% (1/13) interventions on patient-related outcomes. The following implementation and design factors were potentially associated with effectiveness: *a priori problem or performance analyses* (described in 9/13, 69% effective vs 0/5, 0% ineffective interventions), *multifaceted interventions* (8/13, 62% vs 1/5, 20%), and *consideration of the workflow* (9/13, 69% vs 1/5, 20%).

**Conclusions:** CDSS interventions can improve the hospital care of older patients, mostly on process-related outcomes. We identified 2 implementation factors and 1 design factor that were reported more frequently in articles on effective interventions. More studies with strong designs are needed to measure the effect of CDSS on relevant patient-related outcomes, investigate personalized (data-driven) interventions, and quantify the impact of implementation and design factors on CDSS effectiveness.

**Trial Registration:** PROSPERO (International Prospective Register of Systematic Reviews): CRD42019124470; [https://www.crd.york.ac.uk/prospero/display\\_record.php?RecordID=124470](https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=124470).



**KEYWORDS**

aged; clinical decision support systems; geriatrics; hospital; quality of care

## Introduction

### Background

In hospitals, the number and proportion of older patients have increased in the past years and will continue to grow in the following years [1,2]. Hospitalization has a significant impact on the lives of older patients. The incidence of preventable adverse events in a hospital setting is almost twice as high in older patients as in younger patients [3]. In addition, there is a high prevalence of geriatric syndromes and a high risk of functional decline and mortality in older hospitalized patients [4,5]. Geriatric syndromes are described as “common, serious conditions for older persons, holding substantial implications for functioning and quality of life” [6]. In a representative cohort investigating geriatric syndromes in older patients from 3 acute care hospitals, the prevalence of bladder incontinence was 37%, 5% for pressure ulcers, and 18% for delirium [4]. Furthermore, 6% of the patients suffered from one or more falls during the hospital stay [4]. Geriatric syndromes, involvement of multiple health care professionals, and difficulties in communicating with patients complicate hospital care.

Clinical decision support systems (CDSSs) can facilitate and support health professionals in the complex care of older hospitalized patients. CDSSs have the potential to transfer knowledge from guidelines to physicians, pharmacists, and nurses or experts to all hospital physicians, for example, from geriatricians to other specialties. Furthermore, CDSSs can support the implementation of advice in hospital practice by structuring information from different departments or performing calculations [7]. Our previous work indicated that there are several areas where a CDSS is perceived as having the potential to improve geriatric care in the hospital, including falls and delirium [8]. To date, systematic reviews of CDSS for the care of older patients have focused solely on medication and not on other aspects of care [9-11].

Systematic reviews of CDSS interventions, not specifically for older patients, have identified factors that could be associated with CDSS effectiveness, such as providing patient-specific advice [12,13]. Evidence for these factors is low, and further trials are needed to conclude which factors improve effectiveness [13]. A CDSS supporting health care professionals in geriatric care may differ and be more difficult to design and implement because of the complexity of care and the need for hospital-wide interventions. However, the implementation and design factors influencing the effect of CDSS interventions to improve geriatric care have not been studied in a systematic review.

### Objectives

Our study aims to systematically review the effect of CDSS interventions on common problems in the care of older hospitalized patients. The secondary aim is to summarize the implementation and design factors described in the effective or

ineffective interventions and identify gaps in the current literature.

## Methods

### Protocol

The protocol of our systematic review was registered and published on the website of the PROSPERO (International Prospective Register of Systematic Reviews) with the registration number CRD42019124470. [Multimedia Appendix 1](#) contains the completed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist [14].

### Search Strategy

A search strategy combining the categories *older patients*, *geriatric topic*, *hospital*, *CDSS*, and *intervention* was designed and adapted for the databases MEDLINE (via Ovid), Embase (via Ovid), and SCOPUS. The search strategy was based on keywords, medical subject headings, and text words. The search was conducted until April 15, 2020. The full search strategy is shown in [Multimedia Appendix 2](#). Duplicates in the search were detected and deleted in EndNote X9 (Clarivate Analytics), 2019 [15]. In addition, we screened the references of the included studies for missing articles.

### Study Selection

Using a checklist with prespecified eligibility criteria, 2 researchers (BADV and SGR) screened articles for inclusion. These criteria were piloted in the first 200 articles and subsequently adjusted, if necessary. Title and abstract screening was performed using Rayyan [16]. The eligibility criteria were (1) intervention with CDSS, (2) geriatric topic in the care of hospitalized patients aged 65 years or older, (3) evaluation in a controlled trial (including before-after and other quasi-experimental designs), and (4) peer-reviewed journal paper in English. We required that the eligibility criteria were met on the basis of the abstract.

For CDSS, we used the definition of Musen et al [17] of “any computer program designed to help health care professionals to make clinical decisions.” The geriatric topics were derived from our previous study [8], in which we determined which areas of geriatric care CDSS can potentially improve the care of hospitalized older patients and, in addition, the work of Inouye et al [6] describing 5 common geriatric syndromes. The topics included were pressure ulcers, incontinence, falls, functional decline, delirium, medication review, communication with the patient (at discharge), planning (in the hospital), and (communication and collaboration between health care professionals at) discharge and aftercare. For medication review, we used the definition of the Pharmaceutical Care Network Europe, “Medication review is a structured evaluation of a patient’s medicines with the aim of optimising medicines use and improving health outcomes.” This definition entails detecting drug-related problems and recommending

interventions [18]. The geriatric topics had to be part of the inclusion criteria, the aim, or the outcomes of the study.

## Data Extraction and Risk of Bias Assessment

### Overview

Two researchers (BADV and SGR) individually conducted data extraction and risk of bias assessment. We used a data extraction form for data extraction. The form was tested on 2 papers and adjusted as required. If an article referred to another article describing the development or implementation of the intervention, data from this additional article were also extracted. The risk of bias of the included studies was assessed using Cochrane Collaboration's Effective Practice and Organisation of Care (EPOC) risk of bias approach [19]. We extracted all reported outcomes from the included articles: process-related, patient-related, and cost outcomes. Patient-related outcomes could be either clinical or patient-derived outcomes [20]. We extracted data on outcomes measured in both the control and intervention groups. Each step of the inclusion process—data extraction, structuring and mapping of the implementation and design factors, and risk of bias assessment—was conducted independently by 2 researchers (BADV and SGR), and the results were compared. Disagreements were discussed until agreement was achieved and, if necessary, resolved by a third researcher (SM).

### Effectiveness of the Interventions

We used a definition of effectiveness that was previously used in the literature [12]. Interventions were considered effective when the prespecified primary outcome,  $\geq 50\%$  of the prespecified primary outcomes, or, if a primary outcome was not defined,  $\geq 50\%$  of the prespecified outcomes showed significant ( $P < .05$ ) improvement [12]. If an intervention was described in more than one article, the outcomes from all articles assessing the intervention were used to define the effectiveness.

### Implementation and Design Factors

We extracted data on implementation and design factors. The implementation factors were classified according to the Grol

and Wensing Implementation of Change model [21]. Implementation is defined as “a planned process and systematic introduction of innovations and/or changes of proven value” [21]. The model describes the steps for improving patient care with an intervention and summarizes the implementation literature. We extracted any activities that the authors described, which fit one or more steps in this model. Step 4 in this model is the selection of an implementation strategy. To define implementation strategies, we used the classification of implementation strategies in the EPOC taxonomy [22]. Implementation strategies (such as a CDSS or audit and feedback) that fit into the EPOC classification were also extracted from the included studies.

Design factors were classified according to the GUIDES (Guideline Implementation with Decision Support) checklist and the two-stream model [23,24]. The GUIDES checklist is a tool to support the development of successful CDSS and describes 4 groups: content, context, system, and implementation of the CDSS (eg, appropriateness of the information about CDSSs to users). The two-stream model contains elements describing factors that can potentially influence the success of a CDSS. We categorized the two-stream model elements into the 4 groups of the GUIDES checklist to obtain a complete picture of the potential design factors.

### Data Synthesis

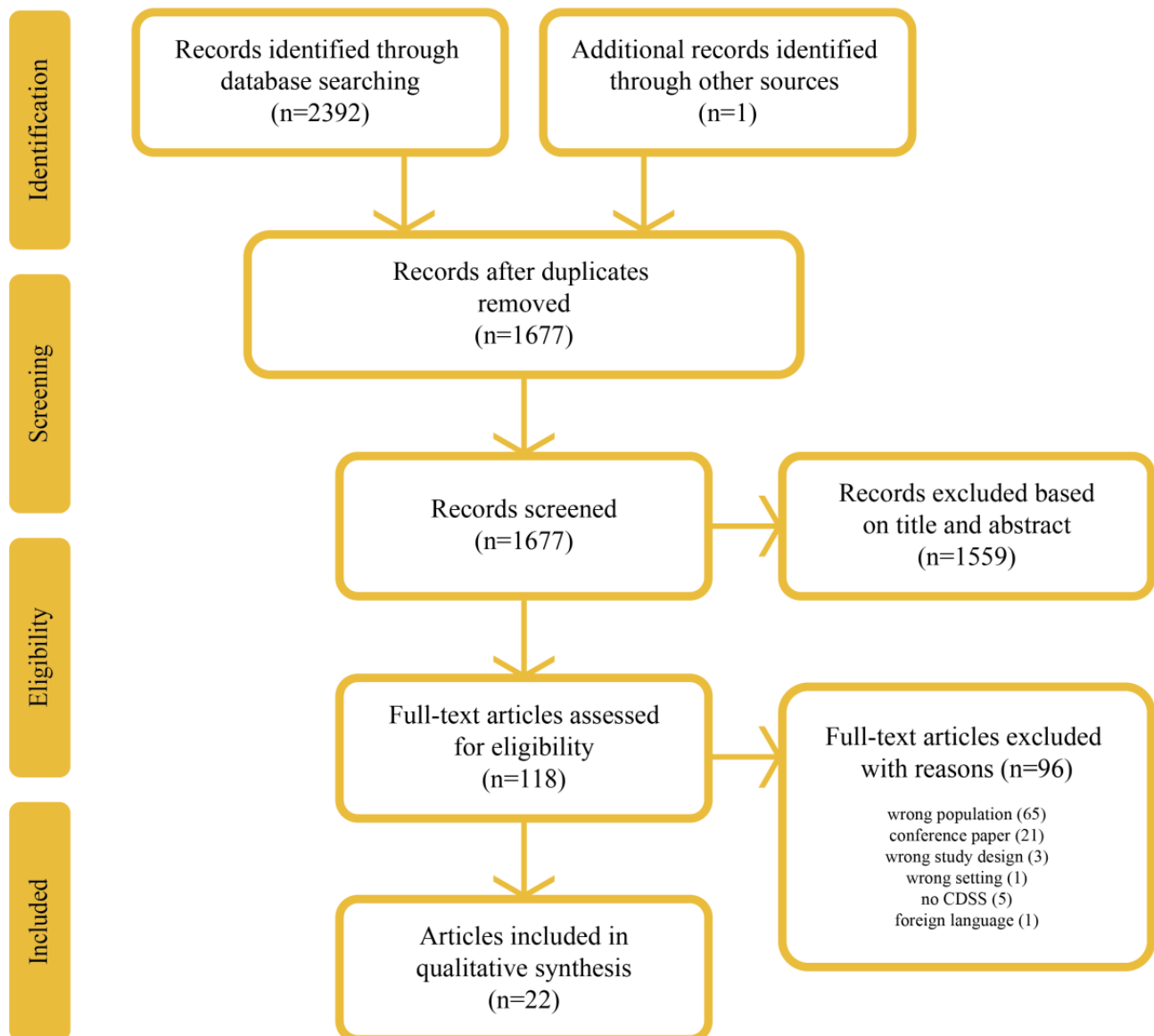
We conducted a narrative synthesis and counted which implementation and design factors were described in more effective interventions than ineffective interventions.

## Results

### Search Results

A total of 2392 articles were identified in the search. Figure 1 shows the PRISMA flow diagram with the number of articles excluded after each screening step and the reasons for excluding the full-text articles. A total of 22 articles were eligible for inclusion in our systematic review.

**Figure 1.** PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of search results. CDSS: Clinical Decision Support Systems.



### Characteristics of Included Studies

All the characteristics of the included articles are shown in [Multimedia Appendix 3](#) [25-46]. The 22 articles described interventions performed in 5 countries: 12 studies in the United States, 5 in Canada, 3 in Ireland, 1 in Italy, and 1 in France.

In total, 18 different CDSS interventions were described in the 22 included articles ([Multimedia Appendix 4](#)). A CDSS intervention was described in 3 articles: 1 article compared prescriptions at admission and discharge of the intervention group, 1 article described the main randomized controlled trial (RCT), and 1 article described the cost-effectiveness of the RCT [25-27]. Another CDSS intervention was described in 2 articles: 1 article evaluated the implementation at the initial site, and 1 article evaluated the implementation at 4 sites [28,29]. Finally, 1 CDSS intervention was linked in 2 articles: 1 article described a subgroup analysis of the earlier RCT [30,31].

Different study designs were selected to evaluate the interventions; 1 article used a cluster-randomized study, 7 articles used an RCT design, 1 article used a stepped wedge trial design, 2 articles used an interrupted time series design, and 11 articles used a before-after design. All RCTs had a registration of a protocol [27,30-34].

### Risk of Bias Assessment

[Multimedia Appendix 5](#) [25-46] shows the results of the risk of bias assessment. In 4 of the 22 articles, all suggested risk of bias criteria were categorized as low or unclear [32,35-37]. Other articles had 1 or more high risks for bias [25-31,33,34,38-46]. We did not find descriptions of the amount of missing data or how missing data were handled in any of the articles. All 7 RCTs had a high or unclear risk for protection against contamination [27,30-34]. The most frequent source of bias was “flawed or absent random sequence generation,” present in 14 studies [25-29,38-46]. This was mainly because

of studies with a nonrandomized design (eg, before-after studies).

### Effectiveness, Outcomes, and Geriatric Topic

In total, 72% (13/18) of interventions were effective in improving care, mainly with regard to process-related outcomes [25,27-29,32,34,35,37,39-42,44-46]. In 77% (10/13) of effective interventions, the effect was based on process-related outcomes, in 15% (2/13) of interventions on both process and patient-related outcomes, and in 8% (1/13) interventions on patient-related outcomes. In 60% (3/5) ineffective interventions, the results were based on both process and patient-related outcomes, in 20% (1/5) of interventions on patient-related outcomes and in 20% (1/5) interventions significance was not calculated; according to the definition we adopted in our review, this intervention was considered ineffective.

Of the 18 interventions, 8 (44%; 6 effective) focused on medication review, 8 (44%; 6 effective) on delirium, 7 (39%; 4 effective) on falls, 5 (28%; 4 effective) on functional decline, 4 (22%; 3 effective) on discharge or aftercare, and 2 (11%; 0 effective) on pressure ulcers. None of the interventions focused on incontinence, planning, or communication with patients at discharge. Part of the interventions on falls (3/7, 43%) and delirium (3/8, 38%) focused on improving drug prescription and not on other risk factors. For discharge, 2 of 4 interventions focused on (and succeeded in) improving prescriptions at emergency department discharge [28,29,32].

We grouped the 81 different outcomes into 6 groups: medication (35), location or duration (11), prevention of geriatric conditions (20), prevalence of geriatric conditions (10), survival (3), and costs (2). Outcomes in the medication and prevention of geriatric conditions groups were mostly process-related. Outcomes in the groups of prevalence of geriatric conditions and survival were mostly patient-related.

Patient-related outcome length of stay was measured in 10 interventions, none of which were primary outcomes, and none of them showed a significant improvement [26,30,31,34,36,38-44]. The 5 interventions measuring 30-day readmission also failed to show an effect on this outcome [30,33,34,39,43]. Other outcomes that did not show an effect in the included studies were survival and cost outcomes, delirium, and orders for consultation [26,30,31,34,36,39-41].

Patient-related outcomes that showed a statistically significant improvement ( $P=.04$ ) were falls, adverse drug reactions, and discharged home (percentage of patients who went home after discharge). Falls or fall rates were measured in 6 interventions and significantly reduced in 2 (primary outcome in 1) [30,36-38,41,42]. Adverse drug reactions or adverse drug events were measured in 2 interventions and significantly reduced in 1 (primary outcome) [26,27,45]. Discharged home was measured in 2 interventions and significantly improved in 1 (no primary outcomes) [30,31,39].

### Implementation Factors

Articles about effective interventions described more often an *a priori problem or performance analyses* and/or included more often *multifaceted interventions* than articles about ineffective

interventions. As [Multimedia Appendix 4](#) shows, in 69% (9/13) effective interventions and 0% (0/5) ineffective interventions, *a priori problem or performance analyses* were conducted before implementation [28,29,32,34,35,37,39,40,44,45]. This was done by reviewing prescribing data, investigating barriers and facilitators, mapping the use of computerized physician order entry, or describing care before implementation. In total, 62% (8/13) effective interventions and 20% (1/5) ineffective interventions were *multifaceted interventions* implying that the intervention had more than one implementation strategy [25-29,34,35,39-41,43,44].

[Multimedia Appendix 6](#) [25-46] shows all implementation and design factors per included article based on the Grol and Wensing Implementation of Change model, the GUIDES checklist, and the two-stream model. None of the included interventions described all 7 steps of the Grol and Wensing Implementation of Change model. All interventions reported an implementation strategy (step 4 in the model). All interventions described a CDSS, which is included in the implementation strategy *reminder*. Aside from *reminder*, the multifaceted interventions used varying strategies: 8 interventions described an educational strategy (7 effective), 2 audit and feedback (2 effective), 2 practice and setting (2 effective), 2 organizational culture (1 effective), and 1 local consensus processes (1 effective).

### CDSS Design Factors

Articles of effective interventions described only 1 design factor more frequently than articles of ineffective interventions: *consideration of the workflow*. The workflow before implementation was described or considered in the CDSS development in 69% (9/13) effective interventions and 20% (1/5) ineffective interventions [25-29,32,36,37,39-42].

The other design factors are shown in [Multimedia Appendix 6](#). Almost all studies described the clinical knowledge of CDSS. None of the studies described clinical knowledge based on prediction models or machine learning. Clinical knowledge was mostly based on the Beers criteria, STOPP (Screening Tool of Older Persons' Prescriptions)/START (Screening Tool to Alert to Right Treatment) criteria, experts, guidelines, or scientific literature [47-51]. In 11 interventions (8 effective), a multidisciplinary team with geriatricians and pharmacists was involved in selecting the clinical knowledge of the CDSS [25-29,32-35,40,42,43,45].

Overall, the presentation of the CDSSs varied and included 6 patient-specific reports (4 effective), 1 in-basket message (0 effective), 7 (non) interruptive alerts (5 effective), 2 default doses in computerized physician order entry (2 effective), and 6 (dynamic) order sets (5 effective). Only 5 interventions, of which 2 were effective, described the use of patient data from multiple parts of the patient record or multiple sources [33,34,43,45,46]. For medication review, 6 of 8 interventions described CDSSs built as stand-alone systems and therefore not integrated into the electronic health record [25-27,34,35,38,45,46]. The users of the systems were physicians in 9 interventions (7 effective), pharmacists in 6 interventions (5 effective), and nurses in 4 interventions (3



effective). Only 3 studies described a CDSS for multiple specialists [40,43-45].

## Discussion

### Principal Findings

In our systematic review, we found 22 articles describing 18 different CDSS interventions for the care of older hospitalized patients evaluated in controlled trials (including before-after and other quasi-experimental designs). These CDSS interventions focused on medication review, falls, delirium, discharge or aftercare, functional decline, and pressure ulcers. In total, 72% (13/18) of the included CDSS interventions effectively improved geriatric care, mainly concerning process-related outcomes. Two implementation factors—*a priori problem or performance analyses* and *multifaceted interventions*—and 1 design factor—*consideration of the workflow*—were described in more articles of effective interventions than ineffective ones. These factors are potentially associated with effectiveness; however, more trials are needed to quantify their impact or assess whether this association is causal in nature. No factors potentially associated with ineffectiveness were identified. We did not find any CDSS interventions for three geriatric problems: incontinence, planning, or communication with patients at discharge. The included interventions had limited effectiveness on patient outcomes. Furthermore, we found no data-driven CDSS in our systematic review.

Most of the 18 included interventions focused on medication review, delirium, and falls. We did not find any CDSS interventions for incontinence, planning, or communication with patients at discharge, and none of the CDSS interventions effectively improved care for pressure ulcers. Of the 8 interventions on medication review, 6 (75%) showed an improvement in prescribing for geriatric patients. This finding aligns with previous systematic reviews, which also stated that computerized support could improve prescribing for older patients [9-11]. For delirium and falls, 75% (6/8) of CDSS interventions improved care for delirium and 57% (4/7) for falls. Our review is the first to assess the effect of CDSS interventions on these common geriatric syndromes in older patients. Notably, even though these geriatric syndromes are multifactorial, almost half of the interventions for falls and delirium addressed only a single risk factor.

We found only 3 factors—2 implementation factors and 1 design factor, which were described in more articles about effective interventions than ineffective ones. In contrast to previously published reviews, no other design factors were identified in our study [12,13]. This could be because of the relatively small number of published CDSS interventions assessing the effect on geriatric care in a controlled trial; 2 of the 3 factors identified in our review were described in previous literature. In line with best practices in implementation science, *a priori analysis of problems and actual performance* was described more often in studies with positive outcomes [21]. The second approach, incorporating CDSS within the workflow, is in accordance with best practices as well [52-54]. However, for the third factor, the literature is inconsistent. We found a potential positive effect

of multifaceted interventions. In the implementation science literature, it is not clear whether multifaceted interventions are more effective than single interventions [55]. For falls, previously published systematic reviews also showed inconsistent results from multifaceted interventions, not specifically with CDSS, in hospitals [56,57].

Scientific literature in geriatrics often has a lower level of evidence because of heterogeneous patient characteristics and the underrepresentation of older patients in clinical trials [58]. Consequently, the clinical knowledge underlying CDSS has a lower level of evidence. The quality of clinical knowledge is important for the impact of the CDSS [59]. For the uptake and acceptance of CDSS in geriatric care, evaluation studies would preferably include patient outcomes not only to contribute to evidence on the effectiveness of the system but also to contribute evidence for the clinical knowledge. Our results showed that patient-related outcomes rarely significantly improved. This can be partly explained by the fact that only 3 interventions were evaluated with a patient-related outcome as the primary outcome, study sample sizes were too small to assess patient outcomes, and/or the choice of patient-related outcomes. In our systematic review, general patient-related outcomes such as length of stay and 30-day readmission did not improve; however, specific patient-related outcomes such as falls and adverse drug events were improved in some of the studies. A paper describing a framework for study designs in patient safety science stated that a common problem is that general patient-related outcomes can be influenced by factors other than the intervention [20]. Other systematic reviews of CDSSs also found sparse evidence for the association of CDSS with patient outcomes [9,12,60,61]. Two systematic reviews mentioned possible reasons: short duration of studies and logistics difficulties measuring the direct effect on patient outcomes and conducting RCTs for CDSS interventions [12,61]. On the contrary, a systematic review of CDSS for inpatients did find an effect on patient-related outcomes [59]. Future studies in geriatric CDSS should include a large enough sample size and duration and select appropriate outcomes directly influenced by the intervention to show significant effects on patient-related outcomes.

In our review, none of the clinical knowledge of the included CDSSs was data-driven; for example, it was based on prediction models or machine learning. Data-driven methods typically analyze large and complex data sets and are promising for CDSS [62,63]. However, evidence of the effectiveness of data-driven CDSS is thus far limited [63]. Challenges for data-driven CDSS include having the models as *black boxes that hamper users' understanding of the clinical knowledge underlying CDSS* [62]. An example of an effective data-driven CDSS without a *black box* is described in the study by Cho et al [64]. In this study, not specifically focused on older patients and therefore not included in our systematic review, a CDSS for pressure ulcers was developed with a Bayesian Network model and linked to the hospital electronic health record. The CDSS effectively reduced the prevalence of pressure ulcers and intensive care unit length of stay [64]. More studies are needed to explore the possibilities of data-driven CDSS for complex populations, such as older hospitalized patients.

The EPOC tool was used to assess the risk of bias in all studies. Nonrandomized study designs (eg, before-after studies) already have a high risk of bias because of their study design. Therefore, the overall bias of the included studies was high, except for 4 studies. Future evaluation studies should use randomized designs where possible or high-quality, nonrandomized designs, such as time series.

### Strengths and Limitations

Our systematic review is the first to provide an overview of the effect of CDSSs in improving care for various common geriatric problems in hospital care for older patients. It is complementary to previously published articles on CDSS for prescribing in this population [9]. CDSSs targeting aspects of care other than medication have not been previously studied in a systematic review. A strength of our study is that we incorporated implementation and design factors in the analysis to contribute to the understanding of CDSS effectiveness in this population. We used previous literature on geriatric care, implementation science, and CDSS to select geriatric topics and structure the implementation and design factors. Another strength of this study is that we used a broad and comprehensive search strategy, including checking the references of the studies. We chose to include all controlled studies; both RCTs and quasi-experimental studies. RCTs are generally considered the highest level of evidence; however, an RCT is often not practical in a CDSS implementation because of contamination issues. Thus, our choice to include other study designs provides a more representative picture of studies conducted with CDSSs.

A limitation of our study is that the included studies and extracted outcomes are heterogeneous and, therefore, not sufficiently comparable for quantitative analysis. More intervention studies are needed to quantify the effects on specific geriatric problems and investigate potential influencing factors

on the effectiveness of these CDSS interventions. Implementation and design factors not described in the articles were not included in the analysis, which may have led to the underrepresentation of these factors. Furthermore, 2 of the 18 included CDSS interventions used almost the same implementation strategy in the same hospital, but at different periods and with a different CDSS design: the first intervention had a manual entry and the second was automatic [34,35]. Our results can be affected by publication bias because, especially with weaker study designs, studies showing an effect are more likely to be published. The inclusion and data extraction processes were performed by 2 individual researchers to minimize potential bias.

### Conclusions

In conclusion, our systematic review shows that CDSS interventions have the potential to improve the hospital care of older patients. In total, 72% (13/18) of the included interventions were effective (mostly on process outcomes). Two implementation factors—*a priori problem or performance analyses* and *multifaceted interventions*—and 1 design factor—*consideration of the workflow*—were reported more frequently in articles of effective interventions. However, more studies are needed to assess the impact of a CDSS intervention on care for older hospitalized patients. Future studies should use a strong study design, such as a randomized trial or interrupted time series. RCTs are often challenging in CDSS research because of the risk of contamination and technical issues in randomizing the intervention. Furthermore, future studies should include a large enough sample size and duration and select specific patient-related outcomes directly affected by the intervention. Future studies should assess the effect on geriatric conditions, quantify the impact of implementation and design factors on CDSS effectiveness, and investigate the potential of personalized (data-driven) interventions.

---

### Acknowledgments

The authors would like to thank the medical information specialist Joost Daams and Master's student Agnieszka van de Leur for their help in building the search strategy.

---

### Conflicts of Interest

None declared.

---

#### Multimedia Appendix 1

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

[PDF File (Adobe PDF File), 118 KB - [medinform\\_v9i7e28023\\_app1.pdf](#) ]

---

#### Multimedia Appendix 2

Search strategy.

[DOCX File , 15 KB - [medinform\\_v9i7e28023\\_app2.docx](#) ]

---

#### Multimedia Appendix 3

Study design, characteristics, and outcomes of the included studies.

[DOCX File , 23 KB - [medinform\\_v9i7e28023\\_app3.docx](#) ]

---

#### Multimedia Appendix 4

Study design, characteristics, outcomes, and main implementation or design factors of the included articles.

[[DOCX File , 21 KB](#) - [medinform\\_v9i7e28023\\_app4.docx](#) ]

Multimedia Appendix 5

Risk of bias.

[[DOCX File , 38 KB](#) - [medinform\\_v9i7e28023\\_app5.docx](#) ]

Multimedia Appendix 6

Implementation and design factors described in the included studies.

[[DOCX File , 51 KB](#) - [medinform\\_v9i7e28023\\_app6.docx](#) ]

## References

1. World Population Ageing. New York, USA: United Nations; 2015.
2. Statistiek Centraal Bureau, Statline. URL: <https://opendata.cbs.nl/statline/#/CBS/nl/?fromstatweb> [accessed 2020-06-18]
3. Thomas E, Brennan TA. Incidence and types of preventable adverse events in elderly patients: population based review of medical records. *Br Med J* 2000 Mar 18;320(7237):741-744 [FREE Full text] [doi: [10.1136/bmj.320.7237.741](https://doi.org/10.1136/bmj.320.7237.741)] [Medline: [10720355](https://pubmed.ncbi.nlm.nih.gov/10720355/)]
4. Lakhan P, Jones M, Wilson A, Courtney M, Hirdes J, Gray LC. A prospective cohort study of geriatric syndromes among older medical patients admitted to acute care hospitals. *J Am Geriatr Soc* 2011 Nov 10;59(11):2001-2008. [doi: [10.1111/j.1532-5415.2011.03663.x](https://doi.org/10.1111/j.1532-5415.2011.03663.x)] [Medline: [22092231](https://pubmed.ncbi.nlm.nih.gov/22092231/)]
5. Ponzetto M, Maero B, Maina P, Rosato R, Ciccone G, Merletti F, et al. Risk factors for early and late mortality in hospitalized older patients: the continuing importance of functional status. *J Gerontol A Biol Sci Med Sci* 2003 Nov 1;58(11):1049-1054. [doi: [10.1093/gerona/58.11.m1049](https://doi.org/10.1093/gerona/58.11.m1049)] [Medline: [14630889](https://pubmed.ncbi.nlm.nih.gov/14630889/)]
6. Inouye S, Studenski S, Tinetti M, Kuchel G. Geriatric syndromes: clinical, research, and policy implications of a core geriatric concept. *J Am Geriatr Soc* 2007 May;55(5):780-791 [FREE Full text] [doi: [10.1111/j.1532-5415.2007.01156.x](https://doi.org/10.1111/j.1532-5415.2007.01156.x)] [Medline: [17493201](https://pubmed.ncbi.nlm.nih.gov/17493201/)]
7. Goud R, van Engen-Verheul M, de Keizer NF, Bal R, Hasman A, Hellemans IM, et al. The effect of computerized decision support on barriers to guideline implementation: a qualitative study in outpatient cardiac rehabilitation. *Int J Med Inform* 2010 Jun;79(6):430-437. [doi: [10.1016/j.ijmedinf.2010.03.001](https://doi.org/10.1016/j.ijmedinf.2010.03.001)] [Medline: [20378396](https://pubmed.ncbi.nlm.nih.gov/20378396/)]
8. Damoiseaux-Volman BA, Medlock S, Ploegmakers KJ, Karapinar-Çarkit F, Krediet CP, de Rooij SE, et al. Priority setting in improving hospital care for older patients using clinical decision support. *J Am Med Dir Assoc* 2019 Aug;20(8):1045-1047. [doi: [10.1016/j.jamda.2019.03.017](https://doi.org/10.1016/j.jamda.2019.03.017)] [Medline: [31056454](https://pubmed.ncbi.nlm.nih.gov/31056454/)]
9. Dalton K, O'Brien G, O'Mahony D, Byrne S. Computerised interventions designed to reduce potentially inappropriate prescribing in hospitalised older adults: a systematic review and meta-analysis. *Age Ageing* 2018 Sep 1;47(5):670-678. [doi: [10.1093/ageing/afy086](https://doi.org/10.1093/ageing/afy086)] [Medline: [29893779](https://pubmed.ncbi.nlm.nih.gov/29893779/)]
10. Clyne B, Bradley MC, Hughes C, Fahey T, Lapane KL. Electronic prescribing and other forms of technology to reduce inappropriate medication use and polypharmacy in older people: a review of current evidence. *Clin Geriatr Med* 2012 May;28(2):301-322. [doi: [10.1016/j.cger.2012.01.009](https://doi.org/10.1016/j.cger.2012.01.009)] [Medline: [22500545](https://pubmed.ncbi.nlm.nih.gov/22500545/)]
11. Yourman L, Concato J, Agostini JV. Use of computer decision support interventions to improve medication prescribing in older adults: a systematic review. *Am J Geriatr Pharmacother* 2008 Jun;6(2):119-129. [doi: [10.1016/j.amjopharm.2008.06.001](https://doi.org/10.1016/j.amjopharm.2008.06.001)] [Medline: [18675770](https://pubmed.ncbi.nlm.nih.gov/18675770/)]
12. Roshanov PS, Fernandes N, Wilczynski JM, Hemens BJ, You JJ, Handler SM, et al. Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. *Br Med J* 2013 Feb 14;346(feb14 1):f657 [FREE Full text] [doi: [10.1136/bmj.f657](https://doi.org/10.1136/bmj.f657)] [Medline: [23412440](https://pubmed.ncbi.nlm.nih.gov/23412440/)]
13. Van de Velde S, Heselmans A, Delvaux N, Brandt L, Marco-Ruiz L, Spitaels D, et al. A systematic review of trials evaluating success factors of interventions with computerised clinical decision support. *Implement Sci* 2018 Aug 20;13(1):114 [FREE Full text] [doi: [10.1186/s13012-018-0790-1](https://doi.org/10.1186/s13012-018-0790-1)] [Medline: [30126421](https://pubmed.ncbi.nlm.nih.gov/30126421/)]
14. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: [10.1371/journal.pmed.1000097](https://doi.org/10.1371/journal.pmed.1000097)] [Medline: [19621072](https://pubmed.ncbi.nlm.nih.gov/19621072/)]
15. Bramer WM, Giustini D, De Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. *J Med Libr Assoc* 2016 Sep 12;104(3). [doi: [10.5195/jmla.2016.24](https://doi.org/10.5195/jmla.2016.24)]
16. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev* 2016 Dec 5;5(1):210 [FREE Full text] [doi: [10.1186/s13643-016-0384-4](https://doi.org/10.1186/s13643-016-0384-4)] [Medline: [27919275](https://pubmed.ncbi.nlm.nih.gov/27919275/)]
17. Shahar Y, Shirliff EH. Clinical Decision Support Systems. In: Shirliff EH, Cimino JJ, editors. *Biomedical Informatics - Computer Applications in Health Care and Biomedicine*. New York, USA: Springer; 2006.



18. Griese-Mammen N, Hersberger KE, Messerli M, Leikola S, Horvat N, van Mil JW, et al. PCNE definition of medication review: reaching agreement. *Int J Clin Pharm* 2018 Oct 2;40(5):1199-1208. [doi: [10.1007/s11096-018-0696-7](https://doi.org/10.1007/s11096-018-0696-7)] [Medline: [30073611](https://pubmed.ncbi.nlm.nih.gov/30073611/)]
19. Cochrane Effective Practice and Organisation of Care. EPOC Resources for Review Authors. 2017. URL: [epoc.cochrane.org/resources/epoc-resources-review-authors](https://epoc.cochrane.org/resources/epoc-resources-review-authors) [accessed 2020-06-01]
20. Brown C, Hofer T, Johal A, Thomson R, Nicholl J, Franklin BD, et al. An epistemology of patient safety research: a framework for study design and interpretation. Part 3. End points and measurement. *Qual Saf Health Care* 2008 Jun;17(3):170-177. [doi: [10.1136/qshc.2007.023655](https://doi.org/10.1136/qshc.2007.023655)] [Medline: [18519622](https://pubmed.ncbi.nlm.nih.gov/18519622/)]
21. Grol R, Wensing R, Eccles M, Davis D. Improving Patient Care. The Implementation of Change in Health Care. Second Edition. Oxford, UK: John Wiley & Sons; 2017.
22. Effective Practice and Organisation of Care. EPOC Taxonomy. 2015. URL: [epoc.cochrane.org/epoc-taxonomy](https://epoc.cochrane.org/epoc-taxonomy) [accessed 2020-06-01]
23. Medlock S, Wyatt J, Patel V, Shortliffe E, Abu-Hanna A. Modeling information flows in clinical decision support: key insights for enhancing system effectiveness. *J Am Med Inform Assoc* 2016 Sep;23(5):1001-1006. [doi: [10.1093/jamia/ocv177](https://doi.org/10.1093/jamia/ocv177)] [Medline: [26911809](https://pubmed.ncbi.nlm.nih.gov/26911809/)]
24. Van de Velde S, Kunnamo I, Roshanov P, Kortteisto T, Aertgeerts B, Vandvik PO, GUIDES expert panel. The GUIDES checklist: development of a tool to improve the successful use of guideline-based computerised clinical decision support. *Implement Sci* 2018 Jun 25;13(1):86 [FREE Full text] [doi: [10.1186/s13012-018-0772-3](https://doi.org/10.1186/s13012-018-0772-3)] [Medline: [29941007](https://pubmed.ncbi.nlm.nih.gov/29941007/)]
25. O'Sullivan D, O'Mahony D, O'Connor MN, Gallagher P, Cullinan S, O'Sullivan R, et al. The impact of a structured pharmacist intervention on the appropriateness of prescribing in older hospitalized patients. *Drugs Aging* 2014 Jun 6;31(6):471-481. [doi: [10.1007/s40266-014-0172-6](https://doi.org/10.1007/s40266-014-0172-6)] [Medline: [24797285](https://pubmed.ncbi.nlm.nih.gov/24797285/)]
26. O'Sullivan D, O'Mahony D, O'Connor MN, Gallagher P, Gallagher J, Cullinan S, et al. Prevention of adverse drug reactions in hospitalised older patients using a software-supported structured pharmacist intervention: a cluster randomised controlled trial. *Drugs Aging* 2016 Jan;33(1):63-73. [doi: [10.1007/s40266-015-0329-y](https://doi.org/10.1007/s40266-015-0329-y)] [Medline: [26597401](https://pubmed.ncbi.nlm.nih.gov/26597401/)]
27. Gallagher J, O'Sullivan D, McCarthy S, Gillespie P, Woods N, O'Mahony D, et al. Structured pharmacist review of medication in older hospitalised patients: a cost-effectiveness analysis. *Drugs Aging* 2016 Apr;33(4):285-294. [doi: [10.1007/s40266-016-0348-3](https://doi.org/10.1007/s40266-016-0348-3)] [Medline: [26861468](https://pubmed.ncbi.nlm.nih.gov/26861468/)]
28. Stevens M, Hastings SN, Markland AD, Hwang U, Hung W, Vandenberg AE, et al. Enhancing quality of provider practices for older adults in the emergency department (equipped). *J Am Geriatr Soc* 2017 Jul;65(7):1609-1614. [doi: [10.1111/jgs.14890](https://doi.org/10.1111/jgs.14890)] [Medline: [28388818](https://pubmed.ncbi.nlm.nih.gov/28388818/)]
29. Stevens MB, Hastings SN, Powers J, Vandenberg AE, Echt KV, Bryan WE, et al. Enhancing the quality of prescribing practices for older veterans discharged from the emergency department (equipped): preliminary results from enhancing quality of prescribing practices for older veterans discharged from the emergency department, a novel multicomponent interdisciplinary quality improvement initiative. *J Am Geriatr Soc* 2015 May 6;63(5):1025-1029. [doi: [10.1111/jgs.13404](https://doi.org/10.1111/jgs.13404)] [Medline: [25945692](https://pubmed.ncbi.nlm.nih.gov/25945692/)]
30. Boustani MA, Campbell NL, Khan BA, Abernathy G, Zawahiri M, Campbell T, et al. Enhancing care for hospitalized older adults with cognitive impairment: a randomized controlled trial. *J Gen Intern Med* 2012 May 3;27(5):561-567 [FREE Full text] [doi: [10.1007/s11606-012-1994-8](https://doi.org/10.1007/s11606-012-1994-8)] [Medline: [22302355](https://pubmed.ncbi.nlm.nih.gov/22302355/)]
31. Khan BA, Calvo-Ayala E, Campbell N, Perkins A, Ionescu R, Tricker J, et al. Clinical decision support system and incidence of delirium in cognitively impaired older adults transferred to intensive care. *Am J Crit Care* 2013 May 1;22(3):257-262 [FREE Full text] [doi: [10.4037/ajcc2013447](https://doi.org/10.4037/ajcc2013447)] [Medline: [23635936](https://pubmed.ncbi.nlm.nih.gov/23635936/)]
32. Terrell K, Perkins A, Dexter P, Hui S, Callahan C, Miller D. Computerized decision support to reduce potentially inappropriate prescribing to older emergency department patients: a randomized, controlled trial. *J Am Geriatr Soc* 2009 Aug;57(8):1388-1394. [doi: [10.1111/j.1532-5415.2009.02352.x](https://doi.org/10.1111/j.1532-5415.2009.02352.x)] [Medline: [19549022](https://pubmed.ncbi.nlm.nih.gov/19549022/)]
33. Gurwitz JH, Field TS, Ogarek J, Tjia J, Cutrona SL, Harrold LR, et al. An electronic health record-based intervention to increase follow-up office visits and decrease rehospitalization in older adults. *J Am Geriatr Soc* 2014 May 29;62(5):865-871 [FREE Full text] [doi: [10.1111/jgs.12798](https://doi.org/10.1111/jgs.12798)] [Medline: [24779524](https://pubmed.ncbi.nlm.nih.gov/24779524/)]
34. Cossette B, Éthier JF, Joly-Mischlich T, Bergeron J, Ricard G, Brazeau S, et al. Reduction in targeted potentially inappropriate medication use in elderly inpatients: a pragmatic randomized controlled trial. *Eur J Clin Pharmacol* 2017 Oct;73(10):1237-1245. [doi: [10.1007/s00228-017-2293-4](https://doi.org/10.1007/s00228-017-2293-4)] [Medline: [28717929](https://pubmed.ncbi.nlm.nih.gov/28717929/)]
35. Cossette B, Bergeron J, Ricard G, Éthier JF, Joly-Mischlich T, Levine M, et al. Knowledge translation strategy to reduce the use of potentially inappropriate medications in hospitalized elderly adults. *J Am Geriatr Soc* 2016 Dec 2;64(12):2487-2494. [doi: [10.1111/jgs.14322](https://doi.org/10.1111/jgs.14322)] [Medline: [27590168](https://pubmed.ncbi.nlm.nih.gov/27590168/)]
36. Holroyd-Leduc J, Abelseth G, Khandwala F, Silvius J, Hogan D, Schmaltz H, et al. A pragmatic study exploring the prevention of delirium among hospitalized older hip fracture patients: Applying evidence to routine clinical practice using clinical decision support. *Implement Sci* 2010 Oct 22;5:81 [FREE Full text] [doi: [10.1186/1748-5908-5-81](https://doi.org/10.1186/1748-5908-5-81)] [Medline: [20969770](https://pubmed.ncbi.nlm.nih.gov/20969770/)]
37. Dykes PC, Carroll DL, Hurley A, Lipsitz S, Benoit A, Chang F, et al. Fall prevention in acute care hospitals: a randomized trial. *J Am Med Assoc* 2010 Nov 3;304(17):1912-1918 [FREE Full text] [doi: [10.1001/jama.2010.1567](https://doi.org/10.1001/jama.2010.1567)] [Medline: [21045097](https://pubmed.ncbi.nlm.nih.gov/21045097/)]

38. Lagrange F, Lagrange J, Bennaga C, Taloub F, Keddi M, Dumoulin B. A context-aware decision-support system in clinical pharmacy: drug monitoring in the elderly. *Le Pharmacien Hospitalier et Clinicien* 2017 Mar;52(1):100-110. [doi: [10.1016/j.phclin.2017.01.004](https://doi.org/10.1016/j.phclin.2017.01.004)]
39. Mattison ML, Catic A, Davis RB, Olveczky D, Moran J, Yang J, et al. A standardized, bundled approach to providing geriatric-focused acute care. *J Am Geriatr Soc* 2014 May 18;62(5):936-942 [FREE Full text] [doi: [10.1111/jgs.12780](https://doi.org/10.1111/jgs.12780)] [Medline: [24749723](https://pubmed.ncbi.nlm.nih.gov/24749723/)]
40. Adeola M, Azad R, Kassie GM, Shirkey B, Taffet G, Liebl M, et al. Multicomponent interventions reduce high-risk medications for delirium in hospitalized older adults. *J Am Geriatr Soc* 2018 Aug 23;66(8):1638-1645. [doi: [10.1111/jgs.15438](https://doi.org/10.1111/jgs.15438)] [Medline: [30035315](https://pubmed.ncbi.nlm.nih.gov/30035315/)]
41. Groshaus H, Boscan A, Khandwala F, Holroyd-Leduc J. Use of clinical decision support to improve the quality of care provided to older hospitalized patients. *Appl Clin Inform* 2017 Dec 16;3(1):94-102. [doi: [10.4338/aci-2011-08-ra-0047](https://doi.org/10.4338/aci-2011-08-ra-0047)]
42. Peterson JF, Kuperman GJ, Shek C, Patel M, Avorn J, Bates DW. Guided prescription of psychotropic medications for geriatric inpatients. *Arch Intern Med* 2005 Apr 11;165(7):802-807. [doi: [10.1001/archinte.165.7.802](https://doi.org/10.1001/archinte.165.7.802)] [Medline: [15824302](https://pubmed.ncbi.nlm.nih.gov/15824302/)]
43. Malone M, Vollbrecht M, Stephenson J, Burke L, Pagel P, Goodwin J. AcuteCare for Elders (ACE) tracker and e-Geriatrician: methods to disseminate ACE concepts to hospitals with no geriatricians on staff. *J Am Geriatr Soc* 2010 Jan;58(1):161-167 [FREE Full text] [doi: [10.1111/j.1532-5415.2009.02624.x](https://doi.org/10.1111/j.1532-5415.2009.02624.x)] [Medline: [20122048](https://pubmed.ncbi.nlm.nih.gov/20122048/)]
44. Booth KA, Simmons EE, Viles AF, Gray WA, Kennedy KR, Biswal SH, et al. Improving geriatric care processes on two medical-surgical acute care units: a pilot study. *J Healthc Qual* 2019;41(1):23-31. [doi: [10.1097/jhq.000000000000140](https://doi.org/10.1097/jhq.000000000000140)]
45. McDonald EG, Wu PE, Rashidi B, Forster AJ, Huang A, Pilote L, et al. The medsafer study: a controlled trial of an electronic decision support tool for deprescribing in acute care. *J Am Geriatr Soc* 2019 Sep;67(9):1843-1850. [doi: [10.1111/jgs.16040](https://doi.org/10.1111/jgs.16040)] [Medline: [31250427](https://pubmed.ncbi.nlm.nih.gov/31250427/)]
46. Ghibelli S, Marengoni A, Djade CD, Nobili A, Tettamanti M, Franchi C, et al. Prevention of inappropriate prescribing in hospitalized older patients using a computerized prescription support system (INTERcheck®). *Drugs Aging* 2013 Oct 14;30(10):821-828. [doi: [10.1007/s40266-013-0109-5](https://doi.org/10.1007/s40266-013-0109-5)] [Medline: [23943248](https://pubmed.ncbi.nlm.nih.gov/23943248/)]
47. O'Mahony D, O'Sullivan D, Byrne S, O'Connor MN, Ryan C, Gallagher P. STOPP/START criteria for potentially inappropriate prescribing in older people: version 2. *Age Ageing* 2015 Mar;44(2):213-218 [FREE Full text] [doi: [10.1093/ageing/afu145](https://doi.org/10.1093/ageing/afu145)] [Medline: [25324330](https://pubmed.ncbi.nlm.nih.gov/25324330/)]
48. By the American Geriatrics Society 2015 Beers Criteria Update Expert Panel. American geriatrics society 2015 updated beers criteria for potentially inappropriate medication use in older adults. *J Am Geriatr Soc* 2015 Nov;63(11):2227-2246. [doi: [10.1111/jgs.13702](https://doi.org/10.1111/jgs.13702)] [Medline: [26446832](https://pubmed.ncbi.nlm.nih.gov/26446832/)]
49. American Geriatrics Society 2012 Beers Criteria Update Expert Panel T. American geriatrics society updated beers criteria for potentially inappropriate medication use in older adults. *J Am Geriatr Soc* 2012 Apr;60(4):616-631 [FREE Full text] [doi: [10.1111/j.1532-5415.2012.03923.x](https://doi.org/10.1111/j.1532-5415.2012.03923.x)] [Medline: [22376048](https://pubmed.ncbi.nlm.nih.gov/22376048/)]
50. Fick DM, Cooper JW, Wade WE, Waller JL, Maclean JR, Beers MH. Updating the Beers criteria for potentially inappropriate medication use in older adults: results of a US consensus panel of experts. *Arch Intern Med* 2003 Dec 8;163(22):2716-2724. [doi: [10.1001/archinte.163.22.2716](https://doi.org/10.1001/archinte.163.22.2716)] [Medline: [14662625](https://pubmed.ncbi.nlm.nih.gov/14662625/)]
51. Gallagher P, Ryan C, Byrne S, Kennedy J, O'Mahony D. STOPP (Screening Tool of Older Person's Prescriptions) and START (Screening Tool to Alert doctors to Right Treatment). Consensus validation. *Int J Clin Pharmacol Ther* 2008 Feb;46(2):72-83. [doi: [10.5414/cpp46072](https://doi.org/10.5414/cpp46072)] [Medline: [18218287](https://pubmed.ncbi.nlm.nih.gov/18218287/)]
52. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. *Implement Sci* 2017 Sep 15;12(1):113 [FREE Full text] [doi: [10.1186/s13012-017-0644-2](https://doi.org/10.1186/s13012-017-0644-2)] [Medline: [28915822](https://pubmed.ncbi.nlm.nih.gov/28915822/)]
53. Gross PA, Bates DW. A pragmatic approach to implementing best practices for clinical decision support systems in computerized provider order entry systems. *J Am Med Informatics Assoc* 2007 Jan 1;14(1):25-28. [doi: [10.1197/jamia.m2173](https://doi.org/10.1197/jamia.m2173)]
54. Moxey A, Robertson J, Newby D, Hains I, Williamson M, Pearson S. Computerized clinical decision support for prescribing: provision does not guarantee uptake. *J Am Med Inform Assoc* 2010;17(1):25-33 [FREE Full text] [doi: [10.1197/jamia.M3170](https://doi.org/10.1197/jamia.M3170)] [Medline: [20064798](https://pubmed.ncbi.nlm.nih.gov/20064798/)]
55. Grol R, Grimshaw J. From best evidence to best practice: effective implementation of change in patients' care. *Lancet* 2003 Oct;362(9391):1225-1230. [doi: [10.1016/s0140-6736\(03\)14546-1](https://doi.org/10.1016/s0140-6736(03)14546-1)]
56. Oliver D, Connelly JB, Victor CR, Shaw FE, Whitehead A, Genc Y, et al. Strategies to prevent falls and fractures in hospitals and care homes and effect of cognitive impairment: systematic review and meta-analyses. *Br Med J* 2007 Jan 13;334(7584):82 [FREE Full text] [doi: [10.1136/bmj.39049.706493.55](https://doi.org/10.1136/bmj.39049.706493.55)] [Medline: [17158580](https://pubmed.ncbi.nlm.nih.gov/17158580/)]
57. Hempel S, Newberry S, Wang Z, Booth M, Shanman R, Johnsen B, et al. Hospital fall prevention: a systematic review of implementation, components, adherence, and effectiveness. *J Am Geriatr Soc* 2013 Apr;61(4):483-494 [FREE Full text] [doi: [10.1111/jgs.12169](https://doi.org/10.1111/jgs.12169)] [Medline: [23527904](https://pubmed.ncbi.nlm.nih.gov/23527904/)]
58. Mooijaart SP, Broekhuizen K, Trompet S, de Craen AJ, Gussekloo J, Oleksik A, et al. Evidence-based medicine in older patients: how can we do better? *Neth J Med* 2015 Jun;73(5):211-218 [FREE Full text] [Medline: [26087800](https://pubmed.ncbi.nlm.nih.gov/26087800/)]

59. Varghese J, Kleine M, Gessner SI, Sandmann S, Dugas M. Effects of computerized decision support system implementations on patient outcomes in inpatient care: a systematic review. *J Am Med Inform Assoc* 2018 May 1;25(5):593-602 [FREE Full text] [doi: [10.1093/jamia/ocx100](https://doi.org/10.1093/jamia/ocx100)] [Medline: [29036406](https://pubmed.ncbi.nlm.nih.gov/29036406/)]
60. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. *Ann Intern Med* 2012 Jul 3;157(1):29-43 [FREE Full text] [doi: [10.7326/0003-4819-157-1-201207030-00450](https://doi.org/10.7326/0003-4819-157-1-201207030-00450)] [Medline: [22751758](https://pubmed.ncbi.nlm.nih.gov/22751758/)]
61. Robertson J, Walkom E, Pearson S, Hains I, Williamsone M, Newby D. The impact of pharmacy computerised clinical decision support on prescribing, clinical and patient outcomes: a systematic review of the literature. *Int J Pharm Pract* 2010 Apr;18(2):69-87. [Medline: [20441116](https://pubmed.ncbi.nlm.nih.gov/20441116/)]
62. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *J Am Med Assoc* 2018 Dec 4;320(21):2199-2200. [doi: [10.1001/jama.2018.17163](https://doi.org/10.1001/jama.2018.17163)] [Medline: [30398550](https://pubmed.ncbi.nlm.nih.gov/30398550/)]
63. Cresswell K, Callaghan M, Khan S, Sheikh Z, Mozaffar H, Sheikh A. Investigating the use of data-driven artificial intelligence in computerised decision support systems for health and social care: A systematic review. *Health Informatics J* 2020 Sep;26(3):2138-2147 [FREE Full text] [doi: [10.1177/1460458219900452](https://doi.org/10.1177/1460458219900452)] [Medline: [31964204](https://pubmed.ncbi.nlm.nih.gov/31964204/)]
64. Cho I, Park I, Kim E, Lee E, Bates DW. Using EHR data to predict hospital-acquired pressure ulcers: a prospective study of a Bayesian Network model. *Int J Med Inform* 2013 Nov;82(11):1059-1067. [doi: [10.1016/j.ijmedinf.2013.06.012](https://doi.org/10.1016/j.ijmedinf.2013.06.012)] [Medline: [23891086](https://pubmed.ncbi.nlm.nih.gov/23891086/)]

## Abbreviations

**CDSS:** clinical decision support system

**EPOC:** Effective Practice and Organisation of Care

**GUIDES:** Guideline Implementation with Decision Support

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

**PROSPERO:** International Prospective Register of Systematic Reviews

**RCT:** randomized controlled trial

**START:** Screening Tool to Alert to Right Treatment

**STOPP:** Screening Tool of Older Persons' Prescriptions

*Edited by C Lovis; submitted 17.02.21; peer-reviewed by T Korteisto, J Klopotoska; comments to author 14.04.21; revised version received 10.05.21; accepted 17.05.21; published 16.07.21.*

*Please cite as:*

*Damoiseaux-Volman BA, van der Velde N, Ruige SG, Romijn JA, Abu-Hanna A, Medlock S*

*Effect of Interventions With a Clinical Decision Support System for Hospitalized Older Patients: Systematic Review Mapping Implementation and Design Factors*

*JMIR Med Inform* 2021;9(7):e28023

URL: <https://medinform.jmir.org/2021/7/e28023>

doi: [10.2196/28023](https://doi.org/10.2196/28023)

PMID: [34269682](https://pubmed.ncbi.nlm.nih.gov/34269682/)

©Birgit A Damoiseaux-Volman, Nathalie van der Velde, Sil G Ruige, Johannes A Romijn, Ameen Abu-Hanna, Stephanie Medlock. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org/>), 16.07.2021. 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

# Contact Tracing Apps: Lessons Learned on Privacy, Autonomy, and the Need for Detailed and Thoughtful Implementation

Katie Hogan<sup>1\*</sup>, BSc; Briana Macedo<sup>2\*</sup>; Venkata Macha<sup>3</sup>, BSc; Arko Barman<sup>4,5</sup>, PhD; Xiaoqian Jiang<sup>6</sup>, PhD

<sup>1</sup>Department of Bioengineering, Rice University, Houston, TX, United States

<sup>2</sup>School of Engineering, Princeton University, Princeton, NJ, United States

<sup>3</sup>School of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

<sup>4</sup>Department of Electrical & Computer Engineering, Rice University, Houston, TX, United States

<sup>5</sup>Data to Knowledge Lab, Rice University, Houston, TX, United States

<sup>6</sup>School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, United States

\*these authors contributed equally

**Corresponding Author:**

Xiaoqian Jiang, PhD

School of Biomedical Informatics

University of Texas Health Science Center at Houston

7000 Fannin St #600

Houston, TX, 77030

United States

Phone: 1 7135003930

Email: [xiaoqian.jiang@uth.tmc.edu](mailto:xiaoqian.jiang@uth.tmc.edu)

## Abstract

The global and national response to the COVID-19 pandemic has been inadequate due to a collective lack of preparation and a shortage of available tools for responding to a large-scale pandemic. By applying lessons learned to create better preventative methods and speedier interventions, the harm of a future pandemic may be dramatically reduced. One potential measure is the widespread use of contact tracing apps. While such apps were designed to combat the COVID-19 pandemic, the time scale in which these apps were deployed proved a significant barrier to efficacy. Many companies and governments sprinted to deploy contact tracing apps that were not properly vetted for performance, privacy, or security issues. The hasty development of incomplete contact tracing apps undermined public trust and negatively influenced perceptions of app efficacy. As a result, many of these apps had poor voluntary public uptake, which greatly decreased the apps' efficacy. Now, with lessons learned from this pandemic, groups can better design and test apps in preparation for the future. In this viewpoint, we outline common strategies employed for contact tracing apps, detail the successes and shortcomings of several prominent apps, and describe lessons learned that may be used to shape effective contact tracing apps for the present and future. Future app designers can keep these lessons in mind to create a version that is suitable for their local culture, especially with regard to local attitudes toward privacy-utility tradeoffs during public health crises.

(*JMIR Med Inform* 2021;9(7):e27449) doi:[10.2196/27449](https://doi.org/10.2196/27449)

**KEYWORDS**

contact tracing; COVID-19; privacy; smartphone apps; mobile phone apps; health information; electronic health; eHealth; pandemic; app; mobile health; mHealth

## Introduction

At the end of 2019, a novel coronavirus was determined to be associated with a group of pneumonia cases in Wuhan, China. In 2020, this novel coronavirus spread rapidly throughout China and the globe, prompting the World Health Organization to name the disease COVID-19 and the virus associated with the disease SARS-CoV-2 [1]. By October 2020, the COVID-19

pandemic had resulted in over 200,000 deaths in the United States [2] and over 1,000,000 deaths globally [3].

Strategies to monitor and control the spread of COVID-19 have hinged around a combination of traditional and nontraditional strategies, including rapid testing, self-quarantine, and contact tracing. Contact tracing has formed a key component of the plans to control the spread of infectious diseases in recent years, yielding a wealth of literature concerning its importance and

efficacy. Contact tracing traditionally involves interviewing infected individuals and following up with any close contacts to communicate increased risk due to exposure and provide information and strategies related to that risk [1]. Alongside face-to-face contact tracing, various digital strategies have been employed to mediate this process and increase efficacy, including the introduction of a wide variety of contact tracing apps. In an article modeling the efficacy of contact tracing for Ebola, Browne et al [4] identified and examined key epidemiological parameters that impact the efficacy of contact tracing, including incubation period, infectious period, and monitoring protocols. During recent Ebola outbreaks, contact tracing apps developed in Guinea and Sierra Leone have provided a means for contact tracers to increase the speed and accuracy of contact tracing and efficient centralization of real-time data, as well as the coordination of resources and interventions [5,6]. These studies suggest that technology may play a key role in increasing the efficacy and timeliness of contact tracing. In contrast to the studies above, which placed technology in the hands of contact tracers, most strategies for COVID-19 contact tracing implementations use available smartphone technology to actively monitor risk for users in the general population.

A study by Ferretti et al [7] modeled the use of contact tracing apps and concluded that widespread use of these apps could be used alongside strategies such as widespread testing and physical distancing to suppress the pandemic successfully. This model suggests that contact tracing apps could allow greater freedom of movement, but low app adoption or incorrect usage could lead to continued spread. Similarly, models developed by Yasaka et al [8] determined that although some disease suppression could be seen with 25% adoption of their app, higher population uptake (75%) would be required for more substantial reductions of the infection curve. A primary concern, therefore, in determining the potential efficacy of a contact

tracing app is the number of people using the app and the efficiency and accuracy of its information distribution. Rowe [9] outlined three conditions necessary for the success of a contact tracing app: (1) correctness of information including diagnosis, (2) high likelihood of smartphone presence during contact, and (3) a high proportion of people using the app. As a result of the myriad technological changes that have occurred in recent years, utilizing digital technology to perform contact tracing is both a truly promising solution and an unprecedented problem. Therefore, extracting lessons from the current pandemic and the role technology can play will serve as crucial components to optimizing public health strategies during current circumstances and future outbreaks.

There has been a wide variety of COVID-19 contact tracing app reviews published over the course of the pandemic across a range of subjects, including technical analysis [10,11], privacy concerns [12-14], and ethics [15-18], each of which present valuable information on a specific aspect of digital contact tracing. This viewpoint is designed to combine information from across the expanding COVID-19 digital contact tracing literature and address key considerations that must be taken into account for developing and refining future contact tracing app design. This paper aims to accomplish this goal by providing background on common strategies for app-based contact tracing, discussing the advantages and limitations of several prominent COVID-19 contact tracing apps, and elucidating the privacy and nonprivacy concerns that have affected their adoption and reliability during the COVID-19 pandemic. Table 1 summarizes the key points of each section. This paper is intended as a hermeneutic literature review, with analysis that provides a specific viewpoint on the subject at hand. As such, a literature review was conducted over the course of July 2020 to March 2021 via PubMed and Google Scholar using terms including COVID-19, contact tracing app, and digital contact tracing, etc.



**Table 1.** Key considerations for the design of contact tracing apps, including those that relate to structure, technology, adoption requirements, and privacy.

App characteristic	Considerations
App structure	<p>Centralized</p> <ul style="list-style-type: none"> <li>Generated user data stored or managed in a central server</li> <li>Direct oversight of user data</li> </ul> <p>Decentralized</p> <ul style="list-style-type: none"> <li>Generated user data stored or managed on user devices</li> <li>User privacy and security advantages</li> </ul>
App technology	<p>GPS and location-based tracking</p> <ul style="list-style-type: none"> <li>Additional potential for privacy concerns</li> <li>Higher noise compared to Bluetooth</li> <li>May be paired with proximity-based tracing for increased accuracy</li> </ul> <p>Bluetooth and proximity-based tracing</p> <ul style="list-style-type: none"> <li>Fewer privacy concerns, particularly when paired with a decentralized structure</li> <li>High noise—signal attenuation and accuracy issues due to environmental signal absorption and reflection</li> </ul>
Adoption requirements	<p>Mandatory</p> <ul style="list-style-type: none"> <li>May be viewed as a privacy violation</li> <li>Higher adoption rates may increase the accuracy and efficacy of the app</li> </ul> <p>Opt-in</p> <ul style="list-style-type: none"> <li>Voluntary use with specific permissions to address privacy concerns</li> <li>May have lower adoption rates that decrease the accuracy and efficacy of the app</li> </ul>
Privacy	<ul style="list-style-type: none"> <li>An app should offer privacy from other users, the app manager, and snoopers</li> <li>The privacy-utility tradeoff of an app must be shaped around the local cultural attitudes</li> <li>Privacy and security have been repeatedly stated as primary concerns for app users: successful apps should adopt a privacy-by-design structure</li> </ul>
Other	<ul style="list-style-type: none"> <li>An app should be easy to use and reduce user fatigue</li> <li>Battery drainage, interference with other medical apps, and incompatibility with some phone models have proved to be barriers for successful global deployment of some contact tracing apps</li> <li>Users must be encouraged to follow all other precautions to limit the spread of disease (ie, recommendations from public health authorities like mask wearing and physical distancing)</li> </ul>

## Key Considerations in App Design and Categories

In this section, we outline some of the common considerations for developing contact tracing apps, namely strategies and technologies employed. We discuss the advantages and disadvantages of each of these strategies, particularly with regard to efficacy and frequency of use of contact tracing apps globally.

### App Parameterization Using Epidemiological Data or Disease and User Characteristics

When developing new apps for contact tracing, several factors must be accounted for within algorithms and interfaces to ensure accurate information and notifications for all parties involved. For instance, rates and types of transmission must be incorporated into the app design in order to determine the number and time frame of contacts that must be notified and tracked. The basic reproduction number ( $R_0$ ) accounts for baseline disease transmissibility without immunity from exposure or vaccination or any intervention to prevent transmission [19]. Transmission for COVID-19 falls into the categories of symptomatic, presymptomatic, asymptomatic, and

environmental transmission [7]. Contact tracing generally most accurately accounts for symptomatic and presymptomatic transmission, as asymptomatic and environmental transmission may not be readily identifiable. However, when contact tracing is paired with large-scale community testing, there is an enhanced ability to model transmissibility, accounting more accurately for asymptomatic and environmental transmission.

Incubation and infection times must also be taken into account to narrow time windows over which contacts should be identified for potential exposure. For COVID-19, the average time to symptom development is 5.1 days, with 99% of symptomatic cases displaying symptoms by 14 days [20]. In a study of infectiousness profiles for COVID-19 infector-infectee transmission pairs, the highest viral load was observed at symptom onset, translating to increased transmission risk [21]. However, 44% of secondary cases developed as a result of exposure during the index patient's presymptomatic stage, allowing COVID-19 to spread rapidly and highlighting a heightened risk when the perceived threat is low. Indeed, the disease's highest transmissibility has been reported to be before or immediately after symptom development [22]. Therefore, when modeling the timescale over which a patient may have

been infectious, contact tracing apps must account for possible transmission up to 2 weeks before symptom development or a positive test result for asymptomatic carriers. With the incorporated passive observation of individual location or contacts, this strategy may help to account and partially compensate for diagnosis delay, which was observed to lead to increased spread, particularly as communities developed and launched containment strategies early in the pandemic [23]. These parameters are essential in initial app development and determining individual risk but should also be considered when determining the amount of time necessary for data storage, particularly given the privacy concerns voiced over the long-term collection of location and health data [14].

Additionally, some thought should be given to the direction of contact tracing with the collected data. Forward tracing, used almost exclusively within current contact tracing apps, works forward from a current positive diagnosis to identify the contacts, particularly during peak infectiousness, who are now at risk of contracting the disease [24]. Backward tracing, however, works back from the current diagnosis to identify the secondary origin of transmission and find additional contacts who are now at risk of having the disease [24,25]. This method of tracing seeks to find the source of an outbreak and has been successful at identifying clusters around an individual known to be contagious. For instance, backward tracing was used to identify clusters and prevent further community spread of COVID-19 in both Japan and Singapore early in the pandemic [24,26]. Likewise, combining these techniques in bidirectional tracing has been shown to be particularly important with a long incubation period, with models showing a 2-fold reduction in effective reproduction number versus forward contact tracing alone [24]. Currently, most contact tracing apps exclusively use forward contact tracing, which fails to take advantage of the full range of data available. Digital tracing allows for easier extension of the tracing window in a bidirectional manner because there is no need to rely on patients' memories. However, if app usage is not high enough within a local population, digital tracing in either direction may be disrupted by network fragmentation and insufficient data [24]. Therefore, in addition to the epidemiological data necessary for app parameterization, due diligence must be given to increasing app usage numbers to increase the efficiency of a particular strategy, even those shown successful in manual tracing and modeling.

When developing app parameterization and settling on technological strategies and techniques for implementation such as those discussed below, app developers must ultimately choose strategies that result in the highest efficacy and accuracy. This strategizing must consider in particular app false positive and false negative rates for the app. False positives (type I error), referring to users incorrectly notified of increased exposure risk, can put undue strain on health care infrastructure by increasing demand for testing, result in increased levels of highly negatively impactful enforced quarantine, and decrease app utility by decreasing user attentiveness to app notifications [10,27]. False negatives (type II error), conversely, refer to individuals who have had close contact and are at high risk but are not identified by the contact tracing app, which may be the result of low app sensitivity or improper tracing [10]. These

high-risk individuals are ignorant of their risk and may contract the disease and spread it further in the community. The best policies and parameters based on this reasoning seek to minimize false negatives first, as these will result in further untracked community spread. These strategies, however, will likely result in a high number of false positives due to prioritization of high-sensitivity, low-specificity methods [27-29].

False positives may result through a variety of means, but initial planning to prevent false positives must begin with the technical strategy chosen and the policy-based design of initial parameters. The definition of close contact (6 feet or 2 meters, 15 minutes) most commonly put forward by public health entities has been argued to be too coarse for mass tracing, as evidenced by the high number of false positives seen with manual contact tracing and that this definition has resulted in decreased accuracy with digital contact tracing as well [27,28]. Likewise, as will be discussed further below, the technical strategies employed, such as GPS or Bluetooth, will directly impact the accuracy and efficacy of an app [11,30]. Signal strength and duration selections for the app, as well as firmware and software compatibility with the app and other users' devices, will play a role [27]. These considerations make it imperative that apps undergo significant real-world testing to determine efficacy, data from which has not yet been revealed for most available COVID-19 contact tracing apps.

### Centralized Versus Decentralized Architecture

The decision to utilize a centralized, decentralized, or hybrid overall structure or strategy is a key initial consideration when designing and implementing contact tracing apps and requires balancing privacy, security, and efficacy concerns. Centralized apps use strategies that employ a main server for data storage and analysis. Conversely, decentralized apps feature data storage that is distributed across the user network, with no individual entity having complete control or information access [31]. A hybrid architecture may have a component of both approaches, with some information handled on individual devices with a central server analyzing data and sending notifications. Contact tracing apps using each of these architectures have been employed for COVID-19, with the choice of structure highly dependent on government and cultural norms in the region of use and the needs of public health officials.

A centralized contact tracing app architecture may require significant trust in the beneficence of government investment and national or regional data infrastructure. In a centralized approach employing technology like Bluetooth LE, for example, a trusted third party (TTP) such as a government or public health entity may assign users an encrypted identifier that is broadcasted during app use [32]. These encrypted identifiers are broadcast to other users, with apps storing identifier lists that may be sent to the main TTP server in the event of a positive diagnosis. Then, users who appear on this list will receive notification of risk from the TTP through their app. Some notable examples of COVID-19 contact tracing apps with a centralized architecture include Singapore's TraceTogether and China's Health Code apps [33,34].

Conversely, decentralized architectures remove the role of data accumulation and analysis from a central server and instead



place these functionalities on user devices. Anonymous user identifiers are generated as random seeds with a short lifetime (chirps) that are exchanged with other user devices. Upon a positive diagnosis, a user may upload seeds and accompanying temporal data to a central server. By analyzing seeds and temporal data from positive users on the central server, exposure notifications are generated on user devices. In this way, the central server does not serve as the primary driver of risk assessment, nor, due to the anonymized or pseudoanonymized nature of seeds and chirps, can identifying details be derived from information available on the central server [10]. There has been a significant shift in interest toward decentralized or hybrid app architectures in recent years for privacy and security reasons, with a model focused on location privacy also indicating lower data retrieval times compared to centralized architectures [35]. Centralized approaches may be at risk of a variety of privacy and security attacks, including deanonymization of personal data through direct breaches of the main server. Likewise, decentralized app structures may also yield privacy and security concerns. For example, while Bluetooth LE and other technologies that have been applied in decentralized structures may record proximity-based data instead of positional data, personal medical information may be extrapolated via a linkage attack that uses data concerning specific, close contact notifications and known positional information from other devices to determine disease status, as explored by Bengio et al [12] in an excellent recent review of security concerns for decentralized app designs. Therefore, the decision to utilize a centralized versus decentralized app strategy centers on tradeoffs between culturally specific privacy concerns and structural trust needed for high population uptake as well as security concerns.

Additional security measures can and should be developed for both centralized and decentralized structures. Alongside encryption and anonymization techniques, blockchain technology has been suggested as a tool for increasing security and data privacy, particularly for decentralized strategies. Blockchain technologies seek to decentralize data by duplicating and distributing information across a global computer systems network, making it difficult to spoof or manipulate data within these distributed databases [36]. Xu et al [37] have developed a blockchain method called BeepTrace, designed as a bridge between users and central servers that reduces the vulnerability of sensitive data such as user identification and location, with the ability to predetermine and regulate the length of data storage. This proposed blockchain would also allow for public accountability of governments and corporations via transparency and ease of information verification, as in the case of manipulated efficacy data. However, in practical implementation, strategies would require refinement to reduce the intensive computational requirement for large populations, particularly dense populations such as in India and China.

## Location- and Proximity-Based Tracking

### GPS

GPS and global information systems (GIS) are routinely used for large-scale disease monitoring and predicting disease spread, as seen with features such as Google Flu Trends, which have

been used to predict the spread of seasonal flu in the United States [38]. Similarly, for COVID-19, online macroscale tracking systems were developed early in the pandemic to allow for up-to-date information on the global and regional spread of the virus. The timely updates of these online GIS and mapping dashboards, including the Johns Hopkins University Center for Systems Science and Engineering dashboard and Who Health Organization Dashboard, provided a means of data sharing concerning outbreak events for the public [39].

Alongside macroscale information, GPS and GIS technology may be employed for individual contact tracing via GPS and social media mapping. The ubiquitous use of GPS-enabled smartphones in many regions of the world provides an opportunity to collect spatiotemporal trajectory data for individuals. One example of this is the STRONG (Spatiotemporal Reporting Over Network and GPS) strategy, proposed and published by Wang et al [40], which analyzed the backend GPS spatiotemporal data collected through the social media app WeChat to trace the close contacts of users, primarily in China. This strategy offers a means of integrating cell phone-based GPS positioning with voluntary real-world transaction data that provides accurate timestamps and position information, although this also requires a substantial relinquishment of individual privacy. Similar to this strategy and adopted soon after the STRONG system was published, the Chinese government began employing a national monitoring system, which, as opposed to GPS data, uses cell phone base station positioning data to evaluate individual exposure and risk [40,41]. Other countries had also employed government-sponsored cell phone location data collection to track individuals and prevent COVID-19 spread, including South Korea, which notified users before entry into “high-risk” zones, and Israel, which used location data to inform contacts of confirmed infected individuals and for quarantine enforcement [42]. Additional apps developed by countries that involve location-based tracking include the Rilevatore Teramoto app released by Italy and apps from Austria (NOVID20), India (Aarogya Setu), Norway (Institute of Public Health app), and Spain (Open Coronavirus), which combine location and proximity-based tracing [32].

Therefore, GPS and location-based mapping have been widely viewed as a potential tool within the context of COVID-19 contact tracing apps. However, further refinement is needed to reduce “noise” within GPS-based systems, which may reduce the efficacy of the system to identify high-risk contacts or areas with sufficient specificity [40]. The granularity of location-based GPS data may not be sufficient to determine a distance of 6 feet (2 meters), with GPS positioning error at approximately 10 meters indoors [11]. This limitation may lead to significantly increased false-positive and false-negative rates, and hence reduce the accuracy and efficacy for GPS-based contact tracing strategies. Bluetooth proximity-based systems are generally considered to be more accurate, with fewer false positives reported [32]. Additionally, privacy concerns arise with the thought of highly specific spatiotemporal tracing, particularly when combined with timestamped real-world interactional data available through social media networks, which may generate problems with adoption. Security concerns must also be

considered, as GPS systems are particularly vulnerable to spoofing attacks, in which a fabricated GPS trail may send incorrect spatiotemporal data to a receiver [43].

One example of a privacy-forward, GIS-based app is the COVI Canada app, initially proposed for government use, although the Canadian government has moved in another direction [44]. The proposed app employs GeoIP services to yield coarse location data that maps risk-stratified zones through which users have moved. Additionally, Bluetooth-based contacts are recorded among individuals. COVI Canada claims a decentralized approach in which pseudonymized personal data such as the number of contacts and diagnosis status are encrypted and sent to the main server where risk is calculated for each user daily via artificial intelligence (AI). The coarse location data yielded by this approach may also provide heat maps of outbreaks to public health authorities.

### **Bluetooth**

In contrast to GPS mapping, Bluetooth records interactions between individuals based on device proximity. When individuals are in close proximity, Bluetooth-based token sharing allows for a precise record of the interaction. In this case, the limited range of Bluetooth-based technologies becomes advantageous for recording only close-range interactions to approximate a 6-foot distance. Additionally, the relative signals' strength can be used to determine the approximate distance between individuals, allowing for proximity tracing of individuals in high-risk settings, such as indoor environments or public transportation [32,45]. However, this strategy requires high adoption for accurate proximity tracking and risk assessment, as it requires a direct interaction between users. Bluetooth has an approximately 10-meter location granularity, with visibility between devices possible up to 30 meters apart [11]. Signal attenuation may be used to indirectly assess distances between devices, although this is not linear. Additionally, high standard deviation in received signal strength between 2 and 6 meters decreases accuracy within this range, and with 2 meters used to indicate close contact, it may lead to decreased app efficacy and increased false positives using Bluetooth technology.

Additionally, although Bluetooth-based tracing is generally considered more accurate than GPS-based strategies, significant issues are present with signal attenuation and accuracy. Signal absorption and reflection by the surrounding environment can lead to inaccuracies in reported distances between users [46,47]. In a recent paper on this topic, Leith and Farrell [48] evaluated the efficacy of Bluetooth LE technology for COVID-19 contact tracing in real-world environments, including users walking in the city, at a meeting table, in a train carriage, and grocery shopping, as well as assessing the impact of device orientation, use of a handbag, and type of indoor wall on signal attenuation. After observing significant impacts from all of these factors and no corresponding decrease in signal strength with increasing distance, the authors called for extensive real-world testing of Bluetooth-based COVID-19 contact tracing apps as well as data to inform the efficacy of such apps compared to manual contact tracing.

Bluetooth-based strategies have become common, with some notable examples, including national apps in several countries. Singapore's TraceTogether app was the first such Bluetooth-based, government-sponsored contact tracing app [33], and since then, the list has grown to include the Pan-European Privacy-Preserving Proximity Tracing (PEPP-PT) app, released by a European consortium, as well as Australia's COVIDSafe [49] and Canada's ABTraceTogether [44]. In contrast, the industry-based joint Apple-Google contact tracing app also relies on Bluetooth proximity tracing, but concerns have arisen around private companies' handling of sensitive health data and this new role for large tech corporations in the public health arena [50].

The PEPP-PT app was initially touted by its creators as having a 90% true-positive and 10% false-negative rate [51]. However, several studies since have indicated that Bluetooth-based proximity tracing may have significantly high error rates. For instance, in a study by Girolami et al [52] involving high school students, a traced interaction accuracy of 81% was obtained. However, in initial planning, 42% of student devices were found to be incompatible with active Bluetooth beaconing and unusable for contact tracing, which would pose a significant issue in the population at large. In a real-world setting, devices will have different versions of Bluetooth technologies and may receive, transmit, and damp signals at different levels. Likewise, at least 50% beacon loss between device dyads was observed, posited to be due to device positioning and unpredictable environmental signal attenuation as discussed above. This study, alongside real-world testing of the Bluetooth-based Google/Apple Exposure Network (GAEN) by Leif et al [48], points to somewhat unreliable and unpredictable efficacy and utility for Bluetooth-based apps and demonstrates the need for further innovation and design.

### **Contact Points**

One alternative to location-based GPS or Bluetooth tracking, which allows for more user privacy, uses an opt-in system of contact points. Yasaka et al [8] proposed the idea of an app that allows users to host or join checkpoints at which other users may also check in by scanning a generated location quick response (QR) code. Ideally, users would generate check-in QR codes for any gathering or public place, which poses a risk for COVID-19 transmission, and when a user tests positive, all users who have checked in at locations with them over the potential infectious period will receive an updated risk level. High user adoption would be necessary for such a system to be efficacious. Additionally, apps that rely on sustained conscious use pose a problem with user fatigue, with drop-offs in check-ins providing the potential for inaccurate risk assessments.

### **Volunteered Health Status**

An alternative to tracking or tracing apps is an app that allows for volunteered health or contact information to be entered to initiate case-based contact tracing outside of the app itself. This tactic of voluntary information entry mitigates some privacy concerns centered around tracking and tracing. The two main categories of apps falling into this category are symptom monitoring apps and case-initiated notification systems. These

apps do not directly perform contact tracing but instead serve to use technology to simplify the contact-tracing process.

Symptom monitoring apps rely upon users to accurately record a log of personal recent health records to assess for COVID-19 risk based on symptomology and may be used most effectively as a screening tool for identifying symptomatic cases. These apps require regular logging of recent symptoms associated with COVID-19, including fever, dry cough, shortness of breath, fatigue, muscle aches, and loss of taste or smell. Yamamoto et al [53] reported an example of this system integrated into an already established health screening app, K-note, which significantly decreased the follow-up burden on health care providers for monitoring close contacts of COVID-19-positive cases, allowing for efficient algorithmic identification of symptomatic users. Additional personal data that might be logged includes COVID-19-positive contacts, travel history, and visits to health care institutions, allowing for further risk stratification. Another such app includes the one implemented by Yap et al [54], COVID-19 Symptom Monitoring and Contact Tracking Record (CoV-SCR), which provides a designated space for users to track symptoms (rated 1-5) as well as keep a 14-day record of travel and close contacts whom the user may notify in the event of a positive diagnosis. Medical and academic institutions have already employed such efforts to keep a regular log of potential symptoms as a means of preventing spread, providing a way to do a first-pass screening of individuals and identify those who may need to be tested or isolated. Additionally, symptom monitoring may be paired with in-person screening, such as temperature checks upon entry to an institution. While many such apps are institution-specific, such symptom logging may be integrated into currently available personal health record apps.

Early identification of potentially COVID-19-positive individuals may enable efficient tracing of contacts and early isolation. However, such apps rely upon the memory and honesty of users to accurately portray health status. Additionally, this strategy fails to account for asymptomatic individuals, who may slip through the holes in this system and further spread the virus. Users may also be unable to account for strangers or individuals encountered in public settings such as public transportation. Passive tracing and tracking apps have therefore become the preferred avenue of exploration and implementation for apps intended for large-scale use, as discussed in the examples below, most of which seek to combine symptom monitoring with location- or proximity-based tracing.

### **Mandatory Versus Voluntary Use**

There are three potential strategies for enforcing such an app at a federal level: opt-in, opt-out, or mandatory use. An opt-in model allows people to download the app if they so choose and has been advocated for on the basis of consent for acquisition, use, and sharing of personal information [23]. An opt-out model would automatically provide the app for all, but users would have the option to delete the app if they preferred not to use it. The final model is a mandatory download of the app for all people without deleting it. Mandatory app usage may be enforced by preventing people from using public services or entering buildings if they do not have the app. China, Hong

Kong, Taiwan, and South Korea have used the mandatory model [55]. Many ethical concerns have been raised about such a design, including rights to personal autonomy and informed consent [16]. However, another concern is widening disparities between those who have access to relevant technology such as the internet and Bluetooth-capable smartphones and those who do not. In addition to an inability to move freely or engage meaningfully in public life, those who do not have smartphones or new operating systems do not have access to incentives to increase app use, such as those Parker et al [18] suggest like funds for charity donations and free mobile phone credits. Overall, the ability of a country to enforce a mandatory download system will depend on the socioeconomic status and cultural values of its citizens, with some nations more likely to experience significant pushback against or widespread inefficiencies in a mandatory system.

Furthermore, once the app is downloaded, behavior on the app may be mandatory/opt-in/opt-out. For example, once the app is downloaded, even if the app itself is optional, location sharing may be nonvoluntary and continuous. Other apps may allow for opt-in for location sharing. Although this may minimize the app's efficacy, this approach allows for greater user privacy if they are venturing somewhere that is private to them. Additionally, apps may require mandatory sharing of infection status, as is the case in Singapore.

### ***Specific App Examples of Successes and Limitations***

In this section, we outline some of the large-scale implementations of contact tracing apps, most notably those created by national governments. We will discuss the successes and downfalls of some of these implementations and touch on how local cultural attitudes influenced the design of each app and how it was received by the local population.

#### **Singapore**

Singapore's TraceTogether app was the first national deployment of a Bluetooth-based contact tracing system in the world. The app was presented as being "for the people," with the ultimate goal of protecting the population [33]. This Singaporean technology provides several lessons for contact tracing, including concerns about Bluetooth, the privacy-utility tradeoff, as well as centralized and decentralized systems. Most importantly, the Singaporean contact tracing strategy reveals how cultural attitudes and norms must be taken into account within a community in order to achieve maximal success.

Aimed at transparency and international cooperation, the Government Technology Agency of Singapore published information about BlueTrace, the protocol that underpins TraceTogether as well as OpenTrace, an open-source repository for other countries that heavily influenced Australia's national app design [56]. The app features a hybrid decentralized-centralized, proximity-based approach and functions through contacts exchanging non-personally identifiable messages, with frequently rotated identifiers for security and privacy. Encounter history is kept on local storage and thus is decentralized. While the app itself is not mandatory,



once a user is tested positive, they are legally required to release their stored data to the government per the Infectious Disease Act [55], after which health officials reach out to contacts based on personal identifiers. At this point, a user's data is stored on a centralized government server [57], and the government retains the right to share this information with other groups, including other governments [55,58]. It should be noted that these requirements would not be the same across countries and communities. The government generally advertised the TraceTogether app as being a privacy-by-design system because of the decentralized setup [59]. However, this is only true for healthy individuals who do not have to share their data to the centralized server for government control.

The BlueTrace authors also note other privacy-by-design implementation choices for the app. The TraceTogether app claims to keep proximity data only for 21 days, therefore limiting the amount of unnecessary data stored [56]. Furthermore, the BlueTrace report claims that users have control over their data, and all of their information is deleted upon request [56,59]. The TraceTogether system requires a health official to confirm that a case is legitimate to avoid false self-reported positives that may stir up panic and decrease the program's legitimacy and trust. However, the additional step of health official approval may be difficult to implement in other nations with high rates of uninsured citizens, who may not be able to seek medical care. This concern will inevitably affect some countries more than others, notably based on the presence or absence of universal health care.

Since the system requires interaction between two users, an increased percentage of people who download the app increases the program's effectiveness quadratically [56]. While Singapore was reaching relatively high usage as compared to other opt-in/opt-out programs, the 17% download rate was still insufficient to reach maximal efficiency [60]. However, the government hesitated to make the app mandatory to prevent pushback regarding surveillance and control concerns. There were some concerns that the Singaporean government was collecting cell phone data through the app, although the government denied these claims [61].

The app itself had several issues in terms of functionality. For example, certain types of phones, including Apple products or older models, could not download the app or experienced severely limited functionality. Further, there were lags between user contacts and device communication and logging. There are also inherent limitations to Bluetooth, including material barriers that attenuate transmission signals. Additionally, Bluetooth technology can interfere with other health-related apps and implantable devices [62]. High variance in transmission power across device types may also lead to difficulty in assigning appropriate thresholds to determine close contact distances relative to signal strength. Battery usage issues have also been reported [56].

To address some of these technical issues and privacy concerns, the government eventually pivoted to distributing a wearable device to all of its citizens, which would complement the TraceTogether app. These devices would not contain any information beyond contact history, thereby protecting mobile

cell phone data. Further, since the devices were identical, they would bypass some of the aforementioned technical issues across phone models and provide access to all. However, this implementation did not reduce concerns about location surveillance. In contrast to GPS, which tracks location, Bluetooth tracks interactions, which means that interactions between lawyers, doctors, and journalists are no longer confidential to the government [55]. Further, even with Bluetooth, it is still possible to narrow down one's location, especially as data accumulates [57].

Chua et al [61] notes that America may be unable to mimic Singapore's system because of cultural attitudes toward privacy-health tradeoffs. This is generally true across all distinct countries, which will each require individual policies that fit their population. The app developers themselves note that this system is specific to Singapore [58].

### Apple/Google

On April 10, 2020, Apple and Google announced a joint effort to create a contact tracing app framework that would serve to facilitate cross-platform monitoring as public health officials globally developed apps for their respective jurisdictions [59]. Phase 1 of this effort focused on releasing an application programming interface (API) for interoperability between iOS and Android systems, which was released to public health officials in May. This release was accompanied by signs that these corporations had reached out to large nations to create some patches between their initial health app designs and the GAEN API, notably Australia [59]. Following this, phase 2 released a means of building access to public health apps into the iOS and Android platforms through an Exposure Notification app provided with software updates. This system allows for Android and iOS users to opt in to an Apple and Google initiative that also ties in local contact tracing apps [59].

The API functions allow Android and iOS devices to exchange data with each other for improved contact tracing. Additionally, this framework features an opt-in, decentralized contact tracing system that relies on Bluetooth technology. The main emphasized advantage of the GAEN API was initially privacy, with no mechanism for recording users' location data. A new Temporary Exposure Key is generated once daily, with Rolling Proximity Identifiers (RPI) generated every 10 minutes. Beacons with RPIs are broadcast every 250 ms, while devices scan for beacons every 4 minutes. The signal strength of received beacons is used to determine distances between users, while the number of beacons may determine the duration of contact exchanged [45,50]. Close contact is generally defined as 15 minutes of contact at distances of less than 2 meters. The exact parameters that qualify a signal as a significant contact may be determined by public health officials, with variable attenuation thresholds chosen by the nation [45]. GAEN API-based apps have been utilized worldwide, particularly within Europe (eg, Germany, Switzerland, and Italy).

The value of Bluetooth-based apps is a precise record with contacts who may be excluded or difficult to trace with traditional contact tracing methods, such as strangers in public settings. However, GAEN API-based apps have recently come under scrutiny for issues with efficacy in these scenarios. Leith

and Farrell's [45] study sought to apply the GAEN API in a real-world public transportation setting, recording measurements on a commuter train between handsets at greater and less than 2 meters for 15 minutes. Using Swiss and German attenuation thresholds from apps released in late May and early June, no close contacts were recorded. The Italian attenuation threshold yielded 50% true positives with a 50% false-positive rate. While this study only used Android devices, it yielded data that suggests significant changes in noise levels based on individual use. Additionally, signal strength was demonstrated to have complex interactions with the environment in which measurements took place, including signal reflection by tram walls and absorption by bodies (at 2.4 Hz), with no clear trend in signal attenuation with respect to distance based on these complex effects. Although more such studies are required, this data suggests that the inherent flaws of Bluetooth-based contact tracing, with small changes such as models of devices and signal absorption and reflection having an outsized impact on outcomes, may significantly decrease the efficacy of the GAEN API. This limitation may not be unique to the GAEN API and is indeed of concern for all Bluetooth-based contact tracing apps.

Although this initiative was initially well received, in addition to issues of efficacy, several concerns surrounding the imposition of large tech corporations' influence in the spheres of public health and politics have arisen. These doubts have centered around a questionable past for both corporations regarding data management as well as the potential for mission creep for large tech companies [50]. The increasing inroads of tech corporations into biomedical fields, as in developing AI medical diagnostics, electronic medical records systems, and software kits for clinical trials, have some worried that in the rush to develop contact tracing apps, traditional medical expertise and professional knowledge has been traded for efficiency and optimization, which are the key values of technological production [50]. In addition to this, upon release of the GAEN API, Apple and Google have refused to work with nations and public health entities with apps in development that feature a centralized approach. This strategy effectively undermined government attempts at app creation in France and Latvia by refusing technical expertise [50]. Agencies have been forced to create workarounds for their apps that are unstable and battery-draining [59]. This lack of good-faith effort in these situations has been seen as evidence that public health and privacy concerns are currently being used as a means of increasing market share through contact tracing app development.

### Ireland

Ireland's COVID Tracker app has been highlighted as a successful adaptation of the GAEN API for national use. As such, the app functions as a decentralized, Bluetooth-based system aimed at preserving individual privacy [62,63]. One of the most significant accomplishments of the opt-in app was its relatively rapid adoption, with 37% of the population having downloaded it within a week of release [62,63]. Over the course of July and early August, 137 users had been notified of their potential exposure to COVID-19 [62]. The success of the Irish

app has led to its government working with other nations to retool the app for their use, including the United States [62].

Prior to launch, Irish health officials requested feedback from academic researchers and data and visualization experts as well as civil societies to evaluate the app, which has been noted as a potential source for public trust and perceived efficacy [64,65]. According to the pre-release report card issued by the Irish Council for Civil Liberties and Digital Rights Ireland, however, there were concerns regarding app privacy and security structure [64]. Most significantly, these groups noted the lack of efficacy data to back up claims of high accuracy, the need for timely deletion of personal data which could be extrapolated to yield user location, and concern over the use of closed-source Apple/Google software and control of health data by foreign private entities.

### China

The Chinese tech group Alibaba released their Alipay contact tracing app Health Code on February 9 in Hangzhou [33,66]. Soon after, Tencent released a similar system on WeChat. The QR code-based Health Code app has a broad audience in China, with the app used in over 300 Chinese cities with at least 900 million users as of August 2020 [33]. Through public-private partnerships and centralized government monitoring, Health Code has now become mandatory for access to public spaces in many areas. The app uses a color-based QR code system that reveals user risk—green, yellow, or red. Those with a green QR code may visit public spaces and others, while those with a yellow or red QR code are subject to self-isolation for 7 to 14 days [33,67]. App data pairs actively collected data such as self-reported symptoms, address, and government ID with passively collected GPS location data, online transaction data, and surveillance via facial recognition technology, CCTV (closed-circuit television), and drones, all of which are monitored and synthesized by a central government server to determine exposure risk and generate the QR codes [66-68]. The users' status is updated daily at midnight to account for the previous day's activities [33].

The mandatory acceptance of this app has been cited for enabling greater effectiveness versus voluntary acceptance for reducing the spread of the virus. Additionally, the app's widespread use has demonstrated a much stronger tracing record versus South Korea, which only requires diagnosed individuals or close contacts to download their app [67]. However, the mandatory, centralized structure of this app may be responsible for considerable user stress. In one study by Joo and Shin [67], users reported that issues with app inaccuracy and errors were less stressful overall than privacy concerns. Because public spaces are restricted, even without a mandate, a de facto mandatory environment would exist for users who want to participate in public life, such as going to work or seeing family members, and could exclude those without access to appropriate technology or reliable internet access, highlighting the complexities of regulating movement based on exposure risk [66].

Likewise, significant concerns have been raised regarding erroneously issued yellow and red QR codes due to incorrect data entry or technical errors, which may necessitate

unwarranted self-isolation [67]. Conversely, false-negative green codes have been reported in Wuhan, with COVID-19–positive individuals given a pass to public areas [33]. Users have reported difficulties in having an incorrect code corrected within the centralized system, adding to technological stress associated with the app [67]. Inconsistencies at the local level have added confusion to this process as users move between regions, with a yellow code requiring 7 days of quarantine in Hangzhou versus 14 days in Shandong [33]. These issues have been paired with a lack of transparency on the part of the government regarding the app's operation [67].

Additional concerns include a reliance on private tech corporations to provide location data, echoing concerns of public-private partnerships that ring similar to those voiced about the Apple/Google GAEN API [67]. Alarms have been raised regarding the potential for Alipay and WeChat to share information with police agencies [33]. Likewise, this public health role provides large tech companies with unprecedented access to individual health information, which concerns some users due to the overall lack of transparency in this process. It remains unclear how exactly public and private entities manage data collected through the Health Code, who owns this data, and how the government regulates the Health Code [33].

### South Korea

The South Korean contact tracing app, Self-quarantine Safety Protection App, was mandatory for all citizens, and data was stored in a centralized database. Citizens could then view the database and see people's whereabouts to determine if they were at risk [68]. One critique of this system was that users had to read through a lot of information daily, such that users stopped actively checking the updated information. Another app, Corona 100 m, created a more streamlined service to identify infection hotspots to avoid information fatigue [69]. However, one of the most common critiques of the South Korean system was the lack of privacy, security, and protection from fellow users/laypeople and protection of the business. Protection from identification is essential to avoid stigmatization of any individual or businesses, especially in the worldly context of economic strain due to the pandemic [70].

Introna and Poulodi define privacy as the protection from judgment from others, a highly relevant concept when considering the case of privacy breaches from contact tracing apps in South Korea [70-72]. In South Korea, businesses were threatened with false reports on site. Further, reporting of cases in or around the region of businesses resulted in economic strain, boycotting, and riots. This relates to Rowe's [9] first condition for a contact tracing system's success: correctness of the information on the app. This first point has two factors to consider: all diagnostic tests have some rate of false negatives and false positives, which may be reduced but not eliminated. Another factor is self- versus physician-reporting of positive cases. If users self-report, security issues arise, including the possibility of false reporting to stir anxiety and fear or even attack a specific location or business, as seen in South Korea. The latter point touches on the issue of privacy for businesses: by sharing the specific location of the infection, businesses around that area are now at risk of financial strain or even

boycott. However, if physicians must report positive cases, the app would acquire less information, with those unable to go into a doctor's office or hospital unaccounted for. Further, whether a business is identified as a hotspot for infection due to blackmail, or it simply was the location of a positive contact point, the business may still be placed at an economic disadvantage. That information about the business is now being shared, potentially without their knowledge or consent, and they may be judged as a result. Methods of contact tracing which use GPS, QR codes, or any location-specific identifiers that are freely shared among users put businesses at risk.

However, businesses were not the only parties affected by South Korea's system. Personal information about individuals was discovered or speculated, leading to the stigmatization of those individuals. This result was especially damaging due to the fact that participation in South Korea's system was mandatory, rather than opt-in or opt-out. Therefore, those infected had no means of privacy from widespread stigmatization from their peers. Even without specific name sharing, individuals experienced online attacks, and events involving collective action against individuals online have been reported [73]. After an outbreak in an area associated with gay clubs, many individuals did not get tested or quarantined out of fear that they would out themselves and be judged by their community. It has been suggested that lottery-style randomized testing may protect against issues such as this [74].

### Norway

Norway's Smittestopp app provides a key example of how rushed development of sensitive technologies can detrimentally backfire. The Smittestopp app used both GPS and Bluetooth, which was stored centrally on a government-controlled cloud platform [75]. The app collected data including mobile phone numbers, age, location data, and contact with infected individuals [75]. The app was said to collect anonymized movement data and would notify users of potential infected close contacts [76]. The app was not open source, which prevented community-based auditing [75]. The hasty development of the app also meant that it had several issues and was not particularly user-friendly, which discouraged its use [75]. Only 1.5 million people (out of 5.3 million) downloaded the app, which was not enough people for useful contact tracing [76]. As such, Norwegian Data Protection Authority deemed Smittestopp illegal because it collected too much personal information without providing a clear data usage policy to its users. The government was advised to shut down the app [76].

Norway's Smittestopp app was forced to undergo significant restructuring in mid-June by the government since the number of cases could not justify surveillance of the people [76]. Norway's rushed development of an app resulted in a less-than-optimal program, with significant security risks. Such an example highlights the need to be careful, thoughtful, and deliberate in the creation of such an app, especially when it comes to privacy concerns. Without doing so, privacy breaches may lead to a public that lacks trust in the app. After that point, it may be difficult, if not impossible, to acquire trust again. It is essential that as governments and organizations move forward,



these apps are prioritized privacy and security prior to distribution to avoid wasted effort, resources, and time.

Despite its limitations and flaws, the app was at first successfully deployed because Norwegian culture places high trust in the government and novel technologies. What may be learned here is that full trust in novel technologies can lead to security and

privacy breaches. However, a full distrust in novel technologies prevents the implementation of innovations that could be beneficial to local and global communities. Therefore, all of these digital contact tracing apps together (summarized in Table 2) show that a healthy skepticism level, as well as standardized and timely auditing, will improve future digital contact tracing technologies

**Table 2.** Summary of key implementation decisions of each described contact tracing app.

App name	Country/company of origin	Centralized/ decentralized/hybrid	GPS/ Bluetooth/ other	Google/ Apple API <sup>a</sup> ?	Mandatory/ opt-in/ opt-out/ other
TraceTogether [33,55,56,58]	Singapore	Hybrid, centralized reporting with a confirmed case	Bluetooth	No	Opt-in
Google/Apple Exposure Notification API [45,50,59]	Apple/Google	Decentralized	Bluetooth	Yes	Opt-in
COVID Tracker [62-64]	Ireland	Decentralized	Bluetooth	Yes	Opt-in
Health Code [33,66-68]	China, Alipay/WeChat	Centralized	GPS, symptom tracking; QR <sup>b</sup> code	No	Mandatory
Self-quarantine Safety Protection App [69-74]	South Korea	Centralized	GPS	No	Mandatory
Smittestopp [75,76]	Norway	Decentralized	Bluetooth	Yes	Opt-in

<sup>a</sup>API: application programming interface.

<sup>b</sup>QR: quick response.

## Primary Limitations and Concerns in App Design

### Privacy and Security

#### Privacy and Security Limitations and Concerns

A high proportion of the local and national population must use a contact tracing app for the app to be successful [9]. This notion is intimately tied with privacy, security, and autonomy. Public perception and acceptance play a key role in app downloads and usage for voluntary opt-in and opt-out systems. However, notions of privacy differ from country to country, and local perceptions will influence which implementation decisions are acceptable. All contact tracing will require some privacy loss, so each community must determine their optimal privacy-utility tradeoff. When public health and wellness are at odds with privacy, this tradeoff may be different than in times of stability. Several surveys have been conducted worldwide to understand opinions about digital contact tracing apps, and the primary cited reasons for not using apps were privacy, security, and surveillance. Given the large concern from users, privacy, security, surveillance, and transparency must be at the forefront of future and current contract tracing app designs.

In Ireland, of those surveyed who reported that they would not download the app, the most common reason was privacy [77]. They found concerns that the app host would use personal data for surveillance purposes, rather than for public health, and that surveillance would continue after the pandemic. Similarly, a survey study in Jordan found that 71.6% agree with the use of contact tracing apps, but only 37.8% used such technology. The

main concerns among survey participants were privacy, voluntary status, and beneficence of the data [78].

In a recent large-scale, multicountry study with 5995 participants, the surveyors found general support for contact tracing apps [79]. However, they note that participants from the United States were generally less supportive of the app than other respondents. They found that this generally corresponded with a lack of trust in the national government. Similar to the Ireland study, the main reasons given for not downloading the app were concerns related to government surveillance (42%) and cybersecurity (35%). Interestingly, they found that 74.8% would definitely or probably download an opt-in app, but only 67.7% would probably or definitely keep an opt-out app. Therefore, concerns of privacy, security, and autonomy are primary concerns for users, especially in the United States.

In a Johns Hopkins study of US citizens, 82% reported that they would use a “perfectly accurate and private” tracing app, but only 24% to 26% would want one with even “a low chance” of a data leak to the government, an employer, a tech company, or a nonprofit organization [80]. Given the benchmark that 80% of smartphone users should download a contact tracing app for it to be most successful, these results highlight the essential nature of a deliberately chosen system, which is private and full transparency by entities with an active role in its execution. Without a promise of privacy, it is far less likely that a contact tracing app will be adopted to the degree necessary for efficacy, making it significantly less useful in public health efforts.

Given that even a low chance of data leak is enough to dissuade over 50% of US users, these apps must be reliable and able to maintain a positive reputation. Even a single privacy leak event or controversy will result in a lack of credibility for the app.



After such an event, the app would unlikely be viewed as “perfectly accurate and private,” preventing it from reaching the >80% margin necessary for maximal efficacy. Therefore, privacy concerns must be at the forefront of contact tracing app design. Contact tracing apps, especially in the United States, should protect users (both individuals and businesses) from fellow users, the authority hosting the app, and snoopers or hackers.

### ***Privacy and Security Recommendations for App Development***

In the time of a pandemic, certain rights may be loosened for public health and safety, for example, requirements such as mandated quarantines and mask wearing [81-84]. While disagreement has arisen with these measures, there is a general sense of compromise for the sake of public health [13,85]. When it comes to contact tracing mobile phone apps, a central question is what privacy-utility tradeoff fits within a local community’s values. Will people allow personal location information and contacts to be shared so that the pandemic could be slowed, lives could be saved, and quarantine restrictions could be relaxed sooner? If so, what are the limits of loss of privacy? Inherently, for these apps to function, some degree of location information must be shared. There is an ethical balance between reducing the havoc of the pandemic and saving lives and risking some loss of privacy [68]. Rowe [9] hypothesized that individuals might be comfortable with the risks of location sharing if that meant increased health and financial protection for themselves and their families. Ghose et al [13] found that Americans increased their rates of location sharing services during the pandemic, suggesting that people were comfortable reducing some levels of personal privacy for the sake of public health. Attitudes in South Korea show that the population is comfortable with privacy leakages for the sake of public health [68], whereas reports in American and many European countries favor privacy over public health interventions [79,80].

Nevertheless, in the design of these apps, the likelihood of an information leak should be minimized and not purposefully done. While individuals may sacrifice privacy for the public, unnecessary risks must be minimized. This requires purposeful security design, transparency about data use that should be exclusively for public health purposes, and a promise of nondiscrimination both at the hand of the app host and fellow users. To ensure these requirements, there must be a balance between minimizing information sharing while maximizing the usefulness of the app itself [18]. Yannakourou et al [86] outlined several guidelines. For example, the minimum amount of individual information must be gathered that still allows for efficacious operation, meaning the removal of outdated data that is older than the incubation period of approximately 2 weeks. Unnecessary additional information is deemed unethical since it serves no purpose for public health but is instead collected for the sake of data collection itself. Singapore’s reported 21-day limit follows this guideline. Furthermore, the app should cease use after its benefits are no longer needed (eg, at the end of the pandemic). If these practices were implemented and advertised, this would also further garner trust and, therefore, increase the population’s usage.

While these concerns relate to privacy and transparency from the central host of the app or from fellow users, the app should also be safe from snoopers. For instance, only 16 of 50 apps reviewed prioritized data encryption and security via data anonymization and aggregate online reporting in one study [87]. This concern reinforces the importance of not rushing to deployment but rather carefully confirming the protection of the app from any type of information leakage. Open-source technology and community-based auditing have been suggested as one means of achieving thorough security [80]. Such transparency encourages trust, allows third parties to confirm the intentions of the app host, and allows the technology to be further improved, including cryptographic methods. The Massachusetts Institute of Technology (MIT) Private Kit: Safe Paths has been commended as an open-source, secure, and decentralized program [88]. The MIT implementation was designed with privacy as a priority, and their transparent, open-source model prevents abuse on the part of the app host. Many others have called for increased levels of transparency, whether that be clear and understandable statements regarding the use of data collection or open-source code available for audit [66].

Since privacy-related concerns were the main cited deterrent to app usage, privacy-by-design implementations are necessary for a successful contact tracing app. An app must be thoroughly vetted and trusted for its ability to maintain its users’ privacy and security. An opt-in system may be best in that case, especially in the United States, as trust in government is lower than in other nations [66]. Regardless, this may require trust in the government or in the authority hosting the app through transparency and decentralization. Open-source implementations allow for outside auditing and clear understanding between the developer and the user to increase transparency and accountability. This is a large concern from those who do not wish to download such apps and feel that the government may be using their data for reasons outside of public health and slowing the virus. If the program is trustworthy by open-source practices and transparency and follows proper ethical guidelines, the app may garner more users.

Several other important ethical aspects must be considered when designing such apps, as has been emphasized by several reviews [15,16,18]. It has also been emphasized that mandatory downloading is not an appropriate solution, as not everyone has access to smartphone technology. Similarly, there are concerns about how contact tracing can be used to free people from quarantine in an equitable and fair manner, given that not everyone has access to this technology [15]. Others have brought up concerns about the use of such apps for surveillance and policing of marginalized populations [89]. There are also concerns that centralized servers may risk individual privacy from state surveillance and third-party data breaches. However, centralized information allows for future epidemiology research or public health service planning.

In sum, there are many ethical concerns surrounding contract tracing apps, many of which pertain to privacy, security, and surveillance. Given that a large number of people must actively use contact tracing apps for them to work, opt-in implementations must be trusted by their user base. There is no

perfect formula for success in implementation, in large part because each community has its own values and will require a unique approach to the privacy-utility tradeoff. Local laws may also place constraints on what is possible. Surveys of local populations can be used as a springboard to app design. App designers must consider local political and cultural attitudes toward technology, privacy, and public health. Politics can determine regulations on app data collection, and so political norms may dictate the limitations of app functionality. Further, cultural attitudes are likely to influence whether local people are willing to download and use the app at all. Of course, political and cultural attitudes can influence one another. Politics are often a reflection of general cultural expectations. We have highlighted several key contact tracing apps designed specifically based on local political and cultural attitudes, such as the apps in Singapore, South Korea, and Norway. These app implementations may not have been successful or possible at all in countries other than their origin. As such, it is of utmost importance to have a good understanding of local attitudes and needs to ensure that the app is most effective. Future contact tracing apps in the United States will require high levels of privacy protection; open-source and transparent design; and a decentralized, opt-in system.

### **Nonprivacy Concerns Related to Apps**

As previously discussed, high and consistent long-term usage is essential for contact tracing app efficacy. Although privacy is a key element limiting the extent of usage for this current generation of COVID-19 apps, several nonprivacy factors remain at play in driving users against consistent long-term usage.

### ***App Efficacy and Perception Issues***

Issues with app performance have resulted in significantly decreased usage. One of the most impactful examples is excessive battery usage due to contact tracing apps. In addition to privacy, Singaporeans have cited depletion of mobile phone battery life as one of the primary reasons against the TraceTogether app [17]. Thus, even in a nation with high levels of digital inclusion and public governance, less than a quarter of Singaporeans downloaded the app initially [55].

Additionally, in response to a software bug in the exposure notifications system developed by Google and Apple [90] caused by the recent update to the GAEN API [91], 83,000 users of Ireland's COVID Tracker app (out of 1.5 million users) deleted it from their mobile devices in an attempt to solve the underlying technical issue [92]. As a result of the Health Service Executive working with Google and NearForm (the Ireland-based company that designed the app) to solve this Android-specific problem, 10,000 users have already reinstalled the app. However, in a recent October 2020 survey, over half of respondents stated that their app's Bluetooth technology adversely affects device battery life [78]. These examples highlight the difficulty in regaining users once challenges are encountered and users are lost. Therefore, it is essential that extensive prerelease testing is performed to determine any underlying issues that could affect widespread usage.

While this technical debacle resulted in a loss of 73,000 app users, rapid responses to such issues may mitigate such nonprivacy factor effects on consistent long-term usage. Other than the bug affecting users of the COVID Tracker in Ireland, there have been no other significant reports of battery draining and overheating issues caused by a COVID-19 app. For example, nations such as Canada have seemingly experienced no battery drainage issues, even though they have also utilized the Apple-Google framework [93]. However, even if there were to be future issues, several organizations have resorted to building their own "localized" platforms with no dependence on the GAEN API [94]. For instance, Virginia's state government reports that they will forgo using the GAEN API for their system, but this may result in the need for significant workarounds for interoperating system communication that results in subpar performance [95]. The lack of transparency and comparability in app design may result in a reduced ability to resolve these issues and decrease app trust in the public.

Another possible factor in limiting the broad usage of COVID-19 apps is a lack of belief in such apps' effectiveness. One element of this distrust is evident in a survey by the Italian polling organization SWG [96]. In a survey asking 800 individuals, "Why did you choose not to download the Immuni app?," the most popular reason was "I do not consider it effective" (44%), after which was "I'm afraid for my privacy" (29%).

Finally, a significant factor against the broad usage of contact apps is an emerging sense of complacency. In a report by the Centers for Disease Control and Prevention, 41% of respondents have faced mental health challenges related to COVID-19 and steps taken to combat the pandemic, including social distancing and stay-at-home orders [97]. The literature has also revealed that stressors such as longer quarantine duration, fears of infection, frustration, boredom, inadequate supplies, inadequate information, financial loss, and stigma have contributed to negative psychological effects, including posttraumatic stress symptoms, confusion, and anger [98]. Downloading an app represents another self-sacrifice citizens have to make alongside activities like social distancing and mask wearing that are already mandated or strongly recommended. The long-term nature of the COVID-19 pandemic may lead to complacency when given the opportunity to make another COVID-19-related personal decision.

The literature has revealed that a decrease in concern, in addition to low political trust, can combine to undermine compliance with governmental restrictions during the pandemic [99]. This complacency can significantly adversely affect the adoption of contact tracing apps, even if the ruling government recommended them.

### ***App Efficacy and Perception Recommendations***

While issues like complacency may stem from the external issues dealing with pandemic fatigue, there are potential solutions for app efficacy and public perception. Bluetooth LE technology, the currently preferred technology for decentralized apps aimed at increasing user privacy, must be demonstrated to consistently and correctly measure user proximity. As of now, Bluetooth LE measurements are subject to discrepancies

in signal attenuation based on the specific device used, the relative orientation of devices, and signal absorption and reflection by bodies, handbags, walls, etc [45]. Additional research must be conducted to increase the accuracy and efficacy of the employed technologies if contact tracing apps are to play an impactful role in infectious disease control moving forward. Toward this end, the National Institute of Standards and Technology, along with the MIT PACT (Private Automated Contact Tracing) project, has issued its “too close for too long” challenge to engage with research organizations worldwide concerning noise reduction and more precise distance and temporal estimation of Bluetooth LE signals [100].

Efforts to reduce unwanted false positives and false negatives have evolved in several directions. The Hamagen app used by Israel, for instance, allows users to see the location and time of potential exposure and indicate if they were not present [10]. Alternatively, data from contact tracing apps could be compared with manual tracing efforts to confirm or refute false positives and negatives, and health care providers could be called upon to report and authenticate data to reduce false negatives [10]. Additionally, big data could be used to filter out false positives in a centralized system.

For proximity-based strategies as well as location-based techniques, simulations and modeling may be used to guide design in terms of efficacy and community tailoring. For instance, Pandl et al [101] recently developed and implemented a spatial proximity simulation, which looked at the effects of both proximity detection range (0.2-10 meters) versus contact tracing app adoption (20%-100%), including simulations of decreased use with increase false-positive rates (25%-100% false positives). At higher adoption rates, longer proximity detection ranges (2 meters, 10 meters) were most effective at reducing disease spread but also resulted in increased false positives, which triggered decreased app utilization in highly reactive scenarios. The authors emphasized that this indicates the need to tailor app parameterization and strategies to match cultural expectations, indicating that less sensitive methods involving Bluetooth, GPS, and QR codes would be more acceptable in countries comfortable with a trade-off of high false positives for high efficacy.

In a recent review discussing digital technologies for contact tracing apps, Trivedi and Vasisht [11] discussed research to improve upon existing Bluetooth LE signals such as time-of-flight measurements using Bluetooth that allowed for accuracy to the foot and hybrid techniques such as those employing Bluetooth LE and acoustic-ranging in conjunction, although these have not been widely validated and are not ubiquitously deployable with today’s smartphone technology. Alternatively, merging technical strategies may result in increased overall specificity, particularly for popular Bluetooth-based approaches. A recent paper from Nguyen et al [46] explored the idea of a multi-smartphone-sensor system for contact tracing using Bluetooth combined with barometer (effected by altitude and winds for indoor or outdoor environments), magnetometer (dynamic time warping used to interpret magnetic field vectors), microphone (high frequency, low amplitude short chirps emitted and time of flight measured), and WiFi data (reliant on a grid of WiFi hotspots). With added

distance-based readings (microphone, WiFi), accuracy was increased from 25% to 65%, and the additional inclusion of environmental readings (barometer, magnetometer) further increased accuracy to 87%. It should be noted, however, that the significant reduction in false positives was accompanied by a small increase in false negatives. However, these results indicate potential for increased efficacy by the synergistic leverage of multiple sensors for proximity-based digital contact tracing.

Finally, in order to increase public trust in contact tracing app efficacy, apps must undergo extensive testing prior to public release. Moreover, this testing must include rigorous evaluations of app efficacy in real-world environments, particularly those in which users are most likely to have encounters that are difficult to trace, such as public spaces. For example, Leith et al [48] have done testing of the GAEN framework in settings such as public transportation. The efficacy of these apps is not as simple as ensuring customer satisfaction, although this is important for maintaining consistent use, but has become a question of public health as such apps are embraced by national governments worldwide. As such, Bhatia et al [102] suggested that mobile health (mHealth) tools such as contact tracing apps may require regulatory intervention and need the introduction of regulatory sandboxes for extensive beta testing among diverse populations in diverse environments [102]. This strategy would ensure that efficacy data were available and rigorous before the release of a public health app in a time such as our current pandemic, which could both increase faith in the intervention and ensure that it is an appropriate allocation of resources. Appropriate and expedient evaluation of digital contact tracing apps will help determine if the benefit of this technology is worth the potential privacy and security risks previously discussed. Colizza et al [103] have called for such evaluation, and they suggest the use of surveys, epidemiological analysis, and experimental studies. At the time of writing, such analyses were minimal. With the lessons gathered from the use of contact tracing apps for COVID-19, better technology should be available for urgent and efficacious deployment in the event of another infectious disease outbreak. Because contact tracing apps call for some surrender of private data, it is necessary first to ensure that the technology used will be effective in the environments where it is most needed.

To our knowledge, the only evaluation of a digital contact tracing app was reported by Salathé et al [104] on the SwissCovid app in Switzerland, which uses the exposure network framework. They demonstrated a proof of principle that the app reached appropriate contacts who later tested positively for SARS-CoV-2. They provided evidence that notified users subsequently sought SARS-COV-2 testing. This suggests that the SwissCovid appropriately notifies people to self-isolate if they are at risk, which can limit transmission as a result. Nevertheless, there has been model-based analysis on contact tracing in general, which has found that contact tracing can reduce the effective reproduction number of a virus if the tracing is done with minimal delay, the tracing is accurate, and those informed of their potential risk follow appropriate isolation protocols [105]. This research suggests that mobile app technology can reduce the tracing delay, and thus, digital contact



tracing apps may minimize viral transmission. Other models have suggested that smartphone-based contact tracing apps can become particularly effective after the first wave of an outbreak. The limitations to effectiveness, as discussed, include the rate of download and use, as well as the accuracy of the app [30]. Still, other than the analysis of the SwissCovid and these theoretical models, evaluations of specific contact tracing apps were not readily available at the time of writing.

## Conclusion

In conclusion, there were many lessons learned from contact tracing apps that were designed to slow the spread of COVID-19. Many of these lessons relate to one core theme: for such an app to work, it is absolutely required that the app be used by a significant portion of the population. For this requirement to be satisfied, the app must offer three main capabilities to its user: (1) a level of trust in the efficacy of the app itself, which requires proper functionality and testing; (2) a level of security and privacy from the app host, fellow users, and malicious entities; and (3) minimal cost or effort from the user, whether that be in the form of battery usage, manual and

frequent use of the app, or a host of other factors. Some of these requirements are obligatorily balanced. Many of the apps that proved unsuccessful so far did not sufficiently meet requirement 1 and potentially requirement 2. Therefore, many of these apps have been unable to gain and retain the number of users needed for accurate contact tracing, rendering them ineffective. It was proven repeatedly that these apps must be properly vetted and tested in multiple aspects prior to deployment, as technical performance issues have led to decreased public trust and usage. Likewise, any legal and ethical considerations concerning privacy must be adequately addressed before releasing an app for public use. To avoid the rushed time constraints from an emergency setting, proper development of such apps will need to happen prior to the emergency. In other words, similar to many other issues, we must be proactive rather than reactive when it comes to the use of contact tracing apps moving forward. Likewise, transparency from all actors involved in the development and management of contact tracing apps is necessary. The use of contact tracing apps during the COVID-19 pandemic will improve contact tracing apps in general by providing these real-world lessons.

## Acknowledgments

This work was conducted in conjunction with the Technology Committee of the American Physician Scientists Association. KH acknowledges the Baylor College of Medicine Medical Scientist Training Program and support from the National Institute of Dental and Craniofacial Research (F31 DE030333). XJ is a CPRIT Scholar in Cancer Research (RR180012), and he was supported in part by the Christopher Sarofim Family Professorship, UT Stars Award, UTHealth Startup, the National Institutes of Health (award number R01AG066749, R01AG066749-01S1, R41HG010978-01S1, and U01TR002062), and the National Science Foundation RAPID #2027790.

## Conflicts of Interest

None declared.

## References

1. Wu Y, Chen C, Chan Y. The outbreak of COVID-19. *Journal of the Chinese Medical Association* 2020;83(3):217-220. [doi: [10.1097/jcma.0000000000000270](https://doi.org/10.1097/jcma.0000000000000270)]
2. Rossen LM, Branum AM, Ahmad FB, Sutton P, Anderson RN. Excess Deaths Associated with COVID-19, by Age and Race and Ethnicity - United States, January 26-October 3, 2020. *MMWR Morb Mortal Wkly Rep* 2020 Oct 23;69(42):1522-1527 [FREE Full text] [doi: [10.15585/mmwr.mm6942e2](https://doi.org/10.15585/mmwr.mm6942e2)] [Medline: [33090978](https://pubmed.ncbi.nlm.nih.gov/33090978/)]
3. Ioannidis JPA. Global perspective of COVID-19 epidemiology for a full-cycle pandemic. *Eur J Clin Invest* 2020 Dec 25;50(12):e13423 [FREE Full text] [doi: [10.1111/eci.13423](https://doi.org/10.1111/eci.13423)] [Medline: [33026101](https://pubmed.ncbi.nlm.nih.gov/33026101/)]
4. Browne C, Gulbudak H, Webb G. Modeling contact tracing in outbreaks with application to Ebola. *J Theor Biol* 2015 Nov 07;384:33-49. [doi: [10.1016/j.jtbi.2015.08.004](https://doi.org/10.1016/j.jtbi.2015.08.004)] [Medline: [26297316](https://pubmed.ncbi.nlm.nih.gov/26297316/)]
5. Sacks JA, Zehe E, Redick C, Bah A, Cowger K, Camara M, et al. Introduction of Mobile Health Tools to Support Ebola Surveillance and Contact Tracing in Guinea. *Glob Health Sci Pract* 2015 Nov 12;3(4):646-659. [doi: [10.9745/ghsp-d-15-00207](https://doi.org/10.9745/ghsp-d-15-00207)]
6. Danquah LO, Hasham N, MacFarlane M, Conteh FE, Momoh F, Tedesco AA, et al. Use of a mobile application for Ebola contact tracing and monitoring in northern Sierra Leone: a proof-of-concept study. *BMC Infect Dis* 2019 Sep 18;19(1):810 [FREE Full text] [doi: [10.1186/s12879-019-4354-z](https://doi.org/10.1186/s12879-019-4354-z)] [Medline: [31533659](https://pubmed.ncbi.nlm.nih.gov/31533659/)]
7. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L, et al. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* 2020 May 08;368(6491):eabb6936 [FREE Full text] [doi: [10.1126/science.abb6936](https://doi.org/10.1126/science.abb6936)] [Medline: [32234805](https://pubmed.ncbi.nlm.nih.gov/32234805/)]
8. Yasaka TM, Lehigh BM, Sahyouni R. Peer-to-Peer Contact Tracing: Development of a Privacy-Preserving Smartphone App. *JMIR Mhealth Uhealth* 2020 Apr 07;8(4):e18936 [FREE Full text] [doi: [10.2196/18936](https://doi.org/10.2196/18936)] [Medline: [32240973](https://pubmed.ncbi.nlm.nih.gov/32240973/)]
9. Rowe F. Contact tracing apps and values dilemmas: A privacy paradox in a neo-liberal world. *Int J Inf Manage* 2020 Dec;55:102178 [FREE Full text] [doi: [10.1016/j.ijinfomgt.2020.102178](https://doi.org/10.1016/j.ijinfomgt.2020.102178)] [Medline: [32836636](https://pubmed.ncbi.nlm.nih.gov/32836636/)]

10. Ahmed N, Michelin RA, Xue W, Ruj S, Malaney R, Kanhere SS, et al. A Survey of COVID-19 Contact Tracing Apps. *IEEE Access* 2020;8:134577-134601. [doi: [10.1109/access.2020.3010226](https://doi.org/10.1109/access.2020.3010226)]
11. Trivedi A, Vasisht D. Digital contact tracing. *SIGCOMM Comput Commun Rev* 2020 Oct 26;50(4):75-81. [doi: [10.1145/3431832.3431841](https://doi.org/10.1145/3431832.3431841)]
12. Bengio Y, Ippolito D, Janda R, Jarvie M, Prud'homme B, Rousseau J, et al. Inherent privacy limitations of decentralized contact tracing apps. *J Am Med Inform Assoc* 2021 Jan 15;28(1):193-195 [FREE Full text] [doi: [10.1093/jamia/ocaa153](https://doi.org/10.1093/jamia/ocaa153)] [Medline: [32584990](https://pubmed.ncbi.nlm.nih.gov/32584990/)]
13. Ghose A, Li B, Macha M, Sun C, Foutz N. Trading Privacy for the Greater Social Good: How Did America React During COVID-19? arXiv. Preprint posted online Jun 10, 2020 [FREE Full text]
14. Bengio Y, Janda R, Yu YW, Ippolito D, Jarvie M, Pilat D, et al. The need for privacy with public digital contact tracing during the COVID-19 pandemic. *The Lancet Digital Health* 2020 Jul;2(7):e342-e344. [doi: [10.1016/s2589-7500\(20\)30133-3](https://doi.org/10.1016/s2589-7500(20)30133-3)]
15. Klenk M, Duijf H. Ethics of digital contact tracing and COVID-19: who is (not) free to go? *Ethics Inf Technol* 2020 Aug 24;1-9 [FREE Full text] [doi: [10.1007/s10676-020-09544-0](https://doi.org/10.1007/s10676-020-09544-0)] [Medline: [32863740](https://pubmed.ncbi.nlm.nih.gov/32863740/)]
16. Mello MM, Wang CJ. Ethics and governance for digital disease surveillance. *Science* 2020 May 29;368(6494):951-954. [doi: [10.1126/science.abb9045](https://doi.org/10.1126/science.abb9045)] [Medline: [32393527](https://pubmed.ncbi.nlm.nih.gov/32393527/)]
17. Pagliari C. The ethics and value of contact tracing apps: International insights and implications for Scotland's COVID-19 response. *J Glob Health* 2020 Dec;10(2):020103 [FREE Full text] [doi: [10.7189/jogh.10.020103](https://doi.org/10.7189/jogh.10.020103)] [Medline: [33110502](https://pubmed.ncbi.nlm.nih.gov/33110502/)]
18. Parker MJ, Fraser C, Abeler-Dörner L, Bonsall D. Ethics of instantaneous contact tracing using mobile phone apps in the control of the COVID-19 pandemic. *J Med Ethics* 2020 Jul 04;46(7):427-431 [FREE Full text] [doi: [10.1136/medethics-2020-106314](https://doi.org/10.1136/medethics-2020-106314)] [Medline: [32366705](https://pubmed.ncbi.nlm.nih.gov/32366705/)]
19. Yuan J, Li M, Lv G, Lu ZK. Monitoring transmissibility and mortality of COVID-19 in Europe. *Int J Infect Dis* 2020 Jun;95:311-315 [FREE Full text] [doi: [10.1016/j.ijid.2020.03.050](https://doi.org/10.1016/j.ijid.2020.03.050)] [Medline: [32234343](https://pubmed.ncbi.nlm.nih.gov/32234343/)]
20. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine* 2020 May 05;172(9):577-582. [doi: [10.7326/m20-0504](https://doi.org/10.7326/m20-0504)]
21. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* 2020 May 15;26(5):672-675. [doi: [10.1038/s41591-020-0869-5](https://doi.org/10.1038/s41591-020-0869-5)] [Medline: [32296168](https://pubmed.ncbi.nlm.nih.gov/32296168/)]
22. Cheng H, Jian S, Liu D, Ng T, Huang W, Lin H, Taiwan COVID-19 Outbreak Investigation Team. Contact Tracing Assessment of COVID-19 Transmission Dynamics in Taiwan and Risk at Different Exposure Periods Before and After Symptom Onset. *JAMA Intern Med* 2020 Sep 01;180(9):1156-1163 [FREE Full text] [doi: [10.1001/jamainternmed.2020.2020](https://doi.org/10.1001/jamainternmed.2020.2020)] [Medline: [32356867](https://pubmed.ncbi.nlm.nih.gov/32356867/)]
23. Rong X, Yang L, Chu H, Fan M. Effect of delay in diagnosis on transmission of COVID-19. *Math Biosci Eng* 2020 Mar 11;17(3):2725-2740 [FREE Full text] [doi: [10.3934/mbe.2020149](https://doi.org/10.3934/mbe.2020149)] [Medline: [32233563](https://pubmed.ncbi.nlm.nih.gov/32233563/)]
24. Bradshaw WJ, Alley EC, Huggins JH, Lloyd AL, Esvelt KM. Bidirectional contact tracing could dramatically improve COVID-19 control. *Nat Commun* 2021 Jan 11;12(1):232 [FREE Full text] [doi: [10.1038/s41467-020-20325-7](https://doi.org/10.1038/s41467-020-20325-7)] [Medline: [33431829](https://pubmed.ncbi.nlm.nih.gov/33431829/)]
25. Kojaku S, Hébert-Dufresne L, Mones E, Lehmann S, Ahn Y. The effectiveness of backward contact tracing in networks. *Nat Phys* 2021 Feb 25;17(5):652-658. [doi: [10.1038/s41567-021-01187-2](https://doi.org/10.1038/s41567-021-01187-2)]
26. Lai SHS, Tang CQY, Kurup A, Thevendran G. The experience of contact tracing in Singapore in the control of COVID-19: highlighting the use of digital technology. *Int Orthop* 2021 Jan 14;45(1):65-69 [FREE Full text] [doi: [10.1007/s00264-020-04646-2](https://doi.org/10.1007/s00264-020-04646-2)] [Medline: [33188602](https://pubmed.ncbi.nlm.nih.gov/33188602/)]
27. Maccari L, Cagno V. Do we need a contact tracing app? *Comput Commun* 2021 Jan 15;166:9-18 [FREE Full text] [doi: [10.1016/j.comcom.2020.11.007](https://doi.org/10.1016/j.comcom.2020.11.007)] [Medline: [33235399](https://pubmed.ncbi.nlm.nih.gov/33235399/)]
28. Cencetti G, Santin G, Longa A, Pigani E, Barrat A, Cattuto C, et al. Digital proximity tracing on empirical contact networks for pandemic control. *Nat Commun* 2021 Mar 12;12(1):1655 [FREE Full text] [doi: [10.1038/s41467-021-21809-w](https://doi.org/10.1038/s41467-021-21809-w)] [Medline: [33712583](https://pubmed.ncbi.nlm.nih.gov/33712583/)]
29. Amann J, Sleigh J, Vayena E. Digital contact-tracing during the Covid-19 pandemic: An analysis of newspaper coverage in Germany, Austria, and Switzerland. *PLoS One* 2021;16(2):e0246524 [FREE Full text] [doi: [10.1371/journal.pone.0246524](https://doi.org/10.1371/journal.pone.0246524)] [Medline: [33534839](https://pubmed.ncbi.nlm.nih.gov/33534839/)]
30. Hernandez-Orallo E, Manzoni P, Calafate CT, Cano J. Evaluating How Smartphone Contact Tracing Technology Can Reduce the Spread of Infectious Diseases: The Case of COVID-19. *IEEE Access* 2020;8:99083-99097. [doi: [10.1109/access.2020.2998042](https://doi.org/10.1109/access.2020.2998042)]
31. Shubina V, Holcer S, Gould M, Lohan ES. Survey of Decentralized Solutions with Mobile Devices for User Location Tracking, Proximity Detection, and Contact Tracing in the COVID-19 Era. *Data* 2020 Sep 23;5(4):87. [doi: [10.3390/data5040087](https://doi.org/10.3390/data5040087)]
32. Zhao Q, Wen H, Lin Z, Xuan D, Shroff N. On the Accuracy of Measured Proximity of Bluetooth-Based Contact Tracing Apps. In: Park N, Sun K, Foresti S, Butler K, Saxena N, editors. *Security and Privacy in Communication Networks. SecureComm 2020. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol 335. Cham: Springer; 2020:49-60.

33. Liang F. COVID-19 and Health Code: How Digital Platforms Tackle the Pandemic in China. *Soc Media Soc* 2020 Jul 11;6(3) [FREE Full text] [doi: [10.1177/2056305120947657](https://doi.org/10.1177/2056305120947657)] [Medline: [34192023](https://pubmed.ncbi.nlm.nih.gov/34192023/)]
34. Dar AB, Lone AH, Zahoor S, Khan AA, Naaz R. Applicability of mobile contact tracing in fighting pandemic (COVID-19): Issues, challenges and solutions. *Comput Sci Rev* 2020 Nov;38:100307 [FREE Full text] [doi: [10.1016/j.cosrev.2020.100307](https://doi.org/10.1016/j.cosrev.2020.100307)] [Medline: [32989380](https://pubmed.ncbi.nlm.nih.gov/32989380/)]
35. Gupta R, Rao UP. Achieving location privacy through CAST in location based services. *J Commun Netw* 2017;19(3):239-249. [doi: [10.1109/jcn.2017.000041](https://doi.org/10.1109/jcn.2017.000041)]
36. Underwood S. Blockchain beyond bitcoin. *Commun ACM* 2016 Oct 28;59(11):15-17. [doi: [10.1145/2994581](https://doi.org/10.1145/2994581)]
37. Xu H, Zhang L, Onireti O, Fang Y, Buchanan WJ, Imran MA. BeepTrace: Blockchain-Enabled Privacy-Preserving Contact Tracing for COVID-19 Pandemic and Beyond. *IEEE Internet Things J* 2021 Mar 1;8(5):3915-3929. [doi: [10.1109/jiot.2020.3025953](https://doi.org/10.1109/jiot.2020.3025953)]
38. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. *Science* 2014 Mar 14;343(6176):1203-1205. [doi: [10.1126/science.1248506](https://doi.org/10.1126/science.1248506)] [Medline: [24626916](https://pubmed.ncbi.nlm.nih.gov/24626916/)]
39. Kamel Boulos MN, Geraghty EM. Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. *Int J Health Geogr* 2020 Mar 11;19(1):8 [FREE Full text] [doi: [10.1186/s12942-020-00202-8](https://doi.org/10.1186/s12942-020-00202-8)] [Medline: [32160889](https://pubmed.ncbi.nlm.nih.gov/32160889/)]
40. Wang S, Ding S, Xiong L. A New System for Surveillance and Digital Contact Tracing for COVID-19: Spatiotemporal Reporting Over Network and GPS. *JMIR Mhealth Uhealth* 2020 Jun 10;8(6):e19457 [FREE Full text] [doi: [10.2196/19457](https://doi.org/10.2196/19457)] [Medline: [32499212](https://pubmed.ncbi.nlm.nih.gov/32499212/)]
41. Kleinman RA, Merkel C. Digital contact tracing for COVID-19. *CMAJ* 2020 Jun 15;192(24):E653-E656 [FREE Full text] [doi: [10.1503/cmaj.200922](https://doi.org/10.1503/cmaj.200922)] [Medline: [32461324](https://pubmed.ncbi.nlm.nih.gov/32461324/)]
42. Abeler J, Bäcker M, Buermeyer U, Zillessen H. COVID-19 Contact Tracing and Data Protection Can Go Together. *JMIR Mhealth Uhealth* 2020 Apr 20;8(4):e19359 [FREE Full text] [doi: [10.2196/19359](https://doi.org/10.2196/19359)] [Medline: [32294052](https://pubmed.ncbi.nlm.nih.gov/32294052/)]
43. Psiaki ML, Humphreys TE, Stauffer B. Attackers can spoof navigation signals without our knowledge. Here's how to fight back GPS lies. *IEEE Spectr* 2016 Aug;53(8):26-53. [doi: [10.1109/mspec.2016.7524168](https://doi.org/10.1109/mspec.2016.7524168)]
44. Alsdurf H, Belliveau E, Bengio Y, Deleu T, Gupta P, Ippolito D, et al. COVI White Paper. arXiv. Preprint posted online May 18, 2020 [FREE Full text]
45. Leith DJ, Farrell S. Measurement-based evaluation of Google/Apple Exposure Notification API for proximity detection in a light-rail tram. *PLoS One* 2020 Sep 30;15(9):e0239943 [FREE Full text] [doi: [10.1371/journal.pone.0239943](https://doi.org/10.1371/journal.pone.0239943)] [Medline: [32997724](https://pubmed.ncbi.nlm.nih.gov/32997724/)]
46. Nguyen KA, Luo Z, Watkins C. Epidemic contact tracing with smartphone sensors. *Journal of Location Based Services* 2020 Sep 01;14(2):92-128. [doi: [10.1080/17489725.2020.1805521](https://doi.org/10.1080/17489725.2020.1805521)]
47. Braithwaite I, Callender T, Bullock M, Aldridge RW. Automated and partly automated contact tracing: a systematic review to inform the control of COVID-19. *The Lancet Digital Health* 2020 Nov;2(11):e607-e621. [doi: [10.1016/s2589-7500\(20\)30184-9](https://doi.org/10.1016/s2589-7500(20)30184-9)]
48. Leith DJ, Farrell S. Coronavirus contact tracing. *SIGCOMM Comput Commun Rev* 2020 Oct 26;50(4):66-74. [doi: [10.1145/3431832.3431840](https://doi.org/10.1145/3431832.3431840)]
49. Currie D, Peng C, Lyle D, Jameson B, Frommer M. Stemming the flow: how much can the Australian smartphone app help to control COVID-19? *Public Health Res Pract* 2020 Jun 30;30(2) [FREE Full text] [doi: [10.17061/phrp3022009](https://doi.org/10.17061/phrp3022009)] [Medline: [32601652](https://pubmed.ncbi.nlm.nih.gov/32601652/)]
50. Sharon T. Blind-sided by privacy? Digital contact tracing, the Apple/Google API and big tech's newfound role as global health policy makers. *Ethics Inf Technol* 2020 Jul 18:1-13 [FREE Full text] [doi: [10.1007/s10676-020-09547-x](https://doi.org/10.1007/s10676-020-09547-x)] [Medline: [32837287](https://pubmed.ncbi.nlm.nih.gov/32837287/)]
51. Chowdhury MJM, Ferdous MS, Biswas K, Chowdhury N, Muthukkumarasamy V. COVID-19 Contact Tracing: Challenges and Future Directions. *IEEE Access* 2020;8:225703-225729. [doi: [10.1109/access.2020.3036718](https://doi.org/10.1109/access.2020.3036718)]
52. Girolami M, Mavilia F, Delmastro F, Distefano E. Detecting Social Interactions through Commercial Mobile Devices. 2018 Presented at: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops); Mar 19-23; Athens, Greece p. 125-130. [doi: [10.1109/percomw.2018.8480397](https://doi.org/10.1109/percomw.2018.8480397)]
53. Yamamoto K, Takahashi T, Urasaki M, Nagayasu Y, Shimamoto T, Tateyama Y, et al. Health Observation App for COVID-19 Symptom Tracking Integrated With Personal Health Records: Proof of Concept and Practical Use Study. *JMIR Mhealth Uhealth* 2020 Jul 06;8(7):e19902 [FREE Full text] [doi: [10.2196/19902](https://doi.org/10.2196/19902)] [Medline: [32568728](https://pubmed.ncbi.nlm.nih.gov/32568728/)]
54. Yap KY, Xie Q. Personalizing symptom monitoring and contact tracing efforts through a COVID-19 web-app. *Infect Dis Poverty* 2020 Jul 13;9(1):93 [FREE Full text] [doi: [10.1186/s40249-020-00711-5](https://doi.org/10.1186/s40249-020-00711-5)] [Medline: [32660568](https://pubmed.ncbi.nlm.nih.gov/32660568/)]
55. Lee T, Lee H. Tracing surveillance and auto-regulation in Singapore: 'smart' responses to COVID-19. *Media International Australia* 2020 Aug 12;177(1):47-60. [doi: [10.1177/1329878x20949545](https://doi.org/10.1177/1329878x20949545)]
56. Bay J, Kek J, Tan A, Hau C, Yongquan L, Tan J, et al. BlueTrace: A privacy-preserving protocol for community-driven contact tracing across borders. Government Technology Agency, Singapore. 2020 Apr 9. URL: <https://bluetrace.io/static/bluetrace-whitepaper-938063656596c104632def383eb33b3c.pdf> [accessed 2021-07-14]



57. Abbas R, Michael K. COVID-19 Contact Trace App Deployments: Learnings From Australia and Singapore. *IEEE Consumer Electron Mag* 2020 Sep 1;9(5):65-70. [doi: [10.1109/mce.2020.3002490](https://doi.org/10.1109/mce.2020.3002490)]
58. Goggin G. COVID-19 apps in Singapore and Australia: reimagining healthy nations with digital technology. *Media International Australia* 2020 Aug 14;177(1):61-75. [doi: [10.1177/1329878x20949770](https://doi.org/10.1177/1329878x20949770)]
59. Michael K, Abbas R. Behind COVID-19 Contact Trace Apps: The Google–Apple Partnership. *IEEE Consumer Electron Mag* 2020 Sep 1;9(5):71-76. [doi: [10.1109/mce.2020.3002492](https://doi.org/10.1109/mce.2020.3002492)]
60. Loi M. How to fairly incentivise digital contact tracing. *J Med Ethics* 2020 Jul 09;medethics-2020-106388 [FREE Full text] [doi: [10.1136/medethics-2020-106388](https://doi.org/10.1136/medethics-2020-106388)] [Medline: [32647047](https://pubmed.ncbi.nlm.nih.gov/32647047/)]
61. Chua AQ, Tan MMJ, Verma M, Han EKL, Hsu LY, Cook AR, et al. Health system resilience in managing the COVID-19 pandemic: lessons from Singapore. *BMJ Glob Health* 2020 Sep 16;5(9):e003317 [FREE Full text] [doi: [10.1136/bmjgh-2020-003317](https://doi.org/10.1136/bmjgh-2020-003317)] [Medline: [32938609](https://pubmed.ncbi.nlm.nih.gov/32938609/)]
62. Leslie M. COVID-19 Fight Enlists Digital Technology: Contact Tracing Apps. *Engineering (Beijing)* 2020 Oct;6(10):1064-1066 [FREE Full text] [doi: [10.1016/j.eng.2020.09.001](https://doi.org/10.1016/j.eng.2020.09.001)] [Medline: [32953197](https://pubmed.ncbi.nlm.nih.gov/32953197/)]
63. Skoll D, Miller JC, Saxon LA. COVID-19 testing and infection surveillance: Is a combined digital contact-tracing and mass-testing solution feasible in the United States? *Cardiovascular Digital Health Journal* 2020 Nov;1(3):149-159. [doi: [10.1016/j.cvdhj.2020.09.004](https://doi.org/10.1016/j.cvdhj.2020.09.004)]
64. HSE Covid Tracker App: Pre-Release Report Card. Irish Council for Civil Liberties, Digital Rights Ireland. 2020 Jul. URL: <https://www.iccl.ie/wp-content/uploads/2020/07/ICCL-DRI-HSE-App-Pre-Release-Report-Card.pdf> [accessed 2021-07-12]
65. Shankar K, Jeng W, Thomer A, Weber N, Yoon A. Data curation as collective action during COVID - 19. *J Assoc Inf Sci Technol* 2020 Sep 02;72(3):280-284. [doi: [10.1002/asi.24406](https://doi.org/10.1002/asi.24406)]
66. Lucivero F, Hallowell N, Johnson S, Prainsack B, Samuel G, Sharon T. COVID-19 and Contact Tracing Apps: Ethical Challenges for a Social Experiment on a Global Scale. *J Bioeth Inq* 2020 Dec 25;17(4):835-839 [FREE Full text] [doi: [10.1007/s11673-020-10016-9](https://doi.org/10.1007/s11673-020-10016-9)] [Medline: [32840842](https://pubmed.ncbi.nlm.nih.gov/32840842/)]
67. Joo J, Shin MM. Resolving the tension between full utilization of contact tracing app services and user stress as an effort to control the COVID-19 pandemic. *Serv Bus* 2020 Sep 01;14(4):461-478. [doi: [10.1007/s11628-020-00424-7](https://doi.org/10.1007/s11628-020-00424-7)]
68. Roche S. Smile, you're being traced! Some thoughts about the ethical issues of digital contact tracing applications. *Journal of Location Based Services* 2020 Aug 24;14(2):71-91. [doi: [10.1080/17489725.2020.1811409](https://doi.org/10.1080/17489725.2020.1811409)]
69. Ryan M. In defence of digital contact-tracing: human rights, South Korea and Covid-19. *IJPC* 2020 Aug 06;16(4):383-407. [doi: [10.1108/ijpcc-07-2020-0081](https://doi.org/10.1108/ijpcc-07-2020-0081)]
70. Park YJ, Choe YJ, Park O, Park SY, Kim Y, Kim J, COVID-19 National Emergency Response Center, Epidemiology Case Management Team. Contact Tracing during Coronavirus Disease Outbreak, South Korea, 2020. *Emerg Infect Dis* 2020 Oct;26(10):2465-2468 [FREE Full text] [doi: [10.3201/eid2610.201315](https://doi.org/10.3201/eid2610.201315)] [Medline: [32673193](https://pubmed.ncbi.nlm.nih.gov/32673193/)]
71. Contact Transmission of COVID-19 in South Korea: Novel Investigation Techniques for Tracing Contacts. *Osong Public Health Res Perspect* 2020 Feb;11(1):60-63 [FREE Full text] [doi: [10.24171/j.phrp.2020.11.1.09](https://doi.org/10.24171/j.phrp.2020.11.1.09)] [Medline: [32149043](https://pubmed.ncbi.nlm.nih.gov/32149043/)]
72. Introna L, Pouloudi A. Privacy in the information age: takeholders, interests and values. *J Bus Ethics* 1999;22:27-38. [doi: [10.1023/A:1006151900807](https://doi.org/10.1023/A:1006151900807)]
73. Park S, Choi GJ, Ko H. Information Technology-Based Tracing Strategy in Response to COVID-19 in South Korea-Privacy Controversies. *JAMA* 2020 Jun 02;323(21):2129-2130 [FREE Full text] [doi: [10.1001/jama.2020.6602](https://doi.org/10.1001/jama.2020.6602)] [Medline: [32324202](https://pubmed.ncbi.nlm.nih.gov/32324202/)]
74. Luciano F. Mind the App-Considerations on the Ethical Risks of COVID-19 Apps. *Philos Technol* 2020 Jun 13;33(2):1-6 [FREE Full text] [doi: [10.1007/s13347-020-00408-5](https://doi.org/10.1007/s13347-020-00408-5)] [Medline: [32837867](https://pubmed.ncbi.nlm.nih.gov/32837867/)]
75. Sandvik KB. “Smittestopp”: If you want your freedom back, download now. *Big Data & Society* 2020 Jul 28;7(2):205395172093998. [doi: [10.1177/2053951720939985](https://doi.org/10.1177/2053951720939985)]
76. Ursin G, Skjesol I, Tritter J. The COVID-19 pandemic in Norway: The dominance of social implications in framing the policy response. *Health Policy Technol* 2020 Dec;9(4):663-672 [FREE Full text] [doi: [10.1016/j.hlpt.2020.08.004](https://doi.org/10.1016/j.hlpt.2020.08.004)] [Medline: [32874857](https://pubmed.ncbi.nlm.nih.gov/32874857/)]
77. O'Callaghan M, Buckley J, Fitzgerald B, Johnson K, Laffey J, McNicholas B, et al. A National Survey of Attitudes to COVID-19 Digital Contact Tracing in the Republic of Ireland Internet. *Research Square*. Preprint posted online July 10, 2020 [FREE Full text] [doi: [10.21203/rs.3.rs-40778/v1](https://doi.org/10.21203/rs.3.rs-40778/v1)]
78. Abuhammad S, Khabour OF, Alzoubi KH. COVID-19 Contact-Tracing Technology: Acceptability and Ethical Issues of Use. *Patient Prefer Adherence* 2020 Sep;14:1639-1647 [FREE Full text] [doi: [10.2147/PPA.S276183](https://doi.org/10.2147/PPA.S276183)] [Medline: [32982188](https://pubmed.ncbi.nlm.nih.gov/32982188/)]
79. Altmann S, Milsom L, Zillesen H, Blasone R, Gerdon F, Bach R, et al. Acceptability of App-Based Contact Tracing for COVID-19: Cross-Country Survey Study. *JMIR Mhealth Uhealth* 2020 Aug 28;8(8):e19857 [FREE Full text] [doi: [10.2196/19857](https://doi.org/10.2196/19857)] [Medline: [32759102](https://pubmed.ncbi.nlm.nih.gov/32759102/)]
80. The app credibility gap. *Nat Biotechnol* 2020 Jul 26;38(7):768-768 [FREE Full text] [doi: [10.1038/s41587-020-0610-4](https://doi.org/10.1038/s41587-020-0610-4)] [Medline: [32591763](https://pubmed.ncbi.nlm.nih.gov/32591763/)]
81. Cheng VC, Wong S, Chuang VW, So SY, Chen JH, Sridhar S, et al. The role of community-wide wearing of face mask for control of coronavirus disease 2019 (COVID-19) epidemic due to SARS-CoV-2. *J Infect* 2020 Jul;81(1):107-114 [FREE Full text] [doi: [10.1016/j.jinf.2020.04.024](https://doi.org/10.1016/j.jinf.2020.04.024)] [Medline: [32335167](https://pubmed.ncbi.nlm.nih.gov/32335167/)]



82. Yeung N, Lai J, Luo J. Face Off: Polarized Public Opinions on Personal Face Mask Usage during the COVID-19 Pandemic. arXiv. Preprint posted online Oct 31, 2020 [FREE Full text] [doi: [10.1109/bigdata50022.2020.9378114](https://doi.org/10.1109/bigdata50022.2020.9378114)]
83. Parmet WE, Sinha MS. Covid-19 — The Law and Limits of Quarantine. *N Engl J Med* 2020 Apr 09;382(15):e28. [doi: [10.1056/nejmp2004211](https://doi.org/10.1056/nejmp2004211)]
84. Haischer MH, Beilfuss R, Hart MR, Opielinski L, Wrucke D, Zirgaitis G, et al. Who is wearing a mask? Gender-, age-, and location-related differences during the COVID-19 pandemic. *PLoS One* 2020 Oct 15;15(10):e0240785 [FREE Full text] [doi: [10.1371/journal.pone.0240785](https://doi.org/10.1371/journal.pone.0240785)] [Medline: [33057375](https://pubmed.ncbi.nlm.nih.gov/33057375/)]
85. Feng S, Shen C, Xia N, Song W, Fan M, Cowling BJ. Rational use of face masks in the COVID-19 pandemic. *The Lancet Respiratory Medicine* 2020 May;8(5):434-436. [doi: [10.1016/s2213-2600\(20\)30134-x](https://doi.org/10.1016/s2213-2600(20)30134-x)]
86. Yannakourou S, Jougleux P, Markou C, Tziayas S, Gregoriou G, Flouri A, et al. Covid-19 and Law: Legal Solutions during Covid-19 Upheavals. *ENTHA* 2020;13:9-29.
87. Sharma T, Bashir M. Use of apps in the COVID-19 response and the loss of privacy protection. *Nat Med* 2020 Aug 26;26(8):1165-1167. [doi: [10.1038/s41591-020-0928-y](https://doi.org/10.1038/s41591-020-0928-y)] [Medline: [32457443](https://pubmed.ncbi.nlm.nih.gov/32457443/)]
88. Rosenkrantz L, Schuurman N, Bell N, Amram O. The need for GIScience in mapping COVID-19. *Health Place* 2021 Jan;67:102389 [FREE Full text] [doi: [10.1016/j.healthplace.2020.102389](https://doi.org/10.1016/j.healthplace.2020.102389)] [Medline: [33526208](https://pubmed.ncbi.nlm.nih.gov/33526208/)]
89. Hendl T, Chung R, Wild V. Pandemic Surveillance and Racialized Subpopulations: Mitigating Vulnerabilities in COVID-19 Apps. *J Bioeth Inq* 2020 Dec 25;17(4):829-834 [FREE Full text] [doi: [10.1007/s11673-020-10034-7](https://doi.org/10.1007/s11673-020-10034-7)] [Medline: [32840858](https://pubmed.ncbi.nlm.nih.gov/32840858/)]
90. O'Brien C. Google to roll out software fix to remedy issue with Covid app. *Irish Times*. 2020 Aug 9. URL: <https://www.irishtimes.com/business/technology/google-to-roll-out-software-fix-to-remedy-issue-with-covid-app-1.4325981> [accessed 2021-07-12]
91. Gorey C. Google to launch fix for battery drain affecting Covid Tracker Ireland app. *Silicon Republic*. 2020 Aug 10. URL: <https://www.siliconrepublic.com/enterprise/battery-drain-covid-tracker-ireland-app-google> [accessed 2021-07-12]
92. O'Brien C. HSE says fix for Covid Tracker app rolled out to all Android users. *Irish Times*. 2020 Aug 10. URL: <https://www.irishtimes.com/business/technology/hse-says-fix-for-covid-tracker-app-rolled-out-to-all-android-users-1.4326646> [accessed 2021-07-12]
93. Gagné C. Everything You Need to Know About the New COVID Alert App. *Chatelaine*. 2020 Aug 10. URL: <https://www.chatelaine.com/health/covid-alert-app/> [accessed 2021-07-12]
94. Leith DJ, Farrell S. Measurement-based evaluation of Google/Apple Exposure Notification API for proximity detection in a commuter bus. *PLoS One* 2021 Apr 29;16(4):e0250826 [FREE Full text] [doi: [10.1371/journal.pone.0250826](https://doi.org/10.1371/journal.pone.0250826)] [Medline: [33914810](https://pubmed.ncbi.nlm.nih.gov/33914810/)]
95. Morrison S. Perhaps months too late, the Apple-Google Covid-19 contact tracing tool comes to America. *Vox*. 2020 Aug 6. URL: <https://www.vox.com/recode/2020/8/6/21357098/apple-google-exposure-notification-virginia-contact-tracing> [accessed 2021-07-12]
96. Immuni: le motivazioni di chi non scarica l'app. *Punto Informatico*. 2020 Jul 9. URL: <https://www.punto-informatico.it/immuni-motivazioni-scarica-app/> [accessed 2021-07-14]
97. Lardieri A. Coronavirus Pandemic Causing Anxiety, Depression in Americans, CDC Finds. *US News*. 2020 Aug 13. URL: <https://www.usnews.com/news/health-news/articles/2020-08-13/coronavirus-pandemic-causing-anxiety-depression-in-americans-cdc-finds> [accessed 2021-07-12]
98. Brooks SK, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *The Lancet* 2020 Mar 14;395(10227):912-920. [doi: [10.1016/S0140-6736\(20\)30460-8](https://doi.org/10.1016/S0140-6736(20)30460-8)] [Medline: [32112714](https://pubmed.ncbi.nlm.nih.gov/32112714/)]
99. Lalot F, Heering MS, Rullo M, Travaglino GA, Abrams D. The dangers of distrustful complacency: Low concern and low political trust combine to undermine compliance with governmental restrictions in the emerging Covid-19 pandemic. *Group Processes & Intergroup Relations* 2020 Oct 30. [doi: [10.1177/1368430220967986](https://doi.org/10.1177/1368430220967986)]
100. Sadjadi S, Greenberg C, Reynolds D. NIST TC4TL Challenge. *NIST*. 2020 Jun. URL: <https://www.nist.gov/itl/iad/mig/nist-tc4tl-challenge> [accessed 2021-07-12]
101. Pandl K, Thiebes S, Schmidt-Kraepelin M, Sunyaev A. How detection ranges and usage stops impact digital contact tracing effectiveness for COVID-19 Internet. *medRxiv*. Preprint posted online Dec 11, 2020 [FREE Full text] [doi: [10.1101/2020.12.08.20246140](https://doi.org/10.1101/2020.12.08.20246140)]
102. Bhatia A, Matthan R, Khanna T, Balsari S. Regulatory Sandboxes: A Cure for mHealth Pilotitis? *J Med Internet Res* 2020 Sep 15;22(9):e21276 [FREE Full text] [doi: [10.2196/21276](https://doi.org/10.2196/21276)] [Medline: [32763889](https://pubmed.ncbi.nlm.nih.gov/32763889/)]
103. Colizza V, Grill E, Mikolajczyk R, Cattuto C, Kucharski A, Riley S, et al. Time to evaluate COVID-19 contact-tracing apps. *Nat Med* 2021 Mar 15;27(3):361-362. [doi: [10.1038/s41591-021-01236-6](https://doi.org/10.1038/s41591-021-01236-6)] [Medline: [33589822](https://pubmed.ncbi.nlm.nih.gov/33589822/)]
104. Salathé M, Althaus C, Anderegg N, Antonioli D, Ballouz T, Bugnon E, et al. Early evidence of effectiveness of digital contact tracing for SARS-CoV-2 in Switzerland. *Swiss Med Wkly* 2020 Dec 14;150:w20457 [FREE Full text] [doi: [10.4414/smw.2020.20457](https://doi.org/10.4414/smw.2020.20457)] [Medline: [33327003](https://pubmed.ncbi.nlm.nih.gov/33327003/)]
105. Kretzschmar ME, Rozhnova G, Bootsma MCJ, van Boven M, van de Wijgert JHHM, Bonten MJM. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *The Lancet Public Health* 2020 Aug;5(8):e452-e459. [doi: [10.1016/s2468-2667\(20\)30157-2](https://doi.org/10.1016/s2468-2667(20)30157-2)]

## Abbreviations

**AI:** artificial intelligence  
**API:** application programming interface  
**CCTV:** closed-circuit television  
**CoV-SCR:** COVID-19 Symptom Monitoring and Contact Tracking Record  
**GAEN:** Google/Apple Exposure Network  
**GIS:** global information system  
**mHealth:** mobile health  
**MIT:** Massachusetts Institute of Technology  
**PACT:** Private Automated Contact Tracing  
**PEPP-PT:** Pan-European Privacy-Preserving Proximity Tracing  
**QR:** quick response  
**R0:** basic reproduction number  
**RPI:** Rolling Proximity Identifier  
**STRONG:** Spatiotemporal Reporting Over Network and GPS  
**TTP:** trusted third party

*Edited by C Lovis; submitted 25.01.21; peer-reviewed by K Pandl, D Wong; comments to author 13.03.21; revised version received 03.04.21; accepted 14.04.21; published 19.07.21.*

*Please cite as:*

*Hogan K, Macedo B, Macha V, Barman A, Jiang X*

*Contact Tracing Apps: Lessons Learned on Privacy, Autonomy, and the Need for Detailed and Thoughtful Implementation*

*JMIR Med Inform 2021;9(7):e27449*

*URL: <https://medinform.jmir.org/2021/7/e27449>*

*doi: [10.2196/27449](https://doi.org/10.2196/27449)*

*PMID: [34254937](https://pubmed.ncbi.nlm.nih.gov/34254937/)*

©Katie Hogan, Briana Macedo, Venkata Macha, Arko Barman, Xiaoqian Jiang. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 19.07.2021. 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 Three Feasibility Tools for Identifying Patient Data and Biospecimen Availability: Comparative Usability Study

Christina Schüttler<sup>1</sup>, MSc; Hans-Ulrich Prokosch<sup>1</sup>, PhD; Martin Sedlmayr<sup>2</sup>, PhD; Brita Sedlmayr<sup>2</sup>, PhD

<sup>1</sup>Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

<sup>2</sup>Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany

**Corresponding Author:**

Christina Schüttler, MSc

Chair of Medical Informatics

Friedrich-Alexander-Universität Erlangen-Nürnberg

Wetterkreuz 15

Erlangen, 91058

Germany

Phone: 49 9131 85 67789

Email: [christina.schuettler@fau.de](mailto:christina.schuettler@fau.de)

**Related Article:**

This is a corrected version. See correction statement: <https://medinform.jmir.org/2021/10/e33105>

## Abstract

**Background:** To meet the growing importance of real-world data analysis, clinical data and biosamples must be timely made available. Feasibility platforms are often the first contact point for determining the availability of such data for specific research questions. Therefore, a user-friendly interface should be provided to enable access to this information easily. The German Medical Informatics Initiative also aims to establish such a platform for its infrastructure. Although some of these platforms are actively used, their tools still have limitations. Consequently, the Medical Informatics Initiative consortium MIRACUM (Medical Informatics in Research and Care in University Medicine) committed itself to analyzing the pros and cons of existing solutions and to designing an optimized graphical feasibility user interface.

**Objective:** The aim of this study is to identify the system that is most user-friendly and thus forms the best basis for developing a harmonized tool. To achieve this goal, we carried out a comparative usability evaluation of existing tools used by researchers acting as end users.

**Methods:** The evaluation included three preselected search tools and was conducted as a qualitative exploratory study with a randomized design over a period of 6 weeks. The tools in question were the MIRACUM i2b2 (Informatics for Integrating Biology and the Bedside) feasibility platform, OHDSI's (Observational Health Data Sciences and Informatics) ATLAS, and the Sample Locator of the German Biobank Alliance. The evaluation was conducted in the form of a web-based usability test (usability walkthrough combined with a web-based questionnaire) with participants aged between 26 and 63 years who work as medical doctors.

**Results:** In total, 17 study participants evaluated the three tools. The overall evaluation of usability, which was based on the System Usability Scale, showed that the Sample Locator, with a mean System Usability Scale score of 77.03 (SD 20.62), was significantly superior to the other two tools (Wilcoxon test; Sample Locator vs i2b2:  $P=.047$ ; Sample Locator vs ATLAS:  $P=.001$ ). i2b2, with a score of 59.83 (SD 25.36), performed significantly better than ATLAS, which had a score of 27.81 (SD 21.79; Wilcoxon test; i2b2 vs ATLAS:  $P=.005$ ). The analysis of the material generated by the usability walkthrough method confirmed these findings. ATLAS caused the most usability problems ( $n=66$ ), followed by i2b2 ( $n=48$ ) and the Sample Locator ( $n=22$ ). Moreover, the Sample Locator achieved the highest ratings with respect to additional questions regarding satisfaction with the tools.

**Conclusions:** This study provides data to develop a suitable basis for the selection of a harmonized tool for feasibility studies via concrete evaluation and a comparison of the usability of three different types of query builders. The feedback obtained from the participants during the usability test made it possible to identify user problems and positive design aspects of the individual tools and compare them qualitatively.

**KEYWORDS**

software tools; user interface; feasibility; evaluation; research

## Introduction

Real-world data analysis in medicine is becoming increasingly important and relies on the timely availability of clinical data and biosamples collected during clinical care processes in university hospitals [1,2]. The exploitation and use of such data are two of the major goals of several national and international initiatives [3-5]. In Germany, this goal is being pursued with a nationwide approach, particularly through the Medical Informatics Initiative (MII), in which all university hospitals have joined forces in four consortia [6]. However, a crucial aspect of this process is not only the allocation and preparation of data from the respective source systems but also their findability for external interest groups such as researchers. For this purpose, the feasibility platforms are a common first contact point before writing a data use request and submitting it to the data provider. At this level, researchers can initially verify whether the affiliated institution has a suitable number of patient records for a planned research project. This usually requires a graphical user interface that can formulate a description of the desired patient cohort based on the study criteria. The MII also aims to establish such a platform as part of its central portal. Although there are various projects that have already implemented such a platform, the tools used have specific limitations, such as single source compatibility, a reduced number of temporal constraints available [7], and limited usability [8]. Consequently, MIRACUM (Medical Informatics in Research and Care in University Medicine) [9], one of the four MII consortia, has set itself the task of carrying out a comparative evaluation of existing tools with regard to their usability to identify the most user-friendly system, and thus forms the best basis for developing a harmonized tool. As the implementation of such a platform was intended to take place as quickly as possible, a preselection of three implementations already used in the consortium (also freely accessible for researchers in Germany) was made for the usability evaluation: the MIRACUM i2b2 (Informatics for Integrating Biology and the Bedside) [10] feasibility platform, OHDSI's (Observational Health Data Sciences and Informatics) ATLAS [11,12], and the Sample Locator [13] of the German Biobank Alliance (GBA) [14,15]. The selection was based on the fact that they differed greatly in terms of complexity and functionality, so good coverage was expected. The usability analysis focused on two questions: (1) which tool offers the best usability (overall) and hence forms the most suitable foundation for creating a balanced tool? and (2) in which areas and with regard to which usability aspects are the tools rated better or worse in comparison and which recommendations can be derived for further development?

This paper describes the procedure used to answer these research questions. As the focus was on the evaluation of user-friendliness, a usability analysis was conducted with potential end users—laypeople, who should be enabled to conduct feasibility studies. To the best of our knowledge, there

has not yet been a study comparing these three query builders—i2b2, ATLAS, and the Sample Locator—in terms of usability. This study should address this gap in the scientific literature. The methodological approach in this study can serve as a model for decision makers and researchers of similar projects. The insights gained from the evaluation of the tools by clinically active researchers will subsequently be used for the further development of a unified tool.

## Methods

### Study Design

To evaluate the previously selected search tools, a qualitative exploratory study with a randomized design over a period of 6 weeks (from August 3, 2020, to September 13, 2020) was conducted. It was carried out in the form of a web-based usability test (usability walkthrough combined with a web-based questionnaire) with female and male participants aged between 26 and 63 years who work as medical doctors that are also engaged in research. In advance of this study, ethical approval was obtained from the Technical University of Dresden (Germany) ethics committee (SR-EK-262062020).

### Recruitment

For a valid evaluation of usability, the study concept called for a study size of 30 subjects. This corresponds to three researchers per MIRACUM site (n=10). Given the number of test persons, it can be assumed that the majority of all usability problems are discovered [16]. A contact person at the respective location identified and approached suitable study participants. In addition to the requirement of being clinically active and engaged in research, the test participants were required to have no experience with the tools to be evaluated, enough time to test all systems and answer a questionnaire, and be willing to record the test. In case of interest in participating in the study, the contact details were forwarded to the study team. At the start of the study, the participants received an email containing all relevant documents for conducting the evaluation. In addition, the study information and a consent form were attached, which needed to be signed and returned to the study team after completion of the study.

### Material

The three tools to be evaluated are the MIRACUM i2b2 feasibility platform (webclient version 1.7.12), OHDSI's ATLAS (version 2.7.7/2.7.8), and GBA's Sample Locator (user interface version 1.3.0-alpha.4 and backend version 6.2.0). The MIRACUM i2b2 feasibility platform is based on proprietary (but internationally widely used) data structures. It is currently based on the six basic modules of the MII core data set and supports participation in international large-scale research [17] (Figure 1). ATLAS is primarily a web interface that allows the use of various OHDSI tools. Functionalities include search and navigation within the OMOP (Observational Medical Outcomes



Partnership) Common Data Model Vocabulary database to identify patient cohorts (Figure 2). The third tool is the Sample Locator, which is designed to search for samples and related data from GBA-affiliated biobanks (Figure 3). Although the

i2b2 and OHDSI ATLAS clients are already heavily applied in international data sharing networks [10,11], the GBA Sample Locator is productive as a first version.

**Figure 1.** Example of a query built using the MIRACUM i2b2. On the right side of the screen, the user can select the appropriate parameters and then drag and drop them into “AND-linked” groups on the left side of the screen. Exclusion criteria are defined by the “Exclude” option in the groups. The search is executed by selecting the button “Run Query.” i2b2: Informatics for Integrating Biology and the Bedside; MIRACUM: Medical Informatics in Research and Care in University Medicine.

The screenshot displays the i2b2 Query & Analysis Tool interface. The top navigation bar includes 'Terms', 'Find Trm', 'Info', 'Workplace', 'Queries', and 'Find Qry'. The left pane shows a tree structure of terms under 'MIRACUM', with 'weiblich' selected. The right pane shows the 'Query Tool' with three groups. Group 1 contains 'weiblich'. The interface includes buttons for 'Run Query', 'Clear', 'New Group', 'Show Query Status', 'Graph Results', and 'Query Report'.

**Figure 2.** Example of a query built using OHDSI’s ATLAS. The criteria in the form of concepts can be selected and linked by selecting the “New Inclusion Criteria” button. The definition of an exclusion criterion is made by defining it as a “noninclusion.” ATLAS requires that an entry and exit event must be defined for the search. The search is executed via the “Generation” tab. OHDSI: Observational Health Data Sciences and Informatics.

The screenshot displays the ATLAS 'New Cohort Definition' interface. On the left is a dark sidebar with navigation options: Home, Data Sources, Search, Concept Sets, Cohort Definitions (highlighted), Characterizations, Cohort Pathways, Incidence Rates, Profiles, Estimation, Prediction, Jobs, Configuration, and Feedback. The main content area is titled 'New Cohort Definition' and includes a search bar and tabs for Definition, Concept Sets, Generation, Reporting, Export, and Messages. The 'Definition' tab is active, showing a text input for a cohort definition description. Below this are three main sections: 'Cohort Entry Events', 'Inclusion Criteria', and 'Cohort Exit'. The 'Cohort Entry Events' section includes options to add initial events and attributes, with a criterion for observation periods greater than 365 days. The 'Inclusion Criteria' section shows a list of criteria, currently containing '1. Gender female', with options to add new criteria or groups. The 'Cohort Exit' section includes settings for event persistence and censoring events. A bottom-left footer mentions 'Apache 2.0 open source software provided by OHDSI join the journey'. A bottom-right dropdown menu lists various criteria types like 'Add Demographic', 'Add Condition Era', 'Add Death', etc.

**Figure 3.** Example of a query built using the German Biobank Alliance Sample Locator. With the Sample Locator, the corresponding criteria are compiled via the selection menus. Input fields within a criterion (eg, diagnosis, as shown in the figure) are linked with “OR,” and input fields between criterion fields are linked with “AND.” An exclusion can be defined using the operator “not equal to.” The search is executed by selecting the “Send” button.

samplelocator.bbmri.de/search

German Biobank Node  
bbmri.de

About Us Negotiator Login

### Donor/Clinical Information

Diagnosis age donor (years) < 80

Sex = Female

Diagnosis ICD-10 = C50.0  
= e.g. C25.1

ADD FIELD

### Sample

ADD FIELD

CLEAR EDIT SEND

Imprint | Privacy Policy | 1.3.0-alpha.4 (UI) | 6.2.0-SNAPSHOT (Backend)

SPONSORED BY THE Federal Ministry

The usability analysis of the examined tools was based on the processing of three tasks. The tasks were structured in such a way that they increased in complexity. Although the first task only required the selection of inclusion criteria (gender, diagnosis, therapy, and laboratory test), the following task also asked for a parameter to be defined as an exclusion criterion.

The final task included a time component, which was queried by specifying a diagnosis period. For the sake of comparability, the respective tasks were coordinated accordingly between the tools, taking into account the tool-specific functionalities (Table 1).

**Table 1.** Queries construction. The users were asked to construct a query according to these specified criteria and find the number of corresponding patients or biosamples.

Criterion type and criterion	i2b2 <sup>a</sup>	Criterion	ATLAS	Criterion	Sample locator
<b>Query 1</b>					
Cohort entry event	N/A <sup>b</sup>	Observation period	Duration: >365 days	N/A	N/A
<b>Inclusion</b>					
Gender	Female	Gender	Female	Sex	Female
Diagnosis	Malignant neoplasm of the brain	Condition occurrence	Malignant neoplasm of the brain	Diagnosis	Carcinoma mammae
Treatment	Temozolomide	Treatment	Temozolomide	Age	<80 years
Lab values	Platelet count: <50.000/uL	Lab values	Platelet count: <50.000/uL	Sample type	Tissue stored in formalin
Cohort exit	N/A	Event will persist until:	End of continuous observation	N/A	N/A
<b>Query 2</b>					
Cohort entry	N/A	Observation Period	Duration: >365 days	N/A	N/A
<b>Inclusion</b>					
Age	>18 years	Age	>18 years	Sex	Male
Diagnosis	Type 2 diabetes mellitus	Diagnosis	Type 2 diabetes mellitus	Diagnosis	Atherosclerotic cardiovascular disease
Lab values	Hemoglobin between 13 and 18 g/dL	Lab values	Hemoglobin between 13 and 18 g/dL	Biosamples	Serum, storage temperature: -70°C OR <sup>c</sup> plasma stabilized, storage temperature: -70°C
<b>Exclusion</b>					
Diagnosis	Myocardial infarction	Diagnosis	Myocardial infarction	N/A	N/A
Cohort exit	N/A	Event will persist until:	End of continuous observation	N/A	N/A
<b>Query 3</b>					
Cohort entry	N/A	Observation Period	Duration: >365 days	N/A	N/A
<b>Inclusion</b>					
Age	>65 years	Age	>65 years	Age	<18 years
Diagnosis	Essential (primary) hypertension	Diagnosis	Hypertensive disease	Diagnosis	Thyroid nodule
Biosamples	Serum	Lab values	LDL <sup>d</sup> cholesterol measurement: value >200	Biosamples	Tissue snap frozen
<b>Temporal constraints</b>					
Diagnosis period	Between 01/01/2020 and 04/30/2020	Diagnosis period	Between 01/01/2020 and 04/30/2020	Diagnosis period	Between 01/01/2020 and 04/30/2020
<b>Exclusion</b>					
Treatment	Lipid-lowering drugs	Treatment	Lipid-lowering drugs	Diagnosis	Concurrent diagnosis of thyroid cancer
Cohort exit	N/A	Event will persist until:	End of continuous observation	N/A	N/A

<sup>a</sup>i2b2: Informatics for Integrating Biology and the Bedside.

<sup>b</sup>N/A: not applicable; the criterion is not applicable for this tool.

<sup>c</sup>The task was to include this criterion with an OR operator.

<sup>d</sup>LDL: low-density lipoprotein.



During the task processing, the participants were asked to record their interactions on video and to express their thoughts (what causes them difficulties and what they like about the system) aloud (so-called *Thinking-Aloud* method) [18]. With the help of the screen recordings as well as the comments of the participants, which were made during the processing of the test, usability problems could be identified and positive or negative aspects of the interaction could be detected.

In addition, a web-based questionnaire was developed for the final assessment of usability. This questionnaire consisted of the following four parts (parts A-D):

- Parts A-C: three question blocks for assessing the usability of each query builder based on the (standardized) System Usability Scale (SUS) [19] and self-developed questions about satisfaction
- Part D: a final question block for a comparative rating of the query builders and for collecting demographic information (eg, age, gender, work experience, previous experience with queries and similar systems, and computer expertise).

The SUS questions and the supplementary questions on satisfaction were to be rated on a five-level rating scale (strongly disagree, disagree, neither agree nor disagree, agree, or strongly agree). For the questions about the person, the corresponding answer options had to be selected or certain blanks had to be filled in.

All test tasks and the web-based questionnaire were pretested in advance. A complete version of the questionnaire is provided in [Multimedia Appendix 1](#).

### Study Flow

The study material included an individualized test manual. It provided the framework and contained all the steps that needed to be taken to successfully conduct the study. The test was designed as an individual session at the workplace of the person (or alternatively in the home office), with a duration of approximately 90 minutes. As the harmonized tool to be developed should primarily address laypeople or casual users and as the evaluation focused on intuitive use, self-descriptiveness, and easy learnability of existing query builders, no training was conducted in advance with the participants. First, the participants were asked to install the screen recording software according to the instructions provided. Once the technology was established, the test subjects evaluated all tools while working through the respective test tasks ([Table 1](#)). The order in which the systems were to be tested was randomized to avoid bias caused by learning effects. The sequence of actions was recorded during the execution of the test tasks. In addition, the testers were asked to verbalize their thoughts about their individual steps in processing. After completion of the task complex of one tool, the subjects were asked to answer related usability and satisfaction questions immediately before continuing with the next system. When the test users encountered difficulties in accomplishing the tasks, the study material contained a rudimentary guide on how to use the tools. Finally, questions about the final and comparative ranking of the tools and about the person (demographic

information) were answered. Once the usability test was completed, the screen recording files had to be loaded into a secured cloud storage by each test subject.

### Data Analysis

#### *Analysis of Screen Recordings (Interactions and Expressions)*

The statements and actions recorded on the screen videos of all test persons were transcribed per system by 2 members of the study team. The protocols were subsequently mutually validated. The transcripts were then scanned by these 2 members for negative aspects or problem areas and positive aspects. Subsequently, all problems or positive statements were collected in an overall list (ie, if a problem was named several times by different test subjects or if it occurred across all test tasks, it was noted as one problem). Each problem was evaluated and rated by 2 independent raters in terms of its severity according to the Nielsen and Mack [20] severity rating, ranging from 0 (no usability problem) to 4 (usability catastrophe). Rating differences between the 2 evaluators were discussed until a consensus was reached.

Furthermore, the problems were classified according to Zapf error taxonomy [21] into use problems (resulting from a lack of fit between user and software) or functional problems (incomplete or missing functionality of a system), as follows:

- Examples of use problems: errors of knowledge, errors of thinking, errors of memory and forgetting, errors of judgment, errors of habit, errors of omission, errors of recognition, and errors of movement
- Examples of functional problems: action blockades, action repetitions, action interruptions, and alternative course of action.

In addition, videos were used to determine how successfully the respective test person completed the tasks. A test task was considered correct if all parameters were entered and if they were correctly linked in the system. A task was considered incorrect if the parameters were incomplete, the link between the parameters was incorrect, or both situations occurred.

#### *Analysis of the Web-Based Questionnaire (Usability and Satisfaction Ratings and Demographic Information)*

The questions of the SUS were analyzed using the scoring method by Brooke [19], which yields possible values from 0 to 100 and allows the values to be compared with the values of a grading scale, where 0 represents an unacceptable usability and 100 represents the best imaginable usability. The additionally formulated questions on satisfaction with the query builder were converted into a numerical scale ranging from 1 (strongly disagree) to 5 (strongly agree). For descriptive analysis, mean scores and SDs were calculated. For the demographic questions (depending on the question type), the percentage was calculated, mean values and SDs were determined, or the free text was analyzed. For open-ended answers (free text), thematic categories were defined, and the answers of the test persons were assigned to these categories. Cases with missing values were deleted from the list. The Wilcoxon rank sum test was used to statistically compare the questionnaire results between

the three query builders. The Pearson correlation was calculated to analyze whether demographic variables had an influence on the evaluation results. The significance level was set at  $P < .05$ . All statistical analyses were performed using SPSS 27.0 (IBM Corporation).

## Results

### Participant Characteristics

Of the 30 potential study participants, 17 (57%) responded. Due to the early termination of the study and the testing of only one query builder, 1 participant had to be excluded. Thus, the data from 16 participants were analyzed. The participants had an average age of 38.13 years (SD 9.68) and about two-thirds of

the subjects were male (10/16, 63%). The average work experience was 10.37 years (SD 10.86). The majority of the participants worked in clinical research or as research assistants (13/16, 81%), whereas 2 participants assigned themselves to other professional groups (professor and quality manager). As far as computer skills are concerned, everyone rated themselves well; either they said that they could handle most systems properly (8/16, 50%) or that they had a significant amount of experience and were technically proficient (7/16, 44%). Only 3 persons stated having previous experience with systems similar to those tested in the usability evaluation (3/16, 19%). In general, less than half of the respondents stated that they had general experience with requesting case numbers for clinical studies (little experience: 5/16, 31% or a lot of experience: 2/16, 13%). The full sample characteristics are presented in [Table 2](#).

**Table 2.** Characteristics of the participants (N=16). Summarized number and percentage per category. For age and work experience mean and SD were calculated.

Variable	Values
<b>Age</b>	
<b>Answered, n (%)</b>	15 (94)
Age (years), mean (SD)	38.13 (9.680)
No answer, n (%)	1 (6)
<b>Gender, n (%)</b>	
Male	10 (63)
Female	5 (31)
No answer	1 (6)
<b>Native language, n (%)</b>	
German	14 (88)
Other: Hungarian	1 (6)
No answer	1 (6)
<b>Difficulties regarding English, n (%)</b>	
Never	7 (44)
Rarely	7 (44)
Sometimes	1 (6)
No answer	1 (6)
<b>Professional group, n (%)</b>	
Clinical researcher	6 (38)
Scientific assistant	7 (44)
Other: professor or quality manager	2 (12)
No answer	1 (6)
<b>Work experience</b>	
<b>Answered, n (%)</b>	13 (81)
Work experience (years), mean (SD)	10.37 (10.861)
No answer, n (%)	3 (19)
<b>Experience with feasibility studies, n (%)</b>	
No experience or little experience	8 (50)
Some experience	5 (31)
Much experience	2 (13)
No answer	1 (6)
<b>Use of similar systems in the past, n (%)</b>	
No	12 (75)
Yes	3 (18)
No answer	1 (6)
<b>Computer skills, n (%)</b>	
Average computer skills	8 (50)
Excellent computer skills	7 (44)
No answer	1 (6)

## Think-Aloud Test Results

### *Negative and Positive Design Aspects*

The evaluation of the material generated by the *Thinking-Aloud* method revealed concrete usability problems. Classification according to the severity scale produced the following result: ATLAS had the most usability problems, with 66 problems noted. These were divided into 21 major problems, 30 minor problems, and 15 cosmetic problems. A major problem was the function for saving:

*That's where it starts: How and where to save? I don't know. Where is it stored here? I have no idea.*

With i2b2, the 48 detected problems were divided into 9 major, 26 minor, and 13 cosmetic problems. Among other things, it was noted that the procedure for defining an exclusion criterion is not clear. The Sample Locator had the lowest number of problems, with 22 problems noted. In contrast to the other tools, however, there are also two problems with the level usability catastrophe. Furthermore, 4 major, 10 minor, and 6 cosmetic problems were identified. One of the usability disasters concerned the *AND* or *OR* combination of the individual criteria. The logic behind this was often not obvious to the users, which became apparent from their comments:

*The question is, how do you represent this "OR" connection here. This is not quite clear now.*

*So, a bit unclear to be honest, whether this is "AND" or "OR."*

In addition to the critical aspects, some positive points and suggestions for improvement could be extracted. In the case of ATLAS, a large number of possible options and settings were highlighted as positive. The same applies to the visualization of the results, which are displayed in the form of a colored square with subareas for the selected criteria. As starting points for the improvement of the handling of a suggestion list for the input of criteria, the specification of units as well as the

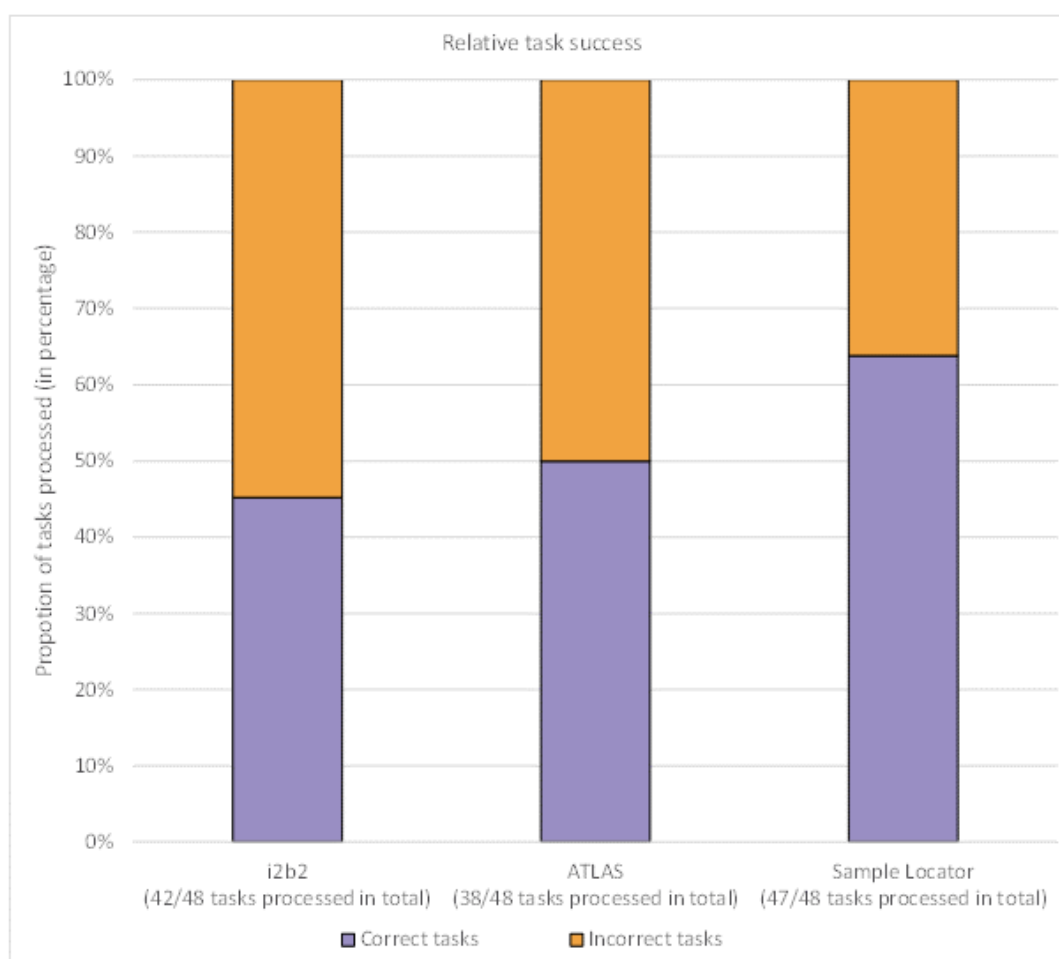
plausibility check before the start of the search was mentioned. i2b2 was able to convince with its intuitive operating concept, where the elements can be easily assigned to the groups via the drag-and-drop function. In addition, the automatically appearing input window for values, as soon as a criterion was selected, was considered supportive. However, it should be possible to select criteria not only from an ontology tree but also via a free-text search. With the Sample Locator, a more comprehensive arrangement of the criteria was desired. For example, the division into *Donor and Clinical Information* and *Sample* and their meaning was not immediately obvious, and the order of the individual criteria could also be improved, for example, thematically related criteria were placed one below the other. However, the Sample Locator was found to be very clear, straightforward, and intuitive to use, so the tasks were "nice and also very easy to implement". [Multimedia Appendix 2](#) shows the most serious usability problems (severity ratings of 3 and 4) of the respective query builders with the corresponding number of participants who have named this problem and the resulting optimization recommendations.

### *Task Success*

Of the 48 tasks evaluated per tool, 47 were completed in Sample Locator, with 30 of them processed correctly and 17 of them not processed correctly. Nineteen tasks were completed successfully for both ATLAS and i2b2. On the contrary, 19 and 23 tasks could not be executed correctly with ATLAS and i2b2, respectively. In the case of i2b2, 6 tasks were not processed, and in the case of ATLAS, 10 tasks were missing. Overall, the Sample Locator scored the best overall test items in terms of absolute correctness. Considering the correctness of the task processing relative to all finished tasks across all respondents, there was no significant difference between the query builders (Wilcoxon test; i2b2 vs Sample Locator:  $P=.07$ ; i2b2 vs ATLAS:  $P=.72$ ; ATLAS vs Sample Locator:  $P=.06$ ). The false or unprocessed tasks were relatively evenly distributed among the three tasks in i2b2 and ATLAS ([Figure 4](#)).



**Figure 4.** Relative task success for the three query builders. Success was denoted when all required parameters were entered and linked correctly. i2b2: Informatics for Integrating Biology and the Bedside.



## Questionnaire Results

### Results of the SUS

The overall evaluation of usability based on the SUS showed that the Sample Locator, with a mean SUS score of 77.03 (SD 20.62), was significantly superior to the other two tools (Wilcoxon test; Sample Locator vs i2b2:  $P=.047$ ; Sample Locator vs ATLAS:  $P=.001$ ). However, i2b2, with a score of 59.83 (SD 25.36), still performed significantly better than ATLAS, which had a score of 27.81 (SD 21.79; Wilcoxon test; i2b2 vs ATLAS:  $P=.005$ ). For reasons of comprehensibility

(positive and negative usability aspects), the individual results of the SUS are presented in Table 3. However, only the overall SUS score can be interpreted as a measure of usability. Using the Pearson correlation, no association between SUS scores and the personal variables age (correlation age-SUS; i2b2:  $P=.87$ ; ATLAS:  $P=.14$ ; Sample Locator:  $P=.66$ ), gender (correlation gender-SUS; i2b2:  $P=.44$ ; ATLAS:  $P=.85$ ; Sample Locator:  $P=.20$ ), work experience (correlation work experience-SUS; i2b2:  $P=.95$ ; ATLAS:  $P=.32$ ; Sample Locator:  $P=.40$ ), and previous experience with cohort research (correlation experience cohort research-SUS; i2b2:  $P=.70$ ; ATLAS:  $P=.58$ ; Sample Locator:  $P=.60$ ) was evident.

**Table 3.** Results of the SUS<sup>a</sup>. Mean ratings from 1 (strongly agree) to 5 (strongly disagree) and SDs are presented.

SUS item	i2b2 <sup>b</sup> (n=15), mean (SD)	ATLAS (n=16), mean (SD)	Sample Locator (n=16), mean (SD)
I think that I would like to use this query builder frequently.	3.60 (1.242)	2.00 (1.000)	3.75 (1.125)
I found this query builder unnecessarily complex. <sup>c</sup>	3.33 (1.234)	1.75 (0.829)	4.44 (0.892)
I thought this query builder was easy to use.	3.13 (1.187)	1.94 (1.029)	4.31 (0.873)
I think that I would need the support of a technical person to be able to use this query builder. <sup>c</sup>	3.53 (1.125)	2.81 (1.333)	4.44 (0.727)
I found the various functions in this query builder were well integrated.	3.27 (1.223)	2.25 (1.031)	3.69 (1.138)
I thought there was too much inconsistency in this query builder. <sup>c</sup>	3.60 (0.737)	2.75 (1.090)	3.75 (1.000)
I would imagine that most people would learn to use this query builder very quickly.	3.27 (1.387)	1.69 (0.982)	4.06 (0.998)
I found this query builder very cumbersome to use. <sup>c</sup>	3.20 (1.424)	1.81 (0.882)	4.13 (1.408)
I felt very confident using this query builder.	3.33 (1.047)	1.88 (0.927)	3.75 (0.856)
I needed to learn a lot of things before I could get going with this query builder. <sup>c</sup>	3.67 (1.113)	2.31 (1.261)	4.50 (0.816)

<sup>a</sup>SUS: System Usability Scale.

<sup>b</sup>i2b2: Informatics for Integrating Biology and the Bedside.

<sup>c</sup>Reverse-coded item.

### **Results of the Additionally Formulated Questions on Satisfaction**

The additional questions regarding the satisfaction with the tools confirm the outcome of the SUS. The Sample Locator achieves the highest ratings, with the exception of the subjectively perceived working speed, which was felt to be the least slowed down with i2b2. The test participants were also satisfied with i2b2, but the Sample Locator was rated

significantly more positively with regard to the presentation of query results (Wilcoxon test;  $P=.03$ ), the ability to undo operating steps (Wilcoxon test;  $P=.01$ ), navigation within the tool (Wilcoxon test;  $P=.005$ ), presentation of information (clarity; Wilcoxon test;  $P=.04$ ), and visual design (Wilcoxon test;  $P=.02$ ). For ATLAS, with the exception of the item *possibility of undoing task steps*, all ratings were generally in the negative range in every aspect of satisfaction (Table 4).

**Table 4.** Results of the satisfaction rating. Mean ratings from 1 (strongly agree) to 5 (strongly disagree) and SDs.

Satisfaction with the query builder	i2b2 <sup>a</sup> (n=15), mean (SD)	ATLAS (n=16), mean (SD)	Sample Locator (n=16), mean (SD)
I am satisfied with the ease with which the tasks can be accomplished.	3.20 (1.265)	1.81 (0.808)	4.06 (1.063)
I am satisfied with the time it takes to complete the tasks.	3.60 (1.242)	1.75 (1.031)	4.31 (0.793)
I am satisfied with the functionality that is provided to complete the tasks.	3.27 (1.280)	2.38 (1.317)	3.69 (1.014)
The terms and designations used in the query builder (eg, for the selection options and for patient characteristics) are immediately understandable to me.	4.00 (0.756)	2.19 (1.130)	4.25 (0.683)
The query builder enables me to complete work steps (eg, the selection of certain clinical or temporal parameters) in the order that seems to make the most sense to me.	3.53 (1.407)	2.88 (1.218)	3.88 (1.147)
The results generated with the query builder are displayed or output in such a way that they meet my requirements (eg, through clear grouping and an attractive visualization).	2.80 (1.265)	2.19 (1.073)	3.56 (1.094)
It is immediately apparent to me which consequences my input in the query builder has.	3.13 (1.302)	1.81 (0.882)	3.19 (1.223)
The query builder offers me the possibility to undo work steps if it is appropriate for my task completion.	3.93 (0.884)	3.88 (0.857)	4.63 (0.619)
I found the navigation within the query builder easy.	3.40 (1.242)	1.75 (0.901)	4.44 (0.892)
I found the information displayed in the query builder to be clear and concise.	3.13 (1.187)	1.81 (0.808)	4.00 (1.317)
The user interface of the query builder is visually appealing.	2.80 (1.424)	2.50 (1.118)	4.19 (1.223)
During my work with the query builder, errors occurred (eg, that options could not be combined and that exclusion criteria did not work). <sup>b</sup>	3.47 (1.552)	2.69 (1.102)	4.19 (1.047)
I sometimes felt slowed down in my work speed by the query builder (eg, by too long waiting times). <sup>b</sup>	4.00 (1.134)	2.63 (1.317)	3.69 (1.352)

<sup>a</sup>i2b2: Informatics for Integrating Biology and the Bedside.

<sup>b</sup>Reverse-coded item.

## Discussion

### Overview

The motivation for this study was to compare three different feasibility platforms and to answer the following questions: (1) which of the systems is best suited as a basis for further development and (2) which positive aspects can be taken over from the other tools for the purpose of more user-friendliness. This paper not only discusses the answers to these questions but also illustrates the approach to achieve this.

### Discussion of Methods

To answer the research questions, a web-based usability test with end users and the established usability methods *Thinking-Aloud* and the questionnaire based on the standardized SUS was chosen as the methodological design.

An advantage of web-based usability testing is the independence of time and place with which such tests can be performed, and no extra test or observation room is required. Web-based testing is a very time-efficient method that allows several people to test a system at the same time. However, this method also has disadvantages: observers have no real-time access to data, and there is no possibility of interacting with the user during data collection [22]. However, studies show that a remote test provides as valid results as a laboratory test: Tullis et al [23], for example, presented results that show high correlations

between laboratory and remote tests for task completion data and task time data. The most critical usability problems were identified using both the techniques. In a study conducted by Andreasen et al [24], three methods for remote usability testing and a traditional laboratory-based *Thinking-Aloud* method were compared. These results also show that the remote method is equivalent to the traditional laboratory method. Therefore, our choice of a web-based test can be considered equivalent to a usability study conducted in the laboratory.

For our web-based usability test, we chose the methods *Thinking-Aloud* and questionnaires. The *Thinking-Aloud* method allowed us to find out what potential users actually think about the query builder. In particular, it enabled us to identify usability problems that could lead to feasible redesign recommendations. We learned why some parts of the user interface are difficult to use and which areas of the tools are easy and intuitive for the user. Advantages of this method are that no special equipment is required for this method, it does not take a lot of time, and data can be collected very quickly, which is sufficient for the most important insights. Furthermore, this method is independent of the level of technical experience of the test persons and can be used for any type of user interface. A disadvantage of the *Thinking-Aloud* method, however, is that it is generally not suitable for detailed statistics. In addition, the situation of constantly expressing thoughts is very unnatural, which makes it difficult for the test person to maintain the

required monolog [25]. Thus, we could also observe that some of our test participants temporarily forgot to verbalize their thoughts. To compensate for the disadvantages of this method, we combined it with a questionnaire on usability and satisfaction with the query builders.

Questionnaires have the advantage that numerous data can be obtained with relatively little effort. The use of standardized questionnaires also supports the objectivity of data collection and allows comparisons between systems [26]. In particular, the SUS is a very reliable questionnaire that detects differences even in small sample sizes. In addition, it has been shown that SUS can effectively distinguish between systems with low and high usability and also correlates to a high degree with other questionnaire-based usability measurement tools [27]. However, questionnaires such as SUS are not suitable for diagnosing usability problems and gathering background information to understand why users evaluate a system in this way. In addition, the relevant aspects of the given questions may be lost. However, we were able to compensate for this disadvantage by using the *Thinking-Aloud* method in combination.

In summary, by combining different methods, the advantages and disadvantages of the respective methods can be balanced and a comprehensive opinion can be obtained. According to Sarodnick and Brau [28], the combination of observation and spontaneous expression of thoughts ensures a high validity of the data.

## Discussion of Results

Our first research question should answer which of the three query builders is the most usable. The results of our study show that each of the evaluated systems has usability shortcomings and thus offers room for improvement. However, overall, the Sample Locator was rated as the tool with the highest usability. The analysis of the screen videos for this tool showed the fewest usability problems, and the questionnaire data showed the highest SUS score (SUS score: 77.03) for this tool. i2b2 is in second place, with a SUS score of 59.83. ATLAS was rated the worst; this system only achieved a SUS score of 27.81, and it was difficult to use from the perspective of the test subjects. The main reason for this difference in evaluation can be found in the complexity of the systems and the target group addressed by these tools: The Sample Locator is aimed at scientists and medical researchers who search for biosamples in academic biobanks. The selection of search criteria for samples is limited in the Sample Locator, so the Sample Locator is a very simple search tool. ATLAS is primarily designed for researchers and experts who need to assemble very complex cohort queries. Therefore, the variety of selection and input options is much higher, which also increases the complexity of operation. i2b2 offers a compromise between these systems. For the goal of the development in MIRACUM, it was asked which tool offers the best integration basis. The answer in this respect is that, of these three tools, the Sample Locator is the most user-friendly from the user's point of view and is therefore considered the best basis for developing a tool for feasibility studies.

The second research question is related to the negative and positive design aspects of the three systems. The strengths of the Sample Locator were mainly its esthetic, minimalist design

and the resulting clarity, the easy input of parameters, and the intuitive navigation. The main disadvantage was that it was not obvious in which logical way the parameters were linked after input. In addition, when entering the age of the donor, it was not clear why an input option for this was available in the two areas *Sample* and *Donor and Clinical Information* and what the difference of the selection option was. In addition, the Sample Locator had only limited functionality. For example, complex periods could not be defined or a concrete storage temperature could not be entered. For further development of the tool, it is recommended to address these usability problems.

i2b2 proved to be very intuitive to use with its *drag-and-drop* operating concept and offered a good and simple way of selecting parameters via the menu tree. However, even with this tool, the parameters were not always linked correctly, despite the short text-bright hints. One reason for this was that users would expect parameters to be linked with *AND* within a field and *OR* between fields. In fact, the opposite was true. Moreover, the display of the results proved to be unfavorable because test persons would not expect to have to select the *Refresh* option actively and repeatedly themselves to get an up-to-date display of the results. It should be noted, however, that the result display inherent in i2b2 is not used in the MIRACUM i2b2 context but in a connected project management tool. Thus, the result display is not a mere usability problem of the i2b2 feasibility tool. Due to the completeness and comparability between the tools, this issue was nevertheless considered in the assessment of i2b2.

From the point of view of the test subjects, ATLAS offered the greatest variety of input and selection options, but this made it difficult to keep relevant information and options clearly arranged and easily recognizable. It was also unclear to the test subjects why the selection of demographic parameters (eg, age and gender) followed a different selection principle than other parameters (eg, diagnoses or therapy). It was also not understandable why an exclusion criterion would have to be defined as a reverse inclusion in the *Inclusion Criteria* area. Most of the test subjects also failed to recognize how to start a search, as they did not link the *Generation* tab with a search option.

To the best of our knowledge, no study has previously compared the usability of query builders for feasibility studies. However, we were able to identify some studies that tested individual query builders with regard to their usability, which can be discussed with the partial results of our study:

- A usability study by Schüttler et al [29] with 27 participants rated a mock-up version during the development phase of the Sample Locator as intuitive and user-friendly. The mean SUS score of the Sample Locator was 80.4, indicating good usability. Our study showed a similarly high SUS score of 77.03, which also indicates good usability of this tool and supports the results of the study by Schüttler et al [29].
- A usability survey of the Criteria2Query tool, which performs queries in the ATLAS web application, revealed that almost half of the participants considered it difficult to perform the task of cohort definition (eg, identifying queryable eligibility concepts) [30]. Our study comes to a



similar conclusion for the web tool ATLAS. The main reason for this is the complexity of the tool.

- A usability study of the EHR4CR (Electronic Health Records for Clinical Research) multisite patient count cohort system with 22 testers resulted in a SUS score of 55.83 (SD 15.37), indicating a low user satisfaction. The authors of the study stated that test subjects had problems, especially with complex queries [8]. We report similar results for all three tested query builders. In particular, queries that asked for *OR* or *NOT* links and a time constraint caused usability problems for the participants.
- An evaluation of a web application for cohort identification and data extraction revealed usability problems such as a missing *undo* function, which means that users could not directly return to the input mask to modify a query [31]. This was also one of the main problems reported with the i2b2 tool.

In summary, it can be said that individual studies come to a similar SUS assessment of the query tools and have reported similar operating problems in individual cases. However, this study, with its comparative design, represents the most comprehensive and systematic usability evaluation to date.

### Limitations

This usability study followed a comprehensive approach to compare the three query builders. However, some limitations must be considered when interpreting the results. Our results refer to a specific target group (researchers in Germany as laypersons or occasional users) and a specific context of use (feasibility queries of medium complexity regarding the general availability of patients or biosamples). The results can, therefore, not be transferred to other people (query experts and trained users who routinely use such tools) or a different objective (complex feasibility queries with the aim of finding very specific patients or biosamples). As other usability problems may arise for different contexts of use, our results are not generalizable. Furthermore, the selection of the three query builders was not based on an extensive analysis of the existing tools. Due to the limited timeframe of the project, this was dispensed with, and the focus was placed on tools that had already been used in other contexts in the project or were known from cooperation with other initiatives. Moreover, it was ensured that the tools are without exception, subject to an open source license, so that the most suitable tool can be used as a foundation for further development, if appropriate. A further limitation concerned the tasks, as the queries were designed for a test environment. To ensure functional comparability between the tools, the complexity of the tasks was based on the simplest tool. Although this did not allow all the functionalities of the other two tools to be fully exploited, care was taken to ensure that the tasks reflected real feasibility queries and thus covered all the required functions. The study procedure required each participant to evaluate all the three tools. Although the order of the test items was randomized to minimize bias due to learning effects, these cannot be completely avoided. However, owing to the strong differences in the basic operating concepts, we assume that such an effect is marginal. As the usability test was not conducted at only one location, it was not feasible to create one identical

test scenario for all participants. The study was carried out by the participants for the most part at their workplace or, in rare cases, in their home office. However, we believe that this corresponds to a more realistic scenario than a laboratory setting, so that the insights gained are more informative. With regard to the tools, it should be mentioned that both ATLAS and the Sample Locator were provided in English only. Here, it must be taken into account that nonnative speakers may find some terms or options difficult to understand. In the case of this study, however, the majority of participants indicated that they had no difficulties with the English language (Table 2), so no bias due to comprehension deficits was expected. Finally, the relatively low number of participants (n=16) might be considered as a methodological weakness. During recruitment, the search for a population of suitable researchers and physicians with no profound knowledge of the tools was a bottleneck; therefore, we were unable to reach the targeted 30 subjects. However, this did not diminish the significance of the results. Kuric et al [32] showed that a sample size of approximately 15 participants yielded good results for the comparative usability evaluation of query builders; therefore, 16 participants were considered sufficient.

### Conclusions

This study provides data to develop a suitable basis for the selection of a harmonized tool for feasibility studies by concrete evaluation and comparison of the usability of three different types of query builders. The feedback of the participants during the usability test made it possible to identify user problems and positive design aspects of the individual tools and to compare them qualitatively. As a result, comparatively, the Sample Locator is the tool with the best usability, that is, the most positive ratings and the lowest number of usability problems. To create a harmonized tool, this tool is therefore considered the most suitable starting point. The Sample Locator outweighs the other tools in terms of the visual design of the user interface, clear and concise presentation of information, navigation, and presentation of results. For further development of the tool, there is a need to revise the display (visibility) of logical links, the provision of selection fields for diagnostic information, a clearer name or area placement of the *Donor Age* selection option, and a clearer presentation of the options for specifying a diagnostic period. Nevertheless, because of the current limitation of supporting only a small set of selection criteria and because, for example, no time constraints are possible, its scalability with respect to much more comprehensive sets of filter criteria (eg, with large possible value lists) and more complex Boolean expressions (including time constraints and dependencies) needs to be carefully considered. The results of our study provide valuable insights for researchers and developers of similar projects, whereby our methodological approach can be used as a blueprint. Our next step will be to apply the findings of this research to develop a harmonized feasibility platform. Although only laypersons were considered in this usability study, this tool could also be expanded in the future to include functions for experts to address an even broader user group. To obtain a holistic picture, experts who are already familiar with the field of feasibility queries will also be consulted.

## Acknowledgments

The authors would like to thank all participating MIRACUM locations—Dresden, Erlangen, Frankfurt am Main, Freiburg, Gießen, Greifswald, Magdeburg, and Mannheim. The authors would like to specially thank all scientists and researchers who participated in this study and provided a valuable insight into the usability of the different query builders through their *loud thoughts* and questionnaire evaluations. The authors would also like to thank Stefanie Schild (Erlangen), Renate Häuslschmid (Freiburg), and Preetha Moorthy (Mannheim) for their support in pretesting the tasks and questionnaires. This study was conducted as part of MIRACUM. MIRACUM is funded by the German Federal Ministry of Education and Research within the Medical Informatics Funding Scheme under the funding codes 01ZZ1801A and 01ZZ1801L. This work is additionally supported by the German Federal Ministry of Education and Research under the funding code 01EY1701.

The present work was performed in fulfillment of the requirements for obtaining the degree “Dr. rer. biol. hum.” from the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (CS).

## Authors' Contributions

CS wrote the first version of this manuscript. BS and CS planned the usability study, which was supervised by HUP and MS. CS and HUP were responsible for the recruitment of the test persons. CS and BS transcribed the statements and actions of all recorded screen videos and analyzed the study data. All authors read the first version of the manuscript and provided valuable suggestions for changes.

## Conflicts of Interest

None declared.

### Multimedia Appendix 1

Web-based questionnaire for usability evaluation (translated from German to English).

[[DOCX File , 24 KB - medinform\\_v9i7e25531\\_app1.docx](#) ]

### Multimedia Appendix 2

Identified usability problems with the highest severity rating (severity of 3 and 4) for the three query builders.

[[DOCX File , 40 KB - medinform\\_v9i7e25531\\_app2.docx](#) ]

## References

1. Maissenhaelter BE, Woolmore AL, Schlag PM. Real-world evidence research based on big data: motivation-challenges-success factors. *Onkologie (Berl)* 2018 Jun 7;24(Suppl 2):91-98 [[FREE Full text](#)] [doi: [10.1007/s00761-018-0358-3](https://doi.org/10.1007/s00761-018-0358-3)] [Medline: [30464373](#)]
2. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W, (on behalf of GetReal Work Package 1). What is real-world data? A review of definitions based on literature and stakeholder interviews. *Value Health* 2017 Jul;20(7):858-865 [[FREE Full text](#)] [doi: [10.1016/j.jval.2017.03.008](https://doi.org/10.1016/j.jval.2017.03.008)] [Medline: [28712614](#)]
3. European Health Data & Evidence Network (EHDEN). URL: <https://www.ehden.eu/> [accessed 2020-10-23]
4. Swiss Personalized Health Network (SPHN). URL: <https://sphn.ch/> [accessed 2020-10-23]
5. Health Research Infrastructure (Health RI). URL: <https://www.health-ri.nl/> [accessed 2020-10-23]
6. Semler SC, Wissing F, Heyder R. German medical informatics initiative. *Methods Inf Med* 2018 Jul;57(S 01):50-56 [[FREE Full text](#)] [doi: [10.3414/ME18-03-0003](https://doi.org/10.3414/ME18-03-0003)] [Medline: [30016818](#)]
7. Bache R, Miles S, Taweel A. An adaptable architecture for patient cohort identification from diverse data sources. *J Am Med Inform Assoc* 2013 Dec;20(e2):327-333 [[FREE Full text](#)] [doi: [10.1136/amiainl-2013-001858](https://doi.org/10.1136/amiainl-2013-001858)] [Medline: [24064442](#)]
8. Soto-Rey I, N'Dja A, Cunningham J, Neue A, Trinczek B, Lafitte C, et al. User satisfaction evaluation of the EHR4CR query builder: a multisite patient count cohort system. *Biomed Res Int* 2015;2015:801436 [[FREE Full text](#)] [doi: [10.1155/2015/801436](https://doi.org/10.1155/2015/801436)] [Medline: [26539525](#)]
9. Prokosch HU, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, et al. MIRACUM: Medical Informatics in Research and Care in University Medicine. *Methods Inf Med* 2018 Jul;57(S 01):82-91 [[FREE Full text](#)] [doi: [10.3414/ME17-02-0025](https://doi.org/10.3414/ME17-02-0025)] [Medline: [30016814](#)]
10. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* 2010;17(2):124-130 [[FREE Full text](#)] [doi: [10.1136/jamia.2009.000893](https://doi.org/10.1136/jamia.2009.000893)] [Medline: [20190053](#)]
11. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015;216:574-578 [[FREE Full text](#)] [Medline: [26262116](#)]

12. Observational Health Data Sciences and Informatics (OHDSI). ATLAS. URL: <https://atlas.ohdsi.org/> [accessed 2020-10-23]
13. German Biobank Alliance (GBA). Sample Locator. URL: <https://samplelocator.bbMRI.de/search> [accessed 2020-10-23]
14. Baber R, Hummel M, Jahns R, von Jagwitz-Biegnitz M, Kirsten R, Klingler C, et al. Position statement from the German biobank alliance on the cooperation between academic biobanks and industry partners. *Biopreserv Biobank* 2019 Aug;17(4):372-374 [FREE Full text] [doi: [10.1089/bio.2019.0042](https://doi.org/10.1089/bio.2019.0042)] [Medline: [31314575](https://pubmed.ncbi.nlm.nih.gov/31314575/)]
15. Schüttler C, Buschhüter N, Döllinger C, Ebert L, Hummel M, Linde J, et al. [Requirements for a cross-location biobank IT infrastructure : survey of stakeholder input on the establishment of a biobank network of the German Biobank Alliance (GBA)]. *Pathologie* 2018 Jul;39(4):289-296. [doi: [10.1007/s00292-018-0435-9](https://doi.org/10.1007/s00292-018-0435-9)] [Medline: [29691676](https://pubmed.ncbi.nlm.nih.gov/29691676/)]
16. Alroobaea R, Mayhew PJ. How many participants are really enough for usability studies? In: Proceedings of the Science and Information Conference. 2014 Presented at: Science and Information Conference; August 27-29, 2014; IEEE, London p. 27-29. [doi: [10.1109/SAI.2014.6918171](https://doi.org/10.1109/SAI.2014.6918171)]
17. Brat GA, Weber GM, Gehlenborg N, Avillach P, Palmer NP, Chiovato L, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *NPJ Digit Med* 2020;3:109 [FREE Full text] [doi: [10.1038/s41746-020-00308-0](https://doi.org/10.1038/s41746-020-00308-0)] [Medline: [32864472](https://pubmed.ncbi.nlm.nih.gov/32864472/)]
18. Jaspers MW, Steen T, van den Bos C, Geenen M. The think aloud method: a guide to user interface design. *Int J Med Inform* 2004 Nov;73(11-12):781-795. [doi: [10.1016/j.ijmedinf.2004.08.003](https://doi.org/10.1016/j.ijmedinf.2004.08.003)] [Medline: [15491929](https://pubmed.ncbi.nlm.nih.gov/15491929/)]
19. Brooke J. SUS: A 'Quick and Dirty' usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. *Usability Evaluation in Industry*. London, England: Taylor and Francis; 1996.
20. Nielsen J, Mack RL. *Usability Inspection Methods*. New York, United States: John Wiley & Sons; 1994:1-448.
21. Zapf D, Brodbeck FC, Prümper J. Handlungsorientierte Fehlertaxonomie in der Mensch - Computer - Interaktion. *Zeitschrift für Arbeits- und Organisationspsychologie*. 1989. URL: [https://people.f3.htw-berlin.de/Professoren/Pruemper/publikation/1989/Zapf%20Brodbeck\\_Pruemper\(1989\).pdf](https://people.f3.htw-berlin.de/Professoren/Pruemper/publikation/1989/Zapf%20Brodbeck_Pruemper(1989).pdf) [accessed 2021-06-26]
22. Bastien JM. Usability testing: a review of some methodological and technical aspects of the method. *Int J Med Inform* 2010 Apr;79(4):18-23. [doi: [10.1016/j.ijmedinf.2008.12.004](https://doi.org/10.1016/j.ijmedinf.2008.12.004)] [Medline: [19345139](https://pubmed.ncbi.nlm.nih.gov/19345139/)]
23. Tullis T, Fleischman S, McNulty M, Cianchette C, Bergel M. An empirical comparison of lab and remote usability testing of web sites. Usability Professional Association Conference, Orlando. 2002. URL: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.457.3080&rep=rep1&type=pdf> [accessed 2021-06-26]
24. Andreasen M, Nielsen H, Schrøder SO, Stage J. What happened to remote usability testing? An empirical study of three methods. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2007 Presented at: CHI07: CHI Conference on Human Factors in Computing Systems; April 28-May 3, 2007; San Jose California USA p. 1405-1414. [doi: [10.1145/1240624.1240838](https://doi.org/10.1145/1240624.1240838)]
25. Nielsen J, Clemmensen T, Yssing C. Getting access to what goes on in people's heads? - Reflections on the think-aloud technique. In: Proceedings of the Second Nordic Conference on Human-computer Interaction. 2002 Presented at: NORDICHI02: NORDICHI 2002, The second Nordic conference on Human-Computer Interaction; October 19-23, 2002; Aarhus Denmark p. 101-110. [doi: [10.1145/572020.572033](https://doi.org/10.1145/572020.572033)]
26. Sauro J, Lewis JR. Standardized usability questionnaires. In: Sauro J, Lewis JR, editors. *Quantifying the User Experience*. New York: Morgan Kaufmann; 2012:185-240.
27. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. *Int J Hum-Comput Int* 2008 Jul 30;24(6):574-594. [doi: [10.1080/10447310802205776](https://doi.org/10.1080/10447310802205776)]
28. Sarodnick F, Brau H. *Methoden Der Usability Evaluation: Wissenschaftliche Grundlagen und Praktische Anwendung*. Bern: Hogrefe; 2016.
29. Schüttler C, Huth V, von Jagwitz-Biegnitz M, Lablans M, Prokosch HU, Griebel L. A federated online search tool for biospecimens (sample locator): usability study. *J Med Internet Res* 2020 Aug 18;22(8):e17739 [FREE Full text] [doi: [10.2196/17739](https://doi.org/10.2196/17739)] [Medline: [32663150](https://pubmed.ncbi.nlm.nih.gov/32663150/)]
30. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. *J Am Med Inform Assoc* 2019 Apr 01;26(4):294-305. [doi: [10.1093/jamia/ocy178](https://doi.org/10.1093/jamia/ocy178)] [Medline: [30753493](https://pubmed.ncbi.nlm.nih.gov/30753493/)]
31. Horvath MM, Winfield S, Evans S, Slopek S, Shang H, Ferranti J. The DEDUCE Guided Query tool: providing simplified access to clinical data for research and quality improvement. *J Biomed Inform* 2011 Apr;44(2):266-276 [FREE Full text] [doi: [10.1016/j.jbi.2010.11.008](https://doi.org/10.1016/j.jbi.2010.11.008)] [Medline: [21130181](https://pubmed.ncbi.nlm.nih.gov/21130181/)]
32. Kuric E, Fernández JD, Drozd O. Knowledge graph exploration: a usability evaluation of query builders for laypeople. In: Acosta M, Cudré-Mauroux P, Maleshkova M, Pellegrini T, Sack H, Sure-Vetter Y, editors. *Semantic Systems. The Power of AI and Knowledge Graphs*. Switzerland: Springer; 2019:326-342.

## Abbreviations

- EHR4CR:** Electronic Health Records for Clinical Research
- FAU:** Friedrich-Alexander-Universität Erlangen-Nürnberg
- GBA:** German Biobank Alliance
- i2b2:** Informatics for Integrating Biology and the Bedside

**MII:** Medical Informatics Initiative

**MIRACUM:** Medical Informatics in Research and Care in University Medicine

**OHDSI:** Observational Health Data Sciences and Informatics

**OMOP:** Observational Medical Outcomes Partnership

**SUS:** System Usability Scale

*Edited by C Lovis; submitted 06.11.20; peer-reviewed by I Soto-Rey, A Mahnke, C Reich, M Schuemie; comments to author 09.01.21; revised version received 18.01.21; accepted 17.05.21; published 21.07.21.*

*Please cite as:*

*Schüttler C, Prokosch HU, Sedlmayr M, Sedlmayr B*

*Evaluation of Three Feasibility Tools for Identifying Patient Data and Biospecimen Availability: Comparative Usability Study*

*JMIR Med Inform 2021;9(7):e25531*

URL: <https://medinform.jmir.org/2021/7/e25531>

doi: [10.2196/25531](https://doi.org/10.2196/25531)

PMID: [34287211](https://pubmed.ncbi.nlm.nih.gov/34287211/)

©Christina Schüttler, Hans-Ulrich Prokosch, Martin Sedlmayr, Brita Sedlmayr. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 21.07.2021. 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

# Measuring the Interactions Between Health Demand, Informatics Supply, and Technological Applications in Digital Medical Innovation for China: Content Mapping and Analysis

Jian Du<sup>1\*</sup>, PhD; Ting Chen<sup>2\*</sup>, PhD; Luxia Zhang<sup>1</sup>, MPH, MD

<sup>1</sup>National Institute of Health Data Science, Peking University, Beijing, China

<sup>2</sup>Institutes of Science and Development, Chinese Academy of Sciences, Beijing, China

\*these authors contributed equally

**Corresponding Author:**

Luxia Zhang, MPH, MD

National Institute of Health Data Science

Peking University

No.38 Xueyuan Road

Beijing, 100191

China

Phone: 86 82806538

Email: [zhanglx@bjmu.edu.cn](mailto:zhanglx@bjmu.edu.cn)

## Abstract

**Background:** There were 2 major incentives introduced by the Chinese government to promote medical informatics in 2009 and 2016. As new drugs are the major source of medical innovation, informatics-related concepts and techniques are a major source of digital medical innovation. However, it is unclear whether the research efforts of medical informatics in China have met the health needs, such as disease management and population health.

**Objective:** We proposed an approach to mapping the interplay between different knowledge entities by using the tree structure of Medical Subject Headings (MeSH) to gain insights into the interactions between informatics supply, health demand, and technological applications in digital medical innovation in China.

**Methods:** All terms under the MeSH tree parent node “Diseases [C]” or node “Health [N01.400]” or “Public Health [N06.850]” were labelled as H. All terms under the node “Information Science [L]” were labelled as I, and all terms under node “Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]” were labelled as T. The H-I-T interactions can be measured by using their co-occurrences in a given publication.

**Results:** The H-I-T interactions in China are showing significant growth and a more concentrated interplay were observed. Computing methodologies, informatics, and communications media (such as social media and the internet) constitute the majority of I-related concepts and techniques used for resolving the health promotion and diseases management problems in China. Generally there is a positive correlation between the burden and informatics research efforts for diseases in China. We think it is not contradictory that informatics research should be focused on the greatest burden of diseases or where it can have the most impact. Artificial intelligence is a competing field of medical informatics research in China, with a notable focus on diagnostic deep learning algorithms for medical imaging.

**Conclusions:** It is suggested that technological transfers, namely the functionality to be realized by medical/health informatics (eg, diagnosis, therapeutics, surgical procedures, laboratory testing techniques, and equipment and supplies) should be strengthened. Research on natural language processing and electronic health records should also be strengthened to improve the real-world applications of health information technologies and big data in the future.

(*JMIR Med Inform* 2021;9(7):e26393) doi:[10.2196/26393](https://doi.org/10.2196/26393)

**KEYWORDS**

medical informatics; Medical Subject Headings (MeSH); health demand; informatics supply; technological applications

## Introduction

### Background

Medical informatics (MI), or biomedical and health informatics, has become an established scientific discipline worldwide. It studies the data, information, and knowledge of biomedicine and health care and their systematic organization, representation, and analysis methods [1]. Basic research scholars in this community adopt quantitative and qualitative methods for understanding and improving the process surrounding the use of information, with the specific goal of advancing biomedical science, whereas applied research scholars leverage information technologies to improve health care outcomes [2]. The application of health information technology (HIT) was proposed as a promising potential solution for improving the productivity, effectiveness, and quality of health care services. The most important benefits of HITs are to reduce medical errors and costs, improve patients' quality of life, and enhance medical decision making. Informatics with big data can be exploited for a wide variety of applications including artificial intelligence (AI), predictive analytics, and point-of-care clinical decision making [3]. The United States has twice promoted HIT through legislation, including the Health Insurance Portability and Accountability Act (HIPAA) in 1996 and Health Information Technology for Economic and Clinical Health (HITECH) in 2009 [4].

In China, there are also 2 major government incentives in the development of MI. One is the launch of the second round of medical reform in 2009 when a substantial investment was put into MI [5]. An important contribution of this health care reform is that MI has been defined as one of the "four constructs and eight pillars," which is the foundation of this reform. As a result of these health policies, the Chinese government and industry have invested heavily in hospital informatics and population health informatics. The second incentive is that in 2016, China released its first health initiative, Healthy China 2030, which guides and coordinates a nationwide strategy for improving China's population health and the national health system. China aims to establish a comprehensive health information system in all public hospitals and primary health care facilities and to develop "Internet + Health initiatives" by using new internet-based technologies to increase access to health care and improve the quality and efficiency of health care delivery. In particular, telemedicine was encouraged as a means toward connecting residents with public hospitals, and its use was viewed as a way to reduce inequity between urban and rural areas apart from improving access to health care. Starting with the experience of fighting COVID-19, China is speeding up its efforts in the use of cutting-edge information technologies in medicine and health care, with the aim of innovating the management and service mode, optimizing the allocated resources, and improving service efficiency [6].

However, evidence shows that there is an imbalance between research and practice of MI in China. This discipline had long been focused on library-oriented informatics instead of hospital-oriented informatics. Academic MI research lags behind HIT applications in China. Current MI in China can be described

as "hot in industrial HIT applications and cold in academic research." This increased focus on HIT applications rather than MI research has hampered the applications of theoretical research to a real-world setting, resulting in repeated HIT construction and huge resource waste in China [7-9].

To characterize the landscape of academic research in MI around the world or in China, previous researchers either used publication data collected in terms of informatics-related concepts, such as the Medical Subject Headings (MeSH) "Public Medical informatics" [10,11], or specialty journals [12]. However, MI is a multidisciplinary field; data from specialty concepts and journals may not reflect the activities outside of the MI communities. Thus, it is difficult to depict a complete picture about the pattern of interactions between informatics concepts or techniques and health or medical needs. To the best of our knowledge, there have been limited systematic investigations of the interactions between health demand and informatics supply in China.

To fill this gap, this paper proposes a new approach to collecting research publications with a broad interpretation of the MI using the hierarchical tree of MeSH terms. We determined whether there is an interaction between health and informatics by examining the proportion of the MeSH terms included in the article that falls into either the health MeSH branch or the informatics MeSH branch. For instance, a given publication can be understood to be included in the field of MI if it is indexed with either of the following MeSH terms from 2 branches: (1) the MeSH tree parent node "Diseases [C]" or node "Health [N01.400]" or "Public Health [N06.850]" and (2) the MeSH tree parent node "Information Science [L]." This paper further demonstrates this interaction between health demand and informatics supply using visualizations such as the ternary map and sankey map.

All terms under the MeSH tree parent node "Diseases [C]" or node "Health [N01.400]" or "Public Health [N06.850]" were labelled as H. All terms under the node "Information Science [L]" were labelled as I, and all terms under node "Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]" were labelled as T.

### Research Questions

The primary goal was to measure the pattern of interactions between health demand and informatics supply in China, as well as the interplay among informatics supply, health demand, and technological applications. This paper mainly uses the United States as a comparator for investigating the development of MI in China, with a particular focus on the H-I-T interaction in each country. The research questions are as follows:

- (1) What is the pattern of interactions between health demand and informatics supply as well as technological applications in China?
- (2) Are informatics research efforts dealing with the major health needs that carry the greatest burden of disease in China?
- (3) What is the pattern of the interplay among informatics supply, health demand, and technological applications in China for important specific areas, such as AI?

## Literature Review

### *Related Research on the Development of MI in China*

To achieve universal access to medical resources, the Chinese government has put HIT as an important technical support tool for the country's health care system. According to a longitudinal study by the China Hospital Information Management Association (CHIMA) Annual Survey, the overall adoption of electronic medical records (EMRs) in China in 2018 has surpassed that of the United States in 2015 and Germany in 2017 on average, yet with only about one-fifth of the required funding and about one-fourth of the required human resources per hospital as compared to the US HITECH project [13].

In addition, China is planning to build a regional medical consortium of hospitals based on regional HITs promoted by health information exchanges in the country. There are several studies exploring the unique contributions and characteristics of HIT development in Chinese hospitals. According to the CHIMA's Annual Survey from 2006 to 2015, the electronic sharing of medical data among Chinese hospitals is growing rapidly. The percentage of hospitals relying solely on paperwork for data interaction declined from 43.3% in 2011 to 8.0% in 2015. There was a strong positive linear correlation between hospitals that join the consortium and the accessibility of electronic medical data exchange plan. The number of hospitals endorsing dual referral systems and appointments, allowing data to be browsed between hospitals and regional information systems, and offering remote consultation services grew to 65.0%, 61.6%, and 81.9% respectively, in 2015, compared to 18.8%, 16.8%, and 10.9% respectively, in 2011 [14]. From 2007 to 2018, 10,954 hospital Chief Information Officers across 32 administrative regions in Mainland China were interviewed in the CHIMA Annual Survey [13]. In terms of funding, the sampled hospitals' annual HIT investment and their average investment per bed increased substantially. With regard to information system development, as of 2018, the average EMR implementation rate of the sampled hospitals exceeded the average of 2015 in US counterparts and 2017 in German counterparts (85.26% vs 83.8% and 68.4%, respectively).

However, academic research in MI lags behind HIT applications within China. Hu et al [15] revealed the notable growth of MI education, a specialty rooted in medical library and information science education, in recent years in China. Although its development has been affected by frequent name changes and an unclear identity, its success has not been entirely ignored. It is recommended that in China, (1) MI treated as a "must-have" discipline be given high priority; (2) independent, balanced degree programs be set up; (3) a specialty of "medical informatics" be established under the "medicine" category; and (4) curricula be integrated with international MI education.

Lei et al [7] argued that Chinese researchers in MI have made insufficient contributions to the global community despite China's substantial HIT market and tremendous investments in hospital information systems. MI has traditionally been focused on medical library or bibliographic information instead of medical (hospital information or patient information) information. Its slow progress is largely due to the misdirected concentration, insufficient teaching staff who have received

formal education of MI, and the incorrect positioning as an undergraduate discipline. Liu et al [16] compared MI education at the top 10 universities in 3 Asian countries. Japan and South Korea have developed modernized educational systems for MI. Universities in Mainland China offer very few curriculum systems in line with international standards and practices. Analysis of the development of MI and the current status of continuing education in China and the United States were presented from the perspective of conferences. Four MI conferences in China and 2 in the United States were conducted for both quantitative and qualitative analyses: China Medical Information Association Annual Symposium (CMIAAS), China Hospital Information Network Annual Conference (CHINC), China Health Information Technology Exchange Annual Conference (CHITEC), China Annual Proceeding of Medical Informatics (CPMI) vs the American Medical Informatics Association (AMIA) and Healthcare Information and Management Systems Society (HIMSS). CMIAAS and CPMI are mainstream academic conferences, while CHINC and CHITEC are industry conferences in China. The results showed that considering China's economy's scale along with the huge investment in HIT, the country is at a low level in terms of the conference output and attendee diversity [8,9]. Moreover, basic MI research funding is inadequate in China compared with the huge investments in HIT applications [7]. As such, the current development of MI in China can be characterized as "hot in industry applications and cold in academic research."

### *Mapping Interactions Between Different Knowledge Entities Using the MeSH Tree*

The MeSH thesaurus is a controlled and hierarchically organized vocabulary produced by the National Library of Medicine. It is used to index, catalog, and search biomedical and health-related information [17]. The MeSH terms are organized in a tree-like network structure consisting of 16 branches coded using A–N, V, and Z. The name of the branches are Anatomy, Organisms, Diseases, Chemicals and Drugs, Analytical, Diagnostic and Therapeutic Techniques and Equipment, Psychiatry and Psychology, Phenomena and Processes, Disciplines and Occupations, Anthropology, Education, Sociology and Social Phenomena, Technology, Industry, Agriculture, Humanities, Information Science, Named Groups, Health Care, Publication Characteristics, and Geographical. Within each branch, MeSH terms with shorter "Tree Number" identification codes are relatively general concepts that branch out into more specific concepts. Each article in PubMed is typically assigned to several MeSH terms.

The MeSH tree is a widely recognized controlled vocabulary thesaurus for information retrieval systems [18]; it has been used to map the interactions between different knowledge entities in the biomedical and health domain.

One is to measure the translational interactions between basic research and applied research reflected by MeSH terms. Weber [19] introduced an approach to mapping PubMed articles onto a graph, called the "Triangle of Biomedicine," by assigning articles in 3 categories (Human, Animal, and Molecular/Cellular Biology [HAC]) based on the number of MeSH terms they have that fall into each of these categories. Each publication is given

a code based on whether it contains MeSH terms from that group (eg, a publication containing 1 or 10 MeSH terms from a cellular group would be given a “C”). Weber defined translation as a movement of a collection of articles, or the articles that cite those articles, toward the human corner. Based on this framework, a data science team at the US National Institutes of Health (NIH) modified the algorithm so that the HAC categories are fractionally counted, which is done for each article by dividing the number of HAC terms in each category by the total number of terms in all 3 categories [20]. In place of the binary variable Weber [19] used, NIH’s development opens up the triangle so an article can appear anywhere on it, instead of just the 7 points in the Weber triangle. Recently, Ke [21] further integrated the elements in this model. He adopted a working definition of cell- and animal-related MeSH terms as basic and human-related as applied. Ke proposed a method to place publications onto the translational spectrum, by learning embeddings of controlled vocabularies. He applied these learning methods on MeSH terms to obtain similarities between human-related terms and the rest, which in total determines the degree of basicness of the articles.

The other is measuring medical innovation through the interplay among the demand, supply, and technology in terms of MeSH terms. Several scholars have taken advantage of the fact that MeSH is organized as a hierarchical tree, and the relevant topic areas that correspond to particular MeSH nodes and their subtrees can be used to measure the process of medical innovation. Agarwal and Searls [22] were the first to conceptualize the medical innovation interaction in terms of “demand” (represented as “diseases” in MeSH terms) versus “supply” (represented as new “drugs and chemicals” in MeSH terms). Focusing on 3 main branches — “diseases,” “drugs and chemicals,” and “techniques and equipment” — Leydesdorff et al [23] used base maps and overlay techniques to investigate the translations and interactions and thus to gain a bibliometric perspective on the dynamics of medical innovations. Based on the study by Agarwal and Searls [22], Petersen et al [24] developed a triple helix model of medical innovation — supply, demand, and technological capabilities — by introducing a third branch of MeSH terms referring to “Analytical, Diagnostic and Therapeutic Techniques and Equipment” (namely, “Techniques and Equipment”), which provides yet another perspective relevant to medical innovation. Compared with only the demand and supply interactions investigated in the study by Agarwal and Searls [22], technological capabilities make it possible to observe the generated innovation in the forms of products, processes, and services.

### ***HIT Innovation in Comparison With the More Well-Established Pharmaceutical Industries***

HIT and evidence-based digital medicine can also be understood as a medical technology (the “supply” side) similar to drugs and devices to meet the needs of health care and disease management (the “demand” side). Worldwide, the chaotic and subpar processes and results of HIT innovation are noted in the wake of tremendous investments in capital and human resources, especially when compared with the more well-established drug and device industries [25]. “Evidence-based medicine” is best suited to deal with the uncertainty surrounding the MI and HIT

applications [26]. Evidence-based MI can be defined as the “conscientious, explicit, and judicious use of the current best evidence” to support a health care decision that employs information technologies [27]. There are few studies on the application of evidence-based medicine to evaluate the effectiveness and safety of HIT and digital health interventions as well as AI algorithms on health [26,28-31]. Evidence-based MI, despite some progress, is still in the early phases of development [1]. It is the responsibility of the whole community to build evidence in MI, providing it is considered to be a scientific discipline [27]. Drug and device innovations must follow a standardized pipeline of production processes, while HIT innovations do not meet the equivalent standards. As a consequence, when it comes to producing effective and reliable products for the public, HIT lags behind the more mature drug and device industries.

As new drugs are the major source of medical innovation, informatics-related concepts and techniques are a major source of digital medical innovation. Inspired by this point, along with the aforementioned framework to measure medical innovation in the “Mapping Interactions Between Different Knowledge Entities Using the MeSH Tree” section, we suggest measuring digital medical innovations (or MI innovations or HIT innovations) by replacing “drugs and chemicals” with “information science”-related MeSH terms.

## ***Methods***

### **H-I-T Model**

#### ***Overview***

We used 3 MeSH branches as representations of *Health* demand, *Informatics* supply, and *Technological* applications (the H-I-T model) and used their co-occurrences to measure the digital medical innovation process in China. The detailed definition of HIT is as follows. For “H,” the entire “Diseases [C]” branch as well as 2 subbranches (ie, “Health [N01.400]” and “Public Health [N06.850]”) are regarded as a representation of health demand for HIT innovations. “Health [N01.400]” and “Public Health [N06.850]” are under the branches “Population Characteristics [N01]” and “Environmental and Public Health [N06],” respectively — with the top root “Health Care [N].” So, we use 2 MeSH terms, “health” and “public health,” to represent the population health demand and the “diseases category” MeSH terms to indicate the individual health demand (specific disease management).

For “I,” the “Information Science [L]” branch is a representation of the supply side in terms of informatics concepts and techniques.

For “T,” the “Analytic, Diagnostic, and Therapeutic Techniques and Equipment [E]” branch is a representation of state-of-the-art technological applications, namely the functions to be realized by informatics (eg, diagnosis, therapeutics, surgical procedures, investigative techniques, equipment, and supplies).

In the H-I-T model, every related article can be classified as health demand (H), informatics supply (I), technological applications (T), or a combination of these 3 using the MeSH



terms and HIT score. MeSH terms are arranged in an alphabetical and hierarchical structure from the most general level to the narrowest level. Table 1 shows the branches of MeSH terms used in distinguishing the H-I-T classification. Note that since the MeSH term “Public Health” has another tree number H02.403.720 (branch of medicine concerned with the prevention and control of disease and disability and the

promotion of physical and mental health of the population on the international, national, state, or municipal level), terms under this branch are also included in the health demand (H) category. These terms include Epidemiology, Molecular Epidemiology, Pharmacoepidemiology, Preventive Medicine, Environmental Medicine, Occupational Medicine, and Preventive Psychiatry.

**Table 1.** Medical Subject Heading (MeSH) terms used in each health demand, information supply, technological applications (H-I-T) category

H-I-T category	MeSH branches	Number of terms
Health demand (H)	Diseases [C], Health [N01.400], Public Health [N06.850]	5331
Informatics supply (I)	Information Science [L]	419
Technological applications (T)	Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]	2985

It is noted that for each publication, only a MeSH major topic (ie, MeSH [primary] terms) are used in our data collection and computation. In PubMed publications, a MeSH term that is one of the main topics discussed in the article is denoted by an asterisk(\*) on the MeSH term or MeSH/Subheading combination and is referred to as a MeSH major topic. The major topic can reveal the most essential research content of an article.

### ***Mathematical Description of the H-I-T Model***

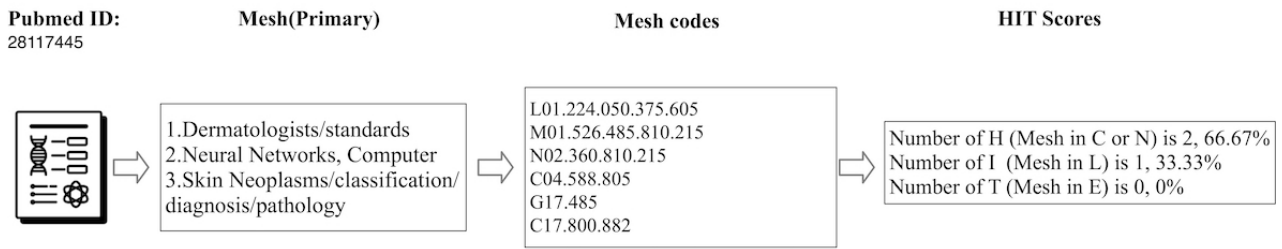
The classification algorithm calculates the percentage for each category by dividing the number of H-I-T MeSH (primary) terms in each category by the total number of terms in all 3 categories. Figure 1 shows 2 examples of calculating HIT scores. The first article with PMID 28117445 was tagged with 3 MeSH (primary) terms. It is noted that 1 MeSH term may belong to 2 or more branches and have 2 or more MeSH codes. In this situation, we marked each MeSH code once. Now, 3 terms became 6 MeSH codes; the codes beginning with C or N06.850

or N01.400 were classified as “H.” The codes beginning with L were classified as “I.” The codes beginning with E were classified as “T.” The final HIT scores were calculated using the codes belonging to the 3 H-I-T categories only, for instance, as indicated in Figure 1 with a total of 3 H-I-T MeSH terms: 2 for H, 1 for I, and 0 for T. The final H-I-T scores for this article are H=2/3, I=1/3, T=0, with only the linkages between H and I and none with T.

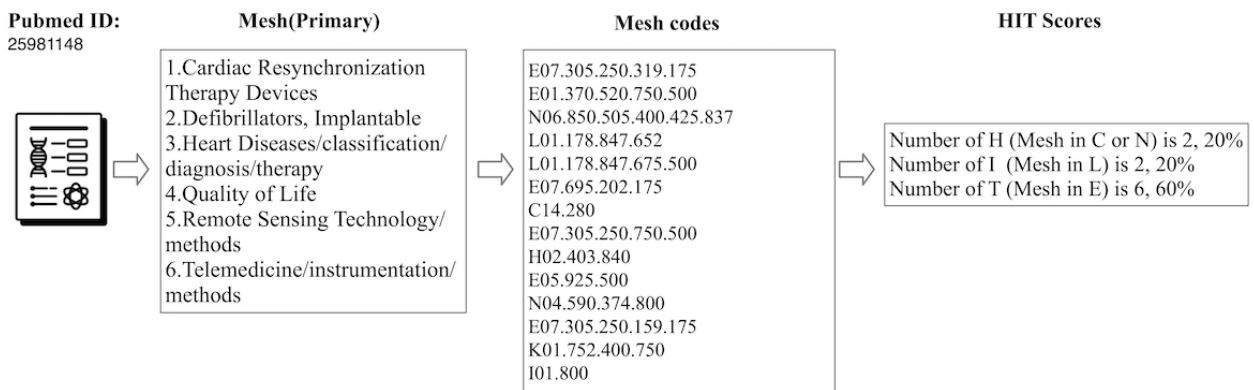
The H-I-T scores for the second article with PMID 25981148 were also calculated using the same algorithm. It has a total of 10 H-I-T MeSH terms: 2 for H, 2 for I, and 6 for T. The techniques and equipment (“Cardiac Resynchronization Therapy Devices,” “Defibrillators, Implantable,” and “Remote Sensing Technology”) have linked the health demand (“Heart Diseases,” “Quality of Life”) with the informatics supply (“Telemedicine”). Without such techniques and equipment as the “Remote Sensing Technology,” it is hard to apply telemedicine to heart disease care.

**Figure 1.** The calculation process of health demand, informatics supply, technological applications (H-I-T) scores for 2 example articles.

**Article: Dermatologist-level classification of skin cancer with deep neural networks**



**Article: HRS Expert Consensus Statement on remote interrogation and monitoring for cardiovascular implantable electronic devices**



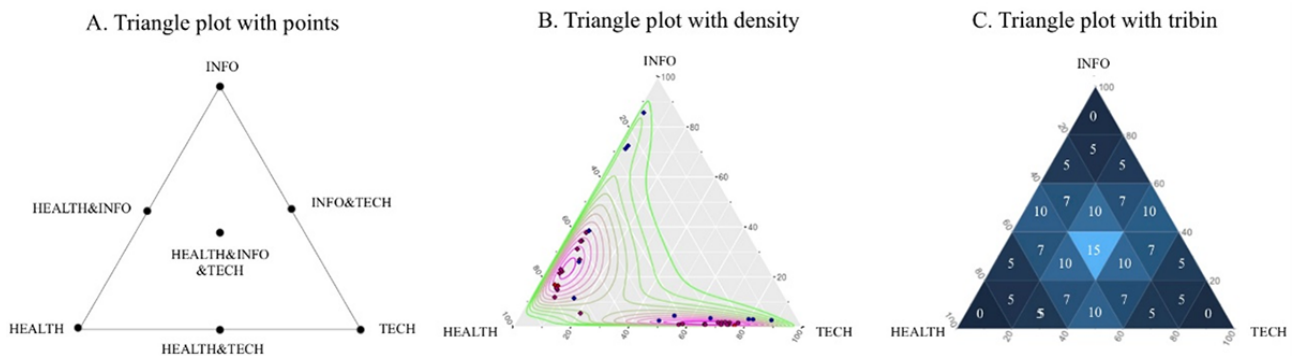
**Visualization of the H-I-T Model**

To maximize the utility of the H-I-T model, we adapted Weber's Triangle of Biomedicine [19] to show the composition of systems composed of H-I-T. Each of the 3 apexes represents health demand (H), informatics supply (I), and technological applications (T). If an article contains a 100% H or I or T, it will be placed at one of the vertices of the triangle as a dot. If an article contains a 50% H and 50% I, then it will be placed in the middle of the left edge of the triangle, as shown in Figure 2A. If an article contains at least 33% H/I/T, then it will be placed in the center of the triangle, and so on.

Usually, there are only 5 to 10 MeSH (primary) terms in a paper, and the percentage repetition rate will be high when calculating H-I-T scores for each article. There will be a large number of

points overlapping on the triangle graph, and the points themselves lose their meaning due to visual clutter. Often, scholars use density contours in triangles (Figure 2B) to improve visualization, but those density markers alone are still difficult to observe quickly by the human eye. This paper further improves on the display details of the triangle by dividing the entire triangle into many mini tribins. We cut the whole triangle into *N* equal parts in 3 directions, and then the *N*\**N* small tribins appear within the original large triangle. It enables us to bin points in small triangular areas; the number of points in each small area can be counted. With large datasets, we are able to display the counted number and color on tribins together (Figure 2C). Perhaps it can add more richness to triangle diagrams and thus enhance the visualization of the HIT triangle diagram. The triangle diagrams in this paper were implemented by using ggtern library in R [32].

**Figure 2.** Three ways to display the health demand, informatics supply, technological applications (H-I-T) triangle: (A) triangle plot with points only on the vertex, middle of the edges, and center; (B) triangle plot with points and a density contour; (C) triangle plot with tribin.



## Mapping ICD-10 to MeSH Terms

To approximate the extent of health needs, we use the World Health Organization (WHO) Global Burden of Disease (GBD) survey as useful information. The WHO provides the corresponding codes of the International Classification of Diseases (ICD-10) in which each of the aforementioned diseases is classified.

Most recently, Yegros et al [33] matched WHO ICD-10 with MeSH terms in a corresponding table to find whether research efforts address global health needs. We reviewed this correspondence table again and used it to map the correlation

between disease burden and rates of informatics-related publications for China. To link publications to diseases, we used not only the MeSH terms assigned to ICD-10 codes but also all MeSH terms located beneath them in the MeSH tree. This, for instance, enabled us to assign publications with the MeSH term “Diabetic Nephropathies” to the disease “Kidney Diseases” even if the MeSH term “Kidney Diseases” was not assigned to these publications. In fact, the term “Diabetic Nephropathies” is the subordinate concept of the term “Kidney Diseases.” Table 2 shows the correspondence table between ICD-10 and MeSH for specific cardiovascular diseases.

**Table 2.** Correspondence table between International Classification of Disease (ICD)-10 and Medical Subject Headings (MeSH) for 2 specific cardiovascular diseases.

Cardiovascular disease	ICD-10 code	Matched		Excluded	
		MeSH Tree Number	MeSH terms	MeSH Tree Number	MeSH terms
Hypertensive heart disease	I10-I15	C14.907.489	Hypertension	C13.703.395; C14.907.489.480	Hypertension, Pregnancy-Induced
Ischemic heart disease	I20-I25	C14.280.647; C14.907.585	Myocardial Ischemia	N/A <sup>a</sup>	N/A

<sup>a</sup>N/A: not applicable.

## Data Collection

In order to systematically collect publications relating to informatics supply and health demand, here, we use a new approach to collect publications that provide a broad interpretation of MI using the hierarchical tree of MeSH 2020 terms. A given publication can be understood to be included in the field of MI if it is indexed with both of the following MeSH (primary) terms from each of the 2 branches: (1) the MeSH tree parent node “Diseases [C]” or node “Health [N01.400]” or “Public Health [N06.850]” and (2) the MeSH tree parent node “Information Science [L].” Note that we restrict our analysis to the “Major Topic Headings” for each article, which are indicated in each PubMed article page by an asterisk \* next to the MeSH term; these MeSH (primary) terms are sufficient to identify the article’s core content.

A total of 213,215 publications during 2010-2020 (till June, 30 2020) were initially collected from MEDLINE using the co-occurrences of the 2 branched MeSH (primary) terms. We excluded publications indexed by such MeSH (primary) terms as (1) “Systematic Reviews as Topic” and (2) “Meta-analysis as Topic.” While located within the parent MeSH tree of “Information Science” and “Public Health,” they do not reflect the informatics supply and health demand, respectively, but represent a secondary research approach. Other exclusion criteria are given to such publications with publication types as Review and Retracted publications, as well as with the keyword “bibliometric” occurring in the title. This process leaves 194,567 global publications for the following analysis.

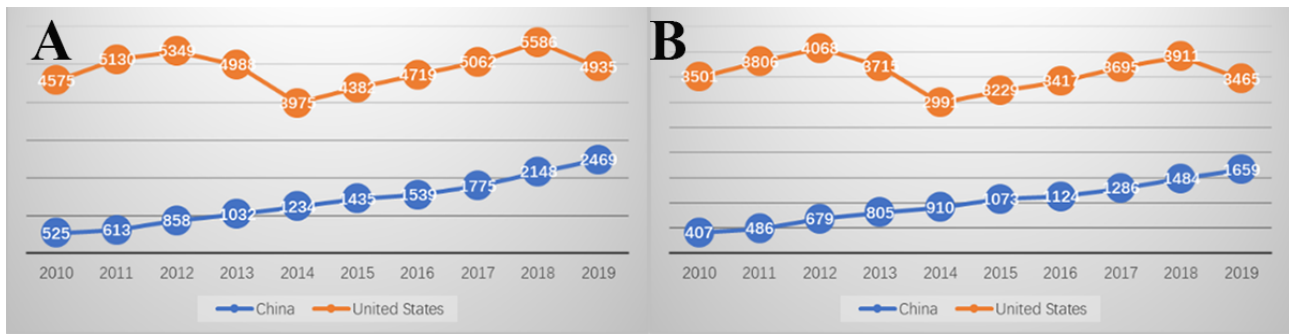
## Results

### Overall Results

The United States ranked first based on the number of publications, accounting for one-quarter of the world’s total publications. China ranked second, with the number of publications basically equal to the third, the United Kingdom. Compared with the United States and United Kingdom, China’s publications increased rapidly since 2010, when the Chinese government launched health reforms for the second time and invested significant funding in HIT (Figure 3, Table 3).

It is noted that of all MI papers published by China, there are 1083 in the Chinese language, published in MEDLINE-indexed Chinese journals, such as *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi* (N=126), *Nan Fang Yi Ke Da Xue Xue Bao* (N=80), and *Zhonghua Liu Xing Bing Xue Za Zhi* (N=77; Table 3). It reflected the interactions among MI with biomedical engineering as well as epidemiology, which is an important subdiscipline of public health and preventive medicine in China. We first performed an exploratory data analysis and found that the pattern of interactions between health demand and informatics supply for the United Kingdom and China is similar. There is a significant difference between the United States and United Kingdom/China. In the following section, we will primarily compare China with the United States and answer the aforementioned research questions.

**Figure 3.** Number of (A) health demand and informatics supply (H-I) and (B) health demand, informatics supply, and technological applications (H-I-T) publications for China and the United States.



**Table 3.** Distribution of global publications on medical informatics (n=194,567).

Ranking	Country/territory	Number of publications	%
1	United States	49,353	25.4
2	China	14,105 (1083 in Chinese, 13,022 in English)	7.2
3	United Kingdom	13,783	7.1
4	Germany	9469	4.9
5	Canada	8326	4.3
6	Australia	7110	3.7
7	Italy	6644	3.4
8	Japan	6443	3.3
9	France	6037	3.1
10	The Netherlands	5596	2.9

### Interactions Between Health Demand, Informatics Supply, and Technological Applications

#### Overall H-I-T Interactions

First, we calculated the average H-I-T scores of publications for the world, the United States, and China. As shown in Table 4, we found that if we only count the average scores, the H-I-T scores for China, the United States, and the world are very similar. In general, the H score is higher than the I and T scores, indicating that the number of H-related MeSH major topic terms is more than those of I- and T-related topics. In other words, this research tends to be health demand-oriented. Then, we counted the H-I and H-I-T interacting publications for China and the United States (Figure 3). Compared with the fluctuating trends for the United States, the interactions of both H-I and H-I-T for China show a notable increasing trend.

Next, we mapped publications from China and the United States on the H-I-T triangle graphs based on the H-I-T model, as shown

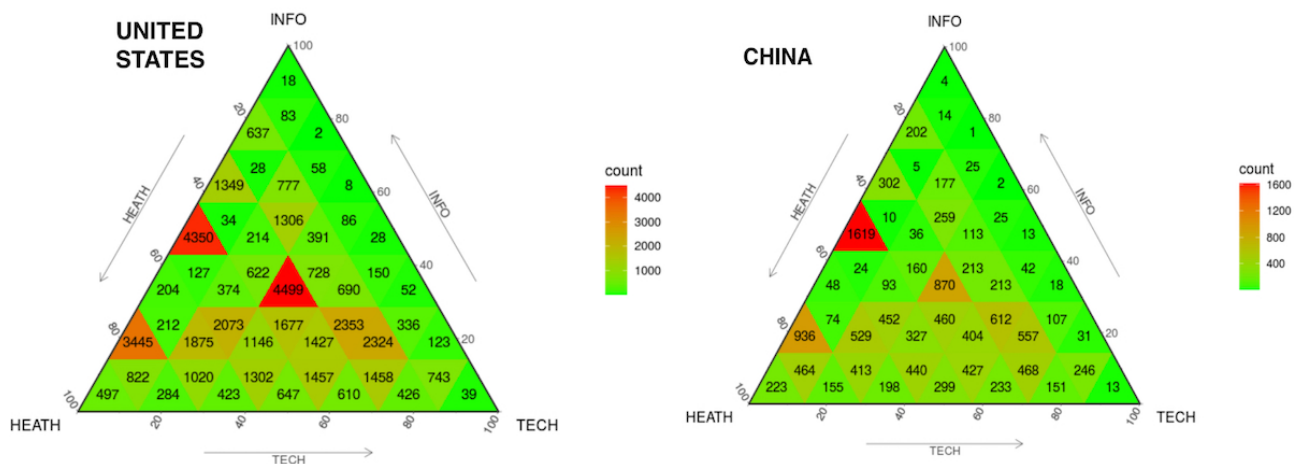
in Figure 4. Since the search strategy was that any given paper will definitely contain the H and I MeSH terms, we would naturally assume that most articles on the triangle would tend towards the edges H and I. However, the distribution of the United States' publications was slightly unexpected. In the US triangle diagram, the reddest subtriangle is in the center (N=4499) instead of the edges of H and I (N=4350). The distribution of Chinese publications in the triangle is quite different from that of the United States; they are much less prone to the T side of the triangle. Most articles from China are located at the edges of H and I. In the just-centered subtriangle, the number of articles with the same percentages of H, I, and T MeSH primary terms (each category accounts for one-third) is much less for China (N=870) than the United States (N=4499). This shows that, while China's existing techniques and equipment have been established to link informatics supply and health demand, the informatics research efforts in the United States have provided a stronger H-I-T link than China.

**Table 4.** The average health demand, information supply, and technological application (H-I-T) scores for the world, the United States, and China.

Country/territory	H	I	T
United States	0.434	0.280	0.287
Global	0.436	0.277	0.288
China	0.450	0.268	0.282



**Figure 4.** Overall layout of publications in the health demand, informatics supply, and technological application (H-I-T) triangle for the United States and China.



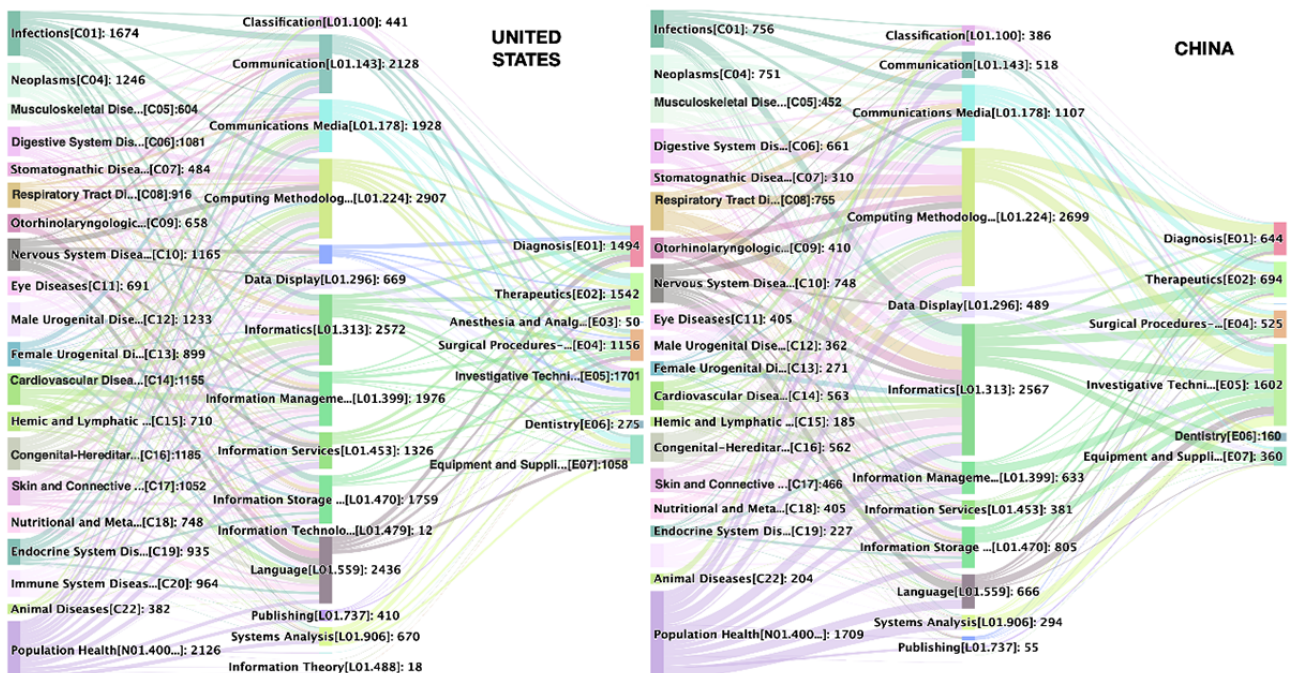
**Detailed H-I-T Interactions**

Such differences between China and the United States are also observed when the detailed H-I-T interactions are considered. We use the first level of disease classification in the MeSH tree and combine “Health [N01.400]” and “Public Health [N06.850]” as “population health.” The first-level concepts in the “Information Science” branch and “Analytic, Diagnostic, and Therapeutic Techniques and Equipment” branch are used to map the interactions between informatics supply, technological applications, and health demand, which is represented by various specific diseases and population health.

Figure 5 depicts the detailed profile of the H-I-T interactions. In general, such interactions in China are relatively weaker compared with those in the United States. According to Figure 5, whether it comes to diseases, informatics, or technology

applications, the United States is more balanced, while China is more concentrated. Computing methodologies, informatics, and communications media (such as social media and the internet) constitute the majority of the informatics domain and are the 3 most common HITs used for resolving the health and disease problems in China. In China, social media, the internet, and other forms of communication media are not only used to solve public health problems but also applied to specific disease problems, especially infections diseases, cardiovascular diseases, respiratory tract diseases, neoplasms, nervous system diseases, and many other chronic diseases. The State Council has released a medium to long-term plan (2017-2025) on the prevention and treatment of chronic diseases and emphasized the roles of internet+health in promoting health. Thus, social media and other internet media are extremely important for health promotion and self-care among patients and their family members and caregivers.

**Figure 5.** Detailed profile of health demand, informatics supply, and technological application (H-I-T) interactions for the United States and China.



As is shown in Figure 5, population health is the largest informatics research target for both the United States and China, since public MI has been an established research area for both of them. In fact, the term “public health informatics” was introduced into the MeSH tree in 2003 and defined as “the systematic application of information and computer sciences to public health practice, research, and learning.” But there were some differences in the health demand domains interacting with informatics between China and the United States. For China, the major diseases supported by informatics research efforts include nervous system diseases, neoplasms, digestive system diseases, cardiovascular diseases, and respiratory tract diseases. The types of diseases that are most interactive in China are ranked rather low in the United States. Such cases were observed for male urogenital diseases, immune system diseases, and nutritional and metabolic diseases. In other words, China and the United States are using similar HITs to solve their own health problems.

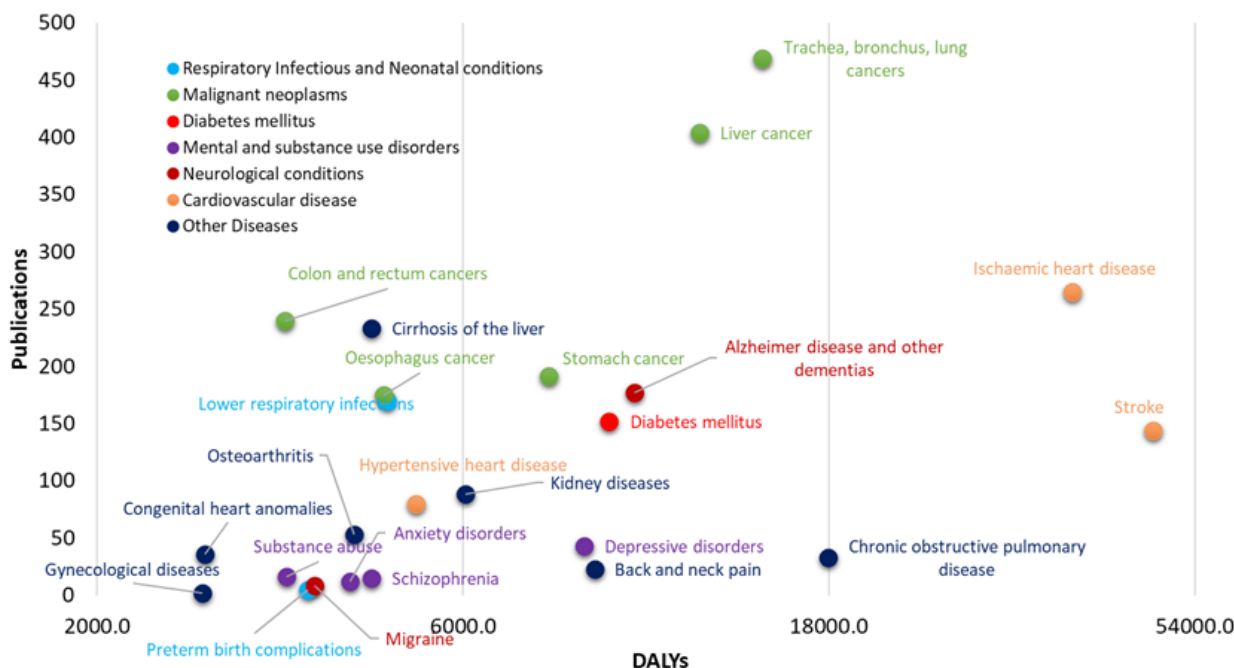
### Informatics Research Efforts vs Burden of Disease in China

The question to be discussed in this subsection is whether China informatics research efforts are dealing with the major health needs, measured by the burden of disease. The Years of Life Lost was used, and Disability-Adjusted Life Years (DALYs) were provided by the WHO as a proxy of the burden of disease. Although DALYs are not free of limitations, they are one of the most established proxies of disease burden. We identified informatics-related publications related to a selection of the diseases considered in the WHO GBD estimates for the year

2016 [34], which is much closer to the publication search window used in this paper. Combined with the rankings of the top 25 leading causes of DALYs in China during 1990-2017 [35], we considered the 24 most specific diseases with high DALYs in China in the “Communicable, maternal, perinatal and nutritional conditions” and “Noncommunicable diseases” groups. We did not consider diseases in the “Injuries” group since it is difficult to match them with MeSH terms, although “road injuries” ranked fifth in the leading causes of DALYs in 2017.

According to Figure 6, generally, there is a positive correlation between the burden of disease and the oriented informatics research publications, when analyzed by specific therapeutic areas classified by the WHO. The linear regression results showed  $y = 0.0079x$ ,  $R^2 = 0.4103$ . Two major cardiovascular diseases, stroke and ischemic heart disease, are leading causes of DALYs in China and in recent years, have attracted a considerable amount of information science research. For several major malignant neoplasms, such as lung cancer, liver cancer, stomach cancer, esophagus cancer, and colon and rectum cancers, although their DALYs vary largely, research has attracted the most informatics research efforts. Among the top 24 diseases related to DALYs in China, there is a gap between the burden of disease and informatics research efforts for chronic obstructive pulmonary disease, back and neck pain, and depressive disorders, which have a higher burden yet lower informatics research efforts. Overall, among mental disorders, beside depressive disorders, schizophrenia, substance abuse, and anxiety disorders have disproportionate informatics research efforts to their disease burden.

Figure 6. Top 24 disease-related Disability-Adjusted Life Years (DALYs) versus informatics-related publications in China.



## Interactions Between Health Demand, AI Supply, and Technological Applications

The “computing methodologies” represent the largest part of information science research for both the United States and China, of which AI is the most dominant area. We use the “Artificial Intelligence” MeSH terms and all terms beneath it in the MeSH tree to construct AI sub-datasets. According to [Table 5](#), the proportion of AI-specific publications across all publications linking health demand and information science supply is significantly higher for China (11.3%) than the global average level (6.8%) and that in the United States (7.4%). The number of AI publications has grown rapidly, especially since 2016 ([Figure 7](#)). In the MeSH tree, “Artificial Intelligence” is defined as the theory and development of computer systems that perform tasks that normally require human intelligence. Such tasks may include speech recognition, learning; visual perception; mathematical computing; reasoning, problem solving, decision making, and the translation of language. It has 8 subbranches: (1) Computer Heuristics, (2) Expert Systems, (3) Fuzzy Logic, (4) Knowledge Bases (ie, Biological Ontologies), (5) Machine Learning, (6) Natural Language Processing, (7) Computer Neural Networks, and (8) Robotics.

As a subfield of information science, AI and related technologies are increasingly prevalent in medical research and are beginning to be applied to health care and medical research. In this section, we specifically analyzed and discussed the H-I-T interactions in AI-related research publications. [Figure 8](#) shows the H-I-T interactions in AI research in China and the United States, and the secondary MeSH terms under MeSH topic “Artificial Intelligence [L01.224.050.375]” were selected to calculate co-occurrence relationships between health demand, AI supply, and technology applications.

As to the connection between health demand and AI supply, the most focused domain of health demand is population health; the most concentrated AI concepts are computer neural networks and machine learning; and the most extensive technology applications are investigative techniques, diagnosis techniques, and therapeutic techniques. Investigative techniques are commonly used in preclinical and clinical research, epidemiology, chemistry, immunology, and genetics, among others. The investigative techniques do not include techniques specifically applied to diagnosis, therapeutics, anesthesia and analgesia, surgical procedures, surgical, and dentistry. After a detailed analysis of the specific topics under “Investigative techniques,” we found the most focused topics are “observation,” “research design,” and “epidemiologic methods” and “models, theoretical.” We concluded the linkage “Population health—machine learning (including deep learning such as

neural networks)—Investigative techniques” shows hot topics such as machine learning—enabled clinical research and disease prediction models based on real-world data. Now, we turn from “population health” to specific diseases. Both countries have used AI technologies for research on all diseases. Among those, nervous system diseases and neoplasms are the most focused AI-targeted diseases for the United States and China, respectively. Diseases that very rarely use AI include infections, otorhinolaryngologic diseases, immune system diseases, and hemic and lymphatic diseases.

On the informatics supply side, machine learning and neural networks are the most commonly used techniques in both Chinese and US publications. In the United States, “machine learning” and “computer neural networks” co-occurred 1169 and 914 times, respectively, with health demand—related terms. China, on the other hand, was counted 829 and 901 times, respectively. In health AI research by US scholars, “computer neural network” technologies have been used mainly in clinical research and real-world studies, with 248 occurrences, whereas it is rarely used for diagnosis [E01], with only 5 occurrences. However, in Chinese scholars’ research, “computer neural network” technology has been mainly supported for “Diagnosis” [E01] with 209 co-occurrences.

Furthermore, another significant difference observed between health AI research by Chinese and US scholars is the use of natural language processing technology. Natural language processing co-occurred 325 times with healthy demand for the articles by US scholars, but only 65 times for China. As it is indicated in the Sankey diagram ([Figure 8](#)), the most common use of natural language processing techniques in the United States focused on “Population Health,” followed by cardiovascular disease, nervous system disease, and many other specific diseases.

The phenomenon that natural language processing—related research seems a gap or a weak point for China may be due to the relative lack of research on electronic health records (EHRs) in China. The MeSH term “Electronic Health Records” is a standardized term that unifies synonyms such as Computerized Medical Records and EMR. It is defined as media that facilitate transportability of pertinent information concerning a patient’s illness across varying providers and geographic locations. Some versions include direct linkages to online consumer health information that is relevant to the health conditions and treatments related to a specific patient. While continuously increasing ([Figure 7](#)), China has insufficient research publications as to its overall health information research publications on EHR (193, 1.4%), compared with the world average (3.3%) and the United States (4.8%).

**Table 5.** Comparison of the number of publications between health demand and artificial intelligence (AI) supply.

Country/territory	Number of AI publications	Number of all H-I <sup>a</sup> -interacted publications	Percentage (%)
Global	13,258	149,567	6.8
China	1592	14,105	11.3
United States	3628	49,353	7.4

<sup>a</sup>H-I: health-related MeSH (Medical Subject Headings) terms co-occurred with information-science-related MeSH terms.



Figure 7. Publications about (A) artificial intelligence and (B) electronic health records for the United States and China.

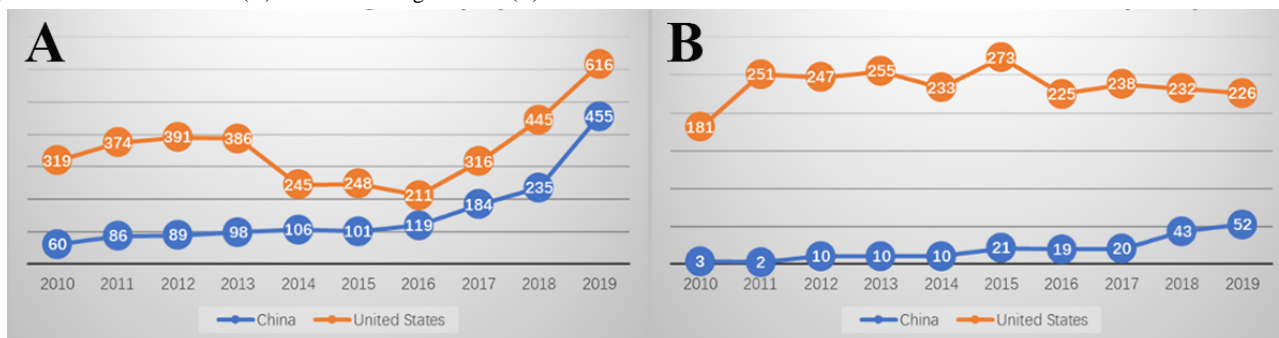
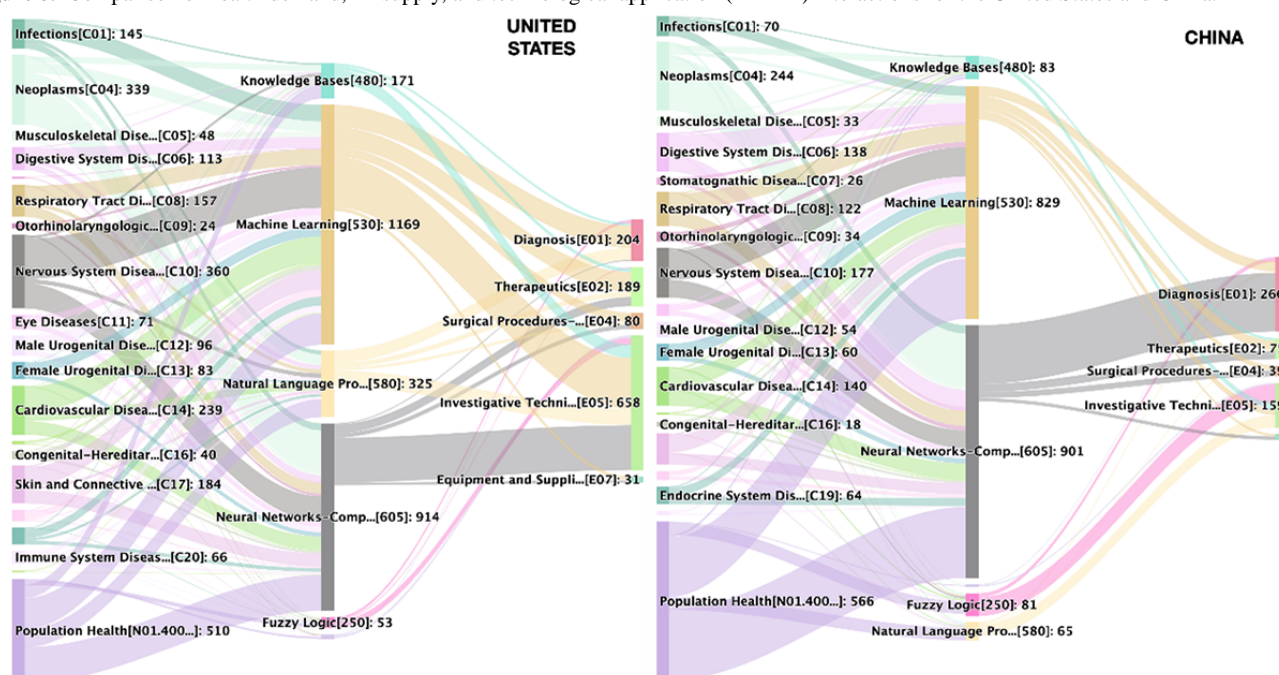


Figure 8. Comparison of health demand, AI supply, and technological application (H-AI-T) interactions for the United States and China.



## Discussion

### Interactions Between Health Demand, Informatics Supply, and Technological Applications

Informatics uses the synergistic “bridging” of electronic data to benefit individual diseases and population health. There has been a consistent increase in the number of publications tagged with both health-related and information science–related MeSH terms since 2010 in China. This is in accordance with the observation that “a significant upward trend particularly after 2011 in the number of articles by Chinese academics in MI based on their publications in 18 international specialty journals” [12]. They also concluded that the global influence of Chinese scholars is growing worldwide; they are making increasingly conscious efforts to enhance their collaborative relationships with international researchers. In their contribution, the hottest and emerging technological fields in MI were examined, such as EMRs, AI, and image processing, whereas the functions these informatics technologies have realized, such as in supporting diagnosis and therapeutics of diseases, and the health needs they are used to address have not been investigated. The focus of our paper is the interplay between knowledge entities through the co-occurrence of the 3 dimensions of health-, information-,

and technology-related terms, instead of macrobibliometric analysis. Our results suggest that the interactions between health demand and informatics supply as well as technological applications in China are showing significant growth. Among the top 3 countries with the highest number of publications, the number of H-I-T publications in China is growing the fastest. In our analysis, population health or public health is the area of greatest demand that interacts with informatics for both the United States and China. Population health can have the most impact on health informatics research. Recently, Bhattarai et al [10] investigated how information and communications technologies (ICTs) were applied to public health and found that “communicable disease monitoring,” “public health policy and research,” and “public health awareness” are the most common public health domains interacting with ICTs. One of the limitations of their study is that, by only using one MeSH tag as a selection criterion, publications without the “public medical informatics” MeSH term were excluded from their dataset. Our research avoids such a limitation.

Despite the increasing output of academic research, the overall interaction between health demand, informatics supply, and technological applications in China is weaker than that in the United States. To some extent, this has affected the technological



transfer of HITs into products and ultimately had an impact on the development of evidence-based digital medicine. The effectiveness and safety of HIT must be evaluated scientifically before it can be used by doctors, patients, and consumers. For example, Bhattarai et al [10] reported that inconsistent results exist regarding the validity of most of the informatics indicators when various predictors are used for disease surveillance and emergency monitoring, such as syndromic surveillance, dispatch calls, over-the-counter drug sales, and school absenteeism. They suggest additional studies should be conducted to further investigate the validity of such predictions. Whereas, for evidence-based medicine, there are clear guidelines on the development and assessment of the effectiveness of biomedical or behavioral health interventions, there is a scarcity of guidelines for the systematic development and assessment of medical and health informatics, and corresponding research has just begun [25].

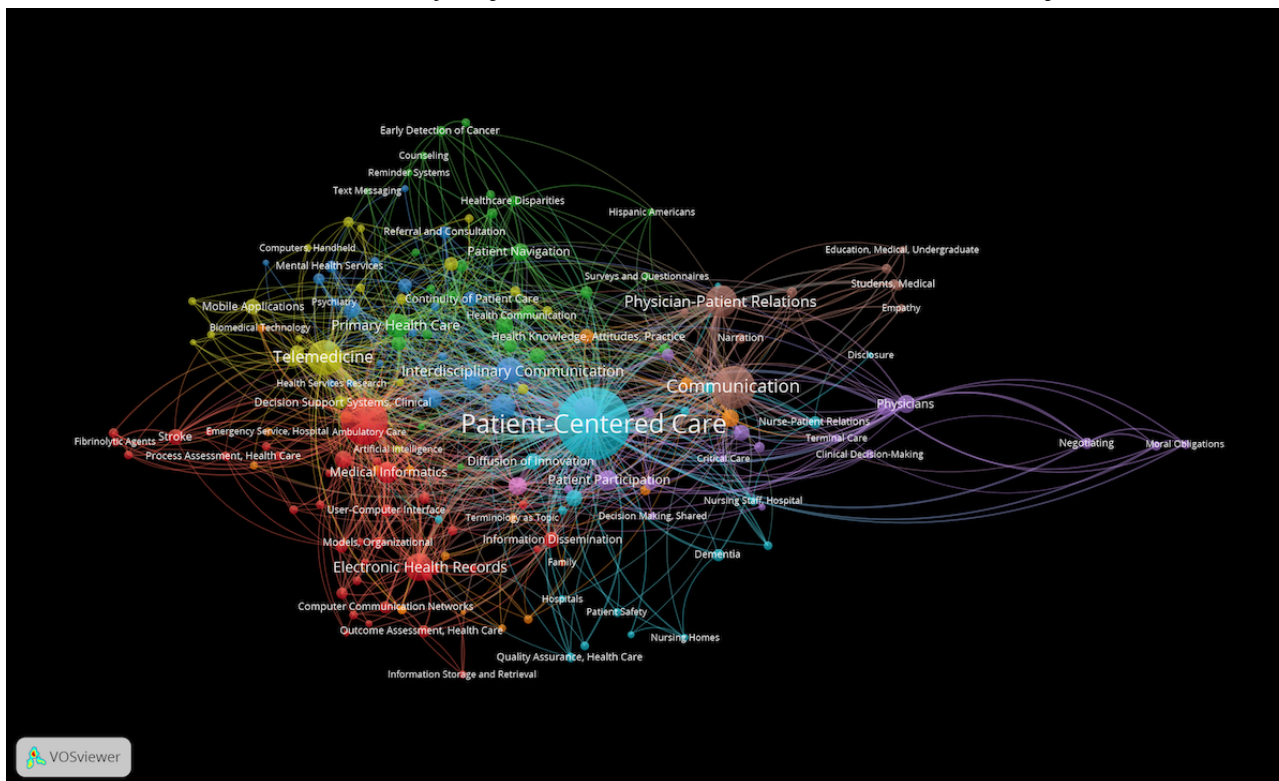
### **Informatics Research to be Focused on the Greatest Burden of Diseases or Where It Can Have the Most Impact**

We are not prepared to emphasize the idea that informatics research efforts must be proportional to the burden of disease. We think it is not contradictory that informatics research should be focused on the greatest burden of diseases or where it can have the most impact. In fact, we have taken into account these 2 viewpoints. The overall results of both the H-I-T interactions and the H-AI-T interactions indicate that population health or public health is the area with the greatest demand that interacts with informatics supply and technological applications for both the United States and China. We think population health can have the greatest impact on health informatics research. Population health and specific disease management are 2 major demanding health domains. While the extent of demand for

specific diseases is easy to measure (eg, using burden of disease data), the extent of the demand for population health is not easily quantified since it is independent of diseases.

We used 2 MeSH terms, “health” and “public health,” to represent population health demand and the “diseases category” MeSH terms to indicate the individual health demand for specific diseases. We noticed that there are 2 MeSH terms, “Delivery of Health Care, Integrated” and “Patient-Centered Care,” that are related to health demands independent of specific diseases. The term “Delivery of Health Care, Integrated” is a health care system that combines physicians, hospitals, and other medical services with a health plan to provide the complete spectrum of medical care for its customers. In a fully integrated system, the 3 key elements — physicians, hospital, and health plan membership — are in balance in terms of matching medical resources with the needs of purchasers and patients. The term “Patient-Centered Care” represents a design of patient care wherein institutional resources and personnel are organized around patients rather than around specialized departments. Unfortunately, these terms are not included in our conceptual framework and data collection. Here, we tried to determine the research landscape between “Patient-Centered Care”/“Delivery of Health Care, Integrated” and “Information Science.” Using such a search strategy, (“Patient-Centered Care” OR “Delivery of Health Care, Integrated”) AND (“information science category”),” with the search field “MeSH Major Topic” from the MEDLINE database, we collected 2071 publications published between 2010 and March 15, 2021. We mapped the co-occurrence clusters of MeSH Major Topics with at least tagged by 10 publications and found that “Patient-Centered Care” tended to link with nursing practices, such as “nursing staff, hospital,” “nursing homes,” “family,” “nurse-patient relations,” “patient participation,” and “patient safety” (Figure 9).

**Figure 9.** The co-occurrence clusters of MeSH Major Topics in “Patient-Centered Care”–related informatics research publications.



Recently, there was an urgent discussion about a shift towards integrated patient-centered models of care [36]. We may conclude from these points that the focus of nursing informatics is oriented around patient-centered care, while MI is disease-oriented, and health informatics is population health-oriented. According to a WHO report published in 2015 [37], integrated people-centered health services mean putting the comprehensive needs of people and communities, not only diseases, at the center of health systems and empowering people to have a more active role in their own health. Traditional models of care, providing hyperspecialized, program-specific delivery to narrowly defined patient cohorts (eg, cardiac programs, diabetes programs), are unable to support integrated complex needs in a manner that achieves optimal outcomes. The focus, therefore, needs to return to the holistic treatment of the whole person, to be culturally and socially sensitive to individual needs, and where appropriate, to include individual, family group, or community models of care. Blending informatics expertise with nursing’s unique perspective on holistic health care ideally situates the profession to inform the integration of emerging models of care in a digital environment. We think that is a great opportunity for the development of nursing informatics.

### Interactions Between Health Demand, AI Supply, and Technological Applications

In our study, “computer neural networks” was one of the hottest informatics techniques in health AI research. China, in general, has only less than half as many health AI publications as the United States, yet has almost the same number of computer neural network publications as the United States. Research from Chinese scholars in the field of health AI primarily focuses on deep learning and is more likely to apply complex deep learning

models (such as deep neural networks) to health and medical diagnoses. This echoes the findings of a recent systematic review that compared the performance of diagnostic deep learning algorithms for medical imaging with that of expert clinicians [29]. Of the 10 randomized trial registrations for deep learning algorithms that were ultimately included, 8 were from China, and 1 was from the United States. Two trials have been completed, both from China, and their results were published in 2019. For the 81 nonrandomized studies that were included, the top 4 countries were as follows: United States (24/81, 30%), China (14/81, 17%), South Korea (12/81, 15%), and Japan (9/81, 11%). Chinese scholars are very active in diagnostic deep learning algorithms for medical imaging. AI-powered imaging became the most mature field within the intelligence and medical science industry, boasting a large market scale, substantial revenue earnings, and a favorable financing environment.

It should be noted that natural language processing is not a fully studied domain in health AI research in terms of H-I-T interactions. After reading the titles of publications, we found that almost all of the US studies in this dataset that used natural language processing were related to EHRs. This is yet further evidence that the level of openness of EMR data in Chinese significantly limits the opportunities for scholars to uncover its value.

The data show that EHRs are not fully studied, especially in China. This coincides with the following evidence from the United States as well as China. Through in-depth qualitative interviews across the United States, Sheikh and colleagues [38] investigated how to improve patient care and population health with HITs and how to reduce health care expenditure. Yet, they found that the following concerns persisted under existing systems: poor usability of EHRs, limited ability to support

multidisciplinary care, and major difficulties with using the health information exchange systems. On the other hand, Zhang and colleagues [39] explored the health applications for big data in China. Although more than 90% of Chinese hospitals use EMRs, sharing data is still difficult with hospital-based systems because they are developed by more than 300 vendors using different data standards. The investigation in the United States revealed that despite the government's substantial investment in information systems, there were barriers to integrated care due to system fragmentation. In China, the National Health Commission hosts the EHRs, the National Healthcare Security Administration hosts insurance claims data, and each hospital has set up a unique medical record system, but none of which are interoperable [5]. As such, it is critical to integrate the myriad of electronic health, medical, and claims records into a unified information system at all provider levels. Even with these challenges, the Chinese government vigorously promoted the development of big data and its application in the health and medical fields. China's State Council has announced that it will establish a national and provincial integrated population health information platform to facilitate data sharing, clinical research, and public health initiatives in the country.

The limitations of this study include that (1) only one database (ie, MEDLINE) was searched; (2) by using the combination of 2 branches of MeSH tags as a selection criterion, publications without or with inaccurately indexed MeSH terms could not be collected; and (3) the mappings of H-I-T interactions are probably not sufficient. A much more complicated and granular MeSH-based classification technique could have been developed. However, keeping the definition of the 3 H-I-T areas simple did not seem to limit our analysis, but rather it made the results easier to interpret. In addition, only using the MeSH Major Topics seems too strict, yet this approach ensures the core content of one given publication. In many countries, researchers need funding, and these will often be determined by research funding bodies or industry, which will determine research thrust and publications. In addition, in some countries where many applied research and developments are within publicly funded health institutions, publication is not prioritized even where there is innovation; in countries where HIT is competitively, commercially produced, research sponsors may limit publication to avoid what is seen as loss of commercial advantage.

## Conclusions

This study proposed a new approach to mapping the interplay between different knowledge entities by using the tree structure of MeSH to gain insights into the interactions between health demand, informatics supply, and technology applications in China. This method can help to collect publication data with a

broader interpretation of medical and health informatics and may also be applied to other interdisciplinary fields, such as medical physics, medical engineering, and medical social sciences.

China's emphasis on medical information technologies or HITs began with the new round of health care reform in 2009. Since then, medical and MI research in China has grown very fast, and the number of publications has exceeded that of the United Kingdom. The United States shows a relatively stable publication trend. While China has made these advances, some institutional and academic gaps still need to be filled in order to fully utilize the advantage of informatics in medical research and health care services. The following observations made throughout the analysis are described in the following paragraphs.

The interplay between health demand and informatics supply for China is slightly sparse, and the interactions between them were mostly observed in cardiovascular diseases, nervous system diseases, neoplasms, and population health, which are studied more with the help of computational methodologies and informatics techniques. Other techniques, such as social media, the internet, and other communication media are mainly used to solve public health problems and are rarely used in other disease research in the United States. While technological applications (T) have been established to link informatics supply (I) and health demand (H), the H-I-T linkages in informatics research in China are weaker than research in the United States. It is suggested that technological transfers, namely the functionality to be realized by medical/health informatics (eg, diagnosis, therapeutics, surgical procedures, laboratory testing techniques, and equipment and supplies) should be strengthened.

There is a positive correlation between the burden of diseases in China and the informatics research efforts for diseases. The major diseases targeted by informatics research efforts are cardiovascular diseases, neoplasms, and respiratory tract diseases, which differ in profile from those in US populations. China and the United States are using similar HITs to solve the different health needs in their respective countries.

China is unbalanced in its use of a combination of information science and medical and health sciences. The overall H-I interactions in China are sparse, focusing on several major diseases and 2 major informatics techniques. Research on EHRs combined with natural language processing should also be strengthened in China to improve the real-world applications of HITs and big data in health and medicine in the future.

All data used to calculate the HIT scores are stored in the Science Data Bank, which includes MeSH terms, MeSH tree list, and the paper list with HIT scores [40].

## Acknowledgments

We thank all the anonymous reviewers for their comments and suggestions. This work was funded by the National Natural Science Foundation of China (71603280, 72074006, 91846101), the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (2017QNRC001), and the Beijing Nova Program Interdisciplinary Cooperation Project (Z191100001119008).

## Conflicts of Interest

None declared.

## References

1. Haux R, Kulikowski C, Bakken S, de Lusignan S, Kimura M, Koch S, et al. Research Strategies for Biomedical and Health Informatics. Some Thought-provoking and Critical Proposals to Encourage Scientific Debate on the Nature of Good Research in Medical Informatics. *Methods Inf Med* 2018 Jan 31;56(S 01):e1-e10. [doi: [10.3414/me16-01-0125](https://doi.org/10.3414/me16-01-0125)] [Medline: [28119991](https://pubmed.ncbi.nlm.nih.gov/28119991/)]
2. D'Avolio L, Farwell WR, Fiore LD. Comparative effectiveness research and medical informatics. *Am J Med* 2010 Dec;123(12 Suppl 1):e32-e37 [FREE Full text] [doi: [10.1016/j.amjmed.2010.10.006](https://doi.org/10.1016/j.amjmed.2010.10.006)] [Medline: [21184865](https://pubmed.ncbi.nlm.nih.gov/21184865/)]
3. Otokiti A. Using informatics to improve healthcare quality. *IJHCQA* 2019 Mar 11;32(2):425-430. [doi: [10.1108/ijhcqa-03-2018-0062](https://doi.org/10.1108/ijhcqa-03-2018-0062)]
4. Simoes E. Health information technology advances health care delivery and enhances research. *Mo Med* 2015;112(1):37-40 [FREE Full text] [Medline: [25812273](https://pubmed.ncbi.nlm.nih.gov/25812273/)]
5. Yip W, Fu H, Chen AT, Zhai T, Jian W, Xu R, et al. 10 years of health-care reform in China: progress and gaps in Universal Health Coverage. *The Lancet* 2019 Sep;394(10204):1192-1204. [doi: [10.1016/s0140-6736\(19\)32136-1](https://doi.org/10.1016/s0140-6736(19)32136-1)]
6. Xi Focus: Xi stresses breaking new ground amid changes through reforms. *XinhuaNet*. 2020 Jul 01. URL: [http://www.xinhuanet.com/english/2020-07/01/c\\_139178782.htm](http://www.xinhuanet.com/english/2020-07/01/c_139178782.htm) [accessed 2021-04-14]
7. Lei J, Meng Q, Li Y, Liang M, Zheng K. The evolution of medical informatics in China: A retrospective study and lessons learned. *Int J Med Inform* 2016 Aug;92:8-14 [FREE Full text] [doi: [10.1016/j.ijmedinf.2016.04.011](https://doi.org/10.1016/j.ijmedinf.2016.04.011)] [Medline: [27318067](https://pubmed.ncbi.nlm.nih.gov/27318067/)]
8. Liang J, Wei K, Meng Q, Chen Z, Zhang J, Lei J. Development of medical informatics in China over the past 30 years from a conference perspective and a Sino-American comparison. *PeerJ* 2017;5:e4082 [FREE Full text] [doi: [10.7717/peerj.4082](https://doi.org/10.7717/peerj.4082)] [Medline: [29177118](https://pubmed.ncbi.nlm.nih.gov/29177118/)]
9. Liang J, Wei K, Meng Q, Chen Z, Zhang J, Lei J. The Gap in Medical Informatics and Continuing Education Between the United States and China: A Comparison of Conferences in 2016. *J Med Internet Res* 2017 Jun 21;19(6):e224 [FREE Full text] [doi: [10.2196/jmir.8014](https://doi.org/10.2196/jmir.8014)] [Medline: [28637638](https://pubmed.ncbi.nlm.nih.gov/28637638/)]
10. Bhattarai AK, Zarrin A, Lee J. Applications of information and communications technologies to public health: A scoping review using the MeSH term: "public health informatics". *Online J Public Health Inform* 2017 Sep 08;9(2):e192 [FREE Full text] [doi: [10.5210/ojphi.v9i2.7985](https://doi.org/10.5210/ojphi.v9i2.7985)] [Medline: [29026457](https://pubmed.ncbi.nlm.nih.gov/29026457/)]
11. 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/)]
12. Deng H, Wang J, Liu X, Liu B, Lei J. Evaluating the outcomes of medical informatics development as a discipline in China: A publication perspective. *Comput Methods Programs Biomed* 2018 Oct;164:75-85 [FREE Full text] [doi: [10.1016/j.cmpb.2018.07.001](https://doi.org/10.1016/j.cmpb.2018.07.001)] [Medline: [30195433](https://pubmed.ncbi.nlm.nih.gov/30195433/)]
13. Liang J, Li Y, Zhang Z, Shen D, Xu J, Yu G, et al. Evaluating the Applications of Health Information Technologies in China During the Past 11 Years: Consecutive Survey Data Analysis. *JMIR Med Inform* 2020 Feb 10;8(2):e17006 [FREE Full text] [doi: [10.2196/17006](https://doi.org/10.2196/17006)] [Medline: [32039815](https://pubmed.ncbi.nlm.nih.gov/32039815/)]
14. Liang J, Zheng X, Chen Z, Dai S, Xu J, Ye H, et al. The experience and challenges of healthcare-reform-driven medical consortia and Regional Health Information Technologies in China: A longitudinal study. *Int J Med Inform* 2019 Nov;131:103954 [FREE Full text] [doi: [10.1016/j.ijmedinf.2019.103954](https://doi.org/10.1016/j.ijmedinf.2019.103954)] [Medline: [31513943](https://pubmed.ncbi.nlm.nih.gov/31513943/)]
15. Hu D, Sun Z, Li H. An overview of medical informatics education in China. *Int J Med Inform* 2013 May;82(5):448-466 [FREE Full text] [doi: [10.1016/j.ijmedinf.2012.04.011](https://doi.org/10.1016/j.ijmedinf.2012.04.011)] [Medline: [22704233](https://pubmed.ncbi.nlm.nih.gov/22704233/)]
16. Liu J, Liu S, Li Y. Comparison of Medical/Health Informatics Education at the Best Global Universities for Clinical Medicine in Mainland China, Japan and South Korea. *Stud Health Technol Inform* 2019 Aug 21;264:1958-1959. [doi: [10.3233/SHTI190733](https://doi.org/10.3233/SHTI190733)] [Medline: [31438427](https://pubmed.ncbi.nlm.nih.gov/31438427/)]
17. Kastrati Z, Imran AS, Yayilgan SY. The impact of deep learning on document classification using semantically rich representations. *Information Processing & Management* 2019 Sep;56(5):1618-1632. [doi: [10.1016/j.ipm.2019.05.003](https://doi.org/10.1016/j.ipm.2019.05.003)]
18. Liu Y, Wacholder N. Evaluating the impact of MeSH (Medical Subject Headings) terms on different types of searchers. *Information Processing & Management* 2017 Jul;53(4):851-870 [FREE Full text] [doi: [10.1016/j.ipm.2017.03.004](https://doi.org/10.1016/j.ipm.2017.03.004)]
19. Weber GM. Identifying translational science within the triangle of biomedicine. *J Transl Med* 2013 May 24;11(1):126 [FREE Full text] [doi: [10.1186/1479-5876-11-126](https://doi.org/10.1186/1479-5876-11-126)] [Medline: [23705970](https://pubmed.ncbi.nlm.nih.gov/23705970/)]
20. Hutchins BI, Davis MT, Meseroll RA, Santangelo GM. Predicting translational progress in biomedical research. *PLoS Biol* 2019 Oct 10;17(10):e3000416 [FREE Full text] [doi: [10.1371/journal.pbio.3000416](https://doi.org/10.1371/journal.pbio.3000416)] [Medline: [31600189](https://pubmed.ncbi.nlm.nih.gov/31600189/)]
21. Ke Q. Identifying translational science through embeddings of controlled vocabularies. *J Am Med Inform Assoc* 2019 Jun 01;26(6):516-523 [FREE Full text] [doi: [10.1093/jamia/ocy177](https://doi.org/10.1093/jamia/ocy177)] [Medline: [30830170](https://pubmed.ncbi.nlm.nih.gov/30830170/)]
22. Agarwal P, Searls DB. Can literature analysis identify innovation drivers in drug discovery? *Nat Rev Drug Discov* 2009 Nov;8(11):865-878 [FREE Full text] [doi: [10.1038/nrd2973](https://doi.org/10.1038/nrd2973)] [Medline: [19876041](https://pubmed.ncbi.nlm.nih.gov/19876041/)]



23. Leydesdorff L, Rotolo D, Rafols I. Bibliometric perspectives on medical innovation using the medical subject Headings of PubMed. *J Am Soc Inf Sci Tec* 2012 Oct 15;63(11):2239-2253. [doi: [10.1002/asi.22715](https://doi.org/10.1002/asi.22715)]
24. Petersen A, Rotolo D, Leydesdorff L. A triple helix model of medical innovation: Supply, demand, and technological capabilities in terms of Medical Subject Headings. *Research Policy* 2016 Apr;45(3):666-681. [doi: [10.1016/j.respol.2015.12.004](https://doi.org/10.1016/j.respol.2015.12.004)]
25. Makhni S, Atreja A, Sheon A, Van Winkle B, Sharp J, Carpenter N. The Broken Health Information Technology Innovation Pipeline: A Perspective from the NODE Health Consortium. *Digit Biomark* 2017 Aug 3;1(1):64-72 [FREE Full text] [doi: [10.1159/000479017](https://doi.org/10.1159/000479017)] [Medline: [32095746](https://pubmed.ncbi.nlm.nih.gov/32095746/)]
26. Rigby M, Ammenwerth E. The Need for Evidence in Health Informatics. *Stud Health Technol Inform* 2016;222:3-13. [Medline: [27198087](https://pubmed.ncbi.nlm.nih.gov/27198087/)]
27. Ammenwerth E. Evidence-based Health Informatics: How Do We Know What We Know? *Methods Inf Med* 2015 Jan 22;54(04):298-307. [doi: [10.3414/me14-01-0119](https://doi.org/10.3414/me14-01-0119)] [Medline: [26196349](https://pubmed.ncbi.nlm.nih.gov/26196349/)]
28. Feldman SS, Buchalter S, Hayes LW. Health Information Technology in Healthcare Quality and Patient Safety: Literature Review. *JMIR Med Inform* 2018 Jun 04;6(2):e10264. [doi: [10.2196/10264](https://doi.org/10.2196/10264)] [Medline: [29866642](https://pubmed.ncbi.nlm.nih.gov/29866642/)]
29. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020 Mar 25;368:m689 [FREE Full text] [doi: [10.1136/bmj.m689](https://doi.org/10.1136/bmj.m689)] [Medline: [32213531](https://pubmed.ncbi.nlm.nih.gov/32213531/)]
30. Séroussi B, Jaulent M, Lehmann C. Looking for the Evidence: Value of Health Informatics. Editorial. *Yearb Med Inform* 2018 Mar 05;22(01):04-06. [doi: [10.1055/s-0038-1638825](https://doi.org/10.1055/s-0038-1638825)]
31. Kowatsch T, Otto L, Harperink S, Cotti A, Schlieter H. A design and evaluation framework for digital health interventions. *Information Technology* 2019;61(5-6):253-263. [doi: [10.1515/itit-2019-0019](https://doi.org/10.1515/itit-2019-0019)]
32. Hamilton NE, Ferry M. ggtern: Ternary Diagrams Using ggplot2. *J. Stat. Soft* 2018;87(1):1-17. [doi: [10.18637/jss.v087.c03](https://doi.org/10.18637/jss.v087.c03)]
33. Yegros A, Van de Klippe W, Abad-Garcia MF, Rafols I. Exploring Why Global Health Needs Are Unmet by Public Research Efforts: The Potential Influences of Geography, Industry, and Publication Incentives. *Health Research Policy and Systems* 2020;18(1):47. [doi: [10.2139/ssrn.3459230](https://doi.org/10.2139/ssrn.3459230)]
34. Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000–2016. World Health Organization. 2018. URL: <https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates> [accessed 2021-02-16]
35. Zhou M, Wang H, Zeng X, Yin P, Zhu J, Chen W, et al. Mortality, morbidity, and risk factors in China and its provinces, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* 2019 Sep;394(10204):1145-1158. [doi: [10.1016/s0140-6736\(19\)30427-1](https://doi.org/10.1016/s0140-6736(19)30427-1)]
36. Hussey P, Kennedy MA. Instantiating informatics in nursing practice for integrated patient centred holistic models of care: a discussion paper. *J Adv Nurs* 2016 May;72(5):1030-1041. [doi: [10.1111/jan.12927](https://doi.org/10.1111/jan.12927)] [Medline: [26890201](https://pubmed.ncbi.nlm.nih.gov/26890201/)]
37. Framework on Integrated People-Centered Health Services: An Overview. World Health Organization. URL: <https://www.who.int/servicedeliverysafety/areas/people-centred-care/fullframe.pdf> [accessed 2021-04-15]
38. Sheikh A, Sood HS, Bates DW. Leveraging health information technology to achieve the "triple aim" of healthcare reform. *J Am Med Inform Assoc* 2015 Jul;22(4):849-856 [FREE Full text] [doi: [10.1093/jamia/ocv022](https://doi.org/10.1093/jamia/ocv022)] [Medline: [25882032](https://pubmed.ncbi.nlm.nih.gov/25882032/)]
39. Zhang L, Wang H, Li Q, Zhao M, Zhan Q. Big data and medical research in China. *BMJ* 2018 Feb 05;360:j5910 [FREE Full text] [doi: [10.1136/bmj.j5910](https://doi.org/10.1136/bmj.j5910)] [Medline: [29437562](https://pubmed.ncbi.nlm.nih.gov/29437562/)]
40. Du J, Chen T, Zhang L. HIT score for 213,215 medical informatics publications during 2010-2020. V2. Science Data Bank. URL: <http://www.doi.org/10.11922/sciencedb.00790> [accessed 2021-06-30]

## Abbreviations

**AI:** artificial intelligence

**AMIA:** American Medical Informatics Association

**CHIMA:** China Hospital Information Management Association

**CHINC:** China Hospital Information Network Annual Conference

**CHITEC:** China Health Information Technology Exchange Annual Conference

**CMIAAS:** China Medical Information Association Annual Symposium

**CPMI:** China Annual Proceeding of Medical Informatics

**DALY:** Disability-Adjusted Life Year

**EHR:** electronic health record

**EMR:** electronic medical record

**GBD:** Global Burden of Disease

**HAC:** Human, Animal, and Molecular/Cellular Biology

**HIMSS:** Healthcare Information and Management Systems Society

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

**HIT:** health information technology

**HITECH:** Health Information Technology for Economic and Clinical Health

**ICD:** International Classification of Diseases

**ICT:** information and communications technology

**MeSH:** Medical Subject Headings

**MI:** medical informatics

**NIH:** National Institutes of Health

**WHO:** World Health Organization

*Edited by G Eysenbach; submitted 09.12.20; peer-reviewed by J Liang, M Rigby, C Smith; comments to author 06.01.21; revised version received 07.05.21; accepted 12.05.21; published 06.07.21.*

*Please cite as:*

*Du J, Chen T, Zhang L*

*Measuring the Interactions Between Health Demand, Informatics Supply, and Technological Applications in Digital Medical Innovation for China: Content Mapping and Analysis*

*JMIR Med Inform 2021;9(7):e26393*

*URL: <https://medinform.jmir.org/2021/7/e26393>*

*doi: [10.2196/26393](https://doi.org/10.2196/26393)*

*PMID: [34255693](https://pubmed.ncbi.nlm.nih.gov/34255693/)*

©Jian Du, Ting Chen, Luxia Zhang. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 06.07.2021. 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

# Assessing the Performance of Clinical Natural Language Processing Systems: Development of an Evaluation Methodology

Lea Canales<sup>1</sup>, PhD; Sebastian Menke<sup>2</sup>, PhD; Stephanie Marchesseau<sup>2</sup>, PhD; Ariel D'Agostino<sup>2</sup>, MD; Carlos del Rio-Bermudez<sup>2</sup>, PhD; Miren Taberna<sup>2</sup>, Dr med; Jorge Tello<sup>2</sup>, MSc

<sup>1</sup>Department of Software and Computing System, University of Alicante, Alicante, Spain

<sup>2</sup>MedSavana SL, Madrid, Spain

**Corresponding Author:**

Jorge Tello, MSc

MedSavana SL

Calle Gran Vía 30, Planta 10

Madrid, 28013

Spain

Phone: 34 627906138

Email: [jtello@savanamed.com](mailto:jtello@savanamed.com)

## Abstract

**Background:** Clinical natural language processing (cNLP) systems are of crucial importance due to their increasing capability in extracting clinically important information from free text contained in electronic health records (EHRs). The conversion of a nonstructured representation of a patient's clinical history into a structured format enables medical doctors to generate clinical knowledge at a level that was not possible before. Finally, the interpretation of the insights gained provided by cNLP systems has a great potential in driving decisions about clinical practice. However, carrying out robust evaluations of those cNLP systems is a complex task that is hindered by a lack of standard guidance on how to systematically approach them.

**Objective:** Our objective was to offer natural language processing (NLP) experts a methodology for the evaluation of cNLP systems to assist them in carrying out this task. By following the proposed phases, the robustness and representativeness of the performance metrics of their own cNLP systems can be assured.

**Methods:** The proposed evaluation methodology comprised five phases: (1) the definition of the target population, (2) the statistical document collection, (3) the design of the annotation guidelines and annotation project, (4) the external annotations, and (5) the cNLP system performance evaluation. We presented the application of all phases to evaluate the performance of a cNLP system called "EHRead Technology" (developed by Savana, an international medical company), applied in a study on patients with asthma. As part of the evaluation methodology, we introduced the Sample Size Calculator for Evaluations (SLiCE), a software tool that calculates the number of documents needed to achieve a statistically useful and resourceful gold standard.

**Results:** The application of the proposed evaluation methodology on a real use-case study of patients with asthma revealed the benefit of the different phases for cNLP system evaluations. By using SLiCE to adjust the number of documents needed, a meaningful and resourceful gold standard was created. In the presented use-case, using as little as 519 EHRs, it was possible to evaluate the performance of the cNLP system and obtain performance metrics for the primary variable within the expected CIs.

**Conclusions:** We showed that our evaluation methodology can offer guidance to NLP experts on how to approach the evaluation of their cNLP systems. By following the five phases, NLP experts can assure the robustness of their evaluation and avoid unnecessary investment of human and financial resources. Besides the theoretical guidance, we offer SLiCE as an easy-to-use, open-source Python library.

(*JMIR Med Inform* 2021;9(7):e20492) doi:[10.2196/20492](https://doi.org/10.2196/20492)

**KEYWORDS**

natural language processing; clinical natural language processing; electronic health records; gold standard; reference standard; sample size

## Introduction

Over the last decades, health care institutions have increasingly abandoned clinical records in paper form and have started to store patients' longitudinal medical information in electronic health records (EHRs). EHRs are widely available and capture large amounts of valuable clinical information from medical backgrounds, examinations, laboratory testing, procedures, and prescriptions [1]. While some clinical data are codified in the structured fields of EHRs, the great majority of relevant clinical information appears embedded within the unstructured narrative free-text [2]. In this free-text section, physicians write down their routine evaluation of the patient and thereby offer a window into real-world clinical practices [3,4].

The resulting exponential growth of digitized data on real-world clinical practice has given rise to specialized research fields such as clinical natural language processing (cNLP) [5,6], which aims at exploring the clinically relevant information contained in EHRs [7-9]. The importance and complexity of improving cNLP systems has given rise to a strong engagement among researchers in developing methods capable of doing so [10-16]. This resulted in improved cNLP systems that have dramatically changed the scale at which information contained in the free-text portion of EHRs can be utilized [17-20] and has provided valuable insights into clinical populations [21-27], epidemiology trends [28-30], patient management [31], pharmacovigilance [32], and optimization of hospital resources [33].

However, there is a lack of guidance on how to evaluate those cNLP systems [34]. Although some ground-breaking work was done by Biber [35] and Paroubek et al [36], who analyzed the representativeness in linguistic corpora and the quantity and quality of annotations needed to establish a representative gold standard, hardly any proposal exists for an end-to-end evaluation methodology of cNLP systems. Criteria for the evaluation of cNLP systems were provided by Friedman and Hripcsak [37] and, 10 years later, Velupillai et al [38]. Those are actionable suggestions to improve the quality of cNLP system evaluations. Based on their judgment, the provision of details about the number of domain experts who participated in the creation of the reference standard, mentions of the sample size, defining the objective of the study, and the presentation of performance measure CIs were deemed relevant aspects that provide robustness to cNLP evaluations [35-37]. Such criteria are key to advancement in cNLP [37] because of the direct and

existential impact these systems have on understanding patients and diseases [39].

A crucial point for the evaluation of a cNLP system is the availability of benchmark data sets in a specific language based on real EHRs. Although many corpora for the medical domain are available in English, they are scarce or nonexistent for other languages. As a consequence, many benchmarks have been designed a priori for clinical publications and are not real EHRs [40]. The downside of this practice is that some important values present in real EHRs are not contained in artificial EHRs. For example, the validation of artificial data sets may not include variables or concepts of the pathology of interest or research objective. Furthermore, real-world data sets entail misspellings, acronyms, and other particularities of the free-text narratives of patients' EHRs, which can be taken into account in the validation process, thereby providing a much more accurate and generalizable evaluation of the cNLP system [41]. Obviously, the use of actual EHRs obliges researchers to implement the necessary steps and tools to guarantee the confidentiality and security of the data, in compliance with hospital ethics committees, national and international regulations, and pharmaceutical industry policies.

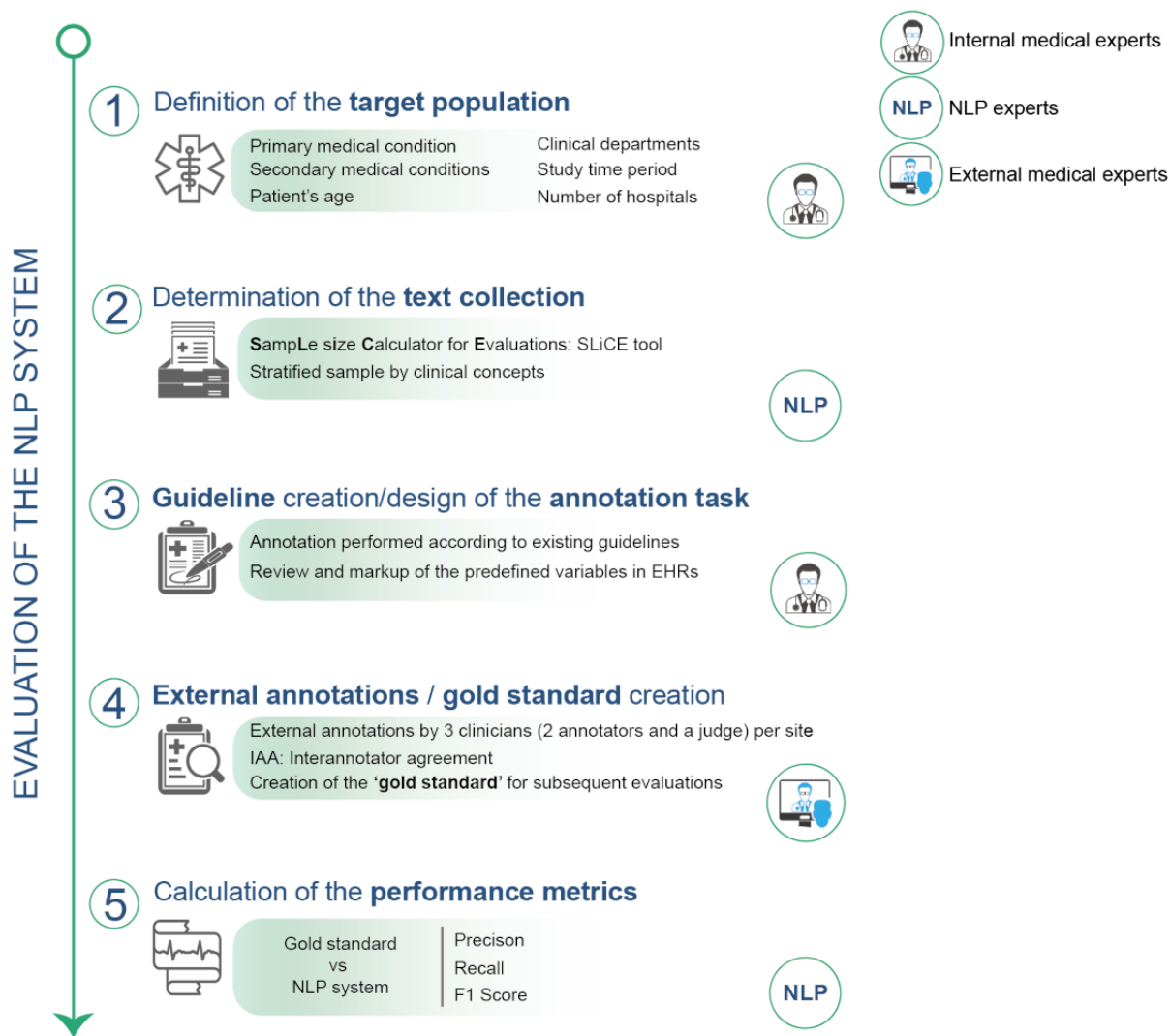
Here, we propose a language-independent evaluation methodology that can help researchers to overcome some of the mentioned obstacles when evaluating their cNLP system. Our objective is to provide a state-of-the-art methodology for the evaluation of cNLP systems, thereby guiding researchers in the field of natural language processing (NLP) in this complex process to ensure the robustness and representativeness of the system's performance metrics. The proposed evaluation methodology is the result of our experiences developing cNLP evaluations in real use-cases dealing with heterogeneous EHRs focusing on a wide range of pathologies from one or several hospitals in different countries.

## Methods

Our evaluation methodology is a set of methods and principles used to perform a cNLP system evaluation, which extends from the establishment of the reference standard to the measurement and presentation of the evaluation metrics. It consists of five phases : (1) definition of the target population, (2) statistical document collection, (3) design of the annotation guidelines and annotation project, (4) external annotations and gold standard creation, and (5) cNLP system performance evaluation (Figure 1).



**Figure 1.** The proposed evaluation methodology consists of five phases that guarantee the evaluation of a clinical natural language processing system against a gold standard providing unbiased performance metrics. NLP: natural language processing, EHR: electronic health record.



In the following paragraphs, we present the five phases of the proposed evaluation methodology in the context of cNLP systems. However, this approach is not limited to cNLP systems and the phases can be adapted to perform equally useful evaluations of nonclinical NLP systems.

**Phase 1: Definition of the Target Population**

The target population is defined by sets of nonlinguistic and linguistic characteristics. Nonlinguistic characteristics of the target population are, for example, the type of hospitals that participate in the evaluation, as this defines the clinical departments commonly in charge of those patients, or factors such as patient age (eg, patients under 18 years of age for a pediatric disease) or gender (eg, men for studying prostate cancer). Linguistic characteristics on the other hand are related to the actual written content in an EHR such as mentions of the primary and secondary medical conditions being evaluated. It is highly recommended to consider secondary medical conditions since they help to determine the criteria of sampling. A list of questions related to the nonlinguistic and linguistic characteristics, which needs to be answered by the responsible

medical experts, helps to identify the scope of the cNLP system evaluation, the requisites for sampling, and the sample size:

- Patient age: is the patient’s age relevant in the studied pathology?
- Hospitals: which hospitals will participate in this evaluation?
- Clinical departments: are there any clinical departments related to the disease that are relevant for this evaluation?
- Time: is there a period of time in which the evaluation should be carried out? (study period)
- Primary medical condition (primary variable): which disease or primary medical condition will be evaluated?
- Secondary medical conditions (secondary variables): which other medical conditions or medical evaluations (eg, symptoms, signs, treatments, or tests) will be considered?

**Phase 2: Statistical Document Collection Using the Sample Size Calculator for Evaluations**

Determining the amount of data needed to capture enough linguistics to be statistically robust as well as selecting the sample to produce consistent performance measures, has been

an open question in NLP research for more than a decade [35-37]. In our evaluation methodology, a linguistic event refers to a particular clinical concept mentioned in EHRs such as a disease, a symptom, or a sign. Thus, the aim of phase 2 is to build a corpus which represents the characteristics of the population as closely as possible by combining an in-house software tool called Sample Size Calculator for Evaluations (SLiCE) and stratified sampling.

### SLiCE

SLiCE is a publicly available software [42] developed by Savana, an international medical company, that enables users to estimate the minimum sample size required to obtain robust metrics of reading performance, whereby robustness is determined by predefining the CI and level. The method was designed using the standard metrics commonly applied in NLP system evaluations: precision (P), recall (R), and F1-score [43]. The input parameters of SLiCE are (1) the desired confidence level (1-alpha), (2) the CI width, (3) expected values of precision and recall, (4) the frequency of the linguistic event to evaluate, and (5) whether this frequency has been calculated “internally” or “externally.” The output of SLiCE contains the sample size as well as the number of positive and negative samples required to ensure the CI for the linguistic events evaluated. The final number of documents to manually annotate is to be shared equally among the participating hospitals in case of a multisite evaluation.

The fundamentals of SLiCE are based on the sample size determination method [44] for proportions [45] and the expected occurrence rate (prevalence) of a linguistic event in the total population. The method consists of fixing a confidence level and a CI to calculate the sample size required to achieve the desired CI. In our proposal, the Clopper-Pearson approach is

employed for CI calculation [45] since it is a common method for calculating binomial CI. Under the Clopper-Pearson approach, the lower and upper confidence limits are determined by:

$$\frac{r}{n}$$

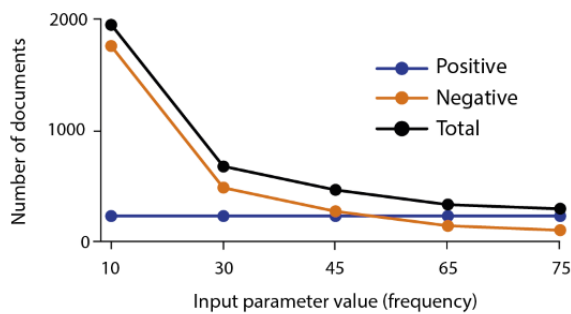
where  $n$  is the number of trials (sample size),  $F$  is the F-Snedecor distribution,  $r$  is the number of successes, and  $\alpha$  is the significance level (eg, 5%).

The proposed method is applicable when the objective is to assess a linguistic event or a set of linguistic events. Consequently, the definition of the target population is key to applying SLiCE since the calculation of the prevalence of the event in the target population is a requirement.

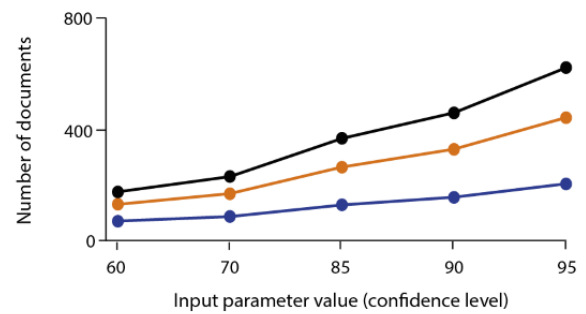
The expected values of precision and recall represent values that are considered achievable by the system. Care should be taken not to overestimate the performance of the system by introducing values higher than 90% when the actual performance is below. This would result in a very small sample size and, consequently, final metrics that are not very robust. If our system achieves values in the evaluation that are far from the expected ones, the probability of complying with the CI is low. Therefore, we recommend applying realistic values of P and R (around 80%) to ensure the robustness of the final metrics. The impact of the frequency of a main variable is the most influential input as more negative examples are needed in case of low frequency to guarantee a representative sample. To achieve a more robust cNLP system evaluation, more documents would need to be annotated. On the contrary, if high recall and precision are expected, the total number of documents to verify this expectation is lower than when low recall and precision are expected (Figure 2).

**Figure 2.** Analysis of SLiCE (Sample Size Calculator for Evaluations) outputs according to changes in input parameters and their impact on the number of documents to be selected for the gold standard.

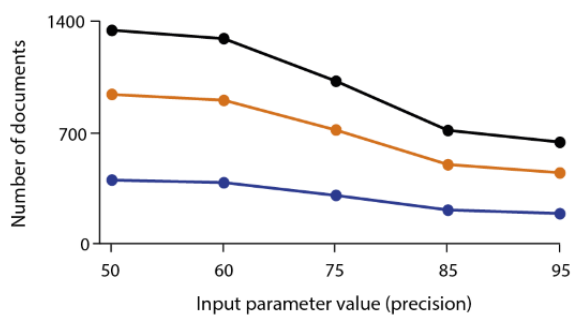
**A. Frequency**



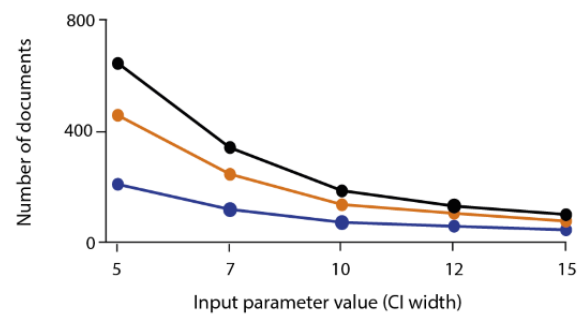
**D. Confidence level**



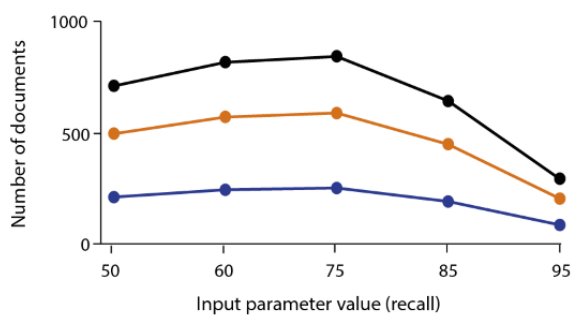
**B. Precision**



**E. CI width**



**C. Recall**



**DEFAULT INPUT PARAMETERS (INTERNAL)**

- Frequency = 30%
- Precision = 95%
- Recall = 85%
- Confidence level = 95%
- CI width = 5%

SLiCE has two additional options which are “internal” and “external.” When “internal” is selected (default), the occurrence rate of the main variable needs to be calculated. To achieve that, the prevalence of this linguistic event can be calculated using the data provided in each hospital. Thus, for each hospital participating in a study, the frequency using the following formula can be calculated:



In the case of developing the evaluation for several hospitals, the final frequency is the average of the occurrence rate of the main variable in each hospital. Poor prevalence variables might require a very large number of documents to be annotated, which is not feasible in practice. However, the prevalence could be measured not from an entire database, but from a subset of relevant EHRs (eg, only a specific department).

When a set of linguistic events (eg, clinical concepts) is evaluated, SLiCE needs to be applied to each clinical concept that defines the target population to ensure the expected CI for all the clinical events. However, this may not always be possible due to time and budget restrictions associated with an evaluation. For this reason, our methodology proposes to apply SLiCE at least for the primary variable defined in the target population. Consequently, for the secondary variables, the sample size does not need to be calculated because they depend on the sample size calculated for the primary variable.

It is important to note that the parameters of the calculator should be decided by the medical experts together with NLP experts in charge of the study based on their expectations regarding the performance of the system. A detailed explanation about the SLiCE algorithm can be found in [Multimedia Appendix 1](#), and a guide on how to use the open-source SLiCE can be found on GitHub [42].

### Stratified Sampling by Clinical Concepts

Once the number of documents needed to create the gold standard as well as the numbers of positive and negative examples needed for the primary variables are calculated using SLiCE, the EHRs to be included in the final validation data set can be selected. In order to stratify clinical concepts, we need to collect the samples of each variable from the subset of documents from the target population. First, the positive and negative examples of the primary variable are selected according to SLiCE. In a second step, negative examples for secondary variables are randomly selected from reports excluding the primary variable. Positive examples for secondary variables are collected using a stratified sampling as a method of probabilistic sampling where the subgroups are identified by each secondary variable to ensure the representativeness of each linguistic event.

### Phase 3: Design of the Annotation Task and Guidelines

The preparation of the annotation project requires the cooperation of NLP experts and the internal medical experts (developers of the study). The annotation task itself is a manual process in which annotators (external medical experts of the participating hospital) review and mark up the predefined variables in the text for each EHR of the gold standard. To guarantee the quality of the resulting annotations [46], it is important to carefully design both the annotation guidelines and the annotation task.

The annotation guidelines consist of a set of instructions that explain what exactly the annotation task consists of. For instance, these guidelines will include the list of variables the annotators are expected to annotate in the free text, as well as resolve possible doubts related to, for example, synonyms or the inclusion of negative concepts. The creation of the guidelines is an iterative process in which NLP experts and internal medical experts participate. Using the initial draft of the annotation guidelines, the annotators are required to perform the annotation task on a small subset of documents in order to validate the design of the annotation project and correct, when applicable, the guidelines. This iterative process ensures that the instructions are clear before the start of the actual annotation task. The final guidelines need to be followed by each participating annotator in order to assure the consistency of annotations, especially across participating institutions. The process described here must always be applied, regardless of the study, the annotation tool (we use Inception at Savana [47]), or the number of documents included in the evaluation.

### Phase 4: External Annotations and Gold Standard Creation

Once the annotation project is prepared for each hospital participating in the study, the external annotation task can start. In this phase, 2 annotators (external medical experts) from each hospital will review independently and blindly (meaning they do not know which document they are annotating compared to their colleague) the whole set of documents selected in the previous steps. It is important to note that the 2 annotators are not allowed to communicate with each other or with the annotation project creators. Their only source of information are the annotation guidelines.

Once all the annotations have been completed by both annotators, a curator (additional external clinical expert) from that same institution is assigned to check every annotation for which the annotators disagree and to make the final decision. This final decision will be the one used for the gold standard creation that later serves to evaluate the cNLP system, while the two previous annotations are used to measure the interannotator agreement (IAA). The IAA is a commonly used approach in cNLP system evaluations [48-50] to identify the upper performance level.

### Phase 5: NLP Performance Evaluation

To measure the quality of annotations and to obtain target metrics for the cNLP system, it is necessary to assess them by measuring the IAA after full completion of the annotation task by the external medical experts. In our methodology, the IAA is calculated using the F1-score [51]. A low agreement can indicate that the annotators might have had difficulties in linguistically identifying the respective variables in the EHRs or that the guidelines are still inadequate in properly describing the annotation task. Thus, the IAA serves as a control mechanism to check the reliability of the annotation and further to establish a target of performance for the cNLP system.

The performance evaluation of the cNLP system is calculated using the standard metrics precision, recall, and their harmonic mean F1-score [43]. P gives us an indicator of the accuracy of information retrieved by the system (equation 3), R gives us an indicator of the amount of information the system retrieves (equation 4), and the F1-score gives us an overall performance indicator of information retrieval (equation 5):



In all cases, true positives are the sum of records correctly retrieved, false negatives are the sum of records not retrieved, and false positives are the sum of records incorrectly retrieved.

In addition to these metrics, the 95% CI for each aforementioned measure can be calculated since this provides information about the range in which the true value lies and thus how robust the metric is. The method employed to calculate the CI is the Clopper-Pearson approach [45], one of the most common methods for calculating binomial CI.

## Results

### Application of the Methodology

The proposed evaluation methodology has been applied for the evaluation of cNLP systems in several clinical research projects at Savana. In this section, we give one example of its application in a project aimed at estimating the prevalence of severe asthma in the Spanish hospital population using Savana's cNLP system "EHRRead Technology".

### Phase 1: Definition of the Target Population

For this study, the population was defined by adult patients with asthma (the primary medical condition), with EHRs available from multiple hospitals and a study period of several years.



## Phase 2: Statistical Document Selection Using SLiCE

### SLiCE

With an average internal frequency for asthma of 48.5% in the target population of the participating hospitals (see subsection “Phase 2: Statistical Document Collection Using the Sample Size Calculator for Evaluations” of the Methods section for details on how this was calculated), an expected precision of 85% and recall of 80% (to be on the safe side, as explained in the previous section) with an interval width of 5% and expected CI of 95%, we obtained the following sample sizes:

- 249 positive examples;
- 270 negative examples;
- 519 total number of documents to annotate;
- 87 documents per hospital (42 positive examples, 45 negative examples).

### Stratified Sampling by Clinical Concepts

In order to ensure the representativeness of the secondary variables of interest to the study, a stratifying approach was applied as explained in the subsection “Phase 2: Statistical Document Collection Using the Sample Size Calculator for Evaluations” of the Methods section (Table 1).

**Table 1.** Study variables detected in selected documents by the clinical natural language processing (cNLP) system compared to the ones obtained by manual annotations.

Variable	Manual annotations	cNLP system detections
Asthma (primary variable)	281	289
Extrinsic asthma	65	49
Bronchodilation test	131	88
Eosinophils in blood	181	164
Gastroesophageal reflux syndrome	181	168
Obesity	54	50
Omalizumab	27	21
Prick test	162	152
Salmeterol + fluticasone	147	80
Total IgE <sup>a</sup>	106	104

<sup>a</sup>IgE: immunoglobulin E.

## Phase 3: Design of the Annotation Task and Guidelines

External medical experts (annotators) were asked to mark the appearance of the clinical variables of interest in the free text of EHRs selected for the gold standard. In this project, we used the annotation tool Inception to facilitate the annotation task [47]. The annotation guidelines can be stored in this annotation tool, and annotators can access them via the user interface at any time during the annotation task.

## Phase 4: External Annotations and Gold Standard Creation

A crucial indicator is the IAA, which describes the difficulty of the external medical experts in evaluating the variables in the free text of EHRs and to set the target for the cNLP system performance. In the asthma study, the validation task appeared to be difficult as suggested by the suboptimal IAA F1-scores of several variables (Table 2). It was noted that the primary variable (asthma) and the first secondary variable (extrinsic asthma) may intersect, leading to confusion among medical experts. Once both annotators finished their annotations and the IAA was calculated, a third external medical expert resolved disagreements for the creation of the gold standard.

**Table 2.** Interannotator agreement (IAA) F1-scores for the primary and secondary variables of the annotation task.

Variable	IAA F1-score (95% CI)
Asthma (primary variable)	0.77 (0.70-0.82)
Extrinsic asthma	0.76 (0.58-0.88)
Bronchodilation test	0.86 (0.78-0.92)
Eosinophils in blood	0.68 (0.57-0.76)
Gastroesophageal reflux syndrome	0.82 (0.73-0.89)
Obesity	0.74 (0.51-0.87)
Omalizumab	0.88 (0.74-0.95)
Prick test	0.72 (0.62-0.80)
Salmeterol + fluticasone	0.81 (0.71-0.88)
Total IgE <sup>a</sup>	0.60 (0.45-0.72)

<sup>a</sup>IgE: immunoglobulin E.

### Phase 5: NLP Performance Evaluation

After the curation of the disagreements between the annotations of the external medical experts, the final gold standard was compared to the cNLP system, leading to higher precision and recall than expected for the primary variable and a CI width of 90-96 for precision and 94-98 for recall. The expected precision and recall used in SLiCE were underestimated compared to the

final metrics, which means that even fewer reports could have been annotated. However, using as little as 519 EHRs, it was possible to evaluate the performance of the cNLP system and obtain performance for the primary variable within the expected CI range (Table 3). Interestingly, the performance metrics of the secondary variables were also high (>0.79) apart from one variable (total immunoglobulin E: F1=0.64) for which the IAA was also low (0.60).

**Table 3.** Performance metrics for primary and secondary variables when comparing the clinical natural language processing system to the gold standard.

Variable	Precision (95% CI)	Recall (95% CI)	F1 value (95% CI)
Asthma (primary variable)	0.94 (0.90-0.96)	0.96 (0.94-0.98)	0.95 (0.92-0.97)
Extrinsic asthma	1.00 (0.93-1.00)	0.75 (0.63-0.85)	0.86 (0.75-1.00)
Bronchodilation test	0.99 (0.94-1.00)	0.66 (0.58-0.74)	0.79 (0.71-0.85)
Eosinophils in blood	0.99 (0.96-1.00)	0.90 (0.84-0.94)	0.94 (0.90-0.97)
Gastroesophageal reflux syndrome	1.00 (0.98-1.00)	0.93 (0.88-0.96)	0.96 (0.93-1.00)
Obesity	1.00 (0.93-1.00)	0.93 (0.82-0.98)	0.96 (0.87-1.00)
Omalizumab	1.00 (0.84-1.00)	0.78 (0.58-0.91)	0.93 (0.68-1.00)
Prick test	0.95 (0.91-0.98)	0.90 (0.84-0.94)	0.92 (0.87-0.96)
Salmeterol + fluticasone	0.98 (0.91-1.00)	0.53 (0.45-0.61)	0.96 (0.60-0.76)
Total IgE <sup>a</sup>	0.64 (0.54-0.74)	0.63 (0.53-0.72)	0.64 (0.54-0.73)

<sup>a</sup>IgE: immunoglobulin E.

## Discussion

### Principal Findings

We developed an easy-to-follow evaluation methodology, based on our experience with cNLP system evaluations using real-world clinical data, to provide guidance on how to evaluate their performance [36,38]. Our motivation was to be able to assure the robustness and representativeness of the performance metrics of evaluations of our cNLP systems, which is crucial for their application in clinical research. We presented the application of our evaluation methodology on a named-entity recognition cNLP system; however, the methodology can easily be adapted to other NLP tasks by adjusting the questions in

phase 1 to the area of the respective NLP system. We routinely apply this methodology in our real-world evidence clinical studies and hope that it is equally useful for other NLP experts to reach a statistically sound evaluation of their own cNLP/NLP systems.

The first phase, the definition of the target population, is crucial for a successful evaluation [35] since it establishes the requisites for sampling and, most importantly, the scope of the cNLP system evaluation. While the linguistic characteristics to be considered for an evaluation are obvious, the questions needed to define the nonlinguistic characteristics are less obvious and require the insights of medical experts. Not properly defining the target population may lead to false expectations [37], a

situation that can be avoided by initializing the cNLP system evaluation answering those questions. In our example, the definition of the target population was straightforward, but depending on the primary medical condition to be studied, this can be much more complex [52].

To ensure that the information extracted by a cNLP system is reliable and accurate, its output must be validated against a corpus of expert-reviewed clinical notes in terms of precision and recall of extracted medical terms. Thus, phase 2 of the evaluation methodology applied SLiCE, a statistical tool that offers guidance for the determination of a gold standard's minimum sample size to ensure the expected levels of CIs of the linguistic events in cNLP system evaluations. The resulting gold standard contains a representative set of EHRs [35] based on the SLiCE output for the main variable in combination with the stratifying approach for the secondary variables, which has shown to lead to a much more representative gold standard than simple random sampling [35]. Frequently, evaluations are carried out using reference standards that are too small to be statistically useful, which might be due to limited resources [53], or that apply, at the other end of the extreme, a resource-wasting "the more, the better" approach [54]. Both scenarios are not satisfactory, and the use of SLiCE can help to avoid them without compromising the robustness of the evaluation or wasting resources.

In situations where the data source lacks predefined categories, other techniques such as discriminant text selection could be applied to limit the population from which to sample (eg, classifying EHRs into clinical services or departments). Thus, the frequency of the primary variable could be calculated over the category of interest, thereby increasing its frequency and lowering the amount of documents to be annotated. In phase 3, the annotation project is prepared with the input of both internal medical experts as well as NLP experts. NLP experts are heavily involved in many aspects of cNLP system evaluations as project leaders, consultants, technical support, providers of performance metrics, and most importantly, creators of the NLP system itself. But generally, NLP experts are not involved in the actual annotation task due to their lack of medical expertise. Nevertheless, the involvement of NLP experts early on in the creation of the annotation project to assist internal medical experts (eg, in the preparation of the guidelines or any NLP preprocessing) can be crucial for the final outcome of the annotation task [52,53].

To assure the quality of annotations and to provide a target for the expected accuracy of the cNLP system, we proposed the calculation of the IAA using the F1-score [51] in phase 4. Although other studies apply the Cohen kappa to measure IAA of mandatory and conditional questions [55], we preferred not to use kappa due to the lack of generalization [56]. Despite the debate about whether IAA really sets the upper limits for an cNLP system [56], we consider IAA to be important information to judge the performance of a cNLP system. In our use-case example, the cNLP system did not perform well for variable "Total IgE"; additionally, the annotators seemed to have had issues as revealed by the low IAA. This confirms the usefulness of the IAA to evaluate the difficulty of the identification of

some variables and hence to better interpret the performance of the cNLP system.

Finally, in phase 5, the cNLP system performance is evaluated. In our use-case study of asthma patients, the performance was actually higher than expected. Although the gold standard could have been even smaller, the amount of documents to be annotated was close to the minimum to still assure the representativeness of the gold standard and robustness of the cNLP system evaluation. As mentioned in the Methods section, Phase 2 subsection, we suggest being conservative with SLiCE's parameters to assure robust performance metrics. When presenting the evaluation results, mentioning the CI width for all performance measures is one of the criteria already defined by Friedman and Hripcsak [37] in their guidelines to improve cNLP system evaluations in the clinical domain. We follow this recommendation and advise every NLP expert to provide the upper and lower CI limits between which the true value lies, so that the robustness of the results can be determined.

### Limitations

Although our methodology offers a thoughtful strategy for cNLP system evaluations that has proven to be very useful, we want to point out some of its limitations. In several steps the methodology requires information that might not be easy to obtain in any project. If an NLP expert does not have access to internal medical experts to jointly work on the project, some of the required information might be difficult to obtain. In this case, external medical experts participating in the study would need to provide this information. In addition, the creation of the annotation project, including the annotation task and guidelines, requires both medical expertise and experience with annotation tasks to anticipate problems that nonexperienced annotators might encounter during the annotation task [53]. However, not all cNLP systems detect variables that can only be annotated correctly by medical experts. Depending on the level of medical knowledge required, and if no medical experts are available, nonexpert annotators can be recruited. Although considered a nonoptimal solution, nonexperts have successfully annotated text corpora in other projects [57-59].

Another problem might be the time required for the annotation task, which needs to be considered in the planning. With a primary variable of low frequency, the amount of documents to annotate can be quite high and external medical experts might not have the time to finish the annotation task in a timely manner or may even become upset with the annotation effort. Therefore, it is important to integrate medical expert knowledge to make sure that all nonlinguistic characteristics are covered to adjust the gold standard in the best possible way. To summarize, many aspects of a successful cNLP system evaluation in the clinical domain result from the essential collaboration between NLP experts and medical experts. The presented evaluation methodology reflects this important collaboration.

### Conclusion

We presented an evaluation methodology to guide NLP experts in cNLP system evaluations. By applying this methodology in a real study, we showed that this methodology is robust and efficient. To base the creation of the gold standard on

performance metrics, results in a statistically useful gold standard which is a huge improvement over studies that do not base their decision on statistical measures. NLP experts who implement such internal controls in their cNLP system evaluation provide a robust evaluation and further respect medical experts' time and economic resources.

## Acknowledgments

The authors would like to acknowledge the Savana Research Group for their contributions to the evaluation methodology.

This research received no grant from any funding agency in the public, commercial, or not-for-profit sector.

## Conflicts of Interest

None declared.

Multimedia Appendix 1

The SLiCE algorithm.

[[DOCX File , 15 KB - medinform\\_v9i7e20492\\_app1.docx](#) ]

## References

1. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. *JAMA* 2014 Jun 25;311(24):2479-2480. [doi: [10.1001/jama.2014.4228](#)] [Medline: [24854141](#)]
2. 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](#)] [Medline: [28323609](#)]
3. 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. *IJIMAI* 2018;4(7):8 [FREE Full text] [doi: [10.9781/ijimai.2017.03.001](#)]
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](#)] [Medline: [33292570](#)]
5. 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 Mar 01;1(2):142-160 [FREE Full text] [doi: [10.1136/jamia.1994.95236145](#)] [Medline: [7719796](#)]
6. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. *J Am Med Inform Assoc* 2020 Mar 01;27(3):457-470 [FREE Full text] [doi: [10.1093/jamia/ocz200](#)] [Medline: [31794016](#)]
7. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012 May 02;13(6):395-405. [doi: [10.1038/nrg3208](#)] [Medline: [22549152](#)]
8. Friedman C, Rindfleisch TC, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *J Biomed Inform* 2013 Oct;46(5):765-773 [FREE Full text] [doi: [10.1016/j.jbi.2013.06.004](#)] [Medline: [23810857](#)]
9. Pérez MN. Mapping of electronic health records in Spanish to the unified medical language system metathesaurus. University of the Basque Country. 2017. URL: <https://addi.ehu.es/handle/10810/23025> [accessed 2019-03-06]
10. Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J Am Med Inform Assoc* 2018 May 01;25(5):530-537 [FREE Full text] [doi: [10.1093/jamia/ocz160](#)] [Medline: [29361077](#)]
11. Yang X, Bian J, Gong Y, Hogan WR, Wu Y. MADEx: A System for Detecting Medications, Adverse Drug Events, and Their Relations from Clinical Notes. *Drug Saf* 2019 Jan;42(1):123-133 [FREE Full text] [doi: [10.1007/s40264-018-0761-0](#)] [Medline: [30600484](#)]
12. 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/ocz132](#)] [Medline: [29186491](#)]
13. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17(5):507-513 [FREE Full text] [doi: [10.1136/jamia.2009.001560](#)] [Medline: [20819853](#)]
14. Friedman C. Towards a comprehensive medical language processing system: methods and issues. *Proc AMIA Annu Fall Symp* 1997:595-599 [FREE Full text] [Medline: [9357695](#)]
15. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak* 2006 Jul 26;6:30 [FREE Full text] [doi: [10.1186/1472-6947-6-30](#)] [Medline: [16872495](#)]
16. Aronson A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001:17-21 [FREE Full text] [Medline: [11825149](#)]



17. Hasan S, Farri O. Clinical natural language processing with deep learning. In: *Data Science for Healthcare: Methodologies and Applications*. New York City, NY: Springer; 2019.
18. Névéol A, Zweigenbaum P. Clinical Natural Language Processing in 2014: Foundational Methods Supporting Efficient Healthcare. *Yearb Med Inform* 2015 Aug 13;10(1):194-198 [[FREE Full text](#)] [doi: [10.15265/IY-2015-035](https://doi.org/10.15265/IY-2015-035)] [Medline: [26293868](https://pubmed.ncbi.nlm.nih.gov/26293868/)]
19. 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](https://pubmed.ncbi.nlm.nih.gov/26293867/)]
20. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. *J Biomed Inform* 2018 Jan;77:34-49 [[FREE Full text](#)] [doi: [10.1016/j.jbi.2017.11.011](https://doi.org/10.1016/j.jbi.2017.11.011)] [Medline: [29162496](https://pubmed.ncbi.nlm.nih.gov/29162496/)]
21. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017 Jan;24(1):198-208 [[FREE Full text](#)] [doi: [10.1093/jamia/ocw042](https://doi.org/10.1093/jamia/ocw042)] [Medline: [27189013](https://pubmed.ncbi.nlm.nih.gov/27189013/)]
22. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep* 2016 May 17;6:26094 [[FREE Full text](#)] [doi: [10.1038/srep26094](https://doi.org/10.1038/srep26094)] [Medline: [27185194](https://pubmed.ncbi.nlm.nih.gov/27185194/)]
23. Jensen K, Soguero-Ruiz C, Oyvind Mikalsen K, Lindsetmo R, Kouskoumvekaki I, Girolami M, et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci Rep* 2017 Apr 07;7(1):46226 [[FREE Full text](#)] [doi: [10.1038/srep46226](https://doi.org/10.1038/srep46226)] [Medline: [28387314](https://pubmed.ncbi.nlm.nih.gov/28387314/)]
24. Zeiberg D, Prahlad T, Nallamothu BK, Iwashyna TJ, Wiens J, Sjoding MW. Machine learning for patient risk stratification for acute respiratory distress syndrome. *PLoS One* 2019;14(3):e0214465 [[FREE Full text](#)] [doi: [10.1371/journal.pone.0214465](https://doi.org/10.1371/journal.pone.0214465)] [Medline: [30921400](https://pubmed.ncbi.nlm.nih.gov/30921400/)]
25. 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):1-9 [[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/)]
26. Ancochea J, Izquierdo JL, Soriano JB. Evidence of Gender Differences in the Diagnosis and Management of Coronavirus Disease 2019 Patients: An Analysis of Electronic Health Records Using Natural Language Processing and Machine Learning. *J Womens Health (Larchmt)* 2021 Mar;30(3):393-404. [doi: [10.1089/jwh.2020.8721](https://doi.org/10.1089/jwh.2020.8721)] [Medline: [33416429](https://pubmed.ncbi.nlm.nih.gov/33416429/)]
27. Graziani D, Soriano J, 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):1-11 [[FREE Full text](#)] [doi: [10.3390/jcm9103259](https://doi.org/10.3390/jcm9103259)] [Medline: [33053774](https://pubmed.ncbi.nlm.nih.gov/33053774/)]
28. Almonacid Sánchez C, Melero Moreno C, Quirce Gancedo S, Sánchez-Herrero MG, Álvarez Gutiérrez FJ, Bañas Conejero D, et al. PAGE Study: Summary of a study protocol to estimate the prevalence of severe asthma in Spain using big-data methods. *J Investig Allergol Clin Immunol* 2020 Jan 23 [[FREE Full text](#)] [doi: [10.18176/jiaci.0483](https://doi.org/10.18176/jiaci.0483)] [Medline: [31983679](https://pubmed.ncbi.nlm.nih.gov/31983679/)]
29. Moon KA, Pollak J, Hirsch AG, Aucott JN, Nordberg C, Heaney CD, et al. Epidemiology of Lyme disease in Pennsylvania 2006-2014 using electronic health records. *Ticks Tick Borne Dis* 2019 Feb;10(2):241-250. [doi: [10.1016/j.ttbdis.2018.10.010](https://doi.org/10.1016/j.ttbdis.2018.10.010)] [Medline: [30420251](https://pubmed.ncbi.nlm.nih.gov/30420251/)]
30. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, CDC Prevention Epicenter Program. Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014. *JAMA* 2017 Oct 03;318(13):1241-1249 [[FREE Full text](#)] [doi: [10.1001/jama.2017.13836](https://doi.org/10.1001/jama.2017.13836)] [Medline: [28903154](https://pubmed.ncbi.nlm.nih.gov/28903154/)]
31. Izquierdo JL, Morena D, González Y, Paredero JM, Pérez B, Graziani D, et al. Clinical Management of COPD in a Real-World Setting. A Big Data Analysis. *Arch Bronconeumol (Engl Ed)* 2021 Feb;57(2):94-100. [doi: [10.1016/j.arbres.2019.12.025](https://doi.org/10.1016/j.arbres.2019.12.025)] [Medline: [32098727](https://pubmed.ncbi.nlm.nih.gov/32098727/)]
32. Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, et al. Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review. *Drug Saf* 2017 Nov;40(11):1075-1089. [doi: [10.1007/s40264-017-0558-6](https://doi.org/10.1007/s40264-017-0558-6)] [Medline: [28643174](https://pubmed.ncbi.nlm.nih.gov/28643174/)]
33. Qiao Z, Sun N, Li X, Xia E, Zhao S, Qin Y. Using Machine Learning Approaches for Emergency Room Visit Prediction Based on Electronic Health Record Data. *Stud Health Technol Inform* 2018;247:111-115. [Medline: [29677933](https://pubmed.ncbi.nlm.nih.gov/29677933/)]
34. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med* 2021 Apr;27(4):582-584. [doi: [10.1038/s41591-021-01312-x](https://doi.org/10.1038/s41591-021-01312-x)] [Medline: [33820998](https://pubmed.ncbi.nlm.nih.gov/33820998/)]
35. Biber D. Representativeness in Corpus Design. *Literary and Linguistic Computing* 1993 Oct 01;8(4):243-257. [doi: [10.1093/lilc/8.4.243](https://doi.org/10.1093/lilc/8.4.243)]
36. Paroubek P, Chaudiron S, Hirschman L. Principles of Evaluation in Natural Language Processing. Association pour le Traitement Automatique des Langues. 2007. URL: <https://www.atala.org/content/principles-evaluation-natural-language-processing> [accessed 2021-07-13]
37. Friedman C, Hripcsak G. Evaluating Natural Language Processors in the Clinical Domain. *Methods Inf Med* 2018 Feb 15;37(04/05):334-344. [doi: [10.1055/s-0038-1634566](https://doi.org/10.1055/s-0038-1634566)]

38. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. *J Biomed Inform* 2018 Dec;88:11-19 [FREE Full text] [doi: [10.1016/j.jbi.2018.10.005](https://doi.org/10.1016/j.jbi.2018.10.005)] [Medline: [30368002](https://pubmed.ncbi.nlm.nih.gov/30368002/)]
39. Becker M, Kasper S, Böckmann B, Jöckel KH, Virchow I. Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation. *Int J Med Inform* 2019 Jul;127:141-146 [FREE Full text] [doi: [10.1016/j.ijmedinf.2019.04.022](https://doi.org/10.1016/j.ijmedinf.2019.04.022)] [Medline: [31128826](https://pubmed.ncbi.nlm.nih.gov/31128826/)]
40. Filannino M, Uzuner. Advancing the State of the Art in Clinical Natural Language Processing through Shared Tasks. *Yearb Med Inform* 2018 Aug;27(1):184-192 [FREE Full text] [doi: [10.1055/s-0038-1667079](https://doi.org/10.1055/s-0038-1667079)] [Medline: [30157522](https://pubmed.ncbi.nlm.nih.gov/30157522/)]
41. 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):1-13 [FREE Full text] [doi: [10.2196/21801](https://doi.org/10.2196/21801)] [Medline: [33090964](https://pubmed.ncbi.nlm.nih.gov/33090964/)]
42. SampLe size Calculator for Evaluations (SLiCE). GitHub. 2021. URL: <https://github.com/Savanamed/slice> [accessed 2021-07-09]
43. Baeza-Yates R, Ribeiro-Neto B. *Modern Information Retrieval*. Boston, MA: Addison-Wesley Longman Publishing Co, Inc; 1999.
44. Lwanga SK, Lemeshow S. *Sample size determination in health studies : a practical manual*. World Health Organization. 1991. URL: <https://apps.who.int/iris/handle/10665/40062> [accessed 2021-07-15]
45. Clopper CJ, Pearson ES. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika* 1934 Dec;26(4):404-413. [doi: [10.2307/2331986](https://doi.org/10.2307/2331986)]
46. Pustejovsky J, Stubbs A. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. Sebastopol, CA: O'Reilly Media; 2012.
47. Klie J, Bugert M, Boullosa B, Eckart DCR, Gurevych I. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. Santa Fe, NM: Association for Computational Linguistics; 2018 Presented at: The 27th International Conference on Computational Linguistics (COLING 2018); 2018; Brussels, Belgium p. 127-132 URL: <https://www.aclweb.org/anthology/C18-2002> [doi: [10.18653/v1/D18-2022](https://doi.org/10.18653/v1/D18-2022)]
48. Weissenbacher D, Sarker A, Magge A, Daughton A, O'Connor K, Paul M, et al. Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Tasks at ACL 2019. In: *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*. 2019 Aug 02 Presented at: 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task; Aug 2; Florence, Italy p. 21-30. [doi: [10.18653/v1/W19-3203](https://doi.org/10.18653/v1/W19-3203)]
49. Trivedi G, Dadashzadeh ER, Handzel RM, Chapman WW, Visweswaran S, Hochheiser H. Interactive NLP in Clinical Care: Identifying Incidental Findings in Radiology Reports. *Appl Clin Inform* 2019 Aug;10(4):655-669 [FREE Full text] [doi: [10.1055/s-0039-1695791](https://doi.org/10.1055/s-0039-1695791)] [Medline: [31486057](https://pubmed.ncbi.nlm.nih.gov/31486057/)]
50. Fu S, Leung LY, Wang Y, Raulli A, Kallmes DF, Kinsman KA, et al. Natural Language Processing for the Identification of Silent Brain Infarcts From Neuroimaging Reports. *JMIR Med Inform* 2019 Apr 21;7(2):1-9 [FREE Full text] [doi: [10.2196/12109](https://doi.org/10.2196/12109)] [Medline: [31066686](https://pubmed.ncbi.nlm.nih.gov/31066686/)]
51. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. *J Am Med Inform Assoc* 2005;12(3):296-298 [FREE Full text] [doi: [10.1197/jamia.M1733](https://doi.org/10.1197/jamia.M1733)] [Medline: [15684123](https://pubmed.ncbi.nlm.nih.gov/15684123/)]
52. Liang JJ, Tsou C, Devarakonda MV. Ground Truth Creation for Complex Clinical NLP Tasks - an Iterative Vetting Approach and Lessons Learned. *AMIA Jt Summits Transl Sci Proc* 2017;2017:203-212 [FREE Full text] [Medline: [28815130](https://pubmed.ncbi.nlm.nih.gov/28815130/)]
53. Xia F, Yetisgen-Yildiz M. *Clinical Corpus Annotation: Challenges and Strategies*. 2020 Presented at: Third Workshop on Building and Evaluating Resources for Biomedical Text Mining; May 26; Istanbul, Turkey.
54. Banko M, Brill E. Scaling to very very large corpora for natural language disambiguation. In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics; 2001 Presented at: 39th Annual Meeting on Association for Computational Linguistics; July 6-11; Toulouse, France p. 26-33. [doi: [10.3115/1073012.1073017](https://doi.org/10.3115/1073012.1073017)]
55. 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/)]
56. Boguslav M, Cohen K. Inter-Annotator Agreement and the Upper Limit on Machine Performance: Evidence from Biomedical Natural Language Processing. *Stud Health Technol Inform* 2017;245:298-302. [Medline: [29295103](https://pubmed.ncbi.nlm.nih.gov/29295103/)]
57. Huang T, Huang C, Ding C, Hsu Y, Giles C. CODA-19: Using a Non-Expert Crowd to Annotate Research Aspects on 10,000+ Abstracts in the COVID-19 Open Research Dataset. *Proc 1st Workshop NLP COVID-19 ACL*. 2020 Sep 17 Presented at: NLP COVID-19 Workshop at ACL 2020; Sept 17; Virtual Meeting URL: <https://www.aclweb.org/anthology/2020.nlpCOVID19-acl.6>
58. Good BM, Nanis M, Wu C, Su AI. Microtask crowdsourcing for disease mention annotation in PubMed abstracts. *Pac Symp Biocomput* 2015:282-293 [FREE Full text] [Medline: [25592589](https://pubmed.ncbi.nlm.nih.gov/25592589/)]
59. Mohan S, Li D. *MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts*. arXiv. Preprint posted online Feb 25, 2019. [FREE Full text]

## Abbreviations

**cNLP:** clinical natural language processing  
**EHR:** electronic health record  
**IAA:** interannotator agreement  
**NLP:** natural language processing  
**P:** precision  
**R:** recall  
**SLiCE:** Sample Size Calculator for Evaluations

*Edited by C Lovis; submitted 20.05.20; peer-reviewed by I Mircheva, M Torii; comments to author 28.06.20; revised version received 31.07.20; accepted 17.06.21; published 23.07.21.*

*Please cite as:*

Canales L, Menke S, Marchesseau S, D'Agostino A, del Rio-Bermudez C, Taberna M, Tello J  
*Assessing the Performance of Clinical Natural Language Processing Systems: Development of an Evaluation Methodology*  
*JMIR Med Inform* 2021;9(7):e20492  
URL: <https://medinform.jmir.org/2021/7/e20492>  
doi: [10.2196/20492](https://doi.org/10.2196/20492)  
PMID: [34297002](https://pubmed.ncbi.nlm.nih.gov/34297002/)

©Lea Canales, Sebastian Menke, Stephanie Marchesseau, Ariel D'Agostino, Carlos del Rio-Bermudez, Miren Taberna, Jorge Tello. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 23.07.2021. 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

# Social Media Insights During the COVID-19 Pandemic: Infodemiology Study Using Big Data

Huyen Thi Thanh Tran<sup>1</sup>, MBA; Shih-Hao Lu<sup>1</sup>, PhD; Ha Thi Thu Tran<sup>2</sup>, LL.M; Bien Van Nguyen<sup>3</sup>, MSc

<sup>1</sup>National Taiwan University of Science and Technology, Taipei, Taiwan

<sup>2</sup>Ho Chi Minh City University of Law, Ho Chi Minh, Vietnam

<sup>3</sup>Dynimlabs, Oulu, Finland

**Corresponding Author:**

Shih-Hao Lu, PhD

National Taiwan University of Science and Technology

No 43, Keelung Rd, Sec 4, Da'an Dist

Taipei, 106335

Taiwan

Phone: 886 2 2737 6735

Fax: 886 2 2737 6360

Email: [shlu@mail.ntust.edu.tw](mailto:shlu@mail.ntust.edu.tw)

## Abstract

**Background:** The COVID-19 pandemic is still undergoing complicated developments in Vietnam and around the world. There is a lot of information about the COVID-19 pandemic, especially on the internet where people can create and share information quickly. This can lead to an infodemic, which is a challenge every government might face in the fight against pandemics.

**Objective:** This study aims to understand public attention toward the pandemic (from December 2019 to November 2020) through 7 types of sources: Facebook, Instagram, YouTube, blogs, news sites, forums, and e-commerce sites.

**Methods:** We collected and analyzed nearly 38 million pieces of text data from the aforementioned sources via SocialHeat, a social listening (infoveillance) platform developed by YouNet Group. We described not only public attention volume trends, discussion sentiments, top sources, top posts that gained the most public attention, and hot keyword frequency but also hot keywords' co-occurrence as visualized by the VOSviewer software tool.

**Results:** In this study, we reached four main conclusions. First, based on changing discussion trends regarding the COVID-19 subject, 7 periods were identified based on events that can be aggregated into two pandemic waves in Vietnam. Second, community pages on Facebook were the source of the most engagement from the public. However, the sources with the highest average interaction efficiency per article were government sources. Third, people's attitudes when discussing the pandemic have changed from negative to positive emotions. Fourth, the type of content that attracts the most interactions from people varies from time to time. Besides that, the issue-attention cycle theory occurred not only once but four times during the COVID-19 pandemic in Vietnam.

**Conclusions:** Our study shows that online resources can help the government quickly identify public attention to public health messages during times of crisis. We also determined the hot spots that most interested the public and public attention communication patterns, which can help the government get practical information to make more effective policy reactions to help prevent the spread of the pandemic.

(*JMIR Med Inform* 2021;9(7):e27116) doi:[10.2196/27116](https://doi.org/10.2196/27116)

**KEYWORDS**

COVID-19; Vietnam; public attention; social media; infodemic; issue-attention cycle; media framing; big data; health crisis management; insight; infodemiology; infoveillance; social listening



## Introduction

### Background

The COVID-19 pandemic situation remains complicated, with nearly 82 million infection cases worldwide as of January 1, 2021 [1]. Due to Vietnam's shared 1350 km land border with China, it was considered at high risk of an uncontrollable outbreak [2]. Yet Vietnam learned many lessons from the severe acute respiratory syndrome (SARS) epidemic when it failed to properly assess the infection risk from patients coming from the epidemic area for treatment at the Vietnamese French hospital, which triggered the SARS outbreak within its borders. Ultimately, Vietnam was the first country that the World Health Organization (WHO) removed from its list of those with community SARS infections [3]. During the first wave of the COVID-19 pandemic in Vietnam, the government quickly evaluated the novel coronavirus as a *strange and dangerous* virus with a high transmission risk that could easily result in an outbreak. The Vietnamese government executed preventive measures early, taking action a month before the WHO declared a Public Health Emergency of International Concern. The outcomes were impressive, as only 415 infections and no deaths were reported between January and June 2020 [4,5]. Following more than 3 consecutive months without cases of community spread infection, the second wave of the COVID-19 outbreak in Vietnam began when the 416th patient was declared infected in Danang on July 25, 2020. Vietnam recorded its first COVID-19 deaths during this period [6].

Pandemics are inherently negative situations; therefore, COVID-19-related news usually includes negative information such as infection rates, deaths, and quarantine information. Being surrounded by negative information can increase negative emotions, thereby driving perceptions of pandemic-related risk [7,8]. Unlike the SARS epidemic in 2003, connecting with potential medical users still mainly relies on email and personal communication (rather than other internet tools) to connect with each other and share information [9]. Many people have actively used the internet as their main source of information about the COVID-19 pandemic. However, the substantial amount of information in cyberspace may confound internet users who are trying to find and correctly evaluate reliable sources. This potentially harmful situation is known as an "infodemic" [10], which the WHO [11] defines as "an overabundance of information—some accurate and some not—that makes it hard for people to find trustworthy sources and reliable guidance when they need it" [12]. Therefore, understanding the dynamics of public attention during a pandemic such as COVID-19 is necessary to help governments, health ministries, or health educators design better guidelines to promote disease prevention and self-protection to return to social life and the "new normal" after resolution of the COVID-19 pandemic [13]. Previous research related to this field focused on analyzing community attention on a specific social media platform that is popular in the researchers' country or region. Abd-Alrazaq et al [14] discussed Twitter users' top concerns during the COVID-19 pandemic by analyzing collected data in English. Ahmad and Murad [15] conducted an online survey on Facebook to determine how social media has affected people during the

COVID-19 pandemic. Another study [16] examined hot search lists on Sina Microblog, China's most popular social media platform, to learn about public attention to COVID-19 in China. Several research papers have focused on public reaction to the COVID-19 pandemic in Vietnam. Trevisan et al [17] examined the country's reaction to and control of the pandemic; another study [18] described the pattern of the pandemic's early stage in Vietnam using a secondary data set provided by the country's Ministry of Health. Other researchers [19] used a survey to understand COVID-19 risk perception from socioeconomic and media attention perspectives.

In this study, we analyze big data collected from popular online sources where people obtain, create, or discuss news and information in Vietnam, including Facebook, news websites, YouTube, forums, blogs, Instagram, and e-commerce sites. Data were collected from December 2019 to November 2020 to offer a wider view from diverse sources and a longer observation period. We analyzed this data to describe a pattern of the social reaction during two different waves of the COVID-19 pandemic in Vietnam using the issue-attention cycle and media framing theories as foundations to develop our research questions.

### Issue-Attention Cycle Theory

The risk of COVID-19 infection is still high worldwide given that vaccination is not yet widely used and some countries are trying to resume normal commercial operations, including commercial flights, in an attempt to recover economically from the consequences of the pandemic. The need to seek and discuss information during a pandemic crisis like COVID-19 is obvious. However, many people try to simplify complex information or rely on their current beliefs; this may create conflict if they must force new information into previous constructs. Facing the risk of illness or death, as in the COVID-19 pandemic, can change people's attitudes toward "accepting information, handling and taking action on it" [20-22]. In 1972, Downs [23] introduced the issue-attention cycle theory that refers to the attention trend line an environmental issue could receive from the public or media, as described in five main stages. In the first stage, only experts or a small number of people interested in the issue are aware of it. In the second stage, the issue captures more attention as awareness of it increases; at this stage, people are optimistic that the problems will be solved one way or another. The third stage is marked by chaos, which peaks when people realize the issues might be far different from their expectations, out of their control, and present with high financial or social benefit costs. A steady drop in public attention to the issue characterizes the fourth stage, which is known as the postproblem phrase. The final stage is marked by replacement of the concerning issues in public attention [23].

However, some researchers argued that the issue-attention cycle can differ depending upon culture [24] and in cases of epidemic hazards [25]. Moreover, the issue-attention cycle is not always fully integrated or fully explanatory in some health-related research, as evidenced by the "Charlie Sheen effect" phenomenon, introduced by Ayers et al [26] in 2016 when they used results from Google's search engine data set to show the correlation between actor Charlie Sheen's disclosure of his

HIV-positive status with the level of public attention to HIV and its prevention.

Our study investigates public attention during the COVID-19 pandemic by examining internet discussion volume to find patterns and determine similarities or differences to the issue-attention cycle theory. The amount of public discussion on social media has changed over time based on the public's response to each real event that occurred during the pandemic. Capturing the amount of public discussion not only helps to point out or compare patterns in issue-attention cycle theory but also shows how the public's attention to specific events is different. From there, it is possible to help the government and stakeholders evaluate the severity of each event to the public and from there learn lessons for possible pandemic prevention in the future. Hence, the research questions related to this theory are:

- RQ1: What is the level (volume) of public attention to COVID-19 in this study?
- RQ2: What does the pattern of public attention to the pandemic look like?

Throughout the COVID-19 pandemic, a concurrent infodemic has bombarded the public, hindering the reception of reliable information sources so citizens can follow recommendations and protect themselves. Therefore, in addition to pointing out patterns of pandemic-related discussions, it is necessary to dig deep into sources that get the most public attention, which can help government and disease control centers stop inaccurate news that has reached a large number of people in a timely manner. These patterns can also help identify popular public channels to help legitimate agencies broadcast disease prevention messages more efficiently. Additionally, analyzing the public sentiment about the pandemic can help governments and the Centers for Disease Control and Prevention deliver more accurate prevention messages to appease public anxiety and insecurity. Therefore, we developed the third and fourth research questions:

- RQ3: Which types of sources gained the most public attention and engagement during the pandemic?
- RQ4: How did people react to the pandemic, as measured by expression of their emotions on social media?

### Media Framing Theory

The mechanism by which individuals create a clear conceptualization or reorient their thoughts about an issue is referred to as framing theory. The concept is based on the acceptance of an issue that can be presented from a number of viewpoints and is perceived as having implications for different principles or factors [27]. Frames matter, especially in communications meant to influence an audience's attitudes and behaviors. Frame use is learned and may be adopted from person to person. Previous studies have shown that politicians have been inspired by the communication styles of other politicians, the media, or even citizens [28-30]. It is understandable that even in conversation and discussion with others, individuals typically adopt the frames that they have learned [30-32].

The explosive growth of information technology and social networking in the digital age has resulted in changes to the

concept of "news," which was once considered the product of a journalist [33]. The concept has broadened now that anyone can create news by uploading it to the internet in the form of pictures, text, video, etc [34]. Sometimes this news is only 140 characters long [35], and its credibility depends on the number of interactions garnered from readers, including likes, shares, and comments [36]. The nature of this news formulation and discourse is dynamic, so it is important to examine how frames used to report on epidemic hazards may change and develop over time [31,37,38]. Understanding the critical role framing plays in communication, scholars have monitored frames over the past decade to detect patterns in problem descriptions, analyze media attention, and investigate differences across forms of media [39]. Thus, in our study, we seek answers to the following questions:

- RQ5: What frames are used and how frequently are they used in communications that occur during the pandemic? What main topics gained the most discussion and attention during the COVID-19 pandemic?
- RQ6: Were different types of frames used during the first and second waves of the COVID-19 pandemic in Vietnam?

### Methods

All information related to COVID-19 in Vietnam was obtained from the Ministry of Health of Vietnam's official COVID-19 disease page [6] and the website [thuvienphapluat.vn](http://thuvienphapluat.vn) [40], an electronic library of legal documents issued by the Vietnamese government. We used this information to create a foundation for collecting data from social platforms and as a basis for comparison with the results obtained after data analysis.

### Data Collection and Processing

This study aims to understand the public reaction to the COVID-19 pandemic via discussions among Vietnamese people on social media. We used SocialHeat, a fee-based social listening tool developed and sponsored by YouNet Group, to crawl data while following the terms of use from 7 types of sources: Facebook, Instagram, news, blogs, forums, e-commerce sites, and YouTube. SocialHeat collected public data on social networks in real time using COVID-19-related keywords (coronavirus, nCoV, SARS-CoV-2, COVID-19, Covid). Counted topics were written in Vietnamese only and pulled from the Facebook application programming interface (API), Instagram API, YouTube API, and Google API (for news, blogs, e-commerce, and forum websites). All data from spam and noise mentions were deleted by applying deep learning and natural language processing in the SocialHeat system (the data set still might contain seeding posts or brand commercial posts, but the numbers of those posts are negligible).

The data set was collected from December 1, 2019, to November 13, 2020, from 63 million Facebook IDs (pages, individual profiles, and groups), 1.2 million YouTube accounts, 9000 news websites, and 300 forums in Vietnam. On account of the amount of data and technology limitations, we divided the timeline into 7 periods to crawl data, then reconnected the data in a complete and continuous timeline. To divide the timeline, we relied on highlighted events that took place during the period observed

(December 1, 2019, to November 13, 2020). Specifically, we used data tracking new daily infections in Vietnam, which was updated by the Ministry of Health of Vietnam [6] and included

the four main stages of the COVID-19 outbreak in Vietnam as described by La et al [41] (as of April 4, 2020) to inform additional development into 7 main phases (Table 1).

**Table 1.** The 7 periods of the COVID-19 pandemic in Vietnam.

Period	Date	Total days, n	Events
1	Before January 23, 2020	54	No confirmed cases in Vietnam
2	January 23 to February 26, 2020	35	First confirmed case in Vietnam; 16th infected case discharged from hospital
3	February 27 to March 5, 2020	8	No new cases in Vietnam
4	March 6 to March 31, 2020	26	17th infected case confirmed and more reported afterward
5	April 1 to April 15, 2020	15	Implementation of social isolation
6	April 16 to July 24, 2020	100	No new cases in the community
7	July 25 to November 13, 2020	112	A new case in the community and the first deaths

## Mention Trend Line

The volume of total mentions (a mention can be an original post, a comment, or a share) about COVID-19–related topics on digital channels, including Facebook, Instagram, news sites, forums, blogs, etc, was tallied and expressed by day to show how Vietnamese citizens reacted to COVID-19 pandemic–related events timeline-by-timeline. This study also integrates the real flow of facts and disease coping measures adopted by the government to analyze the relationship between government policies and peoples' reactions during the pandemic.

## The 500 Most Engaging Sources

To explore which sources attracted the most attention and engagement, we calculated the total interactions on COVID-19–related topics across all ID sources (Facebook, YouTube, and Instagram) and unique links on news, blog, forum, and e-commerce sites, then ranked them in order from highest to lowest. The total interaction with an engaging source equal to the total COVID-19–related posts was posted by observed source, plus total likes, shares, and comments that those posts gained.

Facebook is the most popular social platform in Vietnam [42]. It is not only popular with individuals but also used for official brand fan pages, key opinion leaders (KOLs), TV channels, news, and government departments that use Facebook as a connecting bridge with customers, readers, citizens, etc. Therefore, we categorized Facebook accounts into 8 clusters: community pages, news, TV channels, KOLs, forums, groups, government, and unknown (minor accounts that could not be categorized into any source). Due to the limitations of hand categorization, we chose only the top 500 sources by mentions each period and categorized them for analysis.

## Top 50 Posts by Mentions

We analyzed top posts created during the COVID-19 pandemic to understand which topics attracted the most citizen attention and their associated reactions via discussion sentiment analysis. Top posts were COVID-19–related posts that gained the most mentions (shares, comments) on Facebook, Instagram, YouTube, news sites, blogs, e-commerce sites, and forums.

Previous studies about the information shared on social media by users during crisis events had different ways of classifying content based on real events. For example, Vieweg [43], who studies communications and behavior during mass emergencies, categorized the types of information that users create into three main groups: social environments (eg, caution, advice, medical attention, and offering help), built environment (eg, infrastructure damage), and physical environment (eg, weather forecast and general information about hazards). Based on research of Vieweg [43], Imran et al [44] has inherited and continues to categorize the content collected from researching on social media messages related to disasters into types of content such as caution and advice; casualties and damage; donations of money, goods, or services; people missing, found, or seen; and information source. Meanwhile, Mirbabaie et al [45] studied what happened on social media during the Hurricane Harvey incident to find lessons in dealing with the COVID-19 pandemic and classified the information into seven categories based on the information gathered during the data analysis process: official statement, news and crisis information, personal opinion, personal experience, forwarding message, solicitousness, and humor.

The research on the nature of information spread about the COVID-19 pandemic on Weibo by Li et al [46] classified content into 7 groups based on the previous work of Rudra et al [47] and Vieweg [43], including notifications or measures taken; donating money, goods, or services; emotional support; help seeking; doubt casting and criticizing; counter rumors; and policy reaction. In the process of applying the aforementioned classifications, we identified 5 types of content that appeared frequently but are not suitable for distribution into the 7 existing content groups, including caution and advice; international situation updates; medical issues, treatment, and vaccine; effects of the pandemic on the economy; and entertainment. Thus, in this study, the 50 posts with the most public engagement (interaction) were categorized and sorted into 12 groups of content.

## Sentiment Trend Line, COVID-19–Related Topics' Keyword Frequency, and Social Networks

For all COVID-19 data downloaded from Facebook, Instagram, news sites, forums, blogs, etc, the SocialHeat tool excluded



noise, spam, and advertising posts before using natural language software developed by the YouNet Company for sentiment classification and to extract the top 50 keywords' frequency for the 7 observed periods.

The most frequently mentioned keywords for each period were analyzed and visualized using VOSviewer (Nees Jan van Eck and Ludo Waltman) [48]. A social network and clusters for each period were created using the keywords matrix, in which every two keywords are linked by co-occurrence frequency. In other words, the frequency of occurrence of two keywords in the same article will be shown through the link between two dots. The larger the dot, the more often the keyword appears. The thicker and closer the link between two dots (two keywords), the more frequency the two keywords will appear together.

## Results

### Total Discussions About COVID-19 on Social Media in Vietnam During the First Two Waves of the Pandemic

There was a total of 37,917,631 collectable mentions and 22,652,638 posts about COVID-19 from December 1, 2019, to November 13, 2020. *Collectable mentions* refers to mentions set in public mode on the online channels; therefore, only public data was collected due to privacy settings. The data set was summarized daily and put in chronological order (see [Multimedia Appendix 1](#)). Facebook was the channel that gained the most mentions (27,191,922 mentions, accounting for 96.4% of total mentions), while other channels shared the rest (forums: 232,131 mentions; news sites: 757,582 mentions; blogs: 1058 mentions; reviews: 224 mentions; e-commerce sites: 2231 mentions; YouTube: 20,599 mentions; Instagram: 1857 mentions).

There was a positive correlation between total collectable mentions on social media and daily new COVID-19 infection cases ( $\beta_0=74,451.4$ ;  $\beta_1=9366.9$ ;  $P<.001$ ). In other words, the more new infection cases counted daily, the more posts and mentions of COVID-19 pandemic topics created on social platforms like Facebook, YouTube, Instagram, etc, and on websites including news sites, forums, blogs, etc.

[Multimedia Appendix 1](#) indicates the Vietnamese public's attention and reaction toward the two first waves of the COVID-19 pandemic. Data were divided into 7 periods (the same as those in [Table 1](#)) based on the highlighted events happening in Vietnam. During period 1 (December 1, 2019, to January 22, 2020) and especially before January 12, 2020, Vietnamese people paid little attention to information about the COVID-19 epidemic, although China recorded the first cases in Wuhan [49]. During period 2 (January 23, 2020, to February

26, 2020), public attention increased significantly when Vietnam confirmed that the first COVID-19 case in the country was from a Chinese traveler [50]. Period 3 (February 27, 2020, to March 5, 2020) saw very low public attention when no new infections were confirmed by the government. In period 4 (March 6, 2020, to March 31, 2020), total posts peaked with more than 1.2 million mentions about COVID-19 after the 17th case was confirmed. The highest total collectable mentions (1,255,175 mentions) were made on March 31, one day before the Vietnamese government's implementation of a *social isolation* mandate throughout the country. Period 5 (April 1, 2020, to April 15, 2020) had a deep drop but a stable number of total and collectable mentions about COVID-19-related topics compared to period 4. Period 6 (April 16, 2020, to July 23, 2020) had a significant steady decrease in public attention toward the pandemic when the government removed the social isolation order and simultaneously did not report new community infection cases. However, there was a small fluctuation indicating increased discussions starting on June 22, 2020, and peaking July 1, 2020, with 320,089 discussions, due to information about a COVID-19 vaccine developed in Vietnam that was expected to be clinically tested in humans in October or November 2020. Additionally, a suspected COVID-19 infection had been discovered in Danang; public attention gradually decreased through period 7, when a community infection was confirmed in Danang.

### Sources With the Most Interaction During the First Wave of the COVID-19 Pandemic in Vietnam

#### Total Interactions on the Top 500 Most Engaging Sources

As shown in [Table 2](#), the community page sources remained the most popular throughout the whole period. The remaining interactions were split among other types of sources, including news sites, KOLs, government sites, etc.

During period 1 (December 1, 2019, to January 22, 2020), news sources gained the most public reaction, and TV channel sources followed right after. After period 1, community page sources steadily earned the most engagement. This was especially true during period 2 (January 23, 2020, to February 26, 2020), period 4 (March 6, 2020, to March 31, 2020), period 5 (April 1, 2020, to April 15, 2020), and period 7 (July 25, 2020, to November 13, 2020), when around 50% of Vietnamese citizens' interactions about the pandemic came from community page sources. Meanwhile, TV channel sources (periods 1, 2, 4, and 7) and KOL sources (periods 3, 5, and 6) alternated second place status in terms of engagement on COVID-19-related topics.

In contrast, forum (periods 2, 6, and 7) and government (periods 1 and 3) sources gained less total interaction.



**Table 2.** Total reactions on the top 500 most engaging sources.

Source	Period 1 (n=265,679), n (%)	Period 2 (n=3,217,036), n (%)	Period 3 (n=137,642), n (%)	Period 4 (n=39,200,553), n (%)	Period 5 (n=10,086,791), n (%)	Period 6 (n=567,114), n (%)	Period 7 (n=96,832,404), n (%)
Community page	63,251 (23.81)	1,600,342 (49.75)	50,549 (36.72)	21,392,949 (54.57)	4,734,704 (46.94)	170,131 (30.00)	42,131,669 (43.51)
Forum	5819 (2.19)	16,918 (0.53)	498 (0.36)	540,775 (1.38)	128,025 (1.27)	0 (0.00)	528 (0.00)
Government	346 (0.13)	85,434 (2.66)	0 (0.00)	1,350,699 (3.45)	431,201 (4.27)	98,005 (17.28)	8,340,435 (8.61)
Group	11,934 (4.49)	266,640 (8.29)	15,806 (11.48)	3,367,923 (8.59)	727,409 (7.21)	38,046 (6.71)	7,794,693 (8.05)
Key opinion leaders	39,234 (14.77)	309,160 (9.61)	45,529 (33.08)	3,889,108 (9.92)	1,903,164 (18.87)	165,804 (29.24)	11,732,842 (12.12)
News	73,982 (27.85)	346,683 (10.78)	15,043 (10.93)	3,300,849 (8.42)	724,832 (7.19)	63,080 (11.12)	10,681,121 (11.03)
TV channel	65,330 (24.59)	359,969 (11.19)	3863 (2.81)	4,107,847 (10.48)	1,384,393 (13.72)	26,499 (4.67)	16,057,180 (16.58)
Unknown	5783 (2.18)	231,890 (7.21)	6354 (4.62)	1,250,403 (3.19)	53,063 (0.53)	5549 (0.98)	93,936 (0.10)

### *The Average Interaction on the Top 500 Most Engaging Sources*

Total interactions on the most engaging sources were calculated by summarizing the number of each source's COVID-19-related posts, likes, shares, and comments. We analyzed the average interaction on the top 500 most engaging sources to understand the efficiency of each COVID-19-related post created by each source.

**Table 3.** The average interaction on the top 500 most engaging sources.

Sources	Period 1 (n=10,167), n (%)	Period 2 (n=102,421), n (%)	Period 3 (n=3184), n (%)	Period 4 (n=1,111,044), n (%)	Period 5 (n=645,151), n (%)	Period 6 (n=12,802), n (%)	Period 7 (n=3,745,249), n (%)
Community page	427 (4.20)	5443 (5.31)	468 (14.70)	74,540 (6.71)	16,327 (2.53)	915 (7.15)	861,254 (23.00)
Forum	1940 (19.08)	5639 (5.51)	249 (7.82)	90,129 (8.11)	32,006 (4.96)	0 (0.00)	528 (0.01)
Government	346 (3.40)	42,717 (41.71)	0 (0.00)	450,233 (40.52)	431,201 (66.84)	4900 (38.28)	321,134 (8.57)
Group	385 (3.79)	5442 (5.31)	368 (11.56)	57,083 (5.14)	13,989 (2.17)	865 (6.76)	87,847 (2.35)
Key opinion leaders	162 (1.59)	3964 (3.87)	149 (4.68)	54,015 (4.86)	17,954 (2.78)	825 (6.44)	2,330,684 (62.23)
News	2000 (19.67)	11,556 (11.28)	1003 (31.50)	94,310 (8.49)	25,887 (4.01)	2426 (18.95)	67,245 (1.80)
TV channel	4666 (45.89)	17,998 (17.57)	644 (20.23)	228,214 (20.54)	81,435 (12.62)	2409 (18.82)	75,552 (2.02)
Unknown	241 (2.37)	9662 (9.43)	303 (9.52)	62,520 (5.63)	26,352 (4.08)	462 (3.61)	1005 (0.03)

### **Top Posts About COVID-19 Topics With the Most Comments or Shares**

The type of COVID-19-related content that received the most attention varied from time to time. Starting from phase 2

As can be seen from [Table 3](#), government sources were leading in periods 2, 4, 5, and 6 with nearly 32% to 67% of the average interactions for the top 500 most engaging sources. News and TV channel sources alternated in the second position in periods 1 and 6 and periods 2, 3, 4, and 5 (TV channels). Period 7 was unique in that KOLs received the highest average engagement (2,330,684/3,745,249, 62.23%), followed by news from community sites (861,254/3,745,249, 23%).

onward, the diversity of content types increased to include caution and advice, policy reaction, and international situation updates ([Table 4](#)).

**Table 4.** Top posts with the most comments or shares.

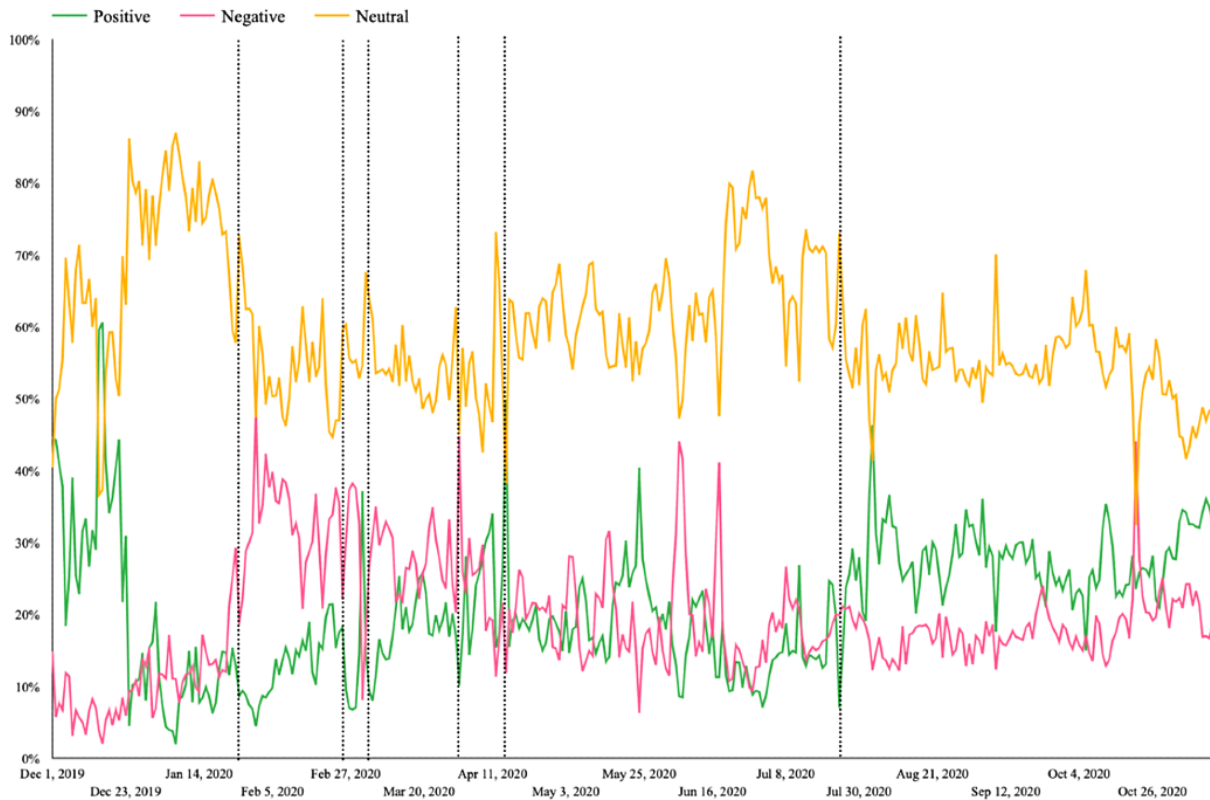
Categories	Period 1 (n=49,480), n (%)	Period 2 (n=337,865), n (%)	Period 3 (n=24,173), n (%)	Period 4 (n=1,375,260), n (%)	Period 5 (n=312,851), n (%)	Period 6 (n=986,403), n (%)	Period 7 (n=9,161,011), n (%)
Caution and advice	27,607 (55.79)	28,869 (8.54)	0 (0.00)	191,017 (13.89)	28,067 (8.97)	129,076 (13.09)	347,096 (3.79)
Notifications or measures have been taken	7116 (14.38)	13,222 (3.91)	2229 (9.22)	162,502 (11.82)	0 (0.00)	23,978 (2.43)	2,120,104 (23.14)
Donation money, goods, or services	0 (0.00)	0 (0.00)	0 (0.00)	12,167 (0.88)	4982 (1.59)	89,919 (9.12)	1,477,020 (16.12)
Emotional support	0 (0.00)	0 (0.00)	4905 (20.29)	328,056 (23.85)	153,652 (49.11)	195,632 (19.83)	970,817 (10.60)
Help seeking	0 (0.00)	0 (0.00)	0 (0.00)	9045 (0.66)	0 (0.00)	0 (0.00)	133,675 (1.46)
Doubt casting and criticizing	0 (0.00)	112,297 (33.24)	3284 (13.59)	186,581 (13.57)	2537 (0.81)	0 (0.00)	666,394 (7.27)
Counter rumors	0 (0.00)	55,397 (16.40)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Policy reaction	1226 (2.48)	116,869 (34.59)	11,642 (48.16)	289,200 (21.03)	112,505 (35.96)	0 (0.00)	1,132,238 (12.36)
International situation updating	13,531 (27.35)	11,211 (3.32)	1678 (6.94)	196,692 (14.30)	11,108 (3.55)	107,243 (10.87)	100,369 (1.10)
Medical issues: treatment, vaccine	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	432,389 (43.83)	0 (0.00)
Effects of the pandemic on the economy	0 (0.00)	0 (0.00)	435 (1.80)	0 (0.00)	0 (0.00)	8166 (0.83)	0 (0.00)
Entertainment	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	2,213,298 (24.16)

In period 1, when the first COVID-19 cases were found in Wuhan and had not yet spread to Vietnam, it is understandable that content concerning *caution and advice* received the most attention, accounting for 55.8% (27,607/49,480). Content about *international updates* was next, accounting for 27.3% (13,531/49,480). In period 2, when Vietnam confirmed that two Chinese tourists were infected with COVID-19 and that these were the first two cases of COVID-19 to appear in Vietnam, articles regarding *policy reaction* were of most interest, accounting for 34.6% (116,869/337,865); content about *doubt casting and criticizing* received almost equal attention, accounting for 33.2% (112,297/337,865). In the third stage when Vietnam had no community cases, people remained interested in topics classified as *policy reaction* (11,642/24,173, 48.2%) and began to pay attention to content on *emotional support* (4905/24,173, 20.3%). During phase 4, as community cases peaked and new infections were recorded, people were most interested in the topic of *emotional support* (328,056/1,375,260, 23.9%) and *policy reaction* (289,200/1,375,260, 21%). This was also the period when interest was shared between the greatest variety of content types with fairly similar distribution. In period 5, when the Vietnamese government applied a social isolation mandate, people were most concerned with *emotional support* (153,652/312,851, 49.17%) and *policy reaction* (112,505/312,851, 36.7%). During period 6, when Vietnam enjoyed 100 days of peace without news of community spread, articles on *medical issues* received

the most attention (432,389/986,403, 43.87%) followed by *emotional support* (195,632/986,403, 19.87%). When community cases reappeared and though there were more COVID-19 deaths in Vietnam, peoples' response was quite optimistic, with attention almost equally divided between the topics of *entertainment* (2,213,298/9,161,011, 24.3%) and *notifications or measures being taken* (2,120,104/9,161,011, 23.2%).

### Sentiment

After all text data related to COVID-19 was crawled, it was processed by a sentiment analysis tool developed by SocialHeat. All discussions were evaluated and sorted into one of three emotional categories, positive, negative, and neutral, based on natural language. In general, we found that people's emotions when discussing COVID-19-related topics fluctuate and are unstable. Emotional neutrality almost always took first place. This is understandable because sources from government organizations and especially television and newspapers are expected to report "independent, reliable, accurate, and comprehensive information" [33]. However, in various time periods, negative and positive emotions alternated in second place. Positive emotions were expressed more often than negative emotions during the first and last periods. Negative emotions were expressed more than positive ones; most appear throughout stages of Vietnam's first COVID-19 wave (periods 2, 3, 4, 5, and 6; see Figure 1).

**Figure 1.** Sentiment trend line from December 1, 2019, to November 13, 2020.

During period 1, when Vietnam had not yet recorded any cases and the pandemic situation had just begun in China, people learned about COVID-19 through information from the Ministry of Health and the press, so their mentality was still stable and optimistic. In period 2, when the first cases were discovered in Vietnam, people become more confused and worried. In period 3, negative emotions exploded when patient 17 was confirmed and there was a risk of community disease spread. Anger, blame, and anxiety were evident through the negative emotions expressed in the text lines discussed on social networks at that time. In period 7 when Vietnam experienced its second wave of COVID-19 with the re-emergence of community infection and the first recorded COVID-19 deaths, the optimism shown through positive emotions overwhelms the negative emotions expressed during this period. People have gradually adapted to the pandemic after experiencing the first wave and have confidence in the government's ability to control the pandemic; positive signals that a Vietnamese COVID-19 vaccine would soon enter the human testing phase may have also contributed to the positive outlook [51].

## Top Keywords' Frequency and Social Network Analysis of Discussions on the Internet During the COVID-19 Pandemic in Vietnam

### Top Keywords

The top 50 keywords were compiled and ranked in order from all discussions on the COVID-19 pandemic topic gathered during the study period. However, of the top 50, many keywords are synonyms, so we have grouped them into 36 keywords. The content of the top keywords was related to 4 main groups, including *COVID-19 pandemic and epidemic outbreaks* expressed through keywords such as epidemic, COVID-19, Vietnam, case, Danang, Hanoi, and Bach Mai hospital; *policy reactions* as shown through words including quarantine, against, prevention, mask, province, and government; *medical issues* expressed through patient, hospital, test, contact, infected, virus, and treatment; and *disease situation in the world* through words like situation, the United States, and money (see Table 5).

**Table 5.** Top 36 keywords for COVID-19–related topics during the COVID-19 pandemic in Vietnam from December 1, 2019, to November 13, 2020.

Rank	Word	Frequency, n
1	epidemic	16,373,688
2	COVID-19	10,720,319
3	patient	9,838,764
4	quarantine	9,783,246
5	medical	9,349,795
6	go	8,428,703
7	hospital	8,257,058
8	Vietnam	7,789,027
9	case	7,643,082
10	disease	6,397,632
11	Danang	5,621,518
12	against	5,208,937
13	city	5,066,441
14	infected	4,734,692
15	virus	4,099,245
16	situation	3,950,498
17	information	3,854,982
18	province	3,850,230
19	prevention	3,692,883
20	citizen	3,357,150
21	mask	3,316,868
22	way	3,310,041
23	test	2,784,690
24	contact	2,764,565
25	government	2,705,459
26	Hanoi	2,422,218
27	family	2,366,785
28	money	2,177,978
29	coronavirus	2,039,959
30	The US	1,903,865
31	vehicle	1,842,155
32	result	1,821,271
33	treatment	1,790,499
34	Bach Mai hospital	1,663,212
35	together	1,654,416
36	get sick	1,475,917

### *Social Network Co-occurrence of the 7 Periods*

To better understand the context behind the most mentioned keywords and to highlight the top concerns about the COVID-19 pandemic expressed in internet discussions in each period in Vietnam, we extracted the top 50 keywords for each stage and visualized the associations between the keywords using

VOSviewer software. The larger the dots, the more weight (frequency) that keyword possessed. The thicker and closer the link between two keywords, the more frequently both keywords appear.

The relationship between the top keywords in period 1 when no infections were found in Vietnam is shown in [Figure 2](#). The





During period 5 (Figure 6), the social isolation stage, the most prominent keywords were “epidemic,” “COVID-19,” “enterprise,” “unanimously,” etc. The green cluster represents the policy reaction aspect, in particular medical issues and economic solutions, such as supporting businesses and people working in production and consumption. Representative

keywords included “prime minister,” “economic,” “solution,” “bank,” “electricity,” “rice,” etc. The pink cluster represents the economic concerns, as expressed by keywords including “enterprise,” “salary,” “business,” “labor,” “working,” “jobs,” “contract,” “society,” “poverty,” etc.

Figure 3. Co-occurrences of the top keywords in period 2. WHO: World Health Organization.

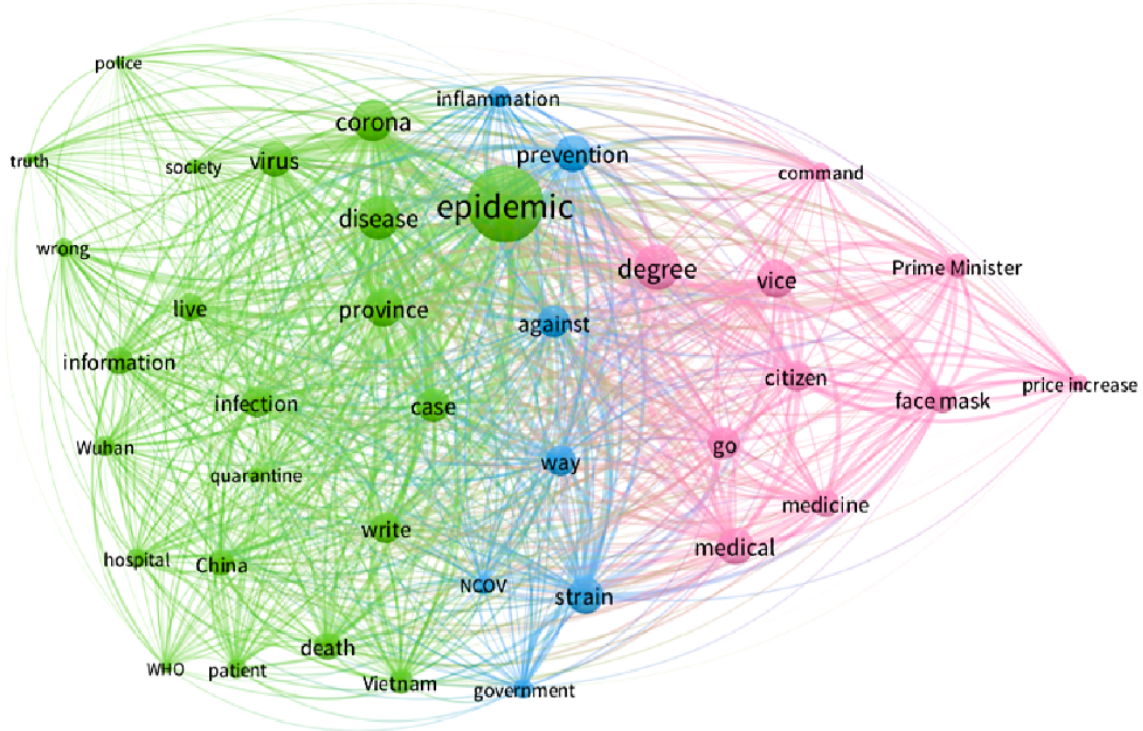
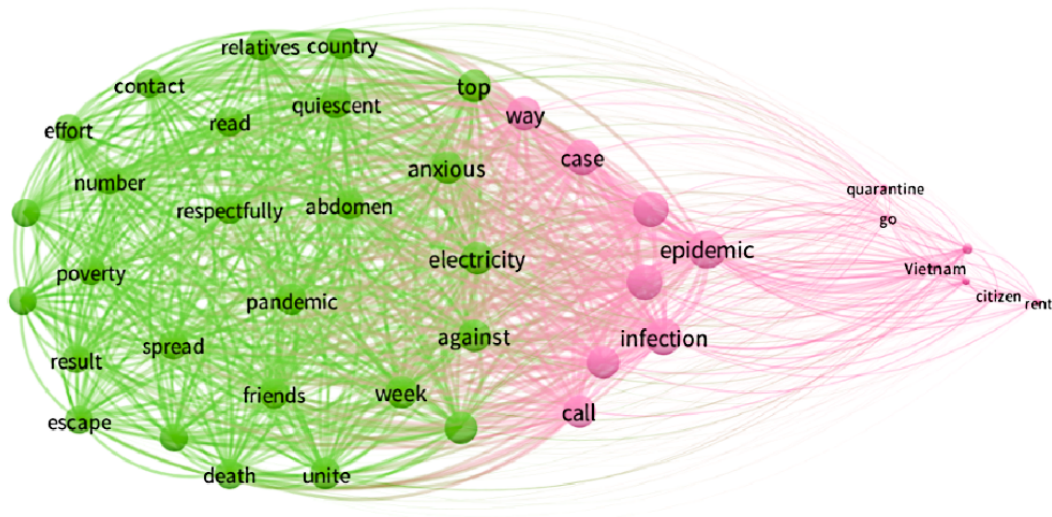


Figure 4. Co-occurrences of the top keywords in period 3.













point is that the issue-attention cycles that occurred during the COVID-19 pandemic did not represent the main issue (the pandemic) but rather showed the subissues (real events) related to the main issue. Moreover, the last stage of each cycle was a transition between cycles. This means the last stage of this cycle may be the first stage of the cycle that occurs after it.

When the first COVID-19 cases were discovered in Wuhan, China, people were not too concerned about this strange disease, despite the attention given to it by the Vietnamese government, especially the Ministry of Health and related agencies. However, when Vietnam saw its first cases of infection, people began to pay more attention. Anxiety peaked when people became aware that this is a dangerous, contagious, potentially fatal disease and that there was no vaccine yet. The situation was eased when the government's pandemic prevention responses were effective.

During the second COVID-19 pandemic wave in Vietnam, people remained interested in the pandemic but discussed it less on social networks. Public attention peaked with the first COVID-19-related deaths in Vietnam. Public attention then quickly dropped and diverted to other issues. This shows that although the second COVID-19 pandemic wave in Vietnam appeared to have a more negative factor (the first recorded deaths), the public's attitude was not as intense as it had been during the first COVID-19 pandemic wave. This may be explained by people's acceptance of the fact that death is a foreseeable outcome for patients infected with COVID-19 and at the same time an expression of *not feeling surprised* after 6 months living through the pandemic.

### ***The Most Engaging Sources During the COVID-19 Pandemic***

Per our data analysis, community pages on Facebook received the most total interaction from the public, likely because these aggregate information for the community with diverse content types. Each of these news sites usually post multiple articles per day on the same COVID-19 topic. However, in terms of average efficiency per article, government-controlled news sites outperformed other news sources. Drawing from this conclusion, we recommend that the government increase the number of articles posted to sources under its control to achieve the greatest dissemination of information to the community. In addition, the government can also coordinate with sources such as community pages and KOLs' pages to quickly, accurately, and easily distribute disease information to the public.

Through our analysis, *the frames of communication* (top posts that gained the most interaction) can be used to explain public sentiment about the COVID-19 pandemic. We categorized COVID-19 topics garnering top public interest into 12 categories, based on the adoption of 7 types of COVID-19 information described by Li et al [46] and simultaneously developed 6 additional content type categories based on our data analysis processing. These categories included caution and advice; notifications or measures taken; donation of money, goods, or services; emotional support; help seeking; doubt casting and criticizing; counter rumors; policy reaction; international situation updates; medical issues, treatment, or vaccines; effects of the pandemic on the economy; and entertainment. When the pandemic first started in China,

information about *cautions and advice* and *international situation updates* got the most attention. Negative emotions were just beginning to be expressed and did not prevail. However, negative emotions gradually increased when cases first appeared in Vietnam, and articles about *policy reaction* gained the most attention, followed by *doubt casting and criticizing* articles. During the peak of the first pandemic wave, negative emotions peaked as well, but the public still paid the most attention to *policy reactions* and the *emotional support* articles. The desire for negative emotions to subside was shown by the public giving the most attention to *emotional support* articles in addition to articles about *policy reaction* and *medical issues*. Although negative emotions persisted in the second wave of the pandemic in Vietnam, articles related to *entertainment* gained the most attention. This shows the public's optimism during the crisis, as they have experienced the first wave of epidemics in the past and have hopes of a *new normal life* to come with the expectation of mass vaccine distribution next year.

### **Limitations**

This study has some limitations. First, despite using big data to analyze the phenomenon of public reaction toward the COVID-19 pandemic, some noise or spam remains in the data set; the SocialHeat tool could not completely filter these out due to technology limitations and the complexities of natural language. Though natural language has been applied and innovated daily in SocialHeat's tool, some texts or paragraphs containing incorrect grammar, teen code, dialects, etc, could not be processed or categorized. It is also important to note that although the data set was pulled from diverse sources like Facebook, YouTube, news sites, etc, the observed format was text only. This means that other formats such as video with text or audio captions or images with textboxes were not analyzed by the SocialHeat tool. Hence, this led to a shortage in the final data set results such as sentiments categorized, extracted top sources, and extracted top posts.

Additionally, due to privacy policies, the data set can only collect data that is installed in public mode, especially for data obtained from social networking platforms like YouTube, Instagram, and Facebook. Moreover, because the data collection time is quite long (11 months), the amount of data poured into the system is large and requires a substantial amount of time for the system to process noise and spam, and give statistical results. This led to a situation in which we wanted to analyze the *top posts by mentions* in depth, but often encountered links that no longer worked because the owner of the post had changed the view mod from *public* to *friends* or *private*, or even deleted the post. This caused difficulties and data deficiencies in our analysis.

Finally, we have almost 38 million data in total, which the system could not process all at once due to technical limitations. Therefore, we could not extract top posts by mentions, top sources by mentions, or overall sentiment of all sources.

### **Future Work**

The topics discussed on the COVID-19 issue are varied. The classification of content groups as we propose in the study is

still limited when it is impossible to analyze the public's emotional index for each type of topic. Understanding the feelings of the community on specific topics related to the COVID-19 topic can help the government and stakeholders come up with precise and meticulous guidance on disease reactions. Therefore, we suggest that researchers focus on analyzing the public's sentiment index for each type of topic that the public is discussing to come up with appropriate ideas and options to support the medical information management in pandemic times.

### Conclusions

Through our research, we found that using different types of information sources can be effective in different pandemic phases. The same goes for pandemic-related content types. We also highlighted hot spots of public concern regarding the COVID-19 pandemic. These results can help governments or health educators communicate pandemic prevention guidelines more effectively to the public. This is significant not only for prevention during the current COVID-19 pandemic but also could serve as a useful reference for the health crisis management field for potential diseases in the future.

### Implications

Applying big data in infodemiology studies opens opportunities for getting better insights into a public reaction toward pandemics and related events. The government should take

advantage of social platforms to effectively communicate health information, quickly address fake news, and give real-time response to the hot issues that the public needs to know during the pandemic. To achieve those goals, we suggest three key points to help government and stakeholders have better communication with the public during crisis events like the COVID-19 pandemic:

- Applying artificial intelligence tools in analyzing big data from social media platforms to collect public insights, determine appropriate cooperation channels in spreading news and guidelines, and effectively communicate about health information and instructions
- Promoting an official account of the Ministry of Health on different social media platforms to form the public's habit of updating news from official sources, avoiding *infodemics* during the pandemic
- Collaborating with popular community and KOLs' fan pages to spread information faster and wider to various reader segments

Big data is also meaningful for infodemiology studies. Applying big data allows researchers to have a wider view and easily compare the results across countries, regions, races, or cultures and lead to more research ideas such as descriptive studies or predicting public sentiments or public reactions about the pandemic.

---

### Acknowledgments

This project received support from YouNet Group in big data crawling and analysis.

---

### Conflicts of Interest

None declared.

---

### Multimedia Appendix 1

Total mention trend line from December 1, 2019, to November 13, 2020.

[PNG File , 271 KB - [medinform\\_v9i7e27116\\_app1.png](#) ]

---

### References

1. WHO Coronavirus (COVID-19) Dashboard. World Health Organization. URL: <https://covid19.who.int/> [accessed 2020-08-25]
2. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020 Mar 28;395(10229):1054-1062 [FREE Full text] [doi: [10.1016/S0140-6736\(20\)30566-3](https://doi.org/10.1016/S0140-6736(20)30566-3)] [Medline: [32171076](https://pubmed.ncbi.nlm.nih.gov/32171076/)]
3. Le DH, Bloom SA, Nguyen QH, Maloney SA, Le QM, Leitmeyer KC, et al. Lack of SARS transmission among public hospital workers, Vietnam. *Emerg Infect Dis* 2004 Feb;10(2):265-268 [FREE Full text] [doi: [10.3201/eid1002.030707](https://doi.org/10.3201/eid1002.030707)] [Medline: [15030695](https://pubmed.ncbi.nlm.nih.gov/15030695/)]
4. COVID-19 Public Health Emergency of International Concern (PHEIC) Global research and innovation forum. World Health Organization. 2020. URL: [https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-\(pheic\)-global-research-and-innovation-forum](https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-(pheic)-global-research-and-innovation-forum) [accessed 2020-05-19]
5. Khuyen cao phong chong viem phoi cap do chung virut moi Coronavirus tai thanh pho Vu Han, tinh Ho Bac, Trung Quoc. General Department of Preventive Medicine. URL: <https://tinyurl.com/96kewpk2> [accessed 2020-05-19]
6. Trang tin ve dich benh viem duong ho hap cap Covid-19. Ministry of Health. 2020. URL: <https://ncov.moh.gov.vn/> [accessed 2020-08-25]
7. Slovic P, Finucane ML, Peters E, MacGregor DG. Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. *Risk Anal* 2004 Apr;24(2):311-322. [doi: [10.1111/j.0272-4332.2004.00433.x](https://doi.org/10.1111/j.0272-4332.2004.00433.x)] [Medline: [15078302](https://pubmed.ncbi.nlm.nih.gov/15078302/)]



8. Loewenstein GF, Weber EU, Hsee CK, Welch N. Risk as feelings. *Psychol Bull* 2001 Mar;127(2):267-286. [doi: [10.1037/0033-2909.127.2.267](https://doi.org/10.1037/0033-2909.127.2.267)] [Medline: [11316014](https://pubmed.ncbi.nlm.nih.gov/11316014/)]
9. Griffith J, Antonio G, Ahuja A. SARS and the modern day pony express (the World Wide Web). *AJR Am J Roentgenol* 2003 Jun;180(6):1736. [doi: [10.2214/ajr.180.6.1801736](https://doi.org/10.2214/ajr.180.6.1801736)] [Medline: [12760954](https://pubmed.ncbi.nlm.nih.gov/12760954/)]
10. Zarocostas J. How to fight an infodemic. *Lancet* 2020 Feb 29;395(10225):676 [FREE Full text] [doi: [10.1016/S0140-6736\(20\)30461-X](https://doi.org/10.1016/S0140-6736(20)30461-X)] [Medline: [32113495](https://pubmed.ncbi.nlm.nih.gov/32113495/)]
11. Coronavirus disease (COVID-19) pandemic. World Health Organization. 2020. URL: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019> [accessed 2020-08-20]
12. Novel Coronavirus (2019-nCoV): situation report, 3. World Health Organization. 2020. URL: <https://apps.who.int/iris/handle/10665/330762> [accessed 2020-08-20]
13. Gu H, Chen B, Zhu H, Jiang T, Wang X, Chen L, et al. Importance of Internet surveillance in public health emergency control and prevention: evidence from a digital epidemiologic study during avian influenza A H7N9 outbreaks. *J Med Internet Res* 2014 Jan 17;16(1):e20 [FREE Full text] [doi: [10.2196/jmir.2911](https://doi.org/10.2196/jmir.2911)] [Medline: [24440770](https://pubmed.ncbi.nlm.nih.gov/24440770/)]
14. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. *J Med Internet Res* 2020 Apr 21;22(4):e19016 [FREE Full text] [doi: [10.2196/19016](https://doi.org/10.2196/19016)] [Medline: [32287039](https://pubmed.ncbi.nlm.nih.gov/32287039/)]
15. Ahmad AR, Murad HR. The impact of social media on panic during the COVID-19 pandemic in Iraqi Kurdistan: online questionnaire study. *J Med Internet Res* 2020 May 19;22(5):e19556 [FREE Full text] [doi: [10.2196/19556](https://doi.org/10.2196/19556)] [Medline: [32369026](https://pubmed.ncbi.nlm.nih.gov/32369026/)]
16. Zhao Y, Cheng S, Yu X, Xu H. Chinese public's attention to the COVID-19 epidemic on social media: observational descriptive study. *J Med Internet Res* 2020 May 04;22(5):e18825 [FREE Full text] [doi: [10.2196/18825](https://doi.org/10.2196/18825)] [Medline: [32314976](https://pubmed.ncbi.nlm.nih.gov/32314976/)]
17. Trevisan M, Le LC, Le AV. The COVID-19 pandemic: a view from Vietnam. *Am J Public Health* 2020 Aug;110(8):1152-1153. [doi: [10.2105/AJPH.2020.305751](https://doi.org/10.2105/AJPH.2020.305751)] [Medline: [32463704](https://pubmed.ncbi.nlm.nih.gov/32463704/)]
18. Hoang VM, Hoang HH, Khuong QL, La NQ, Tran TTH. Describing the pattern of the COVID-19 epidemic in Vietnam. *Glob Health Action* 2020 Dec 31;13(1):1776526 [FREE Full text] [doi: [10.1080/16549716.2020.1776526](https://doi.org/10.1080/16549716.2020.1776526)] [Medline: [32588779](https://pubmed.ncbi.nlm.nih.gov/32588779/)]
19. Huynh TLD. The COVID-19 risk perception: a survey on socioeconomics and media attention. *Economics Bull* 2020;40(1):A.
20. Brashers DE. Communication and uncertainty management. *J Commun* 2001;51(3):477-497. [doi: [10.1111/j.1460-2466.2001.tb02892.x](https://doi.org/10.1111/j.1460-2466.2001.tb02892.x)]
21. Hill D. Why they buy. *Across the Board* 2003;40(6):27-33.
22. Novac A. Traumatic stress and human behavior. *Psychiatric Times*. 2001. URL: <https://www.psychiatrictimes.com/view/traumatic-stress-and-human-behavior> [accessed 2020-08-25]
23. Downs A. Up and down with ecology: the "Issue-Attention Cycle". In: *The Politics of American Economic Policy Making*. Milton Park, Oxfordshire: Taylor & Francis; 1996.
24. Brossard D, Shanahan J, McComas K. Are issue-cycles culturally constructed? A comparison of French and American coverage of global climate change. *Mass Commun Soc* 2004 Jul;7(3):359-377. [doi: [10.1207/s15327825mcs0703\\_6](https://doi.org/10.1207/s15327825mcs0703_6)]
25. Shih T, Wijaya R, Brossard D. Media coverage of public health epidemics: linking framing and issue attention cycle toward an integrated theory of print news coverage of epidemics. *Mass Commun Soc* 2008 Apr 07;11(2):141-160. [doi: [10.1080/15205430701668121](https://doi.org/10.1080/15205430701668121)]
26. Ayers JW, Althouse BM, Dredze M, Leas EC, Noar SM. News and internet searches about human immunodeficiency virus after Charlie Sheen's disclosure. *JAMA Intern Med* 2016 Apr;176(4):552-554. [doi: [10.1001/jamainternmed.2016.0003](https://doi.org/10.1001/jamainternmed.2016.0003)] [Medline: [26902971](https://pubmed.ncbi.nlm.nih.gov/26902971/)]
27. de Vreese CH. News framing: theory and typology. *Inf Design J* 2005 Apr 18;13(1):51-62. [doi: [10.1075/idjdd.13.1.06vre](https://doi.org/10.1075/idjdd.13.1.06vre)]
28. Riker WH. In: Mueller JE, Calvert RL, Wilson RK, editors. *The Strategy of Rhetoric: Campaigning for the American Constitution*. New Haven, CT: Yale University Press; 1996.
29. Edwards GC, Wood BD. Who influences whom? The president, congress, and the media. *Am Political Sci Rev* 2014 Aug 01;93(2):327-344. [doi: [10.2307/2585399](https://doi.org/10.2307/2585399)]
30. Druckman JN. Political preference formation: competition, deliberation, and the (ir)relevance of framing effects. *Am Political Sci Rev* 2004 Nov 01;98(4):671-686. [doi: [10.1017/s0003055404041413](https://doi.org/10.1017/s0003055404041413)]
31. Gamson WA, Modigliani A. Media discourse and public opinion on nuclear power: a constructionist approach. *Am J Sociol* 1989 Jul;95(1):1-37. [doi: [10.1086/229213](https://doi.org/10.1086/229213)]
32. Walsh KC. *Talking About Politics: Informal Groups and Social Identity in American Life*. Chicago, IL: University of Chicago Press; 2004.
33. Kovach B, Rosenstiel T. *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect*. New York City: Three Rivers Press; 2014.
34. Singer JB, Hermida A, Domingo D, Heinonen A, Paulussen S, Quandt T, et al, editors. *Participatory Journalism: Guarding Open Gates at Online Newspapers*. Hoboken, NJ: John Wiley & Sons; 2011.
35. Tandoc Jr EC, Lim ZW, Ling R. Defining "Fake News": a typology of scholarly definitions. *Digital Journalism* 2017 Aug 30;6(2):137-153. [doi: [10.1080/21670811.2017.1360143](https://doi.org/10.1080/21670811.2017.1360143)]



36. Sundar SS. The MAIN Model: a heuristic approach to understanding technology effects on credibility. In: Metzger MJ, Flanagin AJ, editors. *Digital Media, Youth, and Credibility*. Cambridge, MA: The MIT Press; 2008.
37. McComas K, Shanahan J. Telling stories about global climate change. *Commun Res* 2016 Jun 30;26(1):30-57. [doi: [10.1177/009365099026001003](https://doi.org/10.1177/009365099026001003)]
38. Nisbet MC, Hume M. Attention cycles and frames in the plant biotechnology debate. *Harvard Int J Press/Polit* 2016 Sep 14;11(2):3-40. [doi: [10.1177/1081180x06286701](https://doi.org/10.1177/1081180x06286701)]
39. Semetko H, Valkenburg P. Framing European politics: a content analysis of press and television news. *J Commun* 2000;50(2):A. [doi: [10.1111/j.1460-2466.2000.tb02843.x](https://doi.org/10.1111/j.1460-2466.2000.tb02843.x)]
40. Thu Vien Pháp Luật. URL: <https://thuvienphapluat.vn/> [accessed 2020-05-10]
41. La V, Pham T, Ho M, Nguyen M, Nguyen KP, Vuong T, et al. Policy response, social media and science journalism for the sustainability of the public health system amid the COVID-19 outbreak: the Vietnam lessons. *Sustainability* 2020 Apr 07;12(7):2931. [doi: [10.3390/su12072931](https://doi.org/10.3390/su12072931)]
42. Leading active social media platforms among internet users in Vietnam as of 1st quarter of 2021. Statista. 2020. URL: <https://www.statista.com/statistics/941843/vietnam-leading-social-media-platforms/> [accessed 2020-09-20]
43. Vieweg SE. Situational awareness in mass emergency: a behavioral and linguistic analysis of microblogged communications. ProQuest. 2012. URL: <https://www.proquest.com/openview/540ee2ba902309c5ad7314438e06ea42/1?pq-origsite=gscholar&cbl=18750> [accessed 2020-09-16]
44. Imran M, Elbassuoni S, Castillo C, Diaz F, Meier P. Extracting information nuggets from disaster-related messages in social media. 2013 Presented at: 10th International ISCRAM Conference; May 2013; Baden-Baden, Germany.
45. Mirbabaie M, Bunker D, Stieglitz S, Marx J, Ehnis C. Social media in times of crisis: learning from Hurricane Harvey for the coronavirus disease 2019 pandemic response. *J Inf Technol* 2020 Jun 09;35(3):195-213. [doi: [10.1177/0268396220929258](https://doi.org/10.1177/0268396220929258)]
46. Li L, Zhang Q, Wang X, Zhang J, Wang T, Gao T, et al. Characterizing the propagation of situational information in social media during COVID-19 epidemic: a case study on Weibo. *IEEE Trans Comput Soc Syst* 2020 Apr;7(2):556-562. [doi: [10.1109/tcss.2020.2980007](https://doi.org/10.1109/tcss.2020.2980007)]
47. Rudra K, Ghosh S, Ganguly N, Goyal P, Ghosh S. Extracting situational information from microblogs during disaster events: a classification-summarization approach. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 2015 Presented at: CIKM '15; October 18-23, 2015; Melbourne, Australia. [doi: [10.1145/2806416.2806485](https://doi.org/10.1145/2806416.2806485)]
48. Van Eck NJ, Waltman L. *VOSviewer Manual*. Leiden: Univeriteit Leiden; 2013:1-53.
49. Singhal T. A review of coronavirus disease-2019 (COVID-19). *Indian J Pediatr* 2020 Apr;87(4):281-286 [FREE Full text] [doi: [10.1007/s12098-020-03263-6](https://doi.org/10.1007/s12098-020-03263-6)] [Medline: [32166607](https://pubmed.ncbi.nlm.nih.gov/32166607/)]
50. Coleman J. Vietnam reports first coronavirus cases. *The Hill*. 2020. URL: <https://web.archive.org/web/20200218074232/https://thehill.com/policy/healthcare/public-global-health/479542-vietnam-reports-first-coronavirus-cases> [accessed 2020-10-5]
51. Made-in-Vietnam COVID-19 vaccine scheduled for human trial this month. *Tuoi Tre News*. 2020. URL: <https://tuoitrenews.vn/news/society/20201103/madeinvietnam-covid19-vaccine-scheduled-for-human-trial-this-month/57580.html> [accessed 2020-11-15]

## Abbreviations

- API:** application programming interface  
**KOL:** key opinion leader  
**SARS:** severe acute respiratory syndrome  
**WHO:** World Health Organization

*Edited by C Lovis; submitted 12.01.21; peer-reviewed by A Alasmari, R Subramaniyam, M Kolotylo-Kulkarni, D Huang; comments to author 10.02.21; revised version received 25.02.21; accepted 17.06.21; published 16.07.21.*

*Please cite as:*

Tran HTT, Lu SH, Tran HTT, Nguyen BV

*Social Media Insights During the COVID-19 Pandemic: Infodemiology Study Using Big Data*

*JMIR Med Inform* 2021;9(7):e27116

URL: <https://medinform.jmir.org/2021/7/e27116>

doi: [10.2196/27116](https://doi.org/10.2196/27116)

PMID: [34152994](https://pubmed.ncbi.nlm.nih.gov/34152994/)

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

# A Tool for Evaluating Medication Alerting Systems: Development and Initial Assessment

Wu Yi Zheng<sup>1,2</sup>, PhD; Bethany Van Dort<sup>2</sup>, BMS; Romaric Marcilly<sup>3,4</sup>, PhD; Richard Day<sup>5</sup>, MD, PhD; Rosemary Burke<sup>6</sup>, BPharm; Sepehr Shakib<sup>7</sup>, MD, PhD; Young Ku<sup>8</sup>, BPharm; Hannah Reid-Anderson<sup>9</sup>, BNursing; Melissa Baysari<sup>2</sup>, PhD

<sup>1</sup>Black Dog Institute, Randwick, NSW, Australia

<sup>2</sup>The University of Sydney, Faculty of Medicine and Health, School of Medical Sciences, Biomedical Informatics and Digital Health, Sydney, Australia

<sup>3</sup>Univ Lille, CHU Lille, ULR 2694, METRICS: Évaluation des Technologies de santé des Pratiques médicales, Lille, France

<sup>4</sup>INSERM, CHU Lille, CIC-IT/Evalab 1403, Centre d'Investigation Clinique, Lille, France

<sup>5</sup>University of New South Wales, Randwick, Australia

<sup>6</sup>Sydney Local Health District, Sydney, Australia

<sup>7</sup>Royal Adelaide Hospital, Adelaide, Australia

<sup>8</sup>Hunter New England Local Health District, Newcastle, Australia

<sup>9</sup>Macquarie University Hospital, Sydney, Australia

**Corresponding Author:**

Wu Yi Zheng, PhD

Black Dog Institute

Hospital Rd

Prince of Wales Hospital

Randwick, NSW, 2031

Australia

Phone: 61 422510718

Email: [wuyi.zheng@unsw.edu.au](mailto:wuyi.zheng@unsw.edu.au)

## Abstract

**Background:** It is well known that recommendations from electronic medication alerts are seldom accepted or acted on by users. Key factors affecting the effectiveness of medication alerts include system usability and alert design. Thus, human factors principles that apply knowledge of human capabilities and limitations are increasingly used in the design of health technology to improve the usability of systems.

**Objective:** This study aims to evaluate a newly developed evidence-based self-assessment tool that allows the valid and reliable evaluation of computerized medication alerting systems. This tool was developed to be used by hospital staff with detailed knowledge of their hospital's computerized provider order entry system and alerts to identify and address potential system deficiencies. In this initial assessment, we aim to determine whether the items in the tool can measure compliance of medication alerting systems with human factors principles of design, the tool can be consistently used by multiple users to assess the same system, and the items are easy to understand and perceived to be useful for assessing medication alerting systems.

**Methods:** The Tool for Evaluating Medication Alerting Systems (TEMAS) was developed based on human factors design principles and consisted of 66 items. In total, 18 staff members recruited across 6 hospitals used the TEMAS to assess their medication alerting systems. Data collected from participant assessments were used to evaluate the validity, reliability, and usability of the TEMAS. Validity was assessed by comparing the results of the TEMAS with those of prior in-house evaluations. Reliability was measured using Krippendorff  $\alpha$  to determine agreement among assessors. A 7-item survey was used to determine usability.

**Results:** The participants reported mostly negative ( $n=8$ ) and neutral ( $n=7$ ) perceptions of alerts in their medication alerting system. However, the validity of the TEMAS could not be directly tested, as participants were unaware of any results from prior in-house evaluations. The reliability of the TEMAS, as measured by Krippendorff  $\alpha$ , was low to moderate (range 0.26-0.46); however, participant feedback suggests that individuals' knowledge of the system varied according to their professional background. In terms of usability, 61% (11/18) of participants reported that the TEMAS items were generally easy to understand; however, participants suggested the revision of 22 items to improve clarity.

**Conclusions:** This initial assessment of the TEMAS allowed the identification of its components that required modification to improve usability and usefulness. It also revealed that for the TEMAS to be effective in facilitating a comprehensive assessment of a medication alerting system, it should be completed by a multidisciplinary team of hospital staff from both clinical and technical backgrounds to maximize their knowledge of systems.

(*JMIR Med Inform* 2021;9(7):e24022) doi:[10.2196/24022](https://doi.org/10.2196/24022)

## KEYWORDS

medication alerts; decision support; human factors; assessment tool; usability flaws

## Introduction

### Background

Human factors is the scientific discipline that applies knowledge of human capabilities and limitations to improve the usability of systems, while reducing the potential for errors [1,2]. For decades, human factors research has been integral to the continuous improvement and innovation in industries outside of health care, such as aviation and automobile industries, with human performance limitations and human-system interactions taken into account when designing new technology [3-5]. For example, the failure to apply good human factors principles when designing aircraft and in-vehicle displays has been shown to lead to confusion and errors [3,6].

In recent years, the incorporation of human factors principles into the design of technology in health care has received increasing attention. Numerous studies have aimed to assess and improve clinical decision support in the form of electronic medication alerts [7-10], as it is well known that most recommendations from these alerts are not accepted or acted on by prescribers [11-14]. Excessive display of clinically irrelevant alerts can lead to alert fatigue, where important safety-critical information is ignored by clinicians (eg, doctors, pharmacists, and nurses) [14]. Studies have also investigated the factors influencing alert acceptance and found the following key factors affect the effectiveness of medication alerts: the usability of medication alerting systems, display of alerts, textual information included in alerts, and prioritization of alerts [15-21]. Furthermore, compared with poorly designed alerts, well-designed alerts using human factors principles resulted in faster work, fewer prescribing errors, less workload, and improved usability for prescribers [22,23].

However, what constitutes a well-designed medication safety alert and how compliance with human factors principles can be assessed and improved remain unclear. The Instrument for Evaluating Human Factors Principles in Medication-Related Decision Support Alerts (I-MeDeSA) was developed to evaluate compliance of drug-drug interaction alerts with human factors principles of design [10]. Comprising 26 items with binary scoring (ie, a score of 1 assigned to a yes response and 0 for a no response), the I-MeDeSA assesses compliance of electronic medication alerts with nine human factors principles of design, including alarm philosophy, placement, visibility, prioritization, color, learnability and confusability, text-based information, proximity of task components being displayed, and corrective actions [9]. Initially validated in the United States [10] and used in subsequent studies [7-9,24], several flaws with I-MeDeSA have been identified, including ambiguous item wording;

arbitrary allocation of scores to human factors principles; and the need for more concrete definitions, clearer rationale for each item, and more explicit examples [7,8,24]. In our attempt to use I-MeDeSA to evaluate computerized alerts in Australian systems, we found many of the items to be irrelevant to Australian configurations [7], namely, items that assumed systems implemented more than one level of alert severity and multiple alert types. Thus, we set out to develop an evidence-based self-assessment tool that allows the valid and reliable evaluation of computerized medication alerting systems, in terms of their compliance with human factors principles. Our goal was to develop a tool that could be used by hospital staff with detailed knowledge of the hospital's computerized provider order entry (CPOE) system and alerts (eg, a CPOE pharmacist who assisted in the building and configuration of the system) to identify and address deficient areas. This tool can also be used to facilitate the selection of the most user-friendly and functional medication alerting systems during the procurement process. With the increased adoption of digital health technology, a standardized tool using human factors principles to assess clinical decision support alerts, a crucial component of CPOE systems, would maximize alert acceptance and effectiveness and, therefore, broaden the potential safety benefits of medication-related alerts.

### Objectives

In this paper, we report the development of the Tool for Evaluating Medication Alerting Systems (TEMAS) and our initial attempts to assess its validity, reliability, and usability. In particular, we set out to determine whether (1) the items measure the compliance of medication alerting systems with human factors principles of design, (2) the tool can be consistently used by multiple users to assess the same system, and (3) the items are easy to understand and perceived to be useful for assessing medication alerting systems.

## Methods

### Development of the TEMAS

The pioneering work by Marcilly et al [25] identified 168 usability flaws related to general usability principles and medication-related alerting functions. A detailed description of each principle and its derivation can be found in a systematic qualitative review [25]. In summary, flaws specific to medication-related alerting functions were grouped into six categories, including low signal-to-noise ratio (eg, alerts are irrelevant or redundant), problems with alert content (eg, information required to make a decision is missing), nontransparency of alert functions (eg, no information on the



alert severity scale), timing and display issues (eg, alert not displayed at the right moment to support decision-making), alert distribution issues (eg, alert not displayed to the right clinician), and problems with alert features (eg, no feature for reconsidering an alert later) [25]. Usability flaws were then matched with 58 design principles identified in the literature and two additional principles [26]. A usability flaw was matched with a design principle if it was in direct violation of the principle [26].

The TEMAS was developed by transforming each design principle into a checklist item, using usability flaws identified by Marcilly et al [25] to corroborate the accuracy of each item. [Multimedia Appendix 1](#) includes some example design principles and their corresponding items in the TEMAS.

Following this mapping process, the TEMAS consists of 66 items ([Table 1](#)), which fall into six meta-principles: (1) signal-to-noise ratio, (2) ability to support collaborative work, (3) ability to fit clinicians' workflow and mental model, (4) display of relevant data within the alert, (5) transparency of system rules to the user, and (6) the inclusion of actionable tools within the alert. Each TEMAS item has two response options (ie, yes and no), with space provided for free-text comments. Before distributing the TEMAS to study participants, members of the research team, including experts in human factors, medication safety, digital health, and assessment tool development, checked and provided feedback on TEMAS items; however, pilot testing was not conducted with end users.

**Table 1.** Meta-principles assessed by the Tool for Evaluating Medication Alerting Systems (n=66).

Meta-principle	Items, n (%)	Example question
Optimize the signal-to-noise ratio	17 (26)	Does the alerting system use an evidence-based drug knowledge base to trigger alerts?
Support collaborative work	6 (9)	Does the alerting system trigger alerts to the appropriate team member (eg, medication administration alerts are triggered for nurses)?
Fit the clinicians' workflow and mental model	16 (24)	Does the alerting system display alerts instantly (ie, no lag time)?
Display relevant data within the alert	10 (15)	Does the alert include information on the cause of the unsafe event (eg, medication name and dose)?
Ensure the system rules are transparent to the user	6 (9)	Does the alerting system inform users about the customization options available (eg, turning some alerts off)?
Include actionable tools within the alert	11 (17)	Does the alert provide a function for the user to modify an order?

## Participants and Study Sites

To identify potential participants for the initial evaluation of the TEMAS, a member of the research team at each study site nominated staff members at their hospital with relevant knowledge of their CPOE system and alerts (eg, a CPOE pharmacist responsible for maintaining the system). The study intended to recruit at least two participants from each site. Nominated staff members were contacted by email, and those who expressed an interest in taking part in the study were sent a participant information sheet and consent form. After submitting a signed participant information sheet and consent form, participants received a TEMAS pack. This pack included a copy of the TEMAS and a 7-item survey. Participants were

asked to return completed TEMAS packs to the researchers via email or mail.

The study sites are presented in [Table 2](#). In total, 18 participants across the 6 sites used the TEMAS to assess the medication alerting system at their hospital. Participants included pharmacists (n=11), clinical pharmacologists (n=2), nurses (n=2), doctors (n=2), and a business analyst. Participants were part of the CPOE system implementation team at their hospital or were responsible for maintaining or updating the system. On average, participants had 5.1 (SD 2.9) years of experience using their CPOE system, and as shown in [Table 2](#), Cerner Powerchart and DXC Technology's MedChart were the most frequently assessed systems.

**Table 2.** Study sites and number of participants (n=18).

Study site	Participants, n (%)	CPOE <sup>a</sup> system in use
John Hunter Hospital (NSW <sup>b</sup> )	2 (11)	DXC Medchart
St Vincent's Hospital, Sydney (NSW)	2 (11)	DXC Medchart
Macquarie University Hospital (NSW)	2 (11)	TrakCare
Concord Repatriation General Hospital (NSW)	5 (28)	Cerner Powerchart
Royal North Shore Hospital (NSW)	4 (22)	Cerner Powerchart
Queen Elizabeth Hospital (South Australia)	3 (17)	Sunrise EMR <sup>c</sup>

<sup>a</sup>CPOE: computerized provider order entry.

<sup>b</sup>NSW: New South Wales.

<sup>c</sup>EMR: electronic medical record.

### Study Design and Data Analysis

The evaluation consisted of assessing three components: the validity, reliability, and usability of the TEMAS. Participants were asked to independently use the TEMAS to evaluate the medication alerting system in use at their hospital and then complete a 7-item survey.

To assess validity, the survey included a free-text item on the perceived effectiveness of the alerts in the CPOE system and asked for supporting information or evidence with their response (eg, information on alert override rates, any formal or informal feedback received from users, and results from any in-house user surveys). Data collected from this item were analyzed by categorizing responses according to their positive or negative valence. Supporting information provided by participants was compared with TEMAS results to check whether the

shortcomings of the alerting system identified by the TEMAS were consistent with those identified by in-house evaluations carried out by the hospitals.

To assess reliability, we compared the responses of participants working at the same hospital. Krippendorff  $\alpha$  was calculated to determine interrater reliability.

To assess usability, participants were given the opportunity to provide feedback on each TEMAS item to indicate whether an item was difficult to understand or was not useful (Figure 1). In addition, participants completed a usability survey (Multimedia Appendix 2), which collected basic demographic information, data on the ease of use using a five-point Likert scale (eg, item 1: I thought the TEMAS was easy to use), and free-text comments on the tool. The Likert-scale items were adapted from the system usability scale [27].

**Figure 1.** Feedback options for each Tool for Evaluating Medication Alerting Systems item to assess usability.

Please tick if the item is DIFFICULT to understand	Reason(s)	Please tick if the item is NOT useful	Reason(s)
<input type="checkbox"/>		<input type="checkbox"/>	

### Ethical Clearance

This study was approved by the Hunter New England Human Research Ethics Committee (reference no: HREC/18/HNE/237). In addition, research governance approval was obtained from each study site.

## Results

### Validity of the TEMAS

Participants gave mixed responses with regard to the perceived effectiveness of the alerts in their CPOE system. Of the 17 responses to this item (1 participant did not respond to this item), eight were negative, seven were neutral, and two were positive (Textbox 1).

**Textbox 1.** Selected comments of participants on the perceived effectiveness of alerts.

<p><b>Positive</b></p> <ul style="list-style-type: none"> <li>• “I believe they’re reasonably effective, as they target the conditions that are ‘no-nos’” [Participant #1]</li> <li>• “The alerts are coming from MIMS [Monthly Index of Medical Specialties] Australia and I believe their documentation is thorough.” [Participant #8]</li> </ul> <p><b>Neutral</b></p> <ul style="list-style-type: none"> <li>• “Somewhat effective. Pharmacists review quite a number of alerts via verification of medications, whilst there is a theoretical risk, there may not be many actual incidents.” [Participant #3]</li> </ul> <p><b>Negative</b></p> <ul style="list-style-type: none"> <li>• “Not very effective as prescribers have alert fatigue.” [Participant #2]</li> <li>• “Poor; time consuming; click fatigue; alert fatigue; irrelevant alerts (e.g. non-current meds).” [Participant #6]</li> <li>• “Too many alerts, hard to take out after we put in.” [Participant #7]</li> </ul>
---

However, no participant provided evidence to support their personal assessment of alerts in their hospital’s system; that is,

participants were unaware if their hospital collected meaningful data on the effectiveness of medication alerts in their CPOE system:

*Most of the effect i.e. override rates etc. we don't know* [Participant #9]

*Has much room for improvement based on the evaluation factors in TEMAS however have no figures or paper to back it up* [Participant #4]

**Table 3.** Interrater reliability among participants from each study site (n=6).

Site	All responses, Krippendorff $\alpha$ (95% CI)	Valid responses, Krippendorff $\alpha$ (95% CI)
1	.30 (0.06-0.53)	.32 (0.07-0.53)
2	.46 (0.25-0.67)	.49 (0.27-0.68)
3	.39 (0.32-0.45)	.47 (0.39-0.55)
4	.26 (0.17-0.35)	.32 (0.21-0.42)
5	.40 (0.28-0.51)	.49 (0.37-0.62)
6	.38 (0.16-0.60)	.38 (0.16-0.60)

At the individual TEMAS item level, more than 10 items at 3 study sites did not receive a valid response from all participants working in those sites. They commented that they did not have the relevant knowledge to answer some items:

*Not sure whether a doctor is able to make changes to the order and I'm not aware what their interface looks like.* [Participant #4]

*Unsure - medical officer questions.* [Participant #5]

**Table 4.** Usability of the Tool for Evaluating Medication Alerting Systems.

Survey item	Participants who selected strongly agree <sup>a</sup> or agree <sup>b</sup> , n (%)	Participants who selected neutral <sup>c</sup> , n (%)	Participants who selected strongly disagree <sup>d</sup> or disagree <sup>e</sup> , n (%)	Average score	Range (lower limit-upper limit)
I thought the TEMAS <sup>f</sup> was easy to use. (n=18)	7 (39)	8 (44)	3 (17)	3.2	4 (1-5)
I thought the items in the TEMAS were easy to understand. (n=18)	11 (61)	4 (22)	3 (17)	3.5	3 (2-5)
I thought the TEMAS was useful in helping me to identify areas for improvement in my alerting system. (n=17) <sup>g</sup>	7 (41)	8 (47)	2 (12)	3.3	4 (1-5)

<sup>a</sup>Strongly agree was rated 5.

<sup>b</sup>Agree was rated 4.

<sup>c</sup>Neutral was rated 3.

<sup>d</sup>Disagree was rated 2.

<sup>e</sup>Strongly disagree was rated 1.

<sup>f</sup>TEMAS: Tool for Evaluating Medication Alerting Systems.

<sup>g</sup>One participant did not provide a response to this question.

Of the 66 TEMAS items, 33 (50%) were reported by at least one participant as difficult to understand due to item wording. However, only 15% (10/66) of items confused multiple participants (Table 5). Reasons provided by participants on why items were difficult to understand included a lack of clarity in

## Reliability

Table 3 presents Krippendorff  $\alpha$ , which reflect interrater reliability among participants at each study site. To account for the missing data,  $\alpha$  were also calculated for items with valid responses only (ie, a response of yes or no).

## Usability

Approximately 39% (7/18) of participants thought that the TEMAS was easy to use, with roughly 60% (3/5) of participants reporting that it was easy to understand. Approximately 41% (7/17) of participants found it to be a useful tool for identifying areas for improvement in their medication alerting system (Table 4).

the meaning of the item and their inability to provide a yes or no response (Table 5). Furthermore, 20% (13/66) of items were reported to be *not useful* by participants. However, only the item on whether the alerting system provided explanations on

the classification of alert severity was deemed not useful by multiple participants (n=2 participants).

With regard to responses to free-text questions in the usability survey, participants provided additional comments on how the design of TEMAS could be improved, and other possible users of the tool:

*All of the questions are yes/no, either it is or it isn't - whereas in some cases it might be partially implemented. The questions are also worded such*

*that a "no" answer to any question is a negative, and something should be done about it. [Participant #1]*

*Think the target audience is unclear. Only some of these items can be optimised at a hospital level. Most of the issues are hard-coded and would need to be addressed by the vendor. [Participant #9]*

*Needs to be amended as it's unclear who i.e. IT people or clinical staff the TEMAS is aimed at. These groups require very different language [Participant #14]*

**Table 5.** Items in the Tool for Evaluating Medication Alerting Systems reported to be difficult to understand by multiple participants and example participant responses (n=18).

TEMAS <sup>a</sup> item <sup>b</sup>	Participants, n (%) <sup>c</sup>	Example participant response
A4. Does the alerting system overcome missing data and reconcile multiple entries to trigger relevant alerts (eg, does the alerting system avoid using dated or unreliable data?)	6 (33)	"Extremely broad question" [Participant #10]
A9. Does the alerting system refrain from triggering an alert if a corrective action has already been taken?	5 (28)	"Not sure what 'corrective action' means" [Participant #9]
E6. Does the alerting system inform users of the unsafe events that are checked?	4 (22)	"What is an unsafe event, and where would this be defined?" [Participant #11]
A3. Does the alerting system use multiple sources (eg, patient record, laboratory result repository, and pharmacy) to trigger alerts?	3 (17)	"I am not sure what this is asking" [Participant #7]
A13. Does the alerting system group multiple recommendations for patients with comorbidities?	3 (17)	"Don't think our system has this capability" [Participant #11]
A12. Does the alerting system prioritize alerts according to severity?	2 (11)	"This depends on what you mean by 'prioritise'" [Participant #11]
D1. Does the alert include information on the cause of the unsafe event (eg, medication name and dose)?	2 (11)	"Unsure how to answer" [Participant #12]
D5. Does the alert include relevant patient information and provide a link for users to obtain further patient information?	2 (11)	"Example of patient info? Lab results?" [Participant #13]
F1. Does the alert provide a function for the user to modify an order?	2 (11)	"Only doctors can modify orders. Difficult for other professions to answer" [Participant #12]
F10. Does the alerting system allow users to remove alerts that are irrelevant or outdated?	2 (11)	"Difficult to classify as Y or N" [Participant #12]

<sup>a</sup>TEMAS: Tool for Evaluating Medication Alerting Systems.

<sup>b</sup>The letter and number preceding each item indicates section and item number, respectively.

<sup>c</sup>The values do not sum to 100% as they are not mutually exclusive.

## Discussion

### Principal Findings

In this study, we developed a self-assessment tool for medication alerting systems and aimed to evaluate the validity, reliability, and usability of the TEMAS; however, this proved difficult. The validity of the TEMAS could not be directly tested, as participants in the study were not aware of any in-house system evaluations carried out by the hospitals. As a result, participants reported that there was a lack of evaluation data to support their subjective assessment of the system. The reliability of the TEMAS, as measured by Krippendorff  $\alpha$ , was low to moderate; however, feedback from users indicated that their knowledge of systems was highly variable. In terms of usability, according to the responses to a survey item, the majority of participants agreed that TEMAS items were easy to understand, although

participants identified a number of items that needed improvement.

Several methods are used by hospitals to monitor and evaluate alert effectiveness, including the establishment of review committees consisting of pharmacists and doctors [28-30], development of visual analytic dashboards [13], and collection of end user feedback [31]. A key finding from this study was that no participating hospital had a systematic program in place to gather data on the effectiveness of medication alerts in their CPOE system. Although the view of participants on alerts in their systems were mostly negative, there was a lack of evaluation data to support these subjective assessments. Thus, the validity of the TEMAS could not be directly assessed. Future assessments of the TEMAS should consider applying a different participant screening process whereby only hospitals with available evaluation data are included. However, upon



examining the TEMAS items considered to be *not useful* by study participants, only one item was deemed not useful by multiple users (n=2), suggesting that the content of the TEMAS was relevant in assessing medication alerting systems. Further evaluations of the TEMAS should be conducted in hospitals with in-house data on the effectiveness of alerts in their CPOE system.

Less than half of the participants indicated that the TEMAS was easy to use (7/18, 39%) and useful in identifying areas in the system for improvement (7/17, 41%), with more participants selecting *neutral* for these survey questions. This likely reflects that some TEMAS items needed improvement, which prevented respondents from fully endorsing the usability of the TEMAS. In response to the feedback received on individual TEMAS items, 33% (22/66) of items were modified to improve clarity and reduce ambiguities. To avoid confusion and

misunderstanding due to the use of unsuitable terms (eg, corrective action, item A9; Table 6) and poor item wording (eg, item D1, Table 6), edits were made to the original TEMAS, taking into account participant comments on why they were unable to provide a response (eg, “I am not sure what this is asking” [Participant #5]). We also included examples to provide further clarification of the meaning of each item (Table 6). In response to feedback on difficulties in selecting a yes or no response for some items (eg, only some alerts provide clinically appropriate recommendations and suggest alternatives), the revised version of the TEMAS (Multimedia Appendix 3) included *partial* as an additional response option for each item. In addition, a note has been included to advise users that, depending on the local context, a response of *no* or *partial* to TEMAS items does not automatically indicate a weakness in the system.

**Table 6.** Examples of the revised Tool for Evaluating Medication Alerting Systems items.

Original item <sup>a</sup>	Revised item	Example to clarify the meaning of the item
A9. Does the alerting system refrain from triggering an alert if a corrective action has already been taken?	Does the alerting system refrain from triggering more alerts if the alert recommendation has already been followed?	The system refrains from triggering an alert if drug monitoring actions are already in place.
D1. Does the alert include information on the cause of the unsafe event (eg, medication name and dose)?	Does the alert include information on why the alert was triggered?	Medication names, dosages, and severity of interactions are included in drug-drug interaction alerts.
E6. Does the alerting system inform users of the unsafe events that are checked?	Does the alerting system inform users of the types of orders that will trigger alerts?	Clicking on a “more information” link in the help page informs the user that both order sentences and free-text orders can trigger alerts.

<sup>a</sup>The letter and number preceding each item indicates section and item number, respectively.

The reliability of the TEMAS was shown to be poor, likely reflecting the different levels of system knowledge possessed by the participants. Recruiting participants with equivalent, in-depth knowledge of their hospital’s medication alerting system proved difficult. Usually, one staff member possessed extensive knowledge of the hospital’s system (eg, a CPOE pharmacist), whereas other staff members within the same organization had more specialized knowledge of the hospital’s system (eg, a medical officer or clinical pharmacist). It may be that reliability was affected by differences in clinical practice settings, where staff members from different specialties use different functions of the system and have different views and understanding of the system based on their everyday use. Responses received from users suggest that the TEMAS may be more appropriately used by a team instead of an individual. For example, a participant in a pharmacist role was unsure of items related to prescribing medications, thus deferring these items to medical officers. There was also a suggestion to include system vendors in the evaluation process as “most of the issues are hard-coded and would need to be addressed by the vendor” [Participant #9]. Thus, evaluations carried out by a team consisting of representatives of system users from all clinical backgrounds would allow a more comprehensive evaluation of the alerting system. During this process, different parts of the TEMAS could initially be assigned to different team members based on their role and relevant expertise in the hospital (eg, prescribers are assigned to the *fit the clinician’s workflow and mental model* section).

The TEMAS is not dissimilar to a heuristic evaluation, which is a usability inspection method driven by experts to assess a design or product’s usability [32]. In heuristic analysis, a number of usability experts typically conduct an independent assessment of a product or interface and note usability violations, which are then amalgamated into a master list of usability problems. Using this approach, the identification of usability violations is highly dependent on the expertise of raters, with human factors or usability expertise associated with higher number of violations being detected. This is in contrast to the TEMAS, where we suggest users work as a team, not independently, to complete their alert assessment. This is because items cover a range of system aspects that are unlikely to be known to a single individual. We also suggest that the completion of the TEMAS should not be limited to usability experts but rather a multidisciplinary team of hospital end users (eg, pharmacists, doctors, nurses, and information technology professionals), each contributing their unique knowledge in the evaluation of a medication alerting system.

### Limitations

Our initial evaluation of the TEMAS had several limitations. First, we experienced difficulties in recruiting participants with in-depth knowledge of their hospital’s medication alerting system. Knowledge of some participants was role specific, limiting their capacity to complete the TEMAS, which impacted the interrater reliability. Future assessments of the TEMAS could use a team of system experts with varying expertise from

different professional backgrounds. Second, we did not recruit a site with in-house evaluation data of their medication alerting system, thus limiting our ability to assess the validity of the TEMAS. As a result, findings derived from using the TEMAS to assess the strengths and weaknesses of medication alerting systems should be interpreted with caution and within the context of the organization. Furthermore, the TEMAS is designed to assess medication alerting systems in inpatient care and is likely to require some modification if it is to be used in other settings, such as pharmacy or outpatient settings. Finally, the TEMAS was not piloted with prospective end users before distribution to study sites; however, research team members with expertise in human factors, medication safety, and digital health checked and provided feedback on TEMAS items.

## Conclusions

On basis of the usability flaws matched to human factors design principles, the TEMAS was developed for hospitals to self-assess medication alerts in their CPOE system with the goal of improving the effectiveness of these alerts. This initial evaluation allowed the identification of components of the TEMAS that required modification to improve usability and usefulness, leading to changes to items and the addition of examples and a response option. To be effective in facilitating a comprehensive evaluation, we found that the TEMAS should be completed by a team of multidisciplinary hospital staff from both clinical and technical backgrounds. This study was integral to the evolution of the TEMAS and established a revised version ready for use. As a next step, the updated TEMAS will be trialed by teams of users to assess their medication alerting systems and to compare the assessment results of the TEMAS with the I-MeDeSA.

---

## Acknowledgments

The authors would like to thank Professor Sarah Hilmer for her assistance with participant recruitment and Honorary Associate Professor Peter Hibbert for his advice on the development of this assessment tool. The researchers working on this project were funded by the National Health and Medical Research Council Partnership grant 1134824.

---

## Conflicts of Interest

None declared.

---

### Multimedia Appendix 1

Design principles and corresponding Tool for Evaluating Medication Alerting Systems items.

[[DOCX File, 23 KB](#) - [medinform\\_v9i7e24022\\_app1.docx](#) ]

---

### Multimedia Appendix 2

Usability survey.

[[DOCX File, 18 KB](#) - [medinform\\_v9i7e24022\\_app2.docx](#) ]

---

### Multimedia Appendix 3

The Tool for Evaluating Medication Alerting Systems (revised).

[[DOCX File, 41 KB](#) - [medinform\\_v9i7e24022\\_app3.docx](#) ]

---

## References

1. Baysari M, Clay-Williams R, Loveday T. A human factors resource for health professionals and health services staff. In: The Human Factors and Ergonomics Society of Australia, The Australian Institute of Health Innovation. Australia: Macquarie University, The University of Sydney and the NSW Clinical Excellence Commission; 2019:1-70.
2. Phansalkar S, Edworthy J, Hellier E, Seger DL, Schedlbauer A, Avery AJ, et al. A review of human factors principles for the design and implementation of medication safety alerts in clinical information systems. *J Am Med Inform Assoc* 2010 Sep 01;17(5):493-501 [FREE Full text] [doi: [10.1136/jamia.2010.005264](#)] [Medline: [20819851](#)]
3. Akamatsu M, Green P, Bengler K. Automotive technology and human factors research: past, present, and future. *Int J Veh Technol* 2013 Sep 04;2013:1-27. [doi: [10.1155/2013/526180](#)]
4. Safety Behaviours: Human Factors for Pilots, 2nd Edition. Australia: Civil Aviation Safety Authority Australia; 2019.
5. Salas E, Maurino D, Curtis M. Chapter 1 - Human factors in aviation: an overview. In: Salas E, Maurino D, editors. *Human Factors in Aviation* (Second Edition). San Diego, CA: Academic Press; 2010:3-19.
6. Palmer R. Applying human factors principles in aviation displays: a transition from analog to digital cockpit displays in the CP140 Aurora Aircraft. University of Tennessee, Knoxville. 2007. URL: [https://trace.tennessee.edu/cgi/viewcontent.cgi?article=1217&context=utk\\_gradthes](https://trace.tennessee.edu/cgi/viewcontent.cgi?article=1217&context=utk_gradthes) [accessed 2021-06-25]

7. Baysari MT, Lowenstein D, Zheng WY, Day RO. Reliability, ease of use and usefulness of I-MeDeSA for evaluating drug-drug interaction alerts in an Australian context. *BMC Med Inform Decis Mak* 2018 Oct 05;18(1):83 [FREE Full text] [doi: [10.1186/s12911-018-0666-y](https://doi.org/10.1186/s12911-018-0666-y)] [Medline: [30290797](https://pubmed.ncbi.nlm.nih.gov/30290797/)]
8. Lowenstein D, Zheng WY, Burke R, Kenny E, Sandhu A, Makeham M, et al. Do user preferences align with human factors assessment scores of drug-drug interaction alerts? *Health Informatics J* 2020 Mar 11;26(1):563-575 [FREE Full text] [doi: [10.1177/1460458219840210](https://doi.org/10.1177/1460458219840210)] [Medline: [30973280](https://pubmed.ncbi.nlm.nih.gov/30973280/)]
9. Phansalkar S, Zachariah M, Seidling HM, Mendes C, Volk L, Bates DW. Evaluation of medication alerts in electronic health records for compliance with human factors principles. *J Am Med Inform Assoc* 2014 Oct 01;21(e2):332-340 [FREE Full text] [doi: [10.1136/amiajnl-2013-002279](https://doi.org/10.1136/amiajnl-2013-002279)] [Medline: [24780721](https://pubmed.ncbi.nlm.nih.gov/24780721/)]
10. Zachariah M, Phansalkar S, Seidling HM, Neri PM, Cresswell KM, Duke J, et al. Development and preliminary evidence for the validity of an instrument assessing implementation of human-factors principles in medication-related decision-support systems--I-MeDeSA. *J Am Med Inform Assoc* 2011 Dec 01;18 Suppl 1(Supplement 1):62-72 [FREE Full text] [doi: [10.1136/amiajnl-2011-000362](https://doi.org/10.1136/amiajnl-2011-000362)] [Medline: [21946241](https://pubmed.ncbi.nlm.nih.gov/21946241/)]
11. Edrees H, Amato M, Wong A, Seger D, Bates D. High-priority drug-drug interaction clinical decision support overrides in a newly implemented commercial computerized provider order-entry system: override appropriateness and adverse drug events. *J Am Med Inform Assoc* 2020 Jun 01;27(6):893-900 [FREE Full text] [doi: [10.1093/jamia/ocaa034](https://doi.org/10.1093/jamia/ocaa034)] [Medline: [32337561](https://pubmed.ncbi.nlm.nih.gov/32337561/)]
12. Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. *AMIA Annu Symp Proc* 2007 Oct 11:26-30 [FREE Full text] [Medline: [18693791](https://pubmed.ncbi.nlm.nih.gov/18693791/)]
13. Simpaio AF, Ahumada LM, Desai BR, Bonafide CP, Gálvez JA, Rehman MA, et al. Optimization of drug-drug interaction alert rules in a pediatric hospital's electronic health record system using a visual analytics dashboard. *J Am Med Inform Assoc* 2015 Mar 15;22(2):361-369. [doi: [10.1136/amiajnl-2013-002538](https://doi.org/10.1136/amiajnl-2013-002538)] [Medline: [25318641](https://pubmed.ncbi.nlm.nih.gov/25318641/)]
14. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. *J Am Med Inform Assoc* 2006 Mar 01;13(2):138-147. [doi: [10.1197/jamia.m1809](https://doi.org/10.1197/jamia.m1809)]
15. Abramson EL, Patel V, Pfoh ER, Kaushal R. How physician perspectives on e-prescribing evolve over time. A case study following the transition between EHRs in an outpatient clinic. *Appl Clin Inform* 2016 Oct 26;7(4):994-1006 [FREE Full text] [doi: [10.4338/ACI-2016-04-RA-0069](https://doi.org/10.4338/ACI-2016-04-RA-0069)] [Medline: [27786335](https://pubmed.ncbi.nlm.nih.gov/27786335/)]
16. Ash JS. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *J Am Med Inform Assoc* 2003 Nov 21;11(2):104-112. [doi: [10.1197/jamia.m1471](https://doi.org/10.1197/jamia.m1471)]
17. Brown CL, Mulcaster HL, Triffitt KL, Sittig DF, Ash JS, Reygate K, et al. A systematic review of the types and causes of prescribing errors generated from using computerized provider order entry systems in primary and secondary care. *J Am Med Inform Assoc* 2017 Mar 01;24(2):432-440 [FREE Full text] [doi: [10.1093/jamia/ocw119](https://doi.org/10.1093/jamia/ocw119)] [Medline: [27582471](https://pubmed.ncbi.nlm.nih.gov/27582471/)]
18. Khajouei R, Jaspers MW. The impact of CPOE medication systems' design aspects on usability, workflow and medication orders: a systematic review. *Methods Inf Med* 2010;49(1):3-19. [doi: [10.3414/ME0630](https://doi.org/10.3414/ME0630)] [Medline: [19582333](https://pubmed.ncbi.nlm.nih.gov/19582333/)]
19. Kuperman GJ, Bobb A, Payne TH, Avery AJ, Gandhi TK, Burns G, et al. Medication-related clinical decision support in computerized provider order entry systems: a review. *J Am Med Inform Assoc* 2007 Jan 01;14(1):29-40. [doi: [10.1197/jamia.m2170](https://doi.org/10.1197/jamia.m2170)]
20. Seidling HM, Phansalkar S, Seger DL, Paterno MD, Shaykevich S, Haefeli WE, et al. Factors influencing alert acceptance: a novel approach for predicting the success of clinical decision support. *J Am Med Inform Assoc* 2011;18(4):479-484 [FREE Full text] [doi: [10.1136/amiajnl-2010-000039](https://doi.org/10.1136/amiajnl-2010-000039)] [Medline: [21571746](https://pubmed.ncbi.nlm.nih.gov/21571746/)]
21. McCoy AB, Waitman LR, Lewis JB, Wright JA, Choma DP, Miller RA, et al. A framework for evaluating the appropriateness of clinical decision support alerts and responses. *J Am Med Inform Assoc* 2012 May 01;19(3):346-352 [FREE Full text] [doi: [10.1136/amiajnl-2011-000185](https://doi.org/10.1136/amiajnl-2011-000185)] [Medline: [21849334](https://pubmed.ncbi.nlm.nih.gov/21849334/)]
22. Russ AL, Chen S, Melton BL, Johnson EG, Spina JR, Weiner M, et al. A novel design for drug-drug interaction alerts improves prescribing efficiency. *Joint Comm J Qual Patient Saf* 2015 Sep;41(9):396-405. [doi: [10.1016/s1553-7250\(15\)41051-7](https://doi.org/10.1016/s1553-7250(15)41051-7)]
23. Russ A, Zillich A, Melton B, Russell SA, Chen S, Spina JR, et al. Applying human factors principles to alert design increases efficiency and reduces prescribing errors in a scenario-based simulation. *J Am Med Inform Assoc* 2014 Oct;21(e2):287-296 [FREE Full text] [doi: [10.1136/amiajnl-2013-002045](https://doi.org/10.1136/amiajnl-2013-002045)] [Medline: [24668841](https://pubmed.ncbi.nlm.nih.gov/24668841/)]
24. Cho I, Lee J, Han H, Phansalkar S, Bates D. Evaluation of a Korean version of a tool for assessing the incorporation of human factors into a medication-related decision support system: the I-MeDeSA. *Appl Clin Inform* 2017 Dec 21;05(02):571-588. [doi: [10.4338/aci-2014-01-ra-0005](https://doi.org/10.4338/aci-2014-01-ra-0005)]
25. Marcilly R, Ammenwerth E, Vasseur F, Roehrer E, Beuscart-Zéphir MC. Usability flaws of medication-related alerting functions: a systematic qualitative review. *J Biomed Inform* 2015 Jun;55:260-271 [FREE Full text] [doi: [10.1016/j.jbi.2015.03.006](https://doi.org/10.1016/j.jbi.2015.03.006)] [Medline: [25817918](https://pubmed.ncbi.nlm.nih.gov/25817918/)]
26. Marcilly R, Ammenwerth E, Roehrer E, Niès J, Beuscart-Zéphir MC. Evidence-based usability design principles for medication alerting systems. *BMC Med Inform Decis Mak* 2018 Jul 24;18(1):69 [FREE Full text] [doi: [10.1186/s12911-018-0615-9](https://doi.org/10.1186/s12911-018-0615-9)] [Medline: [30041647](https://pubmed.ncbi.nlm.nih.gov/30041647/)]

27. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. *Int J Hum-Comput Int* 2008 Jul 30;24(6):574-594. [doi: [10.1080/10447310802205776](https://doi.org/10.1080/10447310802205776)]
28. Bhakta S, Colavecchia A, Haines L, Varkey D, Garey K. A systematic approach to optimize electronic health record medication alerts in a health system. *Am J Health Syst Pharm* 2019 Apr 08;76(8):530-536. [doi: [10.1093/ajhp/zxz012](https://doi.org/10.1093/ajhp/zxz012)] [Medline: [31361861](https://pubmed.ncbi.nlm.nih.gov/31361861/)]
29. Kawamanto K, Flynn M, Kukhareva P, ElHalta D, Hess R, Gregory T, et al. A pragmatic guide to establishing clinical decision support governance and addressing decision support fatigue: a case study. *AMIA Annu Symp Proc* 2018;2018:624-633 [FREE Full text] [Medline: [30815104](https://pubmed.ncbi.nlm.nih.gov/30815104/)]
30. Zenziper Y, Kurnik D, Markovits N, Ziv A, Shamiss A, Halkin H, et al. Implementation of a clinical decision support system for computerized drug prescription entries in a large tertiary care hospital. *Isr Med Assoc J* 2014 May;16(5):289-294 [FREE Full text] [Medline: [24979833](https://pubmed.ncbi.nlm.nih.gov/24979833/)]
31. van Camp PJ, Kirkendall ES, Hagedorn PA, Minich T, Kouril M, Spooner SA, et al. Feedback at the Point of Care to Decrease Medication Alert Rates in an Electronic Health Record. *Pediatr Emerg Care* 2020 Jul;36(7):e417-e422. [doi: [10.1097/PEC.0000000000001847](https://doi.org/10.1097/PEC.0000000000001847)] [Medline: [31136457](https://pubmed.ncbi.nlm.nih.gov/31136457/)]
32. Lau F, Kuziemyky C. *Handbook of eHealth Evaluation: An Evidence-Based Approach*. Victoria, BC: University of Victoria; 2017.

## Abbreviations

**CPOE:** computerized provider order entry

**I-MeDeSA:** Instrument for Evaluating Human Factors Principles in Medication-Related Decision Support Alerts

**TEMAS:** Tool for Evaluating Medication Alerting Systems

*Edited by G Eysenbach; submitted 01.09.20; peer-reviewed by A Russ, K Huat, R Ologeanu-Taddei, J Bagby; comments to author 08.10.20; revised version received 04.11.20; accepted 03.06.21; published 16.07.21.*

*Please cite as:*

Zheng WY, Van Dort B, Marcilly R, Day R, Burke R, Shakib S, Ku Y, Reid-Anderson H, Baysari M

*A Tool for Evaluating Medication Alerting Systems: Development and Initial Assessment*

*JMIR Med Inform* 2021;9(7):e24022

URL: <https://medinform.jmir.org/2021/7/e24022>

doi: [10.2196/24022](https://doi.org/10.2196/24022)

PMID: [34269680](https://pubmed.ncbi.nlm.nih.gov/34269680/)

©Wu Yi Zheng, Bethany Van Dort, Romaric Marcilly, Richard Day, Rosemary Burke, Sepehr Shakib, Young Ku, Hannah Reid-Anderson, Melissa Baysari. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org/>), 16.07.2021. 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

# Predicting Unscheduled Emergency Department Return Visits Among Older Adults: Population-Based Retrospective Study

Rai-Fu Chen<sup>1\*</sup>, PhD; Kuei-Chen Cheng<sup>2\*</sup>, MA; Yu-Yin Lin<sup>2\*</sup>, MA; I-Chiu Chang<sup>2\*</sup>, PhD; Cheng-Han Tsai<sup>2,3\*</sup>, MA

<sup>1</sup>Department of Information Management, Chia-Nan University of Pharmacy and Science, Tainan City, Taiwan

<sup>2</sup>Department of Information Management, National Chung Cheng University, Chiayi County, Taiwan

<sup>3</sup>Department of Emergency, Chiayi Branch, Taichung Veterans General Hospital, Chiayi City, Taiwan

\* all authors contributed equally

**Corresponding Author:**

I-Chiu Chang, PhD

Department of Information Management

National Chung Cheng University

No168, Sec 1, University Rd

Minhsiung

Chiayi County, 621301

Taiwan

Phone: 886 5 2720411 ext 16850

Email: [misicc@mis.ccu.edu.tw](mailto:misicc@mis.ccu.edu.tw)

## Abstract

**Background:** Unscheduled emergency department return visits (EDRVs) are key indicators for monitoring the quality of emergency medical care. A high return rate implies that the medical services provided by the emergency department (ED) failed to achieve the expected results of accurate diagnosis and effective treatment. Older adults are more susceptible to diseases and comorbidities than younger adults, and they exhibit unique and complex clinical characteristics that increase the difficulty of clinical diagnosis and treatment. Older adults also use more emergency medical resources than people in other age groups. Many studies have reviewed the causes of EDRVs among general ED patients; however, few have focused on older adults, although this is the age group with the highest rate of EDRVs.

**Objective:** This aim of this study is to establish a model for predicting unscheduled EDRVs within a 72-hour period among patients aged 65 years and older. In addition, we aim to investigate the effects of the influencing factors on their unscheduled EDRVs.

**Methods:** We used stratified and randomized data from Taiwan's National Health Insurance Research Database and applied data mining techniques to construct a prediction model consisting of patient, disease, hospital, and physician characteristics. Records of ED visits by patients aged 65 years and older from 1996 to 2010 in the National Health Insurance Research Database were selected, and the final sample size was 49,252 records.

**Results:** The decision tree of the prediction model achieved an acceptable overall accuracy of 76.80%. Economic status, chronic illness, and length of stay in the ED were the top three variables influencing unscheduled EDRVs. Those who stayed in the ED overnight or longer on their first visit were less likely to return. This study confirms the results of prior studies, which found that economically underprivileged older adults with chronic illness and comorbidities were more likely to return to the ED.

**Conclusions:** Medical institutions can use our prediction model as a reference to improve medical management and clinical services by understanding the reasons for 72-hour unscheduled EDRVs in older adult patients. A possible solution is to create mechanisms that incorporate our prediction model and develop a support system with customized medical education for older patients and their family members before discharge. Meanwhile, a reasonably longer length of stay in the ED may help evaluate treatments and guide prognosis for older adult patients, and it may further reduce the rate of their unscheduled EDRVs.

(*JMIR Med Inform* 2021;9(7):e22491) doi:[10.2196/22491](https://doi.org/10.2196/22491)

**KEYWORDS**

classification model; decision tree; emergency department; older adult patients; unscheduled return visits

## Introduction

### Background and Setting

Many countries today face challenges related to the rapidly aging population. Advances in medical technology and the aging of post-World War II baby boomers have led to a greater proportion of adults aged over 65 years in many industrialized nations' populations. This substantive shift in demographics not only increases the overall demand for health care and medical services but also influences economic and social welfare policies. Older adults are more susceptible to diseases and comorbidities than younger adults, and they exhibit unique and complex clinical characteristics that increase the difficulty of clinical diagnosis and treatment [1]. Older adults also use more emergency medical resources than people in other age demographics do [2-8], and approximately 14.9% of emergency department (ED) patients in the United States are aged 65 years or older [9], making them the most frequent visitors to the ED. In Taiwan, 25.5% of all ED visits are made by adults aged 65 years or older [10], a percentage that is approximately two-fold higher than that in the United States. This age group has the highest rate of ED return visits (EDRVs) [11,12].

A high unscheduled EDRV rate implies that the medical services provided by the ED failed to achieve the expected results of accurate diagnosis and effective treatment [11] and is a key indicator for monitoring the quality of emergency medical care [11,13]. EDRVs might contribute to crowding and further diminish the quality of care in the emergency room. As most older individuals have complex clinical characteristics, EDRVs would use more emergency room resources. In addition, EDRVs increase the risk of contracting infectious diseases in older adults, especially during the COVID-19 pandemic.

Many studies have reviewed the causes of EDRVs among general ED patients [11,14-16]; however, few have focused on older adults, even though this is the age group with the highest rate of EDRVs [11,12]. In fact, the risk of EDRVs for adult ED patients aged older than 65 years is approximately 300% higher than that for adults aged less than 30 years and 200% higher than that for adults aged less than 46 years [11,12,17]. Older adults who repeatedly return to the ED to seek medical assistance are at an increased risk of medical errors and contribute to excessive use of emergency medical resources [11,18]. However, the findings of past studies that did not focus on this demographic may not be suitable for predicting older adults' rate of unscheduled EDRVs. In addition, past studies have mainly collected samples from single targets (hospitals) [11,12,14-17]. Sample collection from a single hospital source can lead to underestimation of return rates, because ED patients who return within 72 hours may visit a different hospital's ED.

Taiwan's health care services have been ranked the highest worldwide by The Richest [19] and second ranking worldwide by the Economist Intelligence Unit [20]. Numbeo ranked the health care system of Taiwan's national health insurance first worldwide in its Health Care Index and Health Care Exp Index [21], with an enrollment of approximately 99.68% of the population [22]. The high ranking of Taiwan's Digital Government Program [23] made the population-wide database,

the National Health Insurance Research Database (NHIRD). Therefore, this study uses the NHIRD and develops a simple and useful prediction model to identify critical factors influencing older adult patients' unscheduled EDRVs through a machine learning technique. Such a model could provide useful suggestions for hospital managers and health care professionals in delivering high ED qualities from the considerations of disease, patient, physician, and institution (hospital) factors. Furthermore, it could serve as a valuable reference for future government planning and promotion of medical services and age-friendly policies.

### Related Studies

The factors influencing EDRVs can be categorized into approximately four areas: disease-related, patient-related, physician-related, and medical institution-related factors. One of the major disease-related reasons for ED visits is a pathological condition with unclear symptoms, signs, and diagnoses, and the primary pathological condition responsible for EDRVs, such as abdominal pain [12,15-17,24-26] with a diagnostic error rate of 68%-73% [15]. Fever is another pathological condition that causes EDRVs [24,27]. Other disease-related factors include infectious disease [25] with urinary tract infections, accounting for 35% of all infectious diseases [12]; muscle, bone, or head traumas [14,26]; cancer [12,27]; and alcoholism, depression, and other mental illnesses. Patients with high triage classification (TC) [24,25,28], heart disease or diabetes [16], or chronic illness with comorbidities [17] also exhibit a high likelihood of an EDRV. A high Charlson Comorbidity Index indicates a high risk of EDRV [29,30], particularly for patients aged older than 75 years [25].

EDRVs are known to increase concurrently with age [11,12,17]. The gender effect on the rate of EDRVs is uncertain [16,24,26,31,32]. Other patient-related factors that have a significant influence on EDRVs include personal insistence on using ED services [33]. EDRVs are 25%-30% higher in low-income countries than in high-income countries [24]. Diagnostic errors by medical staff account for the highest percentage (5.7%-9%) of medical errors [27] and are a common cause of unscheduled EDRVs. Prior studies found that physicians' years of practice significantly influenced the rates of EDRVs [15,34]. Kuan and Mahadevan [15] indicated that the reasons for physicians' years of practice significantly influence the rates of EDRVs are related to their experience and training as ED physicians. Improved communication between physicians and staff, patients, and family members can also reduce the likelihood of EDRVs. Inadequate emergency resources, particularly in rural hospitals [31] or in staffing during nighttime and on weekends [13], increase the likelihood of EDRVs. An ED stay of more than 6 hours is uncommon [18] because longer stays might contribute to crowding problems and diminish the quality of providing expeditious triage, workup, and selection of endangered emergency patients.

In summary, the literature confirms that disease-, patient-, physician-, and institution-related factors all influence the rate of unscheduled EDRVs. As older patients' EDRVs are associated with high risks and high impacts, this study focused on older patients and investigated the effects of the

mentioned influencing factors on their unscheduled EDRVs.

## Methods

### Research Procedures

The study was divided into two stages. The first stage entailed data selection and preprocessing. The second stage entailed data analysis. Machine learning techniques are unlikely to be restricted by statistical analysis assumptions or affected by collinear interactions between independent variables, and they demonstrate superior fault tolerance and learning capability. This study focuses on investigating the factors influencing the classification of 72-hour unscheduled EDRVs. The decision tree technique, one of machine learning classification techniques, is easier to interpret by a nonstatistician and is intuitive to follow compared with other methods (eg, random forest and support vector machine) [35,36]. In addition, decision trees have been widely used in various clinical studies for classification and prediction [37-41], and the analyzed results can be easily applied to clinical practice. Therefore, we used the decision tree as the major analysis method in this study. We used Weka (University of Waikato), one of the most popular machine learning tools, to perform in-depth data analysis for verification.

### Data Selection

We used the NHIRD as the data source and selected records of ED visits by patients aged 65 years or older from 1996 to 2010 and had older adult visits of 162,264 records out of 1,425,335 total ED visits. We then excluded 190 records of deaths and 26,912 records hospitalized within 72 hours after the ED visit. In 2010, Taiwan's Ministry of Health and Welfare amended the emergency TC from four to five classes. To prevent data inconsistency, 21,318 records following the new emergency triage reclassification were excluded from the scope of this research. Meanwhile, the Ministry of Health and Welfare that launched improvements in medical technologies in 2005 might significantly influence the number of unscheduled EDRVs; therefore, 44,114 records from 1996 to 2004 were removed. Finally, 20,478 records with incomplete or illogical values were excluded to ensure the accuracy and consistency of the analyzed data. The final sample size was 49,252 records, including 3510 unscheduled EDRV records within 72 hours.

### Variables

To develop a prediction model for older patients' unscheduled EDRVs, we applied the presence or absence of a 72-hour unscheduled EDRV as the dependent variable. Patient-, disease-, hospital-, and physician-related characteristics were applied as independent variables. Patient-related characteristics included sex, age, economic status (ES), major disease or injury, and chronic illness (eg, hypertension, diabetes, heart disease, bowel dysfunction, cerebrovascular disease, chronic kidney inflammation, vestibular disease, mental illness, arthritis, and cancer drug treatment and monitoring). Disease-related characteristics included TC, diagnostic categories (DC), radiography examination (x-ray test, specific angiography, and ultrasound scan), surgery disposition, disease severity (DS), and length of stay in the ED (LOSED). Hospital-related

characteristics comprised the level of hospital and level of urbanization (LU). Physician-related characteristics included gender, years of practice, and specialty.

Among the aforementioned independent variables, only age was a continuous variable; all other variables were categorical variables with a nominal or ordinal scale. Moreover, the variables of chronic illness and radiography examination comprised several subvariables. Detailed information related to the included variables is presented in [Multimedia Appendix 1](#).

### Analyzed Method

We applied the C4.5 technique (ie, J48 in Weka) to create a decision tree for the classifications. Decision trees use a simple tree structure to represent a set of IF-THEN rules between independent and dependent variables. The tree structure consists of multiple internal and leaf nodes. In a decision tree, each internal node represents a single independent variable, each branch of a node represents one possible value or a set of possible values of the independent variable, and each leaf node represents a class label.

A 10-fold cross-validation method was used to randomly partition the data set into 10 subsets. The validation was repeated 10 times. A confusion matrix was established to evaluate the performance of the classification model. Subsequently, we calculated the average accuracy rate of the classification results for the 10 testing sets. The sensitivity and specificity were also examined. Sensitivity refers to the ability of the prediction model to accurately predict the EDRVs among the sampled population, whereas specificity refers to the ability of the prediction model to accurately predict the samples with no return to the ED; accuracy refers to the accuracy of the prediction model regardless of return or nonreturn to the ED.

The final sample size was 49,252 records, including 3510 unscheduled EDRV records within 72 hours. However, the number of unscheduled EDRVs within 72 hours indicated only 7.13% (3510/49,252) of the emergency visits (not unscheduled EDRVs). This raises a class imbalance problem, which may lead the rare class (unscheduled EDRVs) to be ignored in the prediction model. To overcome this problem, we maintained an approximately 1:1 ratio of unscheduled EDRVs and emergency visits randomly selected from the emergency visit samples (3659/45,742, 7.99%). Then, we combined the total samples of the unscheduled EDRVs and emergency visits into a single test data set. This study increases in proportion to the sample sizes of unscheduled EDRVs and emergency visits by older adult patients for test data sets by setting the attribute of supervised resample (biasToUniform=200) in the Weka software. After such a resampling procedure, the average number of unscheduled EDRVs and emergency visits by older patients were 7231 and 7153, respectively. We obtained 30 test data sets after 30 repeated resampling and mixed procedures, and the test data sets were used for further decision tree analysis through tenfold cross-validation. In this study, the decision tree achieved an average sensitivity of 76.65% for accurately predicting the unscheduled EDRVs, an average specificity of 76.95% for accurately predicting nonreturn to the ED, and an average overall prediction accuracy of 76.80%.

### Ethics Statement

This study was approved by the institutional review board No. SE20209B of Taichung Veterans General Hospital. As the NHIRD data set comprises deidentified secondary data for research purposes, written consent from the study participants was not obtained, and the institutional review board of Taichung Veterans General Hospital issued a formal written waiver of the need for consent.

### Results

#### Decision Tree Analysis

According to the results of gain ratio of the decision tree using C4.5 implemented by Weka J48, the decision tree showed that ES, cancer drug treatment and monitoring, LOSED, cerebrovascular disease, DC, physician year of practice, patient age, LU, x-ray, DS, TC, and hospital level are critical variables for data classification and prediction. The top three influencing variables, in descending order, were ES, chronic illness-cancer drug treatment and monitoring (CICDTM), and LOSED. The 72-hour unscheduled EDRVs by older ED patients was negatively correlated with patients' ES, positively correlated with their CICDTM, and negatively correlated with their

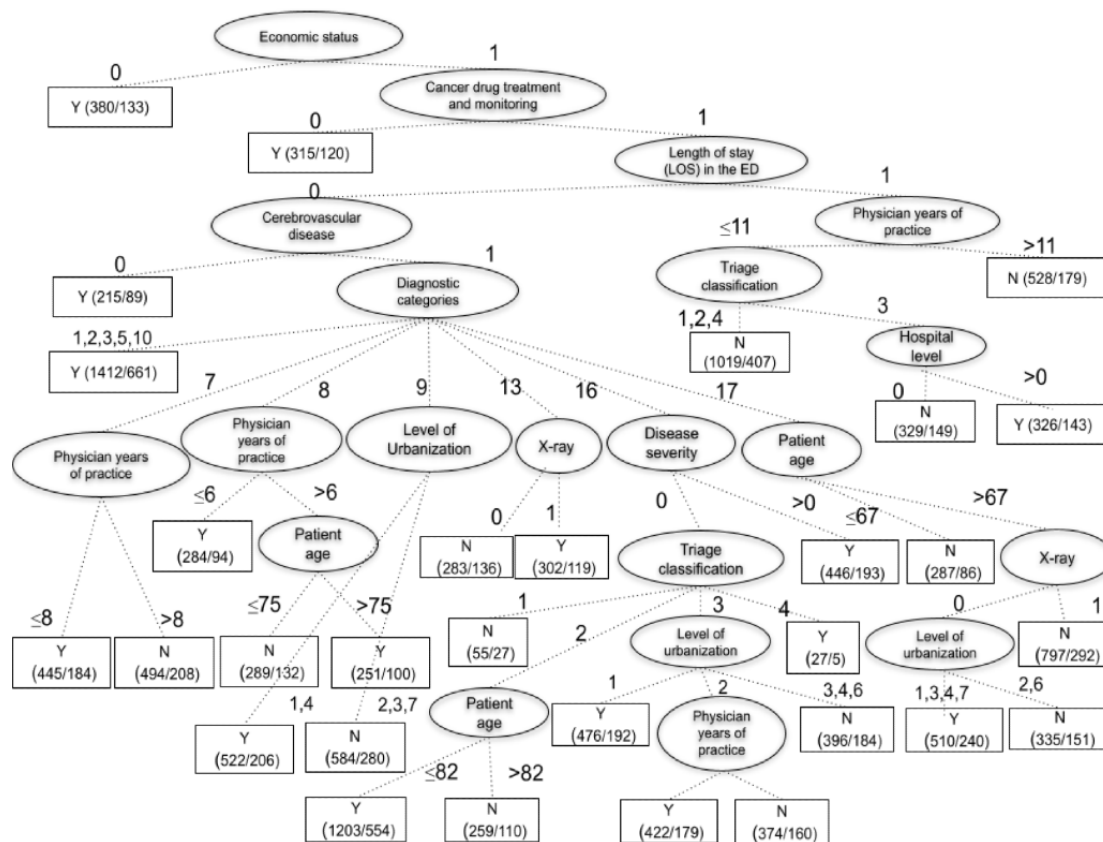
LOSED. This demonstrated that patients from low-income households or those with CICDTM are at a higher risk of unscheduled EDRVs within 72 hours. The likelihood of EDRVs decreased exponentially if older patients had an overnight stay or longer LOSED at their first visit.

#### Decision Criteria for Predicting Older Patients' Unscheduled EDRVs

The decision tree generated 11 prediction patterns (rules) for unscheduled EDRVs in older patients, which are presented in Figure 1.

As shown in the upper section of Figure 1, the top three influencing factors are ES, chronic illness, and LOSED. Each branch of the decision tree represents a decision rule that indicates the decision path of higher possibility or risk within 72-hour unscheduled EDRVs. The front number in the rectangular box represents the EDRV and the other represents the number without EDRV. For example, the left-hand branch of node ES, called Rule 1, represents 380 older adult patients from low-income households with EDRVs and 133 patients without EDRV (Textbox 1). From Rules 3-11, the older patients were not from low-income households and had no cancer drug treatment and monitoring; therefore, these two characteristics are not repeated in the explanation within parentheses.

**Figure 1.** Decision criteria for predicting older patients' unscheduled emergency department return visits. ED: emergency department; LOS: length of stay.





**Textbox 1.** Decision criteria for predicting older patients' unscheduled emergency department return visits.

#### Decision Criteria for Predicting Older Patients' Unscheduled Emergency Department Return Visits

- Rule 1: economic status (ES)=0 (older patients from low-income households)
- Rule 2: ES=1 and chronic illness-cancer drug treatment and monitoring (CICDTM)=0 (older patients from non-low-income households with cancer drug treatment and monitoring)
- Rule 3: ES=1, CICDTM=1, length of stay in the emergency department (LOSED)=0, and chronic illness-cerebrovascular disease (CICD)=0 (older patients stay in the emergency department (ED) for less than 1 d, and with cerebrovascular disease)
- Rule 4: ES=1, CICDTM=1, LOSED=0, CICD=1, and diagnostic categories (DC)=1, 2, 3, 5, and 10 (older patients stay in ED less than 1 day, with no cerebrovascular disease but infectious diseases and parasitic diseases, tumor, endocrine and immune diseases, mental illness, and genito-urinary system diseases)
- Rule 5: ES=1, CICDTM=1, LOSED=0, CICD=1, DC=7, and physician year of practice (PYP)≤8 (older patients stay in the ED for less than 1 day, with no cerebrovascular disease but circulatory system diseases, and treated by physician with 8 or fewer years of practice)
- Rule 6: ES=1, CICDTM=1, LOSED=0, CICD=1, DC=8, and (PYP≤6 or PYP>6 and patient age [PA]>75) (older patients stay in the ED less than 1 day, with no cerebrovascular disease but respiratory diseases, and treated by a physician with 6 or less years; or all the conditions are same but treated by a physician with more than 6 years of practice, and PA is more than 75)
- Rule 7: ES=1, CICDTM=1, LOSED=0, CICD=1, DC=9, and level of urbanization (LU)=1, 4 (older patients stay in the ED for less than 1 day, with no cerebrovascular disease but digestive diseases, and live in a high LU or general town)
- Rule 8: ES=1, CICDTM=1, LOSED=0, CICD=1, DC=13, and x-ray=1 (older patients stay in the ED for less than 1 day, with no cerebrovascular disease but musculoskeletal system diseases, had no x-ray)
- Rule 9: ES=1, CICDTM=1, LOSED=0, CICD=1, DC=16, and (disease severity [DS]=0 and triage classification [TC]=3; LU=1 or LU=2 and PYP>8; or TC=2, PA≤82 or TC=4; or DS=1, 2, 3, 4; older patients stay in ED less than 1 day, with no cerebrovascular disease but have signs, symptoms, and diagnosis less clear, and DS, TC of 3, and live in high LU, and treated by physician with more than 8 years of practice and live in Remote town or all conditions are the same with TC=2 and aged 82 or less, or all conditions are the same as TC=4, or all conditions are the same with DS)
- Rule 10: ES=1, CICDTM=1, LOSED=0, CICD=1, DC=17, PA>67, x-ray=0, and LU=1, 3, 4, 7 (older patients stay in the ED for less than 1 day, with no cerebrovascular disease but injury and poisoning, had no x-ray, and lived in a high LU or an emerging town, general town, or remote town)
- Rule 11: ES=1, CICDTM=1, LOSED=1, PYP≤11, TC=3, and hospital level=1, 2 (older patients stay in ED less than 1 day, treated by physician with 11 or fewer years, with TC and visit Regional Hospital or District Hospital)

## Discussion

### Principal Findings

Among the 28 investigated variables, as shown in [Multimedia Appendix 1](#), in patient-related, disease-related, hospital-related, and physician-related characteristics, only 12 variables were identified as critical criteria in the decision tree prediction model. This study found that ES, age, CICDTM, chronic illness-cerebrovascular disease in patient characteristics and TC, DC, x-ray, DS, and LOSED in disease characteristics were the key factors influencing unscheduled EDRVs. In addition, hospital level and LU in hospital characteristics and years of practice in physician characteristics were key predictive factors of unscheduled EDRVs. The results showed that only a portion of the investigated variables in patient-related, disease-related, hospital-related, and physician-related characteristics were key factors influencing older patients' unscheduled EDRVs. Through the decision tree analysis, this study found 11 useful decision rules for predicting unscheduled EDRVs within 72 hours by the identified factors. The obtained decision rules can be easily applied by physicians and nurses in the ED to evaluate the risk or possibility of unscheduled EDRVs within 72 hours for older patients. ES, CICDTM, and LOSED are highlighted as the top three influencing variables of the prediction model of older patients with unscheduled EDRVs within 72 hours.

In this study, we confirmed that older ED patients with less economic privilege were more likely to return to the ED than those in the opposite group. Furthermore, these findings are consistent with those of a previous study [42]. Possible reasons include that older adults with lower ES have fewer resources to attend to their health and basic preventive health care. They often defer medical treatment and, thus, have a high demand for emergency medical resources. Therefore, the pathological conditions that develop among underprivileged older patients are also more complex than those of their privileged counterparts, which increases the difficulty of treatment and leads to a high rate of EDRVs.

In addition, older patients with chronic symptoms that remain prevalent or frequently relapse may prefer to return to the ED for rapid and convenient treatment, rather than visit an outpatient department. The rates of 72-hour unscheduled EDRVs were higher for older patients who required cancer drug treatment and monitoring or were diagnosed with chronic cerebrovascular diseases. These results confirm the findings of Liaw et al [26], McCusker et al [31], and Wu et al [27]. A possible reason may be that patients with cancer or cerebrovascular disease have a greater need for emergency treatment and hospitalization for pain. In addition, patients diagnosed with chronic cerebrovascular diseases are at a high risk of a second stroke. Active interventions to improve the effectiveness and efficiency of delivering medical education on pain control and stroke

prevention can help patients and their family members manage and alleviate this risk and further reduce their EDRVs.

We also found that older patients with shorter LOSED had higher rates of EDRVs than those who stayed in the ED overnight or longer. Patients older than 65 years are known to have a lower metabolic rate [43], and a longer length of stay (LOS) can enable medical staff to conduct more detailed observations of the effectiveness of the provided treatments. It can also further verify the patient's reaction to the prescribed medicine and modify the types of medicines needed. However, a lack of sufficient ED resources may result in problems such as ED overcrowding, inadequate number of hospital beds, or poor evaluation practices. It may also prevent hospitals from increasing the LOS for older patients in the ED.

In this study, some specific DC (infectious diseases and parasitic diseases, tumors, endocrine and immune diseases, mental illness, circulatory system diseases, respiratory diseases, digestive diseases, genito-urinary system diseases, musculoskeletal system diseases, signs, symptoms and diagnosis less clear, and injury and poisoning) were found to be highly related to unscheduled EDRVs under certain circumstances (patients from non-low-income households and LOS less than 1 day and patients without CICDTM and cerebrovascular disease). The results showed that only a portion of the DC (disease types) [12,25,27] were identified as factors influencing older adult patients' unscheduled EDRVs; however, some DC were not considered as significant factors in this study.

Older patients classified as class 3 or higher on the TC level had a higher likelihood of 72-hour unscheduled EDRVs if they were treated by physicians with less than 11 years of practice, a result partially consistent with a previous study [15,34]. The conventional emergency medicine curriculum does not include geriatrics; curriculum materials focus on the care of adults aged less than 65 years or children. Meanwhile, the challenges in providing medical services to older patients are highly specific and complex. Physicians who have more years of practice may overcome the problems engendered by the lack of formal geriatric training in emergency medicine, whereas lack of experience in the ED increases the difficulty of accurately diagnosing symptoms in older patients.

Moreover, physicians often underestimate the TC of frail older patients because of the absence of prominent symptoms. This increases the risk of delayed treatment and the likelihood of unscheduled EDRVs. Platts-Mills et al [44] have also asserted that the TC is designed specifically for the general adult population and does not have adequate specificity for the older adult population or reflects the severity of their pathological conditions.

### Limitations

As mentioned above, decision trees have been widely used in various clinical studies, and the analyzed results can be easily applied to clinical practice. Our prediction model developed by the decision tree achieved an acceptable rate for sensitivity, specificity, and overall prediction accuracy. Future researchers can use the results of this study as a reference and apply other methods such as random forest or support vector machine to

generate a prediction model and obtain higher accuracy. As data were collected in Taiwan, caution is needed when generalizing the results of this study. Meanwhile, because of the limited content of NHIRD, important variables other than claim-based data cannot be obtained. Furthermore, the insured area and degree of urbanization may be different from the actual area of residence. In addition, the 20,478 records with incomplete or illogical values excluded in this study can cause selection bias. Future studies can use advanced interpolation techniques to explore the characteristics of deleted records and extend the results of this study.

### Conclusions

Compared with previous studies [11,12,14-17], past research samples from a single institution are impossible to discuss patients returning to the ED from different hospitals. This study used the NHIRD to obtain a more comprehensive picture of older adult patients' EDRVs within 72 hours. This study identified 12 key predicting factors out of the 28 investigated factors and provided 11 decision rules for early detection and possible prevention of unscheduled EDRVs within 72 hours. For example, more attention should be paid to patients aged 65 years and older featured with low-income households or those with CICDTM, and have a reasonably longer LOSED at their first visit.

Medical and health care is an important segment of Taiwan's *New Southbound Policy* [45] to engage in partnership with 18 countries of the Southeast Asia-Pacific family. The findings of this study can serve as a reference for those countries in the planning and promotion of medical services and age-friendly policies. In countries with rapidly aging populations, the demographics of EDs have shifted substantively toward older patients, and most of their EDs are not fully prepared for the challenge of caring for the aging population. Although the majority of emergency medical curriculum materials focus on the care of children and adults under 65, some countries have begun to integrate geriatrics and emergency medicine training into a defined and validated geriatric emergency medicine curriculum [46]. The Taiwan Society of Emergency Medicine has included geriatrics in its emergency medicine curriculum, with training and emphasis on acute problems in older patients, including medical and intestate ethics and certification of age-friendly health care institutions.

For physicians, our prediction model can be used as a reference to improve medical management and clinical services to reduce older patients' 72-hour unscheduled EDRVs. Policymakers can use the results of this study to generate incentives for medical institutions to provide appropriate education to older patients and their family members before discharge. Medical institutions may create mechanisms that incorporate our prediction model and develop a decision-making support system for emergency return visits, similar to other clinical decision support systems [47,48]. Such a system can be used in triage procedures for early detection and prevention of unscheduled EDRVs, and it will help health care providers to rapidly identify older patients who are likely to make unscheduled EDRVs and to further reduce the rate of such visits.

In summary, this study is based on large population-based retrospective data from the NHIRD and uses machine learning techniques, which demonstrate superior fault tolerance and learning capability from massive data. The decision tree machine learning technique was further used for data analysis and validation because of its simplicity, interpretability, and applicability of the results compared with other machine learning techniques. Through the decision tree technique, decision rules with important factors influencing the unscheduled EDRV

prediction model from the considerations of patient, disease, hospital, and physician characteristics were obtained. The decision rules may serve as a reference for the early detection of unscheduled EDRV in older adults. Further studies can be based on the findings of this study and integrate hospitals' information systems or electronic medical records to generate appropriate rules for unscheduled EDRVs for older adults in different hospitals.

---

## Acknowledgments

The authors would like to thank Dr Ya Han Hu for his elaboration of the research conception and design stage. In addition, this research was supported in part by the Ministry of Science and Technology of ROC, Taiwan, under contract number MOST109-2410-H-041-001.

---

## Authors' Contributions

In this study, each author has participated sufficiently in the work to take public responsibility for appropriate portions of the content; study conception and design was conducted by RFC and ICC; data acquisition was done by YYL and KCC; analysis and interpretation of data was performed by RFC, CHT, and YYL; ICC, KCC, CHT, and RFC drafted the manuscript; and ICC, KCC, and RFC were the referees for the report ([Multimedia Appendix 2](#)).

---

## Conflicts of Interest

None declared.

---

### Multimedia Appendix 1

Patient characteristics and variables.

[\[DOCX File , 18 KB - medinform\\_v9i7e22491\\_app1.docx \]](#)

---

### Multimedia Appendix 2

Vitae.

[\[DOCX File , 15 KB - medinform\\_v9i7e22491\\_app2.docx \]](#)

---

## References

1. Marengoni A, Angleman S, Melis R, Mangialasche F, Karp A, Garmen A, et al. Aging with multimorbidity: a systematic review of the literature. *Ageing Res Rev* 2011 Sep;10(4):430-439. [doi: [10.1016/j.arr.2011.03.003](#)] [Medline: [21402176](#)]
2. Considine J, Mitchell B, Stergiou HE. Frequency and nature of reported incidents during Emergency Department care. *Emerg Med J* 2011 May;28(5):416-421. [doi: [10.1136/emj.2010.093054](#)] [Medline: [20660904](#)]
3. Downing A, Wilson R. Older people's use of Accident and Emergency services. *Age Ageing* 2005 Jan;34(1):24-30. [doi: [10.1093/ageing/afh214](#)] [Medline: [15496462](#)]
4. Hedges JR, Singal BM, Rousseau EW, Sanders AB, Bernstein E, McNamara RM, et al. Geriatric patient emergency visits. Part II: Perceptions of visits by geriatric and younger patients. *Ann Emerg Med* 1992 Jul;21(7):808-813. [doi: [10.1016/s0196-0644\(05\)81026-1](#)] [Medline: [1610037](#)]
5. Salvi F, Morichi V, Grilli A, Giorgi R, De Tommaso G, Dessì-Fulgheri P. The elderly in the emergency department: a critical review of problems and solutions. *Intern Emerg Med* 2007 Dec;2(4):292-301. [doi: [10.1007/s11739-007-0081-3](#)] [Medline: [18043874](#)]
6. Singal BM, Hedges JR, Rousseau EW, Sanders AB, Bernstein E, McNamara RM, et al. Geriatric patient emergency visits. Part I: Comparison of visits by geriatric and younger patients. *Ann Emerg Med* 1992 Jul;21(7):802-807. [doi: [10.1016/s0196-0644\(05\)81025-x](#)] [Medline: [1610036](#)]
7. Strange GR, Chen EH. Use of emergency departments by elder patients: a five-year follow-up study. *Acad Emerg Med* 1998 Dec;5(12):1157-1162 [FREE Full text] [doi: [10.1111/j.1553-2712.1998.tb02688.x](#)] [Medline: [9864128](#)]
8. Yim VW, Graham CA, Rainer TH. A comparison of emergency department utilization by elderly and younger adult patients presenting to three hospitals in Hong Kong. *Int J Emerg Med* 2009 Apr;2(1):19-24 [FREE Full text] [doi: [10.1007/s12245-009-0087-x](#)] [Medline: [19390913](#)]
9. National Hospital Ambulatory Medical Care Survey: 2011 Emergency Department Summary Tables. Centers for Disease Control and Prevention. URL: [http://www.cdc.gov/nchs/data/ahcd/nhamcs\\_emergency/2011\\_ed\\_web\\_tables.pdf](http://www.cdc.gov/nchs/data/ahcd/nhamcs_emergency/2011_ed_web_tables.pdf) [accessed 2021-07-05]

10. 2014 National Health Insurance Annual Statistical Report - The Rate of Emergency Visits by Sex and Age (Table 23). URL: <https://www.mohw.gov.tw/dl-18857-e5f8512b-8ed3-44eb-92a7-a88190ec808b.html> [accessed 2021-07-05]
11. Nuñez S, Hexdall A, Aguirre-Jaime A. Unscheduled returns to the emergency department: an outcome of medical errors? *Qual Saf Health Care* 2006 Apr;15(2):102-108 [FREE Full text] [doi: [10.1136/qshc.2005.016618](https://doi.org/10.1136/qshc.2005.016618)] [Medline: [16585109](https://pubmed.ncbi.nlm.nih.gov/16585109/)]
12. Martin-Gill C, Reiser RC. Risk factors for 72-hour admission to the ED. *Am J Emerg Med* 2004 Oct;22(6):448-453. [doi: [10.1016/j.ajem.2004.07.023](https://doi.org/10.1016/j.ajem.2004.07.023)] [Medline: [15520938](https://pubmed.ncbi.nlm.nih.gov/15520938/)]
13. Trivedy CR, Cooke MW. Unscheduled return visits (URV) in adults to the emergency department (ED): a rapid evidence assessment policy review. *Emerg Med J* 2015 Apr;32(4):324-329. [doi: [10.1136/emered-2013-202719](https://doi.org/10.1136/emered-2013-202719)] [Medline: [24165201](https://pubmed.ncbi.nlm.nih.gov/24165201/)]
14. Keith KD, Bocka JJ, Kobernick MS, Krome RL, Ross MA. Emergency department revisits. *Ann Emerg Med* 1989 Sep;18(9):964-968. [doi: [10.1016/s0196-0644\(89\)80461-5](https://doi.org/10.1016/s0196-0644(89)80461-5)]
15. Kuan WS, Mahadevan M. Emergency unscheduled returns: can we do better? *Singapore Med J* 2009 Nov;50(11):1068-1071 [FREE Full text] [Medline: [19960161](https://pubmed.ncbi.nlm.nih.gov/19960161/)]
16. White D, Kaplan L, Eddy L. Characteristics of patients who return to the emergency department within 72 hours in one community hospital. *Adv Emerg Nurs J* 2011;33(4):344-353. [doi: [10.1097/TME.0b013e31823438d6](https://doi.org/10.1097/TME.0b013e31823438d6)] [Medline: [22075685](https://pubmed.ncbi.nlm.nih.gov/22075685/)]
17. Minnee D, Wilkinson J. Return visits to the emergency department and related hospital admissions by people aged 65 and over. *N Z Med J* 2011 Mar 25;124(1331):67-74. [Medline: [21725415](https://pubmed.ncbi.nlm.nih.gov/21725415/)]
18. Wolff AM, Bourke J. Detecting and reducing adverse events in an Australian rural base hospital emergency department using medical record screening and review. *Emerg Med J* 2002 Jan;19(1):35-40 [FREE Full text] [doi: [10.1136/emj.19.1.35](https://doi.org/10.1136/emj.19.1.35)] [Medline: [11777869](https://pubmed.ncbi.nlm.nih.gov/11777869/)]
19. Said S. Top 10 best health care systems in the world. The Richest. 2013. URL: <http://www.therichest.com/expensive-lifestyle/lifestyle/top-10-best-health-care-systems-in-the-world/> [accessed 2021-07-05]
20. Lee ML. From recipient to donor: how Taiwan transformed its healthcare system. *Jpn Med Assoc J*. 2012. URL: [https://www.med.or.jp/english/journal/pdf/2012\\_01/023\\_025.pdf](https://www.med.or.jp/english/journal/pdf/2012_01/023_025.pdf) [accessed 2021-07-05]
21. Health care index by country 2020. Numbeo. 2020. URL: [https://www.numbeo.com/health-care/rankings\\_by\\_country.jsp?title=2020](https://www.numbeo.com/health-care/rankings_by_country.jsp?title=2020) [accessed 2021-07-05]
22. National Health Insurance in Taiwan - 2013-2014 Annual Report. URL: [https://www.nhi.gov.tw/Resource/webdata/28139\\_1\\_National%20Health%20Insurance%20in%20Taiwan%202013-2014%20\(bilingual\).pdf](https://www.nhi.gov.tw/Resource/webdata/28139_1_National%20Health%20Insurance%20in%20Taiwan%202013-2014%20(bilingual).pdf) [accessed 2021-07-05]
23. The 2016 Waseda-IAC international e-government ranking. TOSHIO OBI Laboratory. URL: <http://www.e-gov.waseda.ac.jp/ranking2016.htm> [accessed 2021-07-05]
24. Khan NU, Razzak JA, Saleem AF, Khan UR, Mir MU, Aashiq B. Unplanned return visit to emergency department: a descriptive study from a tertiary care hospital in a low-income country. *Eur J Emerg Med* 2011 Oct;18(5):276-278. [doi: [10.1097/MEJ.0b013e3283449100](https://doi.org/10.1097/MEJ.0b013e3283449100)] [Medline: [21326103](https://pubmed.ncbi.nlm.nih.gov/21326103/)]
25. LaMantia MA, Platts-Mills TF, Biese K, Khandelwal C, Forbach C, Cairns CB, et al. Predicting hospital admission and returns to the emergency department for elderly patients. *Acad Emerg Med* 2010 Mar;17(3):252-259 [FREE Full text] [doi: [10.1111/j.1553-2712.2009.00675.x](https://doi.org/10.1111/j.1553-2712.2009.00675.x)] [Medline: [20370757](https://pubmed.ncbi.nlm.nih.gov/20370757/)]
26. Liaw SJ, Bullard MJ, Hu PM, Chen JC, Liao HC. Rates and causes of emergency department revisits within 72 hours. *J Formos Med Assoc* 1999 Jun;98(6):422-425. [Medline: [10443066](https://pubmed.ncbi.nlm.nih.gov/10443066/)]
27. Wu CL, Wang FT, Chiang YC, Chiu YF, Lin TG, Fu LF, et al. Unplanned emergency department revisits within 72 hours to a secondary teaching referral hospital in Taiwan. *J Emerg Med* 2010 May;38(4):512-517. [doi: [10.1016/j.jemermed.2008.03.039](https://doi.org/10.1016/j.jemermed.2008.03.039)] [Medline: [18947963](https://pubmed.ncbi.nlm.nih.gov/18947963/)]
28. Hu KW, Lu YH, Lin HJ, Guo HR, Foo NP. Unscheduled return visits with and without admission post emergency department discharge. *J Emerg Med* 2012 Dec;43(6):1110-1118. [doi: [10.1016/j.jemermed.2012.01.062](https://doi.org/10.1016/j.jemermed.2012.01.062)] [Medline: [22674038](https://pubmed.ncbi.nlm.nih.gov/22674038/)]
29. Friedmann PD, Jin L, Karrison TG, Hayley DC, Mulliken R, Walter J, et al. Early revisit, hospitalization, or death among older persons discharged from the ED. *Am J Emerg Med* 2001 Mar;19(2):125-129. [doi: [10.1053/ajem.2001.21321](https://doi.org/10.1053/ajem.2001.21321)] [Medline: [11239256](https://pubmed.ncbi.nlm.nih.gov/11239256/)]
30. Wang HY, Chew G, Kung CT, Chung KJ, Lee WH. The use of Charlson comorbidity index for patients revisiting the emergency department within 72 hours. *Chang Gung Med J* 2007;30(5):437-444 [FREE Full text] [Medline: [18062175](https://pubmed.ncbi.nlm.nih.gov/18062175/)]
31. McCusker J, Cardin S, Bellavance F, Belzile E. Return to the emergency department among elders: patterns and predictors. *Acad Emerg Med* 2000 Mar;7(3):249-259 [FREE Full text] [doi: [10.1111/j.1553-2712.2000.tb01070.x](https://doi.org/10.1111/j.1553-2712.2000.tb01070.x)] [Medline: [10730832](https://pubmed.ncbi.nlm.nih.gov/10730832/)]
32. Meldon SW, Mion LC, Palmer RM, Drew BL, Connor JT, Lewicki LJ, et al. A brief risk-stratification tool to predict repeat emergency department visits and hospitalizations in older patients discharged from the emergency department. *Acad Emerg Med* 2003 Mar;10(3):224-232 [FREE Full text] [doi: [10.1111/j.1553-2712.2003.tb01996.x](https://doi.org/10.1111/j.1553-2712.2003.tb01996.x)] [Medline: [12615588](https://pubmed.ncbi.nlm.nih.gov/12615588/)]
33. Verelst S, Pierloot S, Desruelles D, Gillet J, Bergs J. Short-term unscheduled return visits of adult patients to the emergency department. *J Emerg Med* 2014 Aug;47(2):131-139. [doi: [10.1016/j.jemermed.2014.01.016](https://doi.org/10.1016/j.jemermed.2014.01.016)] [Medline: [24642045](https://pubmed.ncbi.nlm.nih.gov/24642045/)]
34. Rusnak RA, Stair TO, Hansen K, Fastow JS. Litigation against the emergency physician: common features in cases of missed myocardial infarction. *Ann Emerg Med* 1989 Oct;18(10):1029-1034. [doi: [10.1016/s0196-0644\(89\)80924-2](https://doi.org/10.1016/s0196-0644(89)80924-2)] [Medline: [2802275](https://pubmed.ncbi.nlm.nih.gov/2802275/)]



35. Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomed Signal Process Cont* 2019 Jul;52:456-462. [doi: [10.1016/j.bspc.2017.01.012](https://doi.org/10.1016/j.bspc.2017.01.012)]
36. Wu CT, Lo CL, Tung CH, Cheng HL. Applying data mining techniques for predicting prognosis in patients with rheumatoid arthritis. *Healthcare (Basel)* 2020 Apr 03;8(2):85 [FREE Full text] [doi: [10.3390/healthcare8020085](https://doi.org/10.3390/healthcare8020085)] [Medline: [32260259](https://pubmed.ncbi.nlm.nih.gov/32260259/)]
37. Hammann F, Gutmann H, Vogt N, Helma C, Drewe J. Prediction of adverse drug reactions using decision tree modeling. *Clin Pharmacol Ther* 2010 Jul;88(1):52-59. [doi: [10.1038/clpt.2009.248](https://doi.org/10.1038/clpt.2009.248)] [Medline: [20220749](https://pubmed.ncbi.nlm.nih.gov/20220749/)]
38. Hess EP, Brison RJ, Perry JJ, Calder LA, Thiruganasambandamoorthy V, Agarwal D, et al. Development of a clinical prediction rule for 30-day cardiac events in emergency department patients with chest pain and possible acute coronary syndrome. *Ann Emerg Med* 2012 Feb;59(2):115-125. [doi: [10.1016/j.annemergmed.2011.07.026](https://doi.org/10.1016/j.annemergmed.2011.07.026)] [Medline: [21885156](https://pubmed.ncbi.nlm.nih.gov/21885156/)]
39. Hill JL, Campbell MK, Zou GY, Challis JR, Reid G, Chisaka H, et al. Prediction of preterm birth in symptomatic women using decision tree modeling for biomarkers. *Am J Obstet Gynecol* 2008 Apr;198(4):461-469. [doi: [10.1016/j.ajog.2008.01.007](https://doi.org/10.1016/j.ajog.2008.01.007)] [Medline: [18395044](https://pubmed.ncbi.nlm.nih.gov/18395044/)]
40. Hiramatsu N, Kurosaki M, Sakamoto N, Iwasaki M, Sakamoto M, Suzuki Y, et al. Pretreatment prediction of anemia progression by pegylated interferon alpha-2b plus ribavirin combination therapy in chronic hepatitis C infection: decision-tree analysis. *J Gastroenterol* 2011 Sep;46(9):1111-1119. [doi: [10.1007/s00535-011-0412-z](https://doi.org/10.1007/s00535-011-0412-z)] [Medline: [21681410](https://pubmed.ncbi.nlm.nih.gov/21681410/)]
41. Nishijima DK, Shahlaie K, Echeverri A, Holmes JF. A clinical decision rule to predict adult patients with traumatic intracranial haemorrhage who do not require intensive care unit admission. *Injury* 2012 Nov;43(11):1827-1832 [FREE Full text] [doi: [10.1016/j.injury.2011.07.020](https://doi.org/10.1016/j.injury.2011.07.020)] [Medline: [21839444](https://pubmed.ncbi.nlm.nih.gov/21839444/)]
42. Moore G, Gerdtz MF, Hepworth G, Manias E. Homelessness: patterns of emergency department use and risk factors for re-presentation. *Emerg Med J* 2011 May;28(5):422-427. [doi: [10.1136/emj.2009.087239](https://doi.org/10.1136/emj.2009.087239)] [Medline: [20682956](https://pubmed.ncbi.nlm.nih.gov/20682956/)]
43. Schofield WN. Predicting basal metabolic rate, new standards and review of previous work. *Hum Nutr Clin Nutr* 1985;39 Suppl 1:5-41. [Medline: [4044297](https://pubmed.ncbi.nlm.nih.gov/4044297/)]
44. Platts-Mills TF, Travers D, Biese K, McCall B, Kizer S, LaMantia M, et al. Accuracy of the Emergency Severity Index triage instrument for identifying elder emergency department patients receiving an immediate life-saving intervention. *Acad Emerg Med* 2010 Mar;17(3):238-243 [FREE Full text] [doi: [10.1111/j.1553-2712.2010.00670.x](https://doi.org/10.1111/j.1553-2712.2010.00670.x)] [Medline: [20370755](https://pubmed.ncbi.nlm.nih.gov/20370755/)]
45. New Southbound Promotion Plan. New Southbound Policy. URL: <https://newsouthboundpolicy.trade.gov.tw/PageDetail?pageID=12&nodeID=21> [accessed 2021-07-05]
46. Conroy S, Nickel CH, Jónsdóttir A, Fernandez M, Banerjee J, Mooijaart S, et al. The development of a European curriculum in geriatric emergency medicine. *Eur Geriatr Med* 2016 Jul;7(4):315-321. [doi: [10.1016/j.eurger.2016.03.011](https://doi.org/10.1016/j.eurger.2016.03.011)]
47. Hoot NR, Leblanc LJ, Jones I, Levin SR, Zhou C, Gadd CS, et al. Forecasting emergency department crowding: a prospective, real-time evaluation. *J Am Med Inform Assoc* 2009;16(3):338-345 [FREE Full text] [doi: [10.1197/jamia.M2772](https://doi.org/10.1197/jamia.M2772)] [Medline: [19261948](https://pubmed.ncbi.nlm.nih.gov/19261948/)]
48. Griffey RT, Lo HG, Burdick E, Keohane C, Bates DW. Guided medication dosing for elderly emergency patients using real-time, computerized decision support. *J Am Med Inform Assoc* 2012;19(1):86-93 [FREE Full text] [doi: [10.1136/amiajnl-2011-000124](https://doi.org/10.1136/amiajnl-2011-000124)] [Medline: [22052899](https://pubmed.ncbi.nlm.nih.gov/22052899/)]

## Abbreviations

**CICDTM:** chronic illness-cancer drug treatment and monitoring

**DC:** diagnostic categories

**DS:** disease severity

**ED:** emergency department

**EDRV:** emergency department return visit

**ES:** economic status

**LOS:** length of stay

**LOSED:** length of stay in the emergency department

**LU:** level of urbanization

**NHIRD:** National Health Insurance Research Database

**TC:** triage classification

*Edited by C Lovis; submitted 14.07.20; peer-reviewed by CH Wen, M Yiadom, J Mihanovic, T Goto; comments to author 17.11.20; revised version received 11.01.21; accepted 17.06.21; published 28.07.21.*

*Please cite as:*

*Chen RF, Cheng KC, Lin YY, Chang IC, Tsai CH*

*Predicting Unscheduled Emergency Department Return Visits Among Older Adults: Population-Based Retrospective Study*

*JMIR Med Inform 2021;9(7):e22491*

*URL: <https://medinform.jmir.org/2021/7/e22491>*

*doi: [10.2196/22491](https://doi.org/10.2196/22491)*

*PMID: [34319244](https://pubmed.ncbi.nlm.nih.gov/34319244/)*

©Rai-Fu Chen, Kuei-Chen Cheng, Yu-Yin Lin, I-Chiu Chang, Cheng-Han Tsai. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 28.07.2021. 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 Association Between Using a Mobile Version of an Electronic Health Record and the Well-Being of Nurses: Cross-sectional Survey Study

Tarja Heponiemi<sup>1</sup>, PhD; Anu-Marja Kaihlanen<sup>1</sup>, PhD; Kia Gluschkoff<sup>1</sup>, PhD; Kaija Saranto<sup>2</sup>, PhD; Sari Nissinen<sup>2</sup>, PhD; Elina Laukka<sup>1,3</sup>, MSc; Tuulikki Vehko<sup>1</sup>, PhD

<sup>1</sup>Department of Public Health and Welfare, Finnish Institute for Health and Welfare, Helsinki, Finland

<sup>2</sup>Department of Health and Social Management, University of Eastern Finland, Kuopio, Finland

<sup>3</sup>Research Unit of Nursing Science and Health Management, University of Oulu, Oulu, Finland

**Corresponding Author:**

Tarja Heponiemi, PhD

Department of Public Health and Welfare

Finnish Institute for Health and Welfare

PO Box 30

Helsinki, 00271

Finland

Phone: 358 295247434

Email: [tarja.heponiemi@thl.fi](mailto:tarja.heponiemi@thl.fi)

## Abstract

**Background:** Mobile devices such as tablets and smartphones are increasingly being used in health care in many developed countries. Nurses form the largest group in health care that uses electronic health records (EHRs) and their mobile versions. Mobile devices are suggested to promote nurses' workflow, constant updating of patient information, and improve the communication within the health care team. However, little is known about their effect on nurses' well-being.

**Objective:** This study aimed to examine the association between using a mobile version of the EHR and nurses' perceived time pressure, stress related to information systems, and self-rated stress. Moreover, we examined whether mobile device use modifies the associations of EHR usability (ease of use and technical quality), experience in using EHRs, and number of systems in daily use with these well-being indicators.

**Methods:** This was a cross-sectional population-based survey study among 3610 Finnish registered nurses gathered in 2020. The aforesaid associations were examined using analyses of covariance and logistic regression adjusted for age, gender, and employment sector (hospital, primary care, social service, and other).

**Results:** Nurses who used the mobile version of their EHR had higher levels of time pressure ( $F_{1,3537}=14.96$ ,  $P<.001$ ) and stress related to information systems ( $F_{1,3537}=6.11$ ,  $P=.01$ ), compared with those who did not use mobile versions. Moreover, the interactions of mobile device use with experience in using EHRs ( $F_{1,3581}=14.93$ ,  $P<.001$ ), ease of use ( $F_{1,3577}=10.16$ ,  $P=.001$ ), and technical quality ( $F_{1,3577}=6.45$ ,  $P=.01$ ) were significant for stress related to information systems. Inexperience in using EHRs, low levels of ease of use, and technical quality were associated with higher stress related to information systems and this association was more pronounced among those who used mobile devices. That is, the highest levels of stress related to information systems were perceived among those who used mobile devices as well as among inexperienced EHR users or those who perceived usability problems in their EHRs.

**Conclusions:** According to our results, it seems that at present mobile device use is not beneficial for the nurses' well-being. In addition, mobile device use seems to intensify the negative effects of usability issues related to EHRs. In particular, inexperienced users of EHRs seem to be at a disadvantage when using mobile devices. Thus, we suggest that EHRs and their mobile versions should be improved such that they would be easier to use and would better support the nurses' workflow (eg, improvements to problems related to small display, user interface, and data entry). Moreover, additional training on EHRs, their mobile versions, and workflow related to these should be provided to nurses.

(JMIR Med Inform 2021;9(7):e28729) doi:[10.2196/28729](https://doi.org/10.2196/28729)

**KEYWORDS**

stress related to information systems; time pressure; usability; stress; health and social care

## Introduction

Viewing electronic health records (EHRs) on mobile devices such as tablets and smartphones has increased lately and is becoming more common in the health care sector in many developed countries. Nurses form the largest group in health care that uses EHRs and their mobile versions [1]. A previous study reported that use of mobile versions of EHR is common in hospitals, especially during the handover period and ward hours [1]. Nurses have been found to use the mobile version of EHR mainly for viewing inpatient lists, nursing notes, alerts, and patients' clinical data with high frequency [2].

Mobile device use while working presents many benefits for nurses, as it helps in their workflow and allows real-time updating of patient information [3]. Use of mobile apps in home care has been associated with the promotion of nurse-patient relationship (as it helps both nurses and patients to express their feelings) and reduction in workload and stress, but they are also related to disturbance in personal life among nurses [4]. However, mobile devices are reported to be useful in improving the quality of care and decreasing stress among nurses [5].

In addition, mobile devices are suggested to benefit health care professionals by increasing convenience, accuracy, efficiency, and productivity as well as improving clinical decision making [6]. A previous review noted that mobile devices improved patient documentation through more complete recording, fewer documentation errors, and increased efficiency [7]. Moreover, mobile devices saved time, gave earlier access to new information, and enhanced work patterns [7]. Mobile devices are also regularly used to ensure effective team work within a health care team [8].

However, some downsides also exist related to the use of mobile devices in health care. For example, the problems in overall architecture and user interface may lead to an increased number of input errors, loss of data, or decreased efficiency [9]. During electronic data collection, mobile devices have been found to increase the time of the data entry by twofold and increase the risk of typing errors and missing data compared with electronic data collection on laptops [10]. Moreover, small screens may make information retrieval tasks difficult and increase incorrect choices and scrolling activities [11].

The effect of using mobile versions of EHRs on the well-being of nurses remains unclear. Viewing and updating EHRs on mobile devices might help nurses in their daily work by giving them a chance to use EHR while simultaneously caring for patients. Thus, mobile use might help reduce workload and consequently nurses' work-related stress as well. By contrast, mobile device use may also elicit stress and frustration due to the aforementioned problems (ie, difficult user interface, small screens, and challenges in data entry [9-11]). Previous studies have shown that the usability of EHRs is associated with nurses' well-being [12,13], and thus it could be assumed that mobile use might also have an effect.

Previous studies have shown that usability problems of EHRs, need to use many different systems, and inexperience in using EHRs have been associated with high levels of stress, time pressure, and psychological distress among health care employees [12-16]. However, it is not known whether using EHR with mobile devices has a moderating effect on these associations.

In the light of these previous findings, this study aimed to examine the association between using a mobile version of EHR and perceived time pressure, stress related to information systems, and self-rated stress. Moreover, we examined whether mobile device use modifies the associations of EHR usability (ease of use and technical quality), experience in using EHRs, and number of systems in daily use with the well-being indicators (time pressure, stress related to information systems, and self-rated stress).

## Methods

### Sample

Data were collected in the spring of 2020 via the online survey Webropol [17]. The link to the survey was sent via email by the Finnish Nurses Association, The Union of Health and Social Care Professionals in Finland (Tehy), and the National Professional Association for the Interests of Experts and Managers in Health Care (TAJA) to their members under 65 years of age, which included 58,276/80,622 nurses, midwives, and public health nurses (representing 72.29% of the eligible population) [18]. One reminder was sent to nonresponders to maximize survey responses. A more detailed description of the data collection has been presented previously [18]. Altogether, 10,094 registered nurses opened the link and 3912 responded. Of those who responded, 302 answered that they did not perceive themselves as fit to answer the questionnaire because they had not practiced as registered nurses for a long time. Thus, the final sample included 3610 respondents (n=3340 [92.52%] women) aged between 22 and 65, with a mean age of 45.7 (SD 11.0) [18]. The sample was representative of the eligible population regarding regionality and employment sector. Women were slightly overrepresented and those aged under 40 were slightly underrepresented [18]. Ethical approval for the study was provided by The Finnish Institute for Health and Welfare (THL/482/6.02.01/2020).

### Measures

*Time pressure* was measured with the mean of 2 items measuring how often (during the previous half-year period) a person had been distracted by, worried about, or stressed about (1) being in a constant hurry and time pressure coming from unfinished work tasks and (2) having too little time to do work properly. The items were rated on a 6-point scale ranging from 1 (*never*) to 6 (*constantly*). The scale's reliability (Cronbach  $\alpha$ ) was .94 in the study sample. This measure has been widely used previously and associated, for example, with poor EHR usability among nurses [13].



Stress related to information systems was measured with the mean of 2 items ( $\alpha=.74$ ) framed into a single question that asked how often (during the previous half-year period) the respondent had been distracted by, worried about, or stressed about (1) constantly changing information systems and (2) difficult, poorly performing IT equipment/software. The answers were rated on a 6-point scale ranging from 1 (*never*) to 6 (*constantly*). This measure has previously been used to evaluate and associated with, for example, employees' distress and EHR usability [14,19].

*Self-rated stress* was measured with the widely used single-item self-rated stress measure [20]: "Stress means a situation when a person feels tense, restless, nervous, or anxious, or is unable to sleep at night because his or her mind is troubled all the time. Do you feel that kind of stress these days"? Response options were not at all/just a little/to some extent/quite a lot/very much. For analyses, these were categorized as 0=low stress (not at all/just a little/to some extent) and 1=high stress (quite a lot/very much).

*Mobile device use* was measured by asking respondents whether they also used EHR with mobile devices (such as a smartphone or tablet) with the following answer options: (1) yes (2) no, and (3) not possible to use mobile devices. The measure was coded as 0=no mobile device use (options 2 and 3) and 1=yes, uses mobile devices (option 1).

*Experience in using EHRs* was assessed by asking how experienced the respondent was in using an EHR (ie, as an EHR user) with a 5-point scale ranging from 1 (*beginner*) to 5 (*expert*). For analyses, this variable was coded as 0=*low experience* (answer options 1-3) and 1=*high experience* (answer options 4-5).

*The number of systems in daily use* was assessed by asking about the number of clinical systems that the responder needed to log into daily when working with patients. The response options were 0/1/2/3/4/5 or "more"/"my work does not include clinical work" (coded as *missing*). For analyses, the number of logins was coded as 1=1, 2=2, and 3=3 or more systems in daily use (5 respondents who answered that they had 0 systems in daily use were omitted from the analysis).

The usability measures *ease of use* and *technical quality* were measured with items derived from the validated National Usability-Focused Health Information System Scale (NuHISS) [21]. These measures assessed the usability of the current EHR system in use, not particularly the mobile version of the system. *Ease of use* included 3 items ( $\alpha=.82$ ) assessing the usability of key functionalities of the EHR system such as reading, documenting, and patient data retrieval ("The arrangement of fields and functions is logical on computer screen," "Routine tasks can be performed in a straight forward manner without

the need for extra steps using the system," and "Terminology on the screen is clear and understandable [eg, titles and labels]"). *Technical quality* was measured with 4 items ( $\alpha=.76$ ) assessing reliability and safety aspects of the EHR system ("The systems are stable in terms of technical functionality [does not crash, no downtime]," "The system responds quickly to inputs," "Faulty system function has caused a serious adverse event for the patient [reverse coded]," and "Faulty system function has nearly caused a serious adverse event for the patient [reverse coded]"). The answers were rated on a 5-point Likert scale ranging from 1 (*totally disagree*) to 5 (*totally agree*). The response options also included "Cannot answer," which was coded as missing.

Besides, age, gender, and employment sector were asked in the survey. Employment sector was coded as 1=hospital, 2=primary care, 3=social services, and 4=other.

## Statistical Analysis

The associations of mobile use, experience in using EHRs, number of systems in daily use, ease of use, and technical quality with the time pressure and stress related to information systems were analyzed with analyses of covariance (in separate analyses for each dependent variable). The analyses were adjusted for age, gender, and employment sector. The analyses were conducted in 2 steps. In the first step (Model A), the analysis included mobile use, age, gender, and employment sector. In the second step (Model B) experience in using EHRs, number of systems in daily use, ease of use, and technical quality were added to the former model. Analyses regarding self-rated stress were conducted using logistic regression analyses with the same steps as mentioned above.

Moreover, we examined the interactions of mobile version use with experience in using EHRs, number of systems in daily use, ease of use, and technical quality for the dependent variables with analyses of covariance (for time pressure and SRIS) and logistic regression (for stress) adjusted for age, gender, and main effects (in separate analyses for each interaction and dependent variable).

## Results

### Demographics

The characteristics of the study population are presented in Table 1. A majority of the respondents were women and approximately half worked at hospitals. Using a mobile version of the EHR was not very common, and only 17.70% (639/3610) used mobile devices. Most often, mobile device was used in social services (101/445, 22.7%), after that in hospitals (377/1903, 19.81%) and other sectors (82/467, 17.6%), whereas it was less common in primary care (79/795, 9.9%).

**Table 1.** Social demographics of the study sample (N=3610<sup>a</sup>).

Characteristic	Value
<b>Gender, n (%)</b>	
Women	3340 (92.52)
Men	249 (6.90)
Other (or did not want to report)	21 (0.58)
<b>Employment sector, n (%)</b>	
Hospital	1903 (52.71)
Primary care	795 (22.02)
Social services	445 (12.33)
Other	467 (12.94)
<b>Mobile device use, n (%)</b>	
No	2971 (82.30)
Yes	639 (17.70)
<b>Self-rated stress</b>	
Low	2318 (64.41)
High	1281 (35.59)
<b>Experience in using electronic health records, n (%)</b>	
Low	1135 (31.44)
High	2475 (68.56)
<b>Number of systems in daily use, n (%)</b>	
1	1327 (37.16)
2	1178 (32.99)
3 or more	1066 (29.85)
Age <sup>b</sup> , mean (SD)	45.68 (10.97)
Stress related to information systems <sup>c</sup> , mean (SD)	3.70 (1.13)
Time pressure <sup>c</sup> , mean (SD)	4.54 (1.12)
Ease of use <sup>d</sup> , mean (SD)	3.01 (1.08)
Technical quality <sup>d</sup> , mean (SD)	3.25 (0.98)

<sup>a</sup>Because of missing information in some variables, n varies between 3571 and 3610.

<sup>b</sup>Ranged between 22 and 67.

<sup>c</sup>Ranged between 1 and 6.

<sup>d</sup>Ranged between 1 and 5.

## Main Effects

Age, gender, mobile use, number of systems in daily use, ease of use, and technical quality were associated with time pressure in the fully adjusted model (Model B; Table 2). Younger respondents, women, mobile device users, and those who had a higher number of systems in daily use perceived more time pressure. Higher levels of ease of use and technical quality were associated with less time pressure.

Age, gender, employment sector, mobile use, experience in using EHRs, number of systems in daily use, ease of use, and technical quality were all associated with stress related to information systems in the fully adjusted model (Model B). Older respondents, women, those working in hospitals, mobile device users, less experienced EHR users, and those who had a higher number of systems in daily use perceived more stress related to information systems. Higher levels of ease of use and technical quality were associated with less stress related to information systems.

**Table 2.** The associations of independent variables with stress related to information systems and time pressure (analysis of covariance).

Variable	Time pressure				Stress related to information systems			
	Model A		Model B		Model A		Model B	
	<i>F</i> test	<i>P</i> value	<i>F</i> test	<i>P</i> value	<i>F</i> test	<i>P</i> value	<i>F</i> test	<i>P</i> value
Age	$F_{1,3583}=23.11$	<.001	$F_{1,3537}=24.84$	<.001	$F_{1,3583}=25.23$	<.001	$F_{1,3537}=37.55$	<.001
Gender	$F_{2,3583}=9.13$	<.001	$F_{2,3537}=8.25$	<.001	$F_{2,3583}=8.51$	<.001	$F_{2,3537}=6.00$	.003
Sector	$F_{3,3583}=2.31$	.08	$F_{3,3537}=0.88$	.45	$F_{3,3583}=43.93$	<.001	$F_{3,3537}=24.38$	<.001
Mobile device use	$F_{1,3583}=15.87$	<.001	$F_{1,3537}=14.96$	<.001	$F_{1,3583}=7.41$	.007	$F_{1,3537}=6.11$	.01
Experience			$F_{1,3537}=2.77$	.10			$F_{1,3537}=17.31$	<.001
Number of systems in daily use			$F_{2,3537}=6.90$	.001			$F_{2,3537}=43.17$	<.001
Ease of use			$F_{1,3537}=33.92$	<.001			$F_{1,3537}=269.91$	<.001
Technical quality			$F_{1,3537}=62.83$	<.001			$F_{1,3537}=311.82$	<.001
$R^2$	0.016		0.069		0.043		0.301	

Age, number of systems in daily use, ease of use, and technical quality were associated with self-rated stress in the fully adjusted model (Model B; Table 3). Older respondents were less likely to have self-rated stress. Those who had 3 or more systems in daily use were 1.23 times more likely to have a high level of stress compared with those who had only 1 system in use.

Higher levels of ease of use and technical quality were associated with lower likelihood of stress. Employment sector was significantly associated with stress in Model A ( $P=.02$ ), but after adjusting for number of systems in daily use, experience in using EHRs, ease of use, and technical quality, the association was no longer significant ( $P=.17$ ).

**Table 3.** The results of the logistic regression analysis for self-rated stress.<sup>a</sup>

Demographics	Model A		Model B	
	Odds ratio (95% CI)	<i>P</i> value	Odds ratio (95% CI)	<i>P</i> value
Age	0.99 (0.98-0.99)	.001	0.99 (0.98-0.99)	.001
<b>Gender</b>				
Men	1		1	
Women	1.33 (1.00-1.76)	.05	1.28 (0.96-1.71)	.10
<b>Employment sector</b>		.02		.17
Hospital	1		1	
Primary health care	0.90 (0.75-1.07)	.24	0.94 (0.78-1.12)	.47
Social care	0.87 (0.70-1.08)	.20	1.03 (0.82-1.29)	.82
Other	0.71 (0.57-0.89)	.003	0.78 (0.63-0.98)	.04
<b>Mobile device use</b>				
No	1		1	
Yes, uses mobile device	1.05 (0.87-1.26)	.62	1.01 (0.84-1.22)	.91
<b>Experience in using electronic health records</b>				
Low			1	
High			0.95 (0.81-1.11)	.49
<b>Number of systems in daily use</b>				.04
1			1	
2			1.00 (0.84-1.19)	.97
3 or more			1.23 (1.03-1.46)	.02
Ease of use			0.76 (0.71-0.82)	<.001
Technical quality			0.84 (0.78-0.91)	<.001

<sup>a</sup>For continuous variables, the model odds ratio presented indicate the likelihood of passing from low stress to high stress, compared with 1 SD change in continuous independent variables.

## Interactions

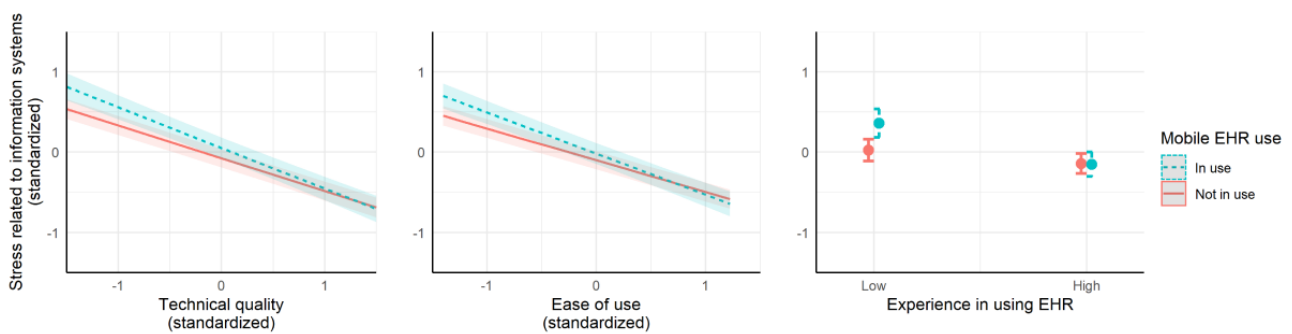
We examined the interactions of mobile device use with experience in using EHRs, number of systems in daily use, ease of use, and technical quality for the dependent variables.

The interaction between mobile device use and experience in using EHRs was significant for stress related to information systems ( $F_{1,3581}=14.93$ ,  $P<.001$ ). As can be seen from [Figure 1](#), the highest levels of stress related to information systems were reported by respondents who used mobile devices and had low experience in using EHRs. Moreover, the interaction between mobile device use and ease of use was significant for stress related to information systems ( $F_{1,3577}=10.16$ ,  $P=.001$ ). The association between ease of use and stress related to

information systems was more pronounced among those who used mobile devices and the highest levels of stress related to information systems was experienced among those who used mobile devices and perceived low levels of ease of use of their EHRs ([Figure 1](#)). Besides, the interaction between mobile device use and technical quality was significant for stress related to information systems ( $F_{1,3577}=6.45$ ,  $P=.01$ ). Similar to ease of use, the association between technical quality and stress related to information systems was more pronounced among those who used mobile devices and the highest levels of stress related to information systems were experienced among those who used mobile devices and perceived low levels of technical quality of their EHRs ([Figure 1](#)). The interaction between mobile device use and number of systems in daily use was nonsignificant for stress related to information systems ( $P=.21$ ).



**Figure 1.** The associations of mobile device use with technical quality, ease of use, and experience in using EHRs for stress related to information systems. EHR: electronic health record.



## Discussion

### Principal Results

According to our results it seems that at present use of mobile versions of EHR is not beneficial to the nurses' well-being. We found that use of mobile versions of EHR was associated with higher levels of time pressure and stress resulting from poorly functioning information systems.

Moreover, use of mobile versions of EHR intensified the association of inexperience and poor usability of EHR with stress related to information systems. More specifically, inexperienced EHR users who used mobile devices had higher levels of stress related to information systems than others. In addition, those who perceived low levels of ease of use or technical quality of their EHRs and used mobile devices had higher levels of stress related to information systems than others.

### Limitations

This was a cross-sectional survey study, and so we cannot draw any causal inferences. Moreover, we used self-reported measures which may have an effect here, given that if a respondent subjectively considers oneself being under stress, it is possible that it affects all responses, including those not related to stress. Therefore, it would be important for future studies to also include objective measures of stress when examining mobile versions. Using self-reported measures may also lead to a possibility of problems related to an inflation of the strengths of relationships and common method variance.

We adjusted our analyses for age, gender, and employment sector, but there may exist a possibility of residual confounding. Although the employment sector was adjusted for, it is possible that our results reflect higher work strain in places where mobile devices are used. In our study mobile device use was most common in social services and it is likely that, for example, in home care mobile devices are more often used and work strain can be high. This should be kept in mind when interpreting our results and future studies are needed on this topic. In addition, our study did not control for the technology used in the mobile version and given the possible wide variation in the usability of mobile versions this might be of importance. We also did not categorize whether the mobile version used was a smartphone or a tablet, which has a big difference in size (and thus their user interface), and is likely to have affected the response of

the nurses. Thus, these issues also need to be considered in the future.

Our sample was rather large ( $n=3610$ ) which also allowed us to examine interactions. Our sample represented the eligible population regarding living region and employment sector, but included slightly more women and those aged over 40 [18]. Data were collected in the spring of 2020 (between March and April) at the time when the COVID-19 epidemic strengthened in Finland (with strict restrictions implemented mid-March). Therefore, only 1 reminder was sent to those who had not answered. This situation may influence the results and especially so in those hospitals that were most strongly affected.

Finland is one of the forerunners in the digitalization of health care [22] and provides a universal health care for all its residents. Therefore, caution should be exercised when generalizing our findings to other dissimilar countries.

### Comparison With Prior Work

Using EHR on the mobile device seems not to be beneficial for the well-being of nurses. We found that it was associated with high stress related to information systems and time pressure. Our finding is not congruent with previous findings suggesting that mobile device use would decrease nurses' stress [4,5]. The discrepancy in the results may be due to different countries, methods, and samples studied. Chiang and Wang [4] performed a qualitative study including 17 community nurses from 6 home care facilities. Johansson et al [5] performed a quantitative study including 398 registered nurses and nursing students attending the undergraduate and graduate nursing programs, and thus their sample was smaller and included younger participants (mean age 34.7). Instead, we used a large ( $n=3610$ ) population-based sample of registered nurses from different employment sectors. Moreover, the study by Johansson et al [5] focused more on advanced mobile devices themselves and on the capacity to locate information from the internet, whereas we focused solely on the mobile version of the EHR.

Although mobile devices are suggested to improve workflow, make continuous updating possible, and improve the communication within the health care team [3], there are many possible reasons as to why they could increase the sense of hurry and strain coming from viewing and updating EHRs on these devices. User interface has been suggested as one of the challenges in the deployment of clinical mobile apps [9]. Lack of visibility on mobile devices may lead to errors; interfaces

are compact and cluttered with information which require full attention; and the change from focusing on real life toward a mobile device may pose problems [9]. Handheld devices have been found to double the data entry time and increase the risks of typing errors and missing data [10]. Small screen is also less effective and increases the number of scrolling activities [11]. Moreover, stability problems related to wireless networks may (negatively) affect nurses' work when using mobile devices [23]. These challenges related to mobile devices may cause extra stress and time pressure to nurses. By contrast, it is also possible that mobile devices are more likely in use at those places where demands and hurry are at high levels.

According to our results, mobile device use intensifies the negative effects of usability problems related to EHRs on the stress that is caused by information systems. Ease of use and technical quality of the EHRs were also directly associated with all examined well-being indicators (ie, time pressure, stress related to information systems, and self-rated stress). Our findings correspond to previous studies showing that usability problems are associated with lower well-being among nurses [12,13] and physicians [14-16]. As mentioned above, mobile devices may have problems with user interface and data entry [9,10]. Thus, it is possible that difficulties in use or technical problems related to EHRs are also reflected in mobile devices causing extra stress and problems for nurses.

To promote the well-being of nurses, EHRs and their mobile versions should be improved, so they would be easy to use and would better support the workflow. Issues related to finance, hardware, communication, security, and user interface have been identified as main challenges regarding the implementation of clinical mobile apps [9]. Using interface that is already familiar to the user and the hierarchical organization of the information have been suggested as means to improve mobile apps [9]. It is important to accurately validate mobile device interfaces which are meant to be used in a clinical setting [24]. Nurses' needs should be fully taken into consideration during the development of mobile versions [25]. For example, the functions that can increase performance and are associated with workflow are suggested to be of importance [2].

Our results suggest that inexperienced users of EHRs seem to be especially at risk when using mobile devices. A previous

study has shown the importance of experience among physicians for managing SRISs and psychological distress levels [14]. Moreover, nurses' low e-care competence has been associated with high time pressure and distress [13]. This calls for more training related to EHRs and their mobile versions. It would be important to provide adequate and systematic support and training for those who are just learning to use the system and who still have skills gaps (ie, focusing especially on new employees and those whose work environment is implementing new systems). For example, continuous educational programs focusing on enhancing nurses' information technology literacy have been suggested [25]. Moreover, in-house information systems support and regular training on acquired information systems would be of importance and could also encourage positive attitudes toward technologies [26].

## Conclusions

According to our findings it seems that a mobile version of EHR is not beneficial for the well-being of nurses. Mobile device use was associated with nurses' perceptions of higher levels of time pressure at work. Moreover, mobile device use was associated with higher stress resulting from poorly functioning and constantly changing information systems. In addition, mobile device use intensified the negative effects of inexperience in using EHRs and poor usability of EHRs on the stress related to information systems. Thus, it seems that at present mobile versions of EHRs need improvements to better support nurses' workflow and well-being. Moreover, more training related to EHRs, their mobile versions, and workflow related to these should be provided to nurses.

It would be important to pay increasing attention to these issues, as nurses are at particular risk of experiencing additional stress and strain resulting from the need to use information systems in their work, and work strain, in turn, has been associated with a higher risk of disability [27]. A significant proportion of nurses' working time is already spent on patient information systems and in addition to this, nurses must constantly learn to use a variety of new electronic services and platforms in their work, which have increased significantly as a result of the COVID-19 epidemic.

---

## Acknowledgments

This study was supported by the Ministry of Social Affairs and Health (project 414919001) and the Strategic Research Council at the Academy of Finland (project 327145). None of them had any role in the design of the study and collection, analysis, and interpretation of the data or in the writing.

---

## Conflicts of Interest

None declared.

---

## References

1. Lee Y, Park YR, Kim J, Kim JH, Kim WS, Lee J. Usage Pattern Differences and Similarities of Mobile Electronic Medical Records Among Health Care Providers. *JMIR Mhealth Uhealth* 2017 Dec 13;5(12):e178 [FREE Full text] [doi: [10.2196/mhealth.8855](https://doi.org/10.2196/mhealth.8855)] [Medline: [29237579](https://pubmed.ncbi.nlm.nih.gov/29237579/)]

2. Kim S, Lee K, Hwang H, Yoo S. Analysis of the factors influencing healthcare professionals' adoption of mobile electronic medical record (EMR) using the unified theory of acceptance and use of technology (UTAUT) in a tertiary hospital. *BMC Med Inform Decis Mak* 2016 Jan 30;16:12 [FREE Full text] [doi: [10.1186/s12911-016-0249-8](https://doi.org/10.1186/s12911-016-0249-8)] [Medline: [26831123](https://pubmed.ncbi.nlm.nih.gov/26831123/)]
3. Schachner MB, Sommer JA, González ZA, Luna DR, Benítez SE. Evaluating the Feasibility of Using Mobile Devices for Nurse Documentation. *Stud Health Technol Inform* 2016;225:495-499. [Medline: [27332250](https://pubmed.ncbi.nlm.nih.gov/27332250/)]
4. Chiang K, Wang H. Nurses' experiences of using a smart mobile device application to assist home care for patients with chronic disease: a qualitative study. *J Clin Nurs* 2016 Jul;25(13-14):2008-2017. [doi: [10.1111/jocn.13231](https://doi.org/10.1111/jocn.13231)] [Medline: [27136280](https://pubmed.ncbi.nlm.nih.gov/27136280/)]
5. Johansson P, Petersson G, Saveman B, Nilsson G. Using advanced mobile devices in nursing practice--the views of nurses and nursing students. *Health Informatics J* 2014 Sep;20(3):220-231 [FREE Full text] [doi: [10.1177/1460458213491512](https://doi.org/10.1177/1460458213491512)] [Medline: [25183609](https://pubmed.ncbi.nlm.nih.gov/25183609/)]
6. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. *P T* 2014 May;39(5):356-364 [FREE Full text] [Medline: [24883008](https://pubmed.ncbi.nlm.nih.gov/24883008/)]
7. Mickan S, Tilson JK, Atherton H, Roberts NW, Heneghan C. Evidence of effectiveness of health care professionals using handheld computers: a scoping review of systematic reviews. *J Med Internet Res* 2013 Oct 28;15(10):e212 [FREE Full text] [doi: [10.2196/jmir.2530](https://doi.org/10.2196/jmir.2530)] [Medline: [24165786](https://pubmed.ncbi.nlm.nih.gov/24165786/)]
8. de Jong A, Donelle L, Kerr M. Nurses' Use of Personal Smartphone Technology in the Workplace: Scoping Review. *JMIR Mhealth Uhealth* 2020 Nov 26;8(11):e18774 [FREE Full text] [doi: [10.2196/18774](https://doi.org/10.2196/18774)] [Medline: [33242012](https://pubmed.ncbi.nlm.nih.gov/33242012/)]
9. Ehrler F, Wipfli R, Teodoro D, Sarrey E, Walesa M, Lovis C. Challenges in the Implementation of a Mobile Application in Clinical Practice: Case Study in the Context of an Application that Manages the Daily Interventions of Nurses. *JMIR Mhealth Uhealth* 2013 Jun 12;1(1):e7 [FREE Full text] [doi: [10.2196/mhealth.2344](https://doi.org/10.2196/mhealth.2344)] [Medline: [25100680](https://pubmed.ncbi.nlm.nih.gov/25100680/)]
10. Haller G, Haller DM, Courvoisier DS, Lovis C. Handheld vs. laptop computers for electronic data collection in clinical research: a crossover randomized trial. *J Am Med Inform Assoc* 2009;16(5):651-659 [FREE Full text] [doi: [10.1197/jamia.M3041](https://doi.org/10.1197/jamia.M3041)] [Medline: [19567799](https://pubmed.ncbi.nlm.nih.gov/19567799/)]
11. Jones M, Marsden G, Mohd-Nasir N, Boone K, Buchanan G. Improving Web interaction on small displays. *Computer Networks*. 1999 May. URL: [https://doi.org/10.1016/S1389-1286\(99\)00013-4](https://doi.org/10.1016/S1389-1286(99)00013-4) [accessed 2020-10-10]
12. Kaihlanen A, Gluschkoff K, Hyppönen H, Kaipio J, Puttonen S, Vehko T, et al. The Associations of Electronic Health Record Usability and User Age With Stress and Cognitive Failures Among Finnish Registered Nurses: Cross-Sectional Study. *JMIR Med Inform* 2020 Nov 18;8(11):e23623 [FREE Full text] [doi: [10.2196/23623](https://doi.org/10.2196/23623)] [Medline: [33206050](https://pubmed.ncbi.nlm.nih.gov/33206050/)]
13. Vehko T, Hyppönen H, Puttonen S, Kujala S, Ketola E, Tuukkanen J, et al. Experienced time pressure and stress: electronic health records usability and information technology competence play a role. *BMC Med Inform Decis Mak* 2019 Aug 14;19(1):160 [FREE Full text] [doi: [10.1186/s12911-019-0891-z](https://doi.org/10.1186/s12911-019-0891-z)] [Medline: [31412859](https://pubmed.ncbi.nlm.nih.gov/31412859/)]
14. Heponiemi T, Kujala S, Vainiomäki S, Vehko T, Lääveri T, Vänskä J, et al. Usability Factors Associated With Physicians' Distress and Information System-Related Stress: Cross-Sectional Survey. *JMIR Med Inform* 2019 Nov 05;7(4):e13466 [FREE Full text] [doi: [10.2196/13466](https://doi.org/10.2196/13466)] [Medline: [31687938](https://pubmed.ncbi.nlm.nih.gov/31687938/)]
15. Vainiomäki S, Heponiemi T, Vänskä J, Hyppönen H. Tailoring EHRs for Specific Working Environments Improves Work Well-Being of Physicians. *Int J Environ Res Public Health* 2020 Jun 30;17(13):4715 [FREE Full text] [doi: [10.3390/ijerph17134715](https://doi.org/10.3390/ijerph17134715)] [Medline: [32630043](https://pubmed.ncbi.nlm.nih.gov/32630043/)]
16. Vainiomäki S, Aalto A, Lääveri T, Sinervo T, Elovainio M, Mäntyselkä P, et al. Better Usability and Technical Stability Could Lead to Better Work-Related Well-Being among Physicians. *Appl Clin Inform* 2017 Oct;8(4):1057-1067 [FREE Full text] [doi: [10.4338/ACI-2017-06-RA-0094](https://doi.org/10.4338/ACI-2017-06-RA-0094)] [Medline: [29241245](https://pubmed.ncbi.nlm.nih.gov/29241245/)]
17. Webropol. URL: <https://webropol.com> [accessed 2021-06-30]
18. Saranto K, Kinnunen U, Koponen S, Kyytsönen M, Hyppönen H, Vehko T. Nurses' competences in information management as well as experiences in health and social care information system support for daily practice. *Finn J EHealth EWelfare* 2020;212-228 [FREE Full text] [doi: [10.23996/fjhw.95711](https://doi.org/10.23996/fjhw.95711)]
19. Heponiemi T, Hyppönen H, Kujala S, Aalto A, Vehko T, Vänskä J, et al. Predictors of physicians' stress related to information systems: a nine-year follow-up survey study. *BMC Health Serv Res* 2018 Dec 13;18(1):284 [FREE Full text] [doi: [10.1186/s12913-018-3094-x](https://doi.org/10.1186/s12913-018-3094-x)] [Medline: [29653530](https://pubmed.ncbi.nlm.nih.gov/29653530/)]
20. Elo A, Leppänen A, Jahkola A. Validity of a single-item measure of stress symptoms. *Scand J Work Environ Health* 2003 Dec;29(6):444-451 [FREE Full text] [Medline: [14712852](https://pubmed.ncbi.nlm.nih.gov/14712852/)]
21. Hyppönen H, Kaipio J, Heponiemi T, Lääveri T, Aalto A, Vänskä J, et al. Developing the National Usability-Focused Health Information System Scale for Physicians: Validation Study. *J Med Internet Res* 2019 May 16;21(5):e12875 [FREE Full text] [doi: [10.2196/12875](https://doi.org/10.2196/12875)] [Medline: [31099336](https://pubmed.ncbi.nlm.nih.gov/31099336/)]
22. PwC. European Hospital Survey: Benchmarking Deployment of e-Health Services (2012?2013): Composite Indicators on eHealth Deployment and on Availability and Use of eHealth Functionalities: Final Report. *JRC Scientific and Policy Reports* 2014:1-310.
23. Shen L, Zang X, Cong J. Nurses' satisfaction with use of a personal digital assistants with a mobile nursing information system in China. *Int J Nurs Pract* 2018 Apr;24(2):e12619. [doi: [10.1111/ijn.12619](https://doi.org/10.1111/ijn.12619)] [Medline: [29356202](https://pubmed.ncbi.nlm.nih.gov/29356202/)]

24. Ehrler F, Haller G, Sarrey E, Walesa M, Wipfli R, Lovis C. Assessing the Usability of Six Data Entry Mobile Interfaces for Caregivers: A Randomized Trial. *JMIR Hum Factors* 2015 Dec 15;2(2):e15 [FREE Full text] [doi: [10.2196/humanfactors.4093](https://doi.org/10.2196/humanfactors.4093)] [Medline: [27025648](https://pubmed.ncbi.nlm.nih.gov/27025648/)]
25. Kuo K, Liu C, Ma C. An investigation of the effect of nurses' technology readiness on the acceptance of mobile electronic medical record systems. *BMC Med Inform Decis Mak* 2013 Aug 12;13:88 [FREE Full text] [doi: [10.1186/1472-6947-13-88](https://doi.org/10.1186/1472-6947-13-88)] [Medline: [23938040](https://pubmed.ncbi.nlm.nih.gov/23938040/)]
26. Ifinedo P. The moderating effects of demographic and individual characteristics on nurses' acceptance of information systems: A canadian study. *Int J Med Inform* 2016 Mar;87:27-35. [doi: [10.1016/j.ijmedinf.2015.12.012](https://doi.org/10.1016/j.ijmedinf.2015.12.012)] [Medline: [26806709](https://pubmed.ncbi.nlm.nih.gov/26806709/)]
27. Mäntyniemi A, Oksanen T, Salo P, Virtanen M, Sjösten N, Pentti J, et al. Job strain and the risk of disability pension due to musculoskeletal disorders, depression or coronary heart disease: a prospective cohort study of 69,842 employees. *Occup Environ Med* 2012 Aug;69(8):574-581. [doi: [10.1136/oemed-2011-100411](https://doi.org/10.1136/oemed-2011-100411)] [Medline: [22573793](https://pubmed.ncbi.nlm.nih.gov/22573793/)]

## Abbreviations

**EHR:** electronic health record

*Edited by C Lovis; submitted 12.03.21; peer-reviewed by J Bagby, J Brooke; comments to author 22.04.21; revised version received 05.05.21; accepted 23.05.21; published 06.07.21.*

*Please cite as:*

*Heponiemi T, Kaihlanen AM, Gluschkoff K, Saranto K, Nissinen S, Laukka E, Vehko T*

*The Association Between Using a Mobile Version of an Electronic Health Record and the Well-Being of Nurses: Cross-sectional Survey Study*

*JMIR Med Inform* 2021;9(7):e28729

URL: <https://medinform.jmir.org/2021/7/e28729>

doi: [10.2196/28729](https://doi.org/10.2196/28729)

PMID: [34255704](https://pubmed.ncbi.nlm.nih.gov/34255704/)

©Tarja Heponiemi, Anu-Marja Kaihlanen, Kia Gluschkoff, Kaija Saranto, Sari Nissinen, Elina Laukka, Tuulikki Vehko. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org>), 06.07.2021. 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

# A Biomedical Knowledge Graph System to Propose Mechanistic Hypotheses for Real-World Environmental Health Observations: Cohort Study and Informatics Application

Karamarie Fecho<sup>1,2</sup>, PhD; Chris Bizon<sup>1</sup>, PhD; Frederick Miller<sup>3</sup>, MD, PhD; Shepherd Schurman<sup>3</sup>, MD; Charles Schmitt<sup>3</sup>, PhD; William Xue<sup>3</sup>, BSc; Kenneth Morton<sup>4</sup>, PhD; Patrick Wang<sup>4</sup>, PhD; Alexander Tropsha<sup>1,5</sup>, PhD

<sup>1</sup>Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

<sup>2</sup>Copperline Professional Solutions, Pittsboro, NC, United States

<sup>3</sup>National Institute of Environmental Health Sciences, Durham, NC, United States

<sup>4</sup>CoVar Applied Technologies, Durham, NC, United States

<sup>5</sup>Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

**Corresponding Author:**

Karamarie Fecho, PhD

Renaissance Computing Institute

University of North Carolina at Chapel Hill

100 Europa Drive, Suite 540

Chapel Hill, NC, 27517

United States

Phone: 1 919 445 9640

Email: [kfecho@copperlineprofessionalsolutions.com](mailto:kfecho@copperlineprofessionalsolutions.com)

## Abstract

**Background:** Knowledge graphs are a common form of knowledge representation in biomedicine and many other fields. We developed an open biomedical knowledge graph-based system termed Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways (ROBOKOP). ROBOKOP consists of both a front-end user interface and a back-end knowledge graph. The ROBOKOP user interface allows users to posit questions and explore answer subgraphs. Users can also posit questions through direct Cypher query of the underlying knowledge graph, which currently contains roughly 6 million nodes or biomedical entities and 140 million edges or predicates describing the relationship between nodes, drawn from over 30 curated data sources.

**Objective:** We aimed to apply ROBOKOP to survey data on workplace exposures and immune-mediated diseases from the Environmental Polymorphisms Registry (EPR) within the National Institute of Environmental Health Sciences.

**Methods:** We analyzed EPR survey data and identified 45 associations between workplace chemical exposures and immune-mediated diseases, as self-reported by study participants (n= 4574), with 20 associations significant at  $P < .05$  after false discovery rate correction. We then used ROBOKOP to (1) validate the associations by determining whether plausible connections exist within the ROBOKOP knowledge graph and (2) propose biological mechanisms that might explain them and serve as hypotheses for subsequent testing. We highlight the following three exemplar associations: carbon monoxide-multiple sclerosis, ammonia-asthma, and isopropanol-allergic disease.

**Results:** ROBOKOP successfully returned answer sets for three queries that were posed in the context of the driving examples. The answer sets included potential intermediary genes, as well as supporting evidence that might explain the observed associations.

**Conclusions:** We demonstrate real-world application of ROBOKOP to generate mechanistic hypotheses for associations between workplace chemical exposures and immune-mediated diseases. We expect that ROBOKOP will find broad application across many biomedical fields and other scientific disciplines due to its generalizability, speed to discovery and generation of mechanistic hypotheses, and open nature.

(JMIR Med Inform 2021;9(7):e26714) doi:[10.2196/26714](https://doi.org/10.2196/26714)

**KEYWORDS**

knowledge graph; knowledge representation; data exploration; generalizability; discovery; open science; immune-mediated disease

## Introduction

“Knowledge graphs” (KGs) have become a common approach for knowledge representation across scientific disciplines, including biomedicine [1]. In a KG, multiple expert-curated “knowledge sources” are integrated into a graph structure, with nodes representing entities and edges providing the relationship between nodes. Within a biomedical KG, the nodes represent biomedical entities, such as *drugs* or *diseases*, and the edges describe relationships that connect the nodes, such as *treats* (eg, *drug treats disease*). The curated knowledge sources that populate a biomedical KG include both databases, such as DrugBank [2] and Comparative Toxicogenomics Database (CTD) [3], and ontologies, such as Monarch Disease Ontology [4] and Human Phenotype Ontology [5]. Various reasoning tools and inferential algorithms are typically applied to KGs [1], thus allowing users to construct complex queries that ask, for example, *if gene X is connected to both chemical exposure Y and disease Z, then the protein product of gene X may represent a potential drug target*. Indeed, successful applications of KGs include drug repurposing [6] and the identification of new drug targets [7].

While KGs, such as Monarch, are openly available, many of the more sophisticated KGs remain proprietary. Perhaps the most well-known proprietary KG is the Freebase-derived Google KG that powers Google’s web search capability [8]. As part of the Biomedical Data Translator program [9-11], we have developed an open KG-based biomedical system termed Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways (ROBOKOP) [12,13]. ROBOKOP is designed to answer questions such as *what genes are associated with recovery from COVID-19 infection? why is imatinib effective in the treatment of asthma? what biological pathways are associated with stroke-related morbidity?* Note that these questions imply complex mechanistic relationships between terms in the query such as *drug* and *disease*. The ROBOKOP KG is designed to provide answers in the form of putative mechanistic pathways connecting the query terms.

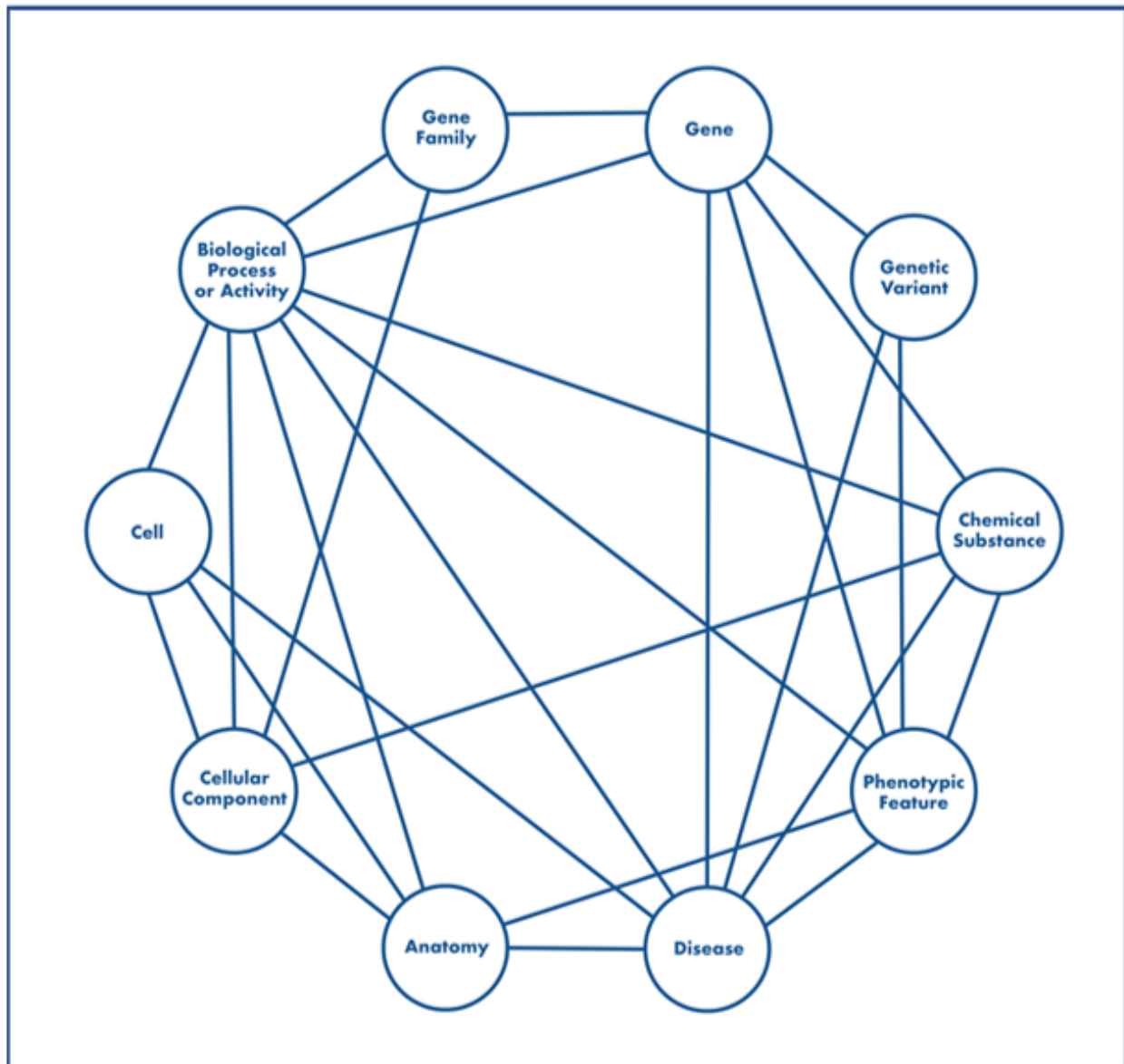
Herein, we provide an overview of ROBOKOP and its real-world application to data derived from the Environmental Polymorphisms Registry (EPR) within the National Institute of Environmental Health Sciences. Specifically, we focused on an EPR study that aimed to explore the impact of workplace exposures on immune-mediated diseases (IMDs) such as asthma, allergy, multiple sclerosis, rheumatoid arthritis, and ulcerative colitis. We first conducted an exploratory analysis of self-reported exposures and IMD symptoms in order to identify significant associations between workplace chemical exposures and IMD. We then used ROBOKOP to (1) validate statistically significant associations by determining whether plausible connections exist within the ROBOKOP KG and (2) propose biological mechanisms that might explain the associations and serve as hypotheses for subsequent testing.

## Methods

### ROBOKOP

ROBOKOP is a biomedical KG-based question-answering system that is comprised of both a front-end user interface (UI) and a back-end KG, both of which are openly available [12-15]. The ROBOKOP KG uses the Biolink model [16] as an upper-level ontology that can be applied to express domain knowledge as a graph of relationships between biomedical entities. The ROBOKOP KG currently contains 6 million nodes and 140 million edges, with nodes representing a wide range of biological entities, such as genes, biological processes, anatomical features, diseases, and phenotypes, and edges representing predicates, such as *is associated with*, *causes*, and *increases expression of*. The ROBOKOP KG is derived from over 30 curated biomedical data sources (Multimedia Appendix 1) that have been integrated into a graph structure (Figure 1). The curated data sources were openly available and accessible via direct import into a local Neo4j instance. Some of the data sources were only partially complete and/or required preprocessing.

**Figure 1.** High-level schema for the Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways (ROBOKOP) knowledge graph, showing classes of nodes as defined by the Biolink model. Note that the schema provides a user guide to assist with the proper construction of queries by providing a visual overview of nodes that are connected in the ROBOKOP knowledge graph.



The ROBOKOP UI allows users to posit queries and quickly explore ranked-and-scored answer sets or subgraphs. ROBOKOP queries are meta-graphs [17-19] or question graphs of the entity types shown in Figure 1. The meta-graph or machine question has a general structure that is defined by the user on the basis of the high-level question of interest. Users can select nodes as named entities specified as a CURIE (Compact Uniform Resource Identifier) or specified only as the entity type. A drop-down menu provides users with choices to free-text node entries. Answers are structured as subgraphs that match the query in topology and type, as well as any desired properties of nodes and edges.

A given query will frequently produce many answers or subgraphs, especially for queries with little specification regarding nodes and edges or with multiple nodes and edges. As such, the ranking of subgraphs by relevance to the query and the strength of the supporting evidence are critical for user

exploration of results. The ROBOKOP answer-ranking algorithm [13] weighs each edge within each subgraph using a metric that is based on the number of PubMed abstracts that cite both the source and target nodes. Publication support and predicate assertions are provided by the curated knowledge sources used to create the KG (Multimedia Appendix 1). Additional publication support is provided by a ROBOKOP service termed OmniCorp [20], which contains a graph of PubMed identifiers linked to concepts (ie, potential nodes in the ROBOKOP KG) referenced within abstracts. OmniCorp is built by processing all PubMed abstracts with the SciGraph Named Entity Recognition application programming interface [21] and matching text in titles and abstracts to concepts from a predetermined set of biological ontologies. The ROBOKOP answer-ranking algorithm calculates a confidence score for each answer subgraph with respect to the distance between leaves of the answer subgraph, considering the edge weights as electrical resistance, as defined by Ohm law [22]. Answer subgraphs with

greater publication counts will be ranked higher, with publication counts derived from the curated knowledge sources treated with greater importance than those from the publication co-occurrences provided by OmniCorp. The confidence score is then augmented with an “informativeness” score, which is inspired by the NAGA scoring model [23] and treats novel more specific assertions (eg, *disease X interacts with geneNPC1*) with greater importance than more generic assertions (eg, *disease X interacts with immune system*).

The ROBOKOP KG can also be queried directly, independent of the UI, using the Cypher query language [24] to find subgraphs within the KG that match the structure of the query. Example Cypher queries can be found online [15].

### Application Use Case Description

The EPR is a study of nearly 20,000 current participants that aims to better understand interactions among environmental exposures and genetic determinants of health and disease [25]. The registry contains survey data on participant exposures and disease history, as well as DNA samples and other biological measurements. As part of the broader effort, investigators have been exploring the impact of workplace exposures on IMDs.

An IMD was defined as a self-reported allergic reaction (allergic rhinitis, hay fever, or seasonal allergies; allergies [other than seasonal]); asthma condition; or autoimmune disorder (psoriasis, thyroid disease [noncancer], hyperthyroidism, hypothyroidism, Crohn disease, multiple sclerosis, celiac disease, Sjogren disease, rheumatoid arthritis, ulcerative colitis, scleroderma or systemic sclerosis, pernicious anemia, myositis, or lupus). EPR survey data were extracted in December 2018. The overall sample size was 4574 participants.

An exploratory analysis was conducted to examine associations between each IMD and specific workplace chemicals classified into one of 18 classes (Multimedia Appendix 2). The survey questions from which the chemicals were obtained were drawn from the EPR “Exposome Survey – Part A: External Exposome, Section B: Chemical and Metal Exposures at Work” and generally structured as follows: “Please select any *heavy metals* you have ever been exposed to for 15 minutes a week or more in any job you have held (CHOOSE ALL THAT APPLY).” Associations between workplace chemical exposures and individual IMDs were examined using chi-square analysis or the Fisher exact test, when sample sizes were small due to missing data or few positive cases. A false discovery rate correction was applied to the association tests. Chemicals were examined individually and also by chemical class. Odds ratios (ORs) with lower and upper bounds were calculated per convention and were not adjusted for small sample sizes. As this was an exploratory analysis, the significance level was set at  $\alpha=.05$  or  $.10$ , and we did not control for potential covariates such as age, sex, and race.

## Results

### Application Use Case Results

A total of 45 exposure-IMD associations were significant at  $P<.10$ , with 20 associations significant at  $P<.05$  after false

discovery rate correction. In all cases, workplace chemical exposures were associated with increased odds of self-reported IMD (Multimedia Appendix 3). Dyes were the most common workplace exposure class associated with IMD conditions. No associations were identified between acids or glues/adhesives and IMD conditions. “Allergies or allergic reaction (other than seasonal allergies)” and “allergic rhinitis, hay fever, or seasonal allergies” were the most common IMD conditions associated with workplace chemical exposures.

### ROBOKOP-Derived Mechanistic Assertions Based on Application Use Case Results

We highlight the ROBOKOP application using three exemplar associations that were chosen because they were significant at  $P<.05$ , evident at both the level of the specific chemical and the chemical class, and representative of different chemical classes and IMDs: (1) carbon monoxide-multiple sclerosis, (2) ammonia-asthma, and (3) isopropanol-allergic rhinitis, hay fever, or seasonal allergies. An overview of ROBOKOP results for each of these examples is provided below, with various functionalities of the interactive UI highlighted in the first example.

#### Carbon Monoxide and Multiple Sclerosis

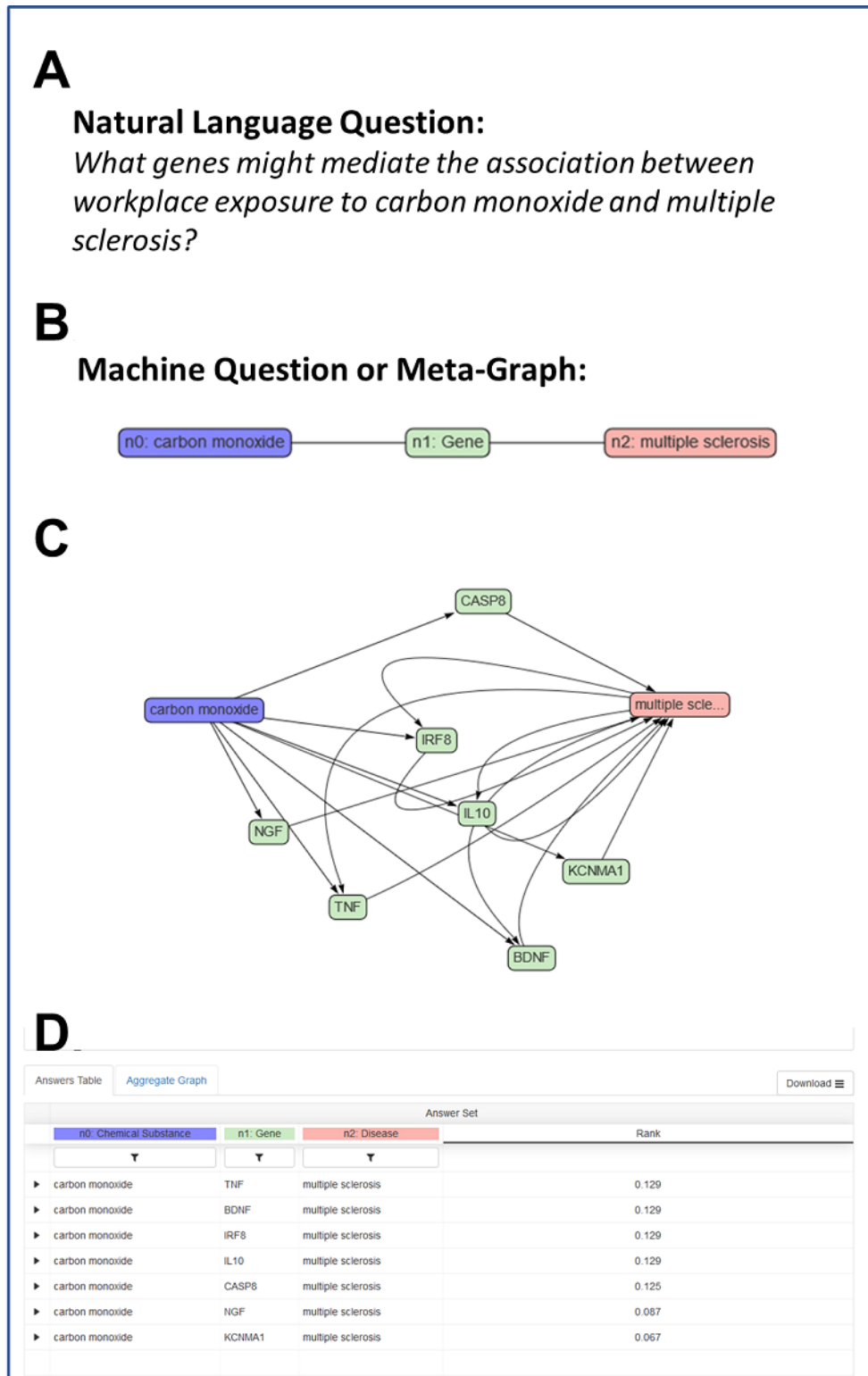
The association between workplace exposure to carbon monoxide and multiple sclerosis was significant at both the chemical level (OR 6.4583, 95% CI 1.8524-18.2844;  $P=.006$ ) and chemical class level (OR 3.8902, 95% CI 1.2521-10.3546;  $P=.03$ ).

We posed the following question to ROBOKOP, but structured as a machine question: *what genes might mediate the association between exposure to carbon monoxide and multiple sclerosis?* The general question, machine question, and answer set are displayed in Figure 2. ROBOKOP identified seven subgraphs and potential intermediary genes as follows: *TNF* (tumor necrosis factor), *BDNF* (brain-derived neurotrophic factor), *IL10* (interleukin-10), *NGF* (nerve growth factor), *IRF8* (interferon regulatory factor 8), *KCNMA1* (potassium calcium-activated channel subfamily M alpha 1), and *CASP8* (caspase 8).

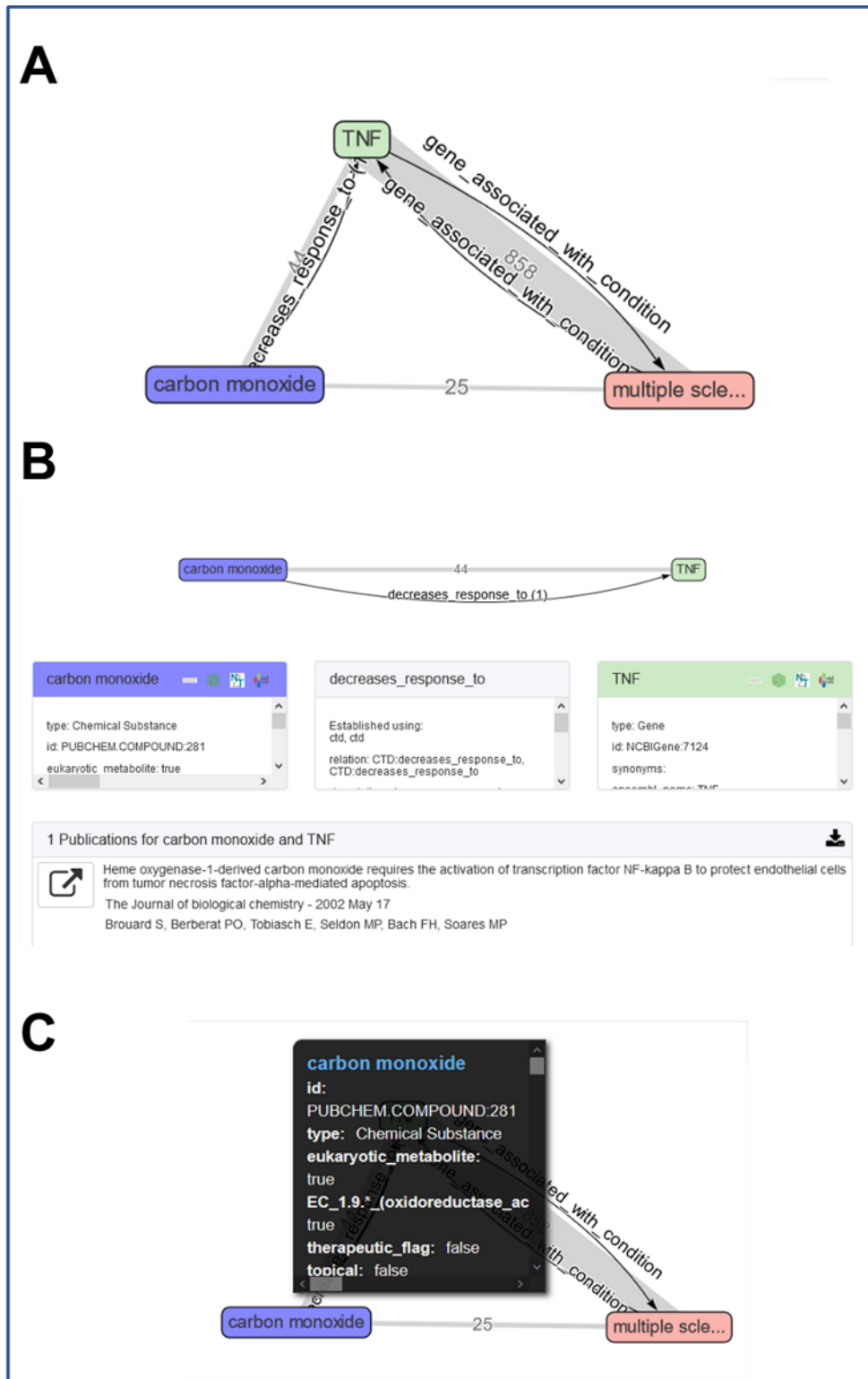
The top-ranked answer set had 858 PubMed publications, contributed by OmniCorp, supporting the association between multiple sclerosis and *TNF*, and 44 PubMed publications, again contributed by OmniCorp, supporting the association between carbon monoxide and *TNF* (Figure 3). Twenty-five additional PubMed publications (from OmniCorp) supported an association between carbon monoxide and multiple sclerosis, with several suggesting the involvement of heme oxygenase-1, which is described as an enzyme that oxidizes heme to bilirubin and carbon monoxide. The multiple sclerosis-*TNF* association was established by both HETIO and Pharos. The carbon monoxide-*TNF* association was established by CTD, with publication support contributed by CTD that again suggested a role for heme oxygenase-1 [26].



**Figure 2.** A Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways (ROBOKOP) high-level question (A) implemented as a machine question or meta-graph (B) designed to explore genes that might mediate the observed association between workplace exposure to carbon monoxide and Environmental Polymorphisms Registry participant self-report of multiple sclerosis. Users select nodes and edges in the ROBOKOP user interface (or by direct Cypher query) to translate the desired natural-language question into an executable machine question, using the schema provided in Figure 1 as a guide. The resultant aggregated answer graph (C) and list of answer subgraphs (D) show six potential mediating genes. Both the answer graph and the list of answer subgraphs are interactive and can be explored by users. For example in (D), users can click on each answer subgraph to explore knowledge sources, predicate assertions, and publication support. TNF: tumor necrosis factor; BDNF: brain-derived neurotrophic factor; IL10: interleukin-10; NGF: nerve growth factor; IRF8: interferon regulatory factor 8; KCNMA1: potassium calcium-activated channel subfamily M alpha 1; CASP8: caspase 8.



**Figure 3.** The top-ranked answer subgraph suggesting the involvement of *TNF* (tumor necrosis factor) (A) as a potential gene that might mediate the observed significant association between workplace exposure to carbon monoxide and multiple sclerosis. The Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways (ROBOKOP) machine question was structured as follows: carbon monoxide – gene – multiple sclerosis. Example publication support for the linkage between carbon monoxide and *TNF* suggesting a role for heme oxygenase-1 (B). Metadata for the carbon monoxide node (C).



**Ammonia and Asthma**

The association between workplace exposure to ammonia and asthma was significant at both the chemical level (OR 2.0422, 95% CI 1.4426-2.8524; *P*=.001) and chemical class level (OR 1.5210, 95% CI 1.1938-1.9311; *P*=.004).

The query that was posed to ROBOKOP aimed to identify potential genes that might mediate the association between ammonia and asthma, and it was structured similarly to the query in Figure 2 (ammonia–gene–asthma). ROBOKOP identified nine answer sets or individual paths through the aggregated answer graph. The intermediary genes were *ADA*

(adenosine deaminase), *PRKG1* (protein kinase cGMP-dependent 1), *S100B* (calcium binding protein B), *TNF*, *MPO* (myeloperoxidase), *IL6* (interleukin-6), *IL1B* (interleukin-1 beta), *PDE4A* (phosphodiesterase 4A), and *PARP1* (poly(ADP-ribose) polymerase 1).

The lowest-ranked answer subgraph identified *ADA* as the intermediary gene. The metadata associated with *ADA* indicated that *ADA* metabolizes adenosine by converting it to inosine and ammonia. For this answer subgraph, the relationship between asthma and ammonia was supported by 93 PubMed publications, identified by OmniCorp. An *ADA*-asthma relationship was established using Monarch, with 32 supporting PubMed publications contributed by OmniCorp. Many of the supporting publications suggested a relationship between *ADA* mutations and asthma, allergy, and immune function [27].

### ***Isopropanol and Allergic Rhinitis, Hay Fever, or Seasonal Allergies***

The association between workplace exposure to isopropanol and allergic rhinitis, hay fever, or seasonal allergies was significant at the chemical level (OR 1.3990, 95% CI 1.1415-1.7155;  $P=.02$ ) and chemical class level (OR 1.2323, 95% CI 1.0287-1.4764;  $P=.09$ ).

As with the other two examples, the query that was posed to ROBOKOP aimed to identify intermediary genes and was structured similarly, except that “allergic disease” was used as the disease node, instead of “allergic rhinitis, hay fever, or seasonal allergies,” because the latter disease node was not available in ROBOKOP.

ROBOKOP identified the following three intermediary genes: *IL6*, *TNF*, and *CSF2* (colony stimulating factor 2). The intermediary gene in the second top-ranked answer set was *IL6*. An isopropanol-*IL6* relationship was established by CTD, with one supporting publication provided by CTD and two additional supporting PubMed publications contributed by OmniCorp. An *IL6*-allergic disease relationship was established by Pharos, with 79 supporting PubMed publications contributed by OmniCorp. Many of those publications [28] suggested the involvement of the innate immune response (eg, basophils, eosinophils, and mast cells), and several publications suggested neuroimmune involvement (eg, “allergic mood” and hippocampal inflammation) [29].

## ***Discussion***

### **Summary of Findings and Related Work**

We identified 45 significant associations between workplace chemical exposures and IMD conditions in an EPR cohort, with 20 of those associations significant at  $P<.05$  and largely unexpected a priori. Statistical evidence for an association, while important, does not establish a causal relationship or provide any insights into underlying mechanisms. Thus, we applied the open ROBOKOP KG system to validate the observed associations by demonstrating plausible connections between exposures and IMD conditions, and we provide mechanistic insights or hypotheses to explain them. We highlighted the

following three use case applications: carbon monoxide-multiple sclerosis, ammonia-asthma, and isopropanol-allergic disease.

ROBOKOP identified plausible answer subgraphs to a query structured as carbon monoxide-gene-multiple sclerosis, thus supporting the statistical association between workplace exposure to carbon monoxide and multiple sclerosis. ROBOKOP further identified the gene that encodes TNF, *TNF*, as one of several potential mediating genes. ROBOKOP metadata and publication support suggested that heme oxygenase-1, which oxidizes heme to bilirubin and carbon monoxide, plays an intermediary role [26]. Levels of heme oxygenase-1 are depressed in patients with multiple sclerosis and further depressed during episodes of disease exacerbation [30], and exogenous carbon monoxide or chemicals that release carbon monoxide or induce heme oxygenase-1 appear to be therapeutic in experimental models of multiple sclerosis [31]. These results suggest that heme oxygenase-1, via carbon monoxide, has a homeostatic or anti-inflammatory role that might protect against multiple sclerosis or suppress disease exacerbation. However, more recent evidence suggests that a chronic heme oxygenase-1 response in glial cells may promote neurodegeneration and thereby exacerbate multiple sclerosis and other neurodegenerative diseases [32].

The statistical association between workplace exposure to ammonia and asthma was supported by multiple answer subgraphs that were returned by ROBOKOP. The gene that encodes *ADA*, *ADA*, was identified by ROBOKOP as one of several potential intermediates in this relationship. ROBOKOP metadata indicated that *ADA* metabolizes adenosine by converting it to inosine and ammonia. ROBOKOP identified publications suggesting an association between *ADA* mutations and asthma, allergy, and immune function, including one publication suggesting an association with aspirin-intolerant asthma [27]. A recent review indicated that *ADA* deficiency may have detrimental effects on multiple organ systems, including the pulmonary system [33]. Several older publications suggest that adenosine acts as a bronchoconstrictor in persons with asthma [34], and a more recent publication suggested an association between exposure to ammonia and asthma exacerbations [35]. These results support a relationship between *ADA* mutations and pulmonary complications, as well as an association between ammonia and pulmonary complications in persons with established respiratory disease.

ROBOKOP provided several answer subgraphs to support the statistical association between workplace exposure to isopropanol and allergic disease. The gene encoding *IL6*, *IL6*, was identified by ROBOKOP as one of several potential intermediary genes, with publication support suggesting the involvement of the innate immune response [28]. In addition, ROBOKOP identified publication support suggesting the involvement of *IL6* in neuroimmune and neurobehavioral correlates of allergic disease [29]. While *IL1* is widely recognized as playing a prominent role in “sickness behavior,” *IL6* also appears to play a role [36,37]. These findings suggest that the association between exposure to isopropanol and allergic disease might actually reflect a relationship between isopropanol and neurological/neurobehavioral correlates of allergic disease. Specifically, the results suggest that isopropanol might trigger

an innate neuroimmune response that results in elevated levels of IL6, which then might trigger neurobehavioral symptoms of allergic disease.

### Limitations

ROBOKOP has several limitations that should be considered when interpreting the results here or using the application. First, many of the associations between workplace chemical exposures and IMD conditions involved complex chemical mixtures such as “toner,” “transmission fluid,” and “motor oil.” ROBOKOP currently maps these entities to Medical Subject Heading terms, but the application does not have an approach in place for mapping such mixtures to the individual chemicals that constitute a given mixture. We are considering approaches to overcome this limitation. Second, ROBOKOP, like all KG-based applications, is limited by the challenge of KG completion. For example, if an exposure-gene relationship has not been established by one of the curated data sources underlying the ROBOKOP KG, then this relationship will not be identified. We are developing algorithms to overcome this limitation by inferring edges in the KG, but for now, this remains a limitation. Third, evidence for ROBOKOP assertions is derived from the following two main sources: (1) the curated knowledge sources used to create the KG and (2) the co-occurrence of terms in PubMed abstracts. The ranking and scoring algorithm that is used to rank answer subgraphs is based on these two sources of evidence, with the first source treated with greater importance than the second. Other relevant factors, such as date of publication and number of studies on a topic (versus publications), are not considered at present, but may be incorporated into future versions of the application. Fourth, as a prototype system, ROBOKOP does not yet support natural language processing capabilities or other sophisticated approaches to aid users. We encourage users to contact the developers and/or post GitHub issues should they encounter any challenges when generating queries and/or evaluating answer subgraphs. Fifth, while not a limitation of ROBOKOP itself, the workplace chemical exposure-IMD associations reported here were derived from survey data (ie, participant self-report) and were not confirmed by clinical record review or expert judgement. As such, the potential exists for misclassification and/or bias in both the reported exposures and the reported IMDs. Moreover, the timing between workplace exposure and the onset of IMD cannot be determined due to limitations of the survey design and participant recollection at the time the survey was administered. Finally, as this was an exploratory analysis, we did not adjust the ORs for small sample sizes or control for potential covariates such as age, sex, and race.

### Conclusion

ROBOKOP demonstrated potential in its use to support real-world observations and generate mechanistic hypotheses. In this paper, we focused on significant associations between

workplace chemical exposures and IMDs, identified as part of a larger EPR study. We note, however, that ROBOKOP has other applications. Indeed, one key feature of ROBOKOP is its generalizability across biomedical domains as a general question-answering system, with capabilities to support a variety of machine questions. For instance, we are using ROBOKOP to explore associations between medications and clinical outcomes, including adverse events, using data derived from electronic health records. We are also testing whether ROBOKOP can be used to support human reasoning on Medical College Admission Tests, and we have promising preliminary results [38]. While the ROBOKOP KG is currently built from biomedical knowledge sources, the general approach is not restricted to the biomedical space. Indeed, on behalf of our institution’s leadership, we are developing a new version of the ROBOKOP KG that is focused on exploring relationships between research proposals and investigator characteristics.

A second key feature of ROBOKOP is its ability to support speed to discovery. For instance, the example use cases presented here took little time to construct and execute, and the interactive UI allows even novice users to posit questions and explore answer subgraphs. The mechanistic insights that were gleaned from the ROBOKOP answer subgraphs were quickly realized, thus allowing for a rapid first-pass analysis of the results and evaluation of the supporting evidence. Moreover, ROBOKOP revealed all potential genes that might mediate the observed exposure-IMD associations via a single query; in effect, the user did not need to spend hours, days, or longer reading through the available literature. This speed to discovery afforded by ROBOKOP also allows investigators to quickly refute associations that may be nothing more than spurious findings (eg, if no answer subgraphs are returned by ROBOKOP). In the examples highlighted herein, we focus on three example exposure-IMD associations. While this may not seem like many, a full literature review to identify potentially novel insights, eliminate spurious findings, and explore supporting evidence would take abundantly more time than was needed to conduct the initial first-pass analysis using ROBOKOP. We plan to leverage ROBOKOP’s speed to discovery in a large-scale analysis of associations between single nucleotide polymorphisms and phenotypes as part of a broader EPR effort and in several other studies.

A third key feature of ROBOKOP is its open nature. Indeed, access to ROBOKOP, whether via the UI or by direct Cypher query of the underlying KG, does not require login authentication or an account; rather, anyone with the URL can access the system. Moreover, the ROBOKOP KG can be downloaded independently of the application [15]. The open nature of ROBOKOP and the multiple routes to access it democratize science and, when coupled with the speed to discovery afforded by the application, should accelerate progress in biomedicine and many other fields.



## Acknowledgments

The authors wish to thank the Biomedical Data Translator Consortium for their support and intellectual input. We also acknowledge and appreciate the contributions of the broader Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways (ROBOKOP) team, including Vinicius Alves, Jim Balhoff, Steven Cox, Yaphet Kabede, Daniel Korn, Eugene Muratov, and Max Wang. In addition, we thank the members of the Environmental Polymorphisms Registry (EPR) Executive Leadership Committee and others in the EPR IMD Study Group, including Farida Akhtari, Montserrat Ayala-Ramirez, Perry Blackshear, Askia Dunnon, David Fargo, Michael Fessler, Stavros Garantziotis, Janet E Hall, Nathaniel MacNell, John McGrath, Alison Motsinger-Reif, and Christine Parks for useful discussions, and Jeremy Erickson, Andy Rooney, and Vickie Walker for helpful comments on the manuscript. This work was supported by the National Center for Advancing Translational Sciences, National Institutes of Health (OT2TR002514) and by the Clinical Research Branch, Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences (ZID ES103354-01).

The corresponding author (KF) can be reached at [kfecho@copperlineprofessionalsolutions.com](mailto:kfecho@copperlineprofessionalsolutions.com) or [kfecho@renci.org](mailto:kfecho@renci.org).

## Authors' Contributions

All authors contributed in a substantive manner to the work described herein and approved the manuscript for journal submission.

## Conflicts of Interest

None declared.

### Multimedia Appendix 1

Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways (ROBOKOP) knowledge graph data sources. [[PDF File \(Adobe PDF File\), 96 KB - medinform\\_v9i7e26714\\_app1.pdf](#)]

### Multimedia Appendix 2

Workplace chemical exposures explored for their association with immune-mediated diseases. [[PDF File \(Adobe PDF File\), 68 KB - medinform\\_v9i7e26714\\_app2.pdf](#)]

### Multimedia Appendix 3

Significant associations between workplace exposures and immune-mediated diseases, identified as part of the Environmental Polymorphisms Registry. [[PDF File \(Adobe PDF File\), 153 KB - medinform\\_v9i7e26714\\_app3.pdf](#)]

## References

1. Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. *Comput Struct Biotechnol J* 2020;18:1414-1428 [[FREE Full text](#)] [doi: [10.1016/j.csbj.2020.05.017](https://doi.org/10.1016/j.csbj.2020.05.017)] [Medline: [32637040](https://pubmed.ncbi.nlm.nih.gov/32637040/)]
2. Wishart D, Feunang Y, Guo A, Lo E, Marcu A, Grant J, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018 Jan 04;46(D1):D1074-D1082 [[FREE Full text](#)] [doi: [10.1093/nar/gkx1037](https://doi.org/10.1093/nar/gkx1037)] [Medline: [29126136](https://pubmed.ncbi.nlm.nih.gov/29126136/)]
3. Davis A, Grondin C, Johnson R, Sciaky D, McMorran R, Wieggers J, et al. The Comparative Toxicogenomics Database: update 2019. *Nucleic Acids Res* 2019 Jan 08;47(D1):D948-D954 [[FREE Full text](#)] [doi: [10.1093/nar/gky868](https://doi.org/10.1093/nar/gky868)] [Medline: [30247620](https://pubmed.ncbi.nlm.nih.gov/30247620/)]
4. Mondo Disease Ontology. URL: <http://www.obofoundry.org/ontology/mondo.html> [accessed 2021-06-25]
5. Köhler S, Carmody L, Vasilevsky N, Jacobsen J, Danis D, Gouridine J, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res* 2019 Jan 08;47(D1):D1018-D1027 [[FREE Full text](#)] [doi: [10.1093/nar/gky1105](https://doi.org/10.1093/nar/gky1105)] [Medline: [30476213](https://pubmed.ncbi.nlm.nih.gov/30476213/)]
6. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* 2017 Sep 22;6:e26726 [[FREE Full text](#)] [doi: [10.7554/eLife.26726](https://doi.org/10.7554/eLife.26726)] [Medline: [28936969](https://pubmed.ncbi.nlm.nih.gov/28936969/)]
7. Chen B, Ding Y, Wild DJ. Assessing drug target association using semantic linked data. *PLoS Comput Biol* 2012 Jul 5;8(7):e1002574 [[FREE Full text](#)] [doi: [10.1371/journal.pcbi.1002574](https://doi.org/10.1371/journal.pcbi.1002574)] [Medline: [22859915](https://pubmed.ncbi.nlm.nih.gov/22859915/)]
8. Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: a collaboratively created graph database for structuring human knowledge. In: SIGMOD '08: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. 2008 Presented at: 2008 ACM SIGMOD International Conference on Management of Data; June 9-12, 2008; Vancouver, Canada p. 1247-1250. [doi: [10.1145/1376616.1376746](https://doi.org/10.1145/1376616.1376746)]
9. Austin CP, Colvis CM, Southall NT. Deconstructing the Translational Tower of Babel. *Clin Transl Sci* 2019 Mar;12(2):85 [[FREE Full text](#)] [doi: [10.1111/cts.12595](https://doi.org/10.1111/cts.12595)] [Medline: [30412342](https://pubmed.ncbi.nlm.nih.gov/30412342/)]

10. Biomedical Data Translator Consortium. The Biomedical Data Translator Program: Conception, Culture, and Community. *Clin Transl Sci* 2019 Mar 09;12(2):91-94 [FREE Full text] [doi: [10.1111/cts.12592](https://doi.org/10.1111/cts.12592)] [Medline: [30412340](https://pubmed.ncbi.nlm.nih.gov/30412340/)]
11. Biomedical Data Translator Consortium. Toward A Universal Biomedical Data Translator. *Clin Transl Sci* 2019 Mar 09;12(2):86-90 [FREE Full text] [doi: [10.1111/cts.12591](https://doi.org/10.1111/cts.12591)] [Medline: [30412337](https://pubmed.ncbi.nlm.nih.gov/30412337/)]
12. Bizon C, Cox S, Balhoff J, Kebede Y, Wang P, Morton K, et al. ROBOKOP KG and KGB: Integrated Knowledge Graphs from Federated Sources. *J Chem Inf Model* 2019 Dec 23;59(12):4968-4973. [doi: [10.1021/acs.jcim.9b00683](https://doi.org/10.1021/acs.jcim.9b00683)] [Medline: [31769676](https://pubmed.ncbi.nlm.nih.gov/31769676/)]
13. Morton K, Wang P, Bizon C, Cox S, Balhoff J, Kebede Y, et al. ROBOKOP: an abstraction layer and user interface for knowledge graphs to support question answering. *Bioinformatics* 2019 Dec 15;35(24):5382-5384 [FREE Full text] [doi: [10.1093/bioinformatics/btz604](https://doi.org/10.1093/bioinformatics/btz604)] [Medline: [31410449](https://pubmed.ncbi.nlm.nih.gov/31410449/)]
14. Robokop: Reasoning Over Biomedical Objected linked in Knowledge Oriented Pathways. URL: <https://robokop.renci.org> [accessed 2021-06-25]
15. Connect to Neo4j. URL: <https://robokopkg.renci.org> [accessed 2021-06-25]
16. Biolink Model. URL: <https://biolink.github.io/biolink-model/> [accessed 2021-06-25]
17. Huang Z, Zheng Y, Cheng R, Sun Y, Mamoulis N, Li X. Meta Structure: Computing Relevance in Large Heterogeneous Information Networks. In: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 1595-1604. [doi: [10.1145/2939672.2939815](https://doi.org/10.1145/2939672.2939815)]
18. Fang Y, Lin W, Zheng V, Wu M, Chang K, Li X. Semantic proximity search on graphs with metagraph-based learning. 2016 Presented at: 32nd International Conference on Data Engineering (ICDE); May 16-20, 2016; Helsinki, Finland p. 277-288. [doi: [10.1109/icde.2016.7498247](https://doi.org/10.1109/icde.2016.7498247)]
19. Zhao H, Yao Q, Li J, Song Y, Lee D. Meta-Graph Based Recommendation Fusion over Heterogeneous Information Networks. In: *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017 Presented at: 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2017; New York, NY p. 635-644. [doi: [10.1145/3097983.3098063](https://doi.org/10.1145/3097983.3098063)]
20. NCATS-gamma/omnicorp. GitHub. URL: <https://github.com/NCATS-Gamma/omnicorp> [accessed 2021-06-25]
21. SciGraph/SciGraph. GitHub. URL: <https://github.com/SciGraph/SciGraph/> [accessed 2021-06-25]
22. Klein DJ, Randić M. Resistance distance. *J Math Chem* 1993 Dec;12(1):81-95. [doi: [10.1007/bf01164627](https://doi.org/10.1007/bf01164627)]
23. Kasneci G, Suchanek F, Ifrim G, Ramanath M, Weikum G. NAGA: Searching and Ranking Knowledge. 2008 Presented at: 24th International Conference on Data Engineering; April 7-12, 2008; Cancun, Mexico. [doi: [10.1109/icde.2008.4497504](https://doi.org/10.1109/icde.2008.4497504)]
24. Cypher Query Language. Neo4j. URL: <https://neo4j.com/developer/cypher-query-language/> [accessed 2021-06-25]
25. Exploring How Your Genes Interact with the Environment. NIEHS, Environmental Polymorphisms Registry, North Carolina DNA Bank. URL: <https://dnaregistry.niehs.nih.gov/> [accessed 2021-06-25]
26. Brouard S, Berberat PO, Tobiasch E, Seldon MP, Bach FH, Soares MP. Heme Oxygenase-1-derived Carbon Monoxide Requires the Activation of Transcription Factor NF- $\kappa$ B to Protect Endothelial Cells from Tumor Necrosis Factor- $\alpha$ -mediated Apoptosis. *Journal of Biological Chemistry* 2002 May;277(20):17950-17961. [doi: [10.1074/jbc.m108317200](https://doi.org/10.1074/jbc.m108317200)]
27. Kim S, Kim Y, Park H, Kim S, Kim S, Ye Y, et al. Adenosine deaminase and adenosine receptor polymorphisms in aspirin-intolerant asthma. *Respir Med* 2009 Mar;103(3):356-363 [FREE Full text] [doi: [10.1016/j.rmed.2008.10.008](https://doi.org/10.1016/j.rmed.2008.10.008)] [Medline: [19019667](https://pubmed.ncbi.nlm.nih.gov/19019667/)]
28. Yang S, Wu J, Zhang Q, Li X, Liu D, Zeng B, et al. Allergic Rhinitis in Rats Is Associated with an Inflammatory Response of the Hippocampus. *Behav Neurol* 2018 Apr 16;2018:8750464 [FREE Full text] [doi: [10.1155/2018/8750464](https://doi.org/10.1155/2018/8750464)] [Medline: [29849816](https://pubmed.ncbi.nlm.nih.gov/29849816/)]
29. Trikojat K, Luksch H, Rösen-Wolff A, Plessow F, Schmitt J, Buske-Kirschbaum A. "Allergic mood" - Depressive and anxiety symptoms in patients with seasonal allergic rhinitis (SAR) and their association to inflammatory, endocrine, and allergic markers. *Brain Behav Immun* 2017 Oct;65:202-209. [doi: [10.1016/j.bbi.2017.05.005](https://doi.org/10.1016/j.bbi.2017.05.005)] [Medline: [28495610](https://pubmed.ncbi.nlm.nih.gov/28495610/)]
30. Fagone P, Patti F, Mangano K, Mammanna S, Coco M, Touil-Boukoffa C, et al. Heme oxygenase-1 expression in peripheral blood mononuclear cells correlates with disease activity in multiple sclerosis. *J Neuroimmunol* 2013 Aug 15;261(1-2):82-86. [doi: [10.1016/j.jneuroim.2013.04.013](https://doi.org/10.1016/j.jneuroim.2013.04.013)] [Medline: [23714423](https://pubmed.ncbi.nlm.nih.gov/23714423/)]
31. Fagone P, Mangano K, Coco M, Perciavalle V, Garotta G, Romao C, et al. Therapeutic potential of carbon monoxide in multiple sclerosis. *Clin Exp Immunol* 2012 Feb;167(2):179-187 [FREE Full text] [doi: [10.1111/j.1365-2249.2011.04491.x](https://doi.org/10.1111/j.1365-2249.2011.04491.x)] [Medline: [22235993](https://pubmed.ncbi.nlm.nih.gov/22235993/)]
32. Schipper HM, Song W, Tavitian A, Cressatti M. The sinister face of heme oxygenase-1 in brain aging and disease. *Prog Neurobiol* 2019 Jan;172:40-70. [doi: [10.1016/j.pneurobio.2018.06.008](https://doi.org/10.1016/j.pneurobio.2018.06.008)] [Medline: [30009872](https://pubmed.ncbi.nlm.nih.gov/30009872/)]
33. Flinn AM, Gennery AR. Adenosine deaminase deficiency: a review. *Orphanet J Rare Dis* 2018 Apr 24;13(1):65 [FREE Full text] [doi: [10.1186/s13023-018-0807-5](https://doi.org/10.1186/s13023-018-0807-5)] [Medline: [29690908](https://pubmed.ncbi.nlm.nih.gov/29690908/)]
34. Ng W, Polosa R, Church M. Adenosine bronchoconstriction in asthma: investigations into its possible mechanism of action. *Br J Clin Pharmacol* 1990;30 Suppl 1:89S-98S [FREE Full text] [doi: [10.1111/j.1365-2125.1990.tb05474.x](https://doi.org/10.1111/j.1365-2125.1990.tb05474.x)] [Medline: [2268511](https://pubmed.ncbi.nlm.nih.gov/2268511/)]

35. Holst G, Thygesen M, Pedersen C, Peel R, Brand J, Christensen J, et al. Ammonia, ammonium, and the risk of asthma: a register-based case-control study in Danish children. *Environmental Epidemiology* 2018;2(3):e019. [doi: [10.1097/ee9.000000000000019](https://doi.org/10.1097/ee9.000000000000019)]
36. Dantzer R. Cytokine, sickness behavior, and depression. *Immunol Allergy Clin North Am* 2009 May;29(2):247-264 [FREE Full text] [doi: [10.1016/j.iac.2009.02.002](https://doi.org/10.1016/j.iac.2009.02.002)] [Medline: [19389580](https://pubmed.ncbi.nlm.nih.gov/19389580/)]
37. Burton MD, Sparkman NL, Johnson RW. Inhibition of interleukin-6 trans-signaling in the brain facilitates recovery from lipopolysaccharide-induced sickness behavior. *J Neuroinflammation* 2011 May 19;8(1):54 [FREE Full text] [doi: [10.1186/1742-2094-8-54](https://doi.org/10.1186/1742-2094-8-54)] [Medline: [21595956](https://pubmed.ncbi.nlm.nih.gov/21595956/)]
38. Fecho K, Bizon C, Cox S, Balhoff J, Kebede Y, Wang P, et al. Application of a biomedical question-answering system to support reasoning on MCAT questions. 2020 Presented at: AMIA 2020 Virtual Annual Symposium; November 14-18, 2020; Virtual.

## Abbreviations

**ADA:** adenosine deaminase

**CTD:** Comparative Toxicogenomics Database

**EPR:** Environmental Polymorphisms Registry

**IL6:** interleukin-6

**IMD:** immune-mediated disease

**KG:** knowledge graph

**OR:** odds ratio

**ROBOKOP:** Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways

**TNF:** tumor necrosis factor

**UI:** user interface

*Edited by C Lovis; submitted 22.12.20; peer-reviewed by J Yang, J Chen, M Devarakonda; comments to author 25.02.21; revised version received 26.04.21; accepted 27.04.21; published 20.07.21.*

*Please cite as:*

*Fecho K, Bizon C, Miller F, Schurman S, Schmitt C, Xue W, Morton K, Wang P, Tropsha A*

*A Biomedical Knowledge Graph System to Propose Mechanistic Hypotheses for Real-World Environmental Health Observations: Cohort Study and Informatics Application*

*JMIR Med Inform* 2021;9(7):e26714

URL: <https://medinform.jmir.org/2021/7/e26714>

doi: [10.2196/26714](https://doi.org/10.2196/26714)

PMID: [34283031](https://pubmed.ncbi.nlm.nih.gov/34283031/)

©Karamarie Fecho, Chris Bizon, Frederick Miller, Shepherd Schurman, Charles Schmitt, William Xue, Kenneth Morton, Patrick Wang, Alexander Tropsha. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org>), 20.07.2021. 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

# Frequency of Participation in External Quality Assessment Programs Focused on Rare Diseases: Belgian Guidelines for Human Genetics Centers

Joséphine Lantoine<sup>1</sup>, MSc, PhD; Anne Brysse<sup>2</sup>, MSc, PhD; Vinciane Dideberg<sup>2</sup>, MD, PhD; Kathleen Claes<sup>3</sup>, MD, DPhil; Sofie Symoens<sup>3</sup>, MSc, PhD; Wim Coucke<sup>4</sup>, MSTAT, PhD; Valérie Benoit<sup>5</sup>, PhD, PharmD; Sonia Rombout<sup>5</sup>, MSc, PhD; Martine De Rycke<sup>6</sup>, MSc, PhD; Sara Seneca<sup>6</sup>, MSc, PhD; Lut Van Laer<sup>7</sup>, MSc, PhD; Wim Wuyts<sup>7</sup>, MSc, PhD; Anniek Corveleyn<sup>8</sup>, IR, PhD; Kris Van Den Bogaert<sup>8</sup>, MSc, PhD; Catherine Rydlewski<sup>9</sup>, MSc, PhD; Françoise Wilkin<sup>9</sup>, MSc, PhD; Marie Ravoet<sup>10</sup>, MSc, PhD; Elodie Fastré<sup>10</sup>, MSc, PhD; Arnaud Capron<sup>4</sup>, MSc, PhD; Nathalie Monique Vandeveldel<sup>1</sup>, PhD, PharmD

<sup>1</sup>Rare Diseases Unit, Department of Quality of Laboratories, Sciensano, Brussels, Belgium

<sup>2</sup>Center of Human Genetics, CHU of Liège, University of Liège, Liège, Belgium

<sup>3</sup>Center for Medical Genetics, Ghent University Hospital, Gent, Belgium

<sup>4</sup>Department of Quality of Laboratories, Sciensano, Brussels, Belgium

<sup>5</sup>Center of Human Genetics, Institut de Pathologie et de Génétique, Gosselies, Belgium

<sup>6</sup>Center for Medical Genetics, Universitair Ziekenhuis Brussel, Vrije Universiteit Brussel, Brussels, Belgium

<sup>7</sup>Center of Medical Genetics, Antwerp University Hospital and University of Antwerp, Edegem, Belgium

<sup>8</sup>Center for Human Genetics, University Hospitals Leuven, Leuven, Belgium

<sup>9</sup>Center of Human Genetics, Hôpital Erasme, Université Libre de Bruxelles, Brussels, Belgium

<sup>10</sup>Center for Human Genetics, Cliniques universitaires Saint-Luc, Université catholique de Louvain, Brussels, Belgium

**Corresponding Author:**

Nathalie Monique Vandeveldel, PhD, PharmD

Rare Diseases Unit

Department of Quality of Laboratories

Sciensano

Juliette Wuytsman street, 14

Brussels, 1050

Belgium

Phone: 32 2 642 55 89

Fax: 32 2 642 56 45

Email: [nathalie.vandeveldel@sciensano.be](mailto:nathalie.vandeveldel@sciensano.be)

## Abstract

**Background:** Participation in quality controls, also called external quality assessment (EQA) schemes, is required for the ISO15189 accreditation of the Medical Centers of Human Genetics. However, directives on the minimal frequency of participation in genetic quality control schemes are lacking or too heterogeneous, with a possible impact on health care quality.

**Objective:** The aim of this project is to develop Belgian guidelines on the frequency of participation in quality controls for genetic testing in the context of rare diseases.

**Methods:** A group of experts analyzed 90 EQA schemes offered by accredited providers and focused on analyses used for the diagnosis of rare diseases. On that basis, the experts developed practical recommendations about the minimal frequencies of participation of the Medical Centers of Human Genetics in quality controls and how to deal with poor performances and change management. These guidelines were submitted to the Belgian Accreditation Body and then reviewed and approved by the Belgian College of Human Genetics and Rare Diseases and by the National Institute for Health and Disability Insurance.

**Results:** The guidelines offer a decisional algorithm for the minimal frequency of participation in human genetics EQA schemes. This algorithm has been developed taking into account the scopes of the EQA schemes, the levels of experience, and the annual volumes of the Centers of Human Genetics in the performance of the tests considered. They include three key principles: (1) the



recommended annual assessment of all genetic techniques and technological platforms, if possible through EQAs covering the technique, genotyping, and clinical interpretation; (2) the triennial assessment of the genotyping and interpretation of specific germline mutations and pharmacogenomics analyses; and (3) the documentation of actions undertaken in the case of poor performances and the participation to quality control the following year. The use of a Bayesian statistical model has been proposed to help the Centers of Human Genetics to determine the theoretical number of tests that should be annually performed to achieve a certain threshold of performance (eg, a maximal error rate of 1%). Besides, the guidelines insist on the role and responsibility of the national public health authorities in the follow-up of the quality of analyses performed by the Medical Centers of Human Genetics and in demonstrating the cost-effectiveness and rationalization of participation frequency in these quality controls.

**Conclusions:** These guidelines have been developed based on the analysis of a large panel of EQA schemes and data collected from the Belgian Medical Centers of Human Genetics. They are applicable to other countries and will facilitate and improve the quality management and financing systems of the Medical Centers of Human Genetics.

(*JMIR Med Inform* 2021;9(7):e27980) doi:[10.2196/27980](https://doi.org/10.2196/27980)

## KEYWORDS

human genetics; external quality assessment; quality control; proficiency testing; frequency; genetic testing; rare diseases; cost-effectiveness; surveillance, public health authorities; public health; health informatics; medical informatics; genetics; human genetics; algorithm

## Introduction

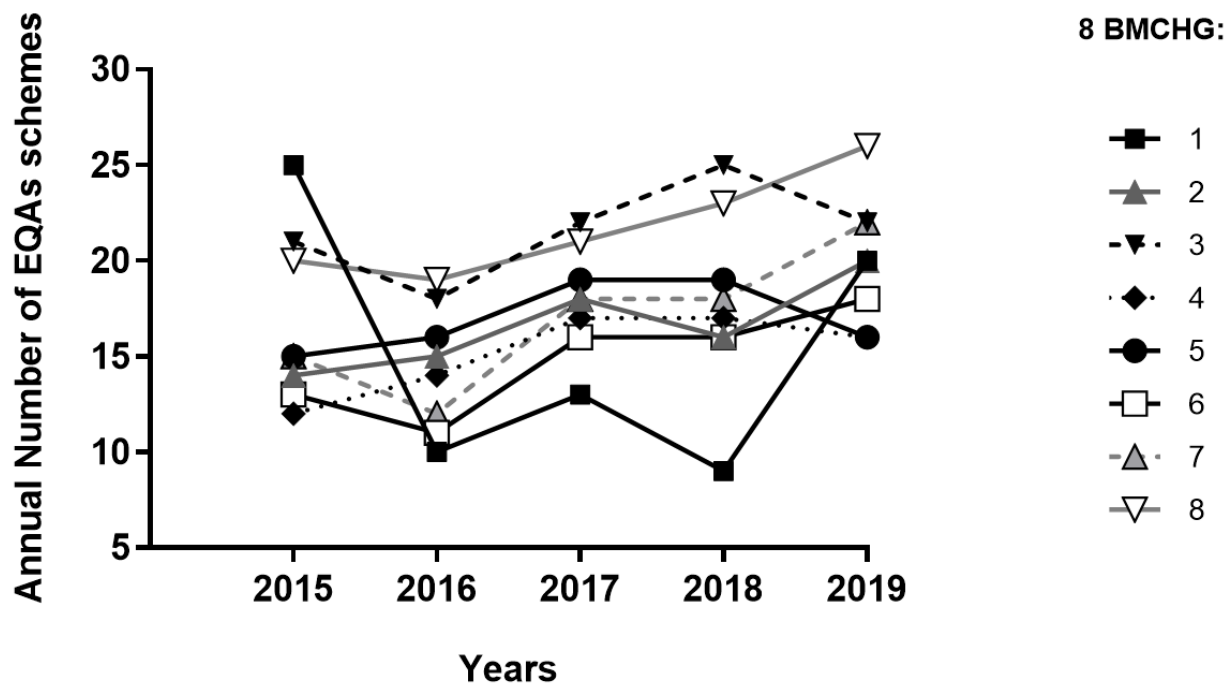
Rare diseases are life-threatening or chronically debilitating conditions affecting less than 5 per 10,000 people [1]. At least 80% of them have a genetic origin and 50% to 75% affect children [2,3]. Despite the discovery of more than 200 new genes every year, the diagnosis of rare diseases often remains delayed because of their complexity and low prevalence [1,4-6].

There is a willingness of European governments to develop harmonized guidelines to improve the quality of genetic testing, particularly by stimulating medical laboratories to acquire accreditation and to participate in external quality assessments (EQAs) [7,8]. Indeed, the participation in EQA schemes is efficient for assessing and improving health care quality, as it allows performance comparisons and the identification of specific problems, areas for improvement, and training needs [9-11]. Besides, it enables monitoring of the compliance with best practice guidelines and is required for the accreditation of medical laboratories according to the ISO 15189 standard [12-14]. It has been shown that the quality of genetic services in Europe can still improve [10,15]. Nevertheless, several recent studies performed in European laboratories for cancer testing have pointed out the positive influence of participation in EQAs on laboratories' performance [16-18]. In several countries such

as in Belgium, the accreditation of the genetic laboratories is a requisite for reimbursement of the diagnostic tests. However, the EQA of the laboratories is still hampered by a lack of a harmonized European framework (numerous and heterogeneous quality schemes, lack of reference systems, and different Member State regulations) [19,20]. Similar concerns have been raised in a recent Belgian study focusing on the frequency of participation in EQA schemes in the fields of molecular microbiology, hematology, and pathology [20]. The authors proposed to harmonize the frequency of participation to quality controls [20]. Indeed, the ISO 15189 standard states that "the laboratory shall participate in an inter-laboratory comparison program (such as an EQA program or proficiency testing program) appropriate to the examination and interpretations of examination results," but does not give precise instructions [13]. This lack of clear national and international directives leads to uneven participation of the Medical Centers of Human Genetics in quality controls [21]. [Figure 1](#) illustrates this phenomenon with the participation of the Belgian Medical Centers of Human Genetics (BMCHG) in EQAs between 2015 and 2019.

To address this lack, we have developed Belgian guidelines about the minimal frequency of participation in EQA schemes for hereditary rare diseases, with reference to international recommendations and national laboratory practices.

**Figure 1.** Evolution of the participation of the 8 BMCHG to the inventoried EQA schemes between 2015 and 2019. BMCHG: Belgian Medical Centers of Human Genetics; EQA: external quality assessment.



## Methods

### Context of the Study

In the context of the Belgian National Plan for Rare Diseases, the Belgian National Institute for Health (Sciensano [22]) is responsible for the harmonization of the quality management system for rare disease diagnostics within the BMCHG.

### Data Collection

In 2018, Sciensano performed a preliminary inventory of 90 EQA schemes related to rare diseases and that the BMCHG participate in. Of note, in the case of cancers, only EQA schemes for rare hereditary cancers were considered, while schemes for somatic mutation detection were excluded. In 2019, Sciensano collected retrospective data about the annual participation of the BMCHG in the inventoried EQA schemes between 2015 and 2019.

### Guidelines for the Participation to Genetic EQA Schemes

To structure and harmonize the frequency of participation of the BMCHG in EQA schemes focused on the genetic diagnosis of rare diseases, a working group composed of two representatives for each of the 8 BMCHG was established by Sciensano in 2019, in consultation with the Belgian College of Human Genetics and Rare Diseases [23].

The working group developed recommendations about the minimal frequency of participation of the BMCHG in quality controls. These recommendations were accompanied by a decisional algorithm to help the BMCHG to plan their future participations in quality controls based on their own experience in the performance of the tests considered and the scopes of the available EQA schemes. Besides, attention was paid to recommendations on actions that should be undertaken in case of poor performance to EQA schemes and to the continuous follow-up and surveillance of the participation of the BMCHG to EQA schemes.

### Validation of the Guidelines

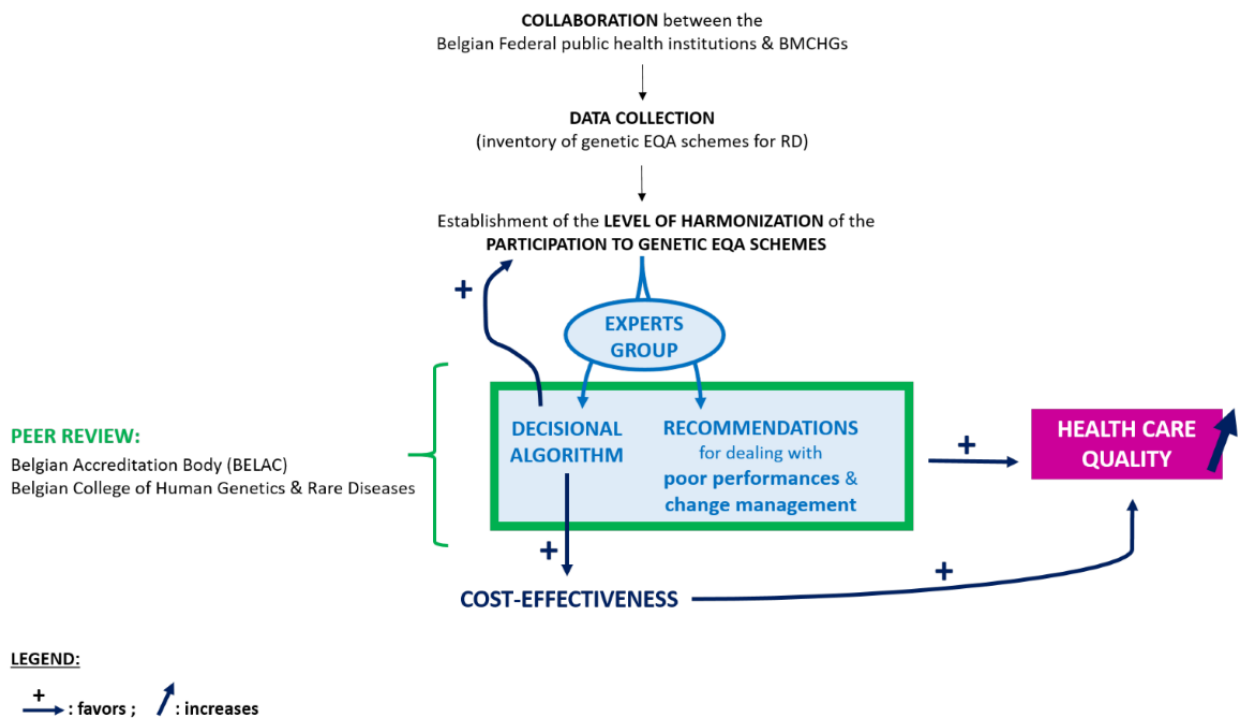
The inventory of 90 genetic EQA schemes focused on rare diseases was used by the working group for the validation of the decisional algorithm based on the routine BMCHG practice.

Besides, the opinion of three accreditation managers from the Belgian Accreditation Body regarding the whole guidelines' draft was requested. The final version of the guidelines was submitted in 2020 to the Belgian College of Human Genetics and Rare Diseases for evaluation and endorsement.

### Study Workflow

The global study workflow is illustrated in [Figure 2](#).

**Figure 2.** Workflow and achievements of the study. BMCHG: Belgian Medical Centers of Human Genetics; EQA: external quality assessment; RD: rare diseases.



**Statistical Analysis**

For some rare disorders, only a few requests are obtained on an annual basis. Due to this low number, the question now arises whether the number of routine analyses can be used as an indicator of performance. Indeed, in genetic testing, errors have a great impact on the patients' and their relatives' lives. It is therefore important to maintain a high quality level, even if every laboratory is subject to underlying errors. To this aim, an error rate of 1% has been set by the working group as the maximal error threshold to define the quality of the performance of genetic tests. This threshold has been determined based on the following:

- The error rate reported in May 2020 by the EQA provider European Molecular Quality Network [24] for its global data for the germline schemes organized by this provider during the past 5 years (2016-2020); mean analytical error rate 1.37% (unpublished data): This percentage is based on

the assessment of more than 33,132 genotypes during 11,044 participations in fully operational EQA schemes for germline mutation testing (technical, molecular pathology, and pilot schemes were excluded; each scheme assesses 3-4 samples). Of note, the mean analytical error rate is defined as any genotyping error that would lead to patient harm.

- Data published in the scientific literature: Indeed, error rates between 0.1% and 1% have been reported for high-throughput DNA sequencing technologies (eg, next generation sequencing) [25,26]. Raw data about error rate percentages published in 5 other peer-reviewed scientific papers for different types of situations (diseases, techniques, etc) [27-31] were also analyzed. Mean error rates and SDs with the number of scenarios investigated by the authors and confidence intervals are reported in Table 1. Based on this analysis it appears that the mean error rates fluctuate approximately between 0% and 4%.

**Table 1.** Data analysis about error rate percentages published in peer-reviewed papers for different types of situations.

Bibliographic references	Investigated scenarios, n <sup>a</sup>	Error rate (%), mean (SD)	95% CIs of the mean (%)
Hofgartner et al, 1999 [27]	8	0.38 (0.34)	0.10 to 0.67
Ewen et al, 2000 [28]	7	0.89 (0.82)	0.13 to 1.65
Bonin et al, 2004 [29]	4	2.20 (1.10)	0.46 to 3.94
Hoffman and Amos, 2005 [30]	8	0.35 (0.19)	0.19 to 0.50
Gilles et al, 2011 [31]	2	0.80 (0.38)	-2.63 to 4.23

<sup>a</sup>Number of investigated scenarios for different types of situations reported in the literature.

To determine the sufficient number of analyses needed to have a maximal error rate of 1%, assuming that the laboratory is performing well, the distribution of possible error rates for a

certain performance statistic was modeled using the *proportion* library of R software (version 3.6.1; R Foundation for Statistical Computing).

A Bayesian model with noninformative prior was used for having a rate of 100% correct analyses for a certain number of analyses and a rate of  $(n - 1) / n$  correct analyses for a certain number of analyses  $n$ . Details concerning the statistical model can be found in [Multimedia Appendix 1](#).

## Results

### Scope of Guidelines

As this study was funded by a grant dedicated to the improvement of the quality of the genetic testing in the BMCHG in the context of rare diseases, all developed guidelines are related to human genetics EQA schemes for rare hereditary diseases, including germline predispositions to cancers and adverse drug effects resulting from pharmacogenomic variants [32].

### EQA Schemes Inventory

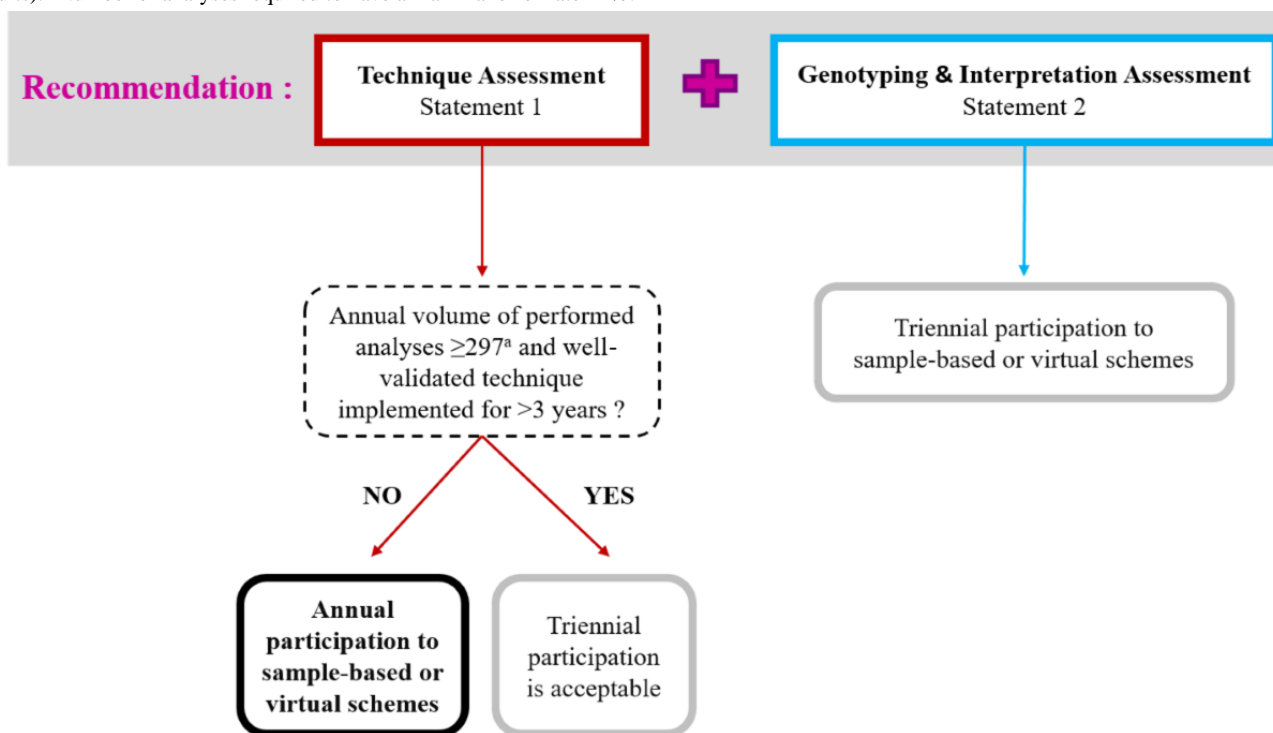
The EQA schemes inventoried during the preliminary phase of the study and considered by the working group for the establishment of the guidelines are mentioned in [Multimedia Appendix 2](#) [24,33-38]. They are focused on the diagnosis of 72 different rare diseases or specific genetic variants involved

in rare diseases. For each scheme, we have reported the aspects that are assessed (technique, genotyping, and interpretation of the results). The majority ( $n=65, 72\%$ ) of the EQA schemes are assessing the technique, genotyping, and interpretation. A total of 21 (23%) of the schemes are assessing both the technique and genotyping. A few of the schemes are covering both the genotyping and interpretation ( $n=1, 1\%$ ), only the technique ( $n=1, 1\%$ ), only the genotyping ( $n=1, 1\%$ ), or only the interpretation ( $n= 1, 1\%$ ). Of note, for 15 selected rare diseases or genetic variants, EQA schemes are offered by 2 or 3 providers, which increases the total number of inventoried EQA schemes ( $n=90$ ) used for the percentage calculations described here. These EQA schemes are marked with an asterisk in [Multimedia Appendix 2](#).

### Guidelines

These guidelines on the minimal frequency of participation in EQA schemes can be divided into three pillars: (1) general recommendations that constitute the backbone of the guidelines and cover the majority of situations encountered (statements 1-2 and [Figure 3](#)), (2) how to address poor performances (statements 3-4), and (4) follow-up and surveillance (statements 5-6).

**Figure 3.** Decisional algorithm for the minimal frequency of participation to external quality assessment schemes (cf to statements 1 and 2 of the Results). <sup>a</sup>Number of analyses required to have a maximal error rate  $\leq 1\%$ .



### General Recommendations

The quality of all diagnostic tests offered by the BMCHG should be frequently assessed. Indeed, the annual rate of some specific analyses (eg, performed in the context of rare diseases) may be very low. However, given the large number of genetic diseases and the low frequency of most of them, an annual participation in all possible EQA schemes is neither feasible nor economically defensible. It is therefore important to run a quality assessment in a comprehensive, rather than an exhaustive way, to assess

the quality of services offered to the patients. We also took into account that the same techniques are applied for analyses of different rare disorders.

### Statement 1: Annual Assessment of All Techniques

The quality of all techniques and technological platforms used by the BMCHG should be annually assessed, even with EQA schemes only based on mock clinical cases, virtual images or variant call format files, or raw data sets. As mentioned by Brookman et al [39], if visual inspections are needed in daily



practice, virtual schemes are useful to test postanalytical performances, notably in the case of fluorescence in-situ hybridization and karyotype analysis.

If different EQA schemes exist to cover the same technique, the centers are free to implement a turnover between those EQAs as long as the technique is covered every year and the clinical indications at least every 3 years (cf statement 2).

An exception is made for well-validated methods implemented for more than 3 years (*in-house* or using a commercial CE [European Conformity]-labelled kit) and for which at least 297 tests per year are performed, meaning that the maximal error rate of the analysis is between 0% and 1%. If those specific conditions are met, a triennial participation to an EQA is sufficient to evaluate a specific technique, as long as the methodology does not change. Of note, the number of tests have been deducted from the Bayesian statistical model performed for the distribution of the maximal error rates (see [Multimedia Appendix 1](#)). The rationale for triennial participation has to be properly documented in the center's quality management system [40].

#### Statement 2: Genotyping and Interpretation Assessment

A triennial assessment of the genotyping and interpretation for the detection of specific germline mutation diseases and pharmacogenomics is considered sufficient as long as the technique involved is covered (cf statement 1). This is also true when EQA schemes only assess the clinical interpretation based on virtual clinical cases or images.

#### How to Address Poor Performances

##### Statement 3: Identification of Errors' Origins

Poor EQA performances because of analytical or clerical errors (eg, copy and paste mistakes) have to be discussed internally. All actions (cause analysis, corrective, and preventive actions) carried out in response to the poor evaluation must be properly documented according to the center's quality management system procedures.

##### Statement 4: Poor Performances With an Impact on the Diagnosis

In case of poor EQA performance due to genotyping or critical interpretation errors that impacts the diagnosis, the center has to participate in an EQA the following year. Actions taken to avoid future errors have to be documented in the quality management system of the centers.

#### Follow-up and Surveillance

##### Statement 5: Management of Changes in Activities and EQA Schemes' Availability

It is the responsibility of the Medical Centers of Human Genetics to regularly review and adapt their participation to EQA schemes based on the present guidelines, changes in activities or infrastructure (eg, significant changes in the annual volume of tests and gene panels or modifications in the technique or analytical equipment), and new schemes introduced on the market. This should be notified in their quality management system.

##### Statement 6: Implication of Public Health Authorities

Public health authorities can play a key role in the improvement and follow-up of the activities, quality, and cost-effectiveness of medical laboratories such as the Medical Centers of Human Genetics. For instance, the Belgian National Institute for Health, called Sciensano, will annually coordinate the participation of the BMCHG to EQA schemes focused on rare diseases and hereditary cancers, ensure the reimbursement of participation fees, and monitor the outcomes. To provide this service, the data regarding the participation of BMCHG will be used to forecast the annual global budget dedicated to the reimbursement of participation fees. This information will then be communicated to the Belgian health care authorities. Besides, Sciensano and the working group will also regularly review and update the Belgian guidelines according to the evolution of the centers' activities, scientific developments, and EQAs' availability.

In the coming years, the collected data about the participation frequencies of the BMCHG in EQA schemes will be included into the Belgian genetic tests database, developed by Sciensano, in collaboration with the BMCHG.

##### Impact of the Guidelines on Health Care Costs

We have studied the impact that the establishment of harmonized guidelines on the minimal frequency of participation of the BMCHG in EQA schemes may have on national health care and genetic centers' expenditures. To this aim, three different scenarios have been compared:

1. The cost estimation if the BMCHG would annually participate to all EQAs included in their assessment scope among the inventoried EQA schemes focused on 72 rare diseases or genetic variants (fictitious scenario)
2. The participation costs of the BMCHG to the same EQAs as in 2019 (in absence of guidelines)
3. The prediction of the annual BMCHGs participation costs (mean over 2020, 2021, and 2022) for the EQAs included in their assessment scope, following the participation frequencies proposed in the guidelines

Based on the costs of the different EQA schemes, the estimated annual expenditures in these three scenarios were €17,400 (~US \$140,444), €82,000 (~US \$98,096) and €70,600 (~US \$84,458), respectively.

These estimations show that the rationalization of the frequency of participation proposed in these guidelines (third scenario), based on the types of EQA schemes and results of previous participation, enables a reduction in global annual participation costs of 14% for the 8 BMCHG.

Based on the developed guidelines on the minimal frequency of participation and current commercial EQA prices, we were able to estimate that a mean annual budget of €9000 (~US \$10,900) is required for each BMCHG to cover the fees requested by the provider to participate in the EQA schemes included in their assessment scope.

## Discussion

### Principal Results and Strengths

A regular participation in quality controls is mandatory for the accreditation of medical laboratories under the ISO 15189 standard [12,14,40]. Accreditation itself is a requisite for the reimbursement of genetic tests in Belgium. However, no Belgian instructions on the required frequency of participation in rare diseases diagnostic and genetic testing EQAs were available prior to this study. This study can be considered as the first Belgian harmonized quality update in terms of frequency of participation in proficiency testing in the field of human genetics.

These guidelines present six main strengths. First, they are based on European recommendations [41-43] and on the clinical and laboratory practice to make them as broad and consistent as possible. Second, they have been developed by a working group composed of representatives of all BMCHG to ensure a harmonization at the national level. Besides, these members have different professional backgrounds and tasks that enabled us to collect the opinions of all stakeholders involved in the performance of different types of genetic tests (molecular, cytogenetic, and biochemical), quality management, and in the interaction with the Belgian health care authorities. Third, a distinction was made based on the aspects assessed by the EQA schemes (technique, analysis, or interpretation) to draft guidelines as relevant as possible. Fourth, a large number of available genetic EQA schemes from accredited providers has been considered. This emphasizes the importance of assessing the quality of highly specific tests performed at a relatively low annual volume in the context of rare diseases. Fifth, a statistical model was used to estimate the probability of a laboratory to make a mistake according to the number of analyses that are performed per year. This new model may help other laboratories to define the minimal number of analyses required to indicate that the experience of a laboratory can be taken into account as a reliable performance indicator. Finally, the guidelines have been approved by the Belgian College of Human Genetics and Rare Diseases and are in accordance with the statements of the ISO 15189 standard referring to the validation of analytical methods [13]. This ensures their clinical relevance and legal accreditation aspects.

Participating in a large number of different EQAs for rare diseases is worthwhile, as it has a role in controlling performance and guarantees permanent education. Furthermore, participating in international EQA schemes enables the performances of a large number of the Centers of Human Genetics to be compared and evaluated by a wide range of international experts. However, taking part in a large number of EQAs is a lot of work and time-consuming. Hence, a balance had to be sought between usefulness and burden. These new Belgian guidelines will improve the harmonization and structuring of the BMCHG quality management system and help the laboratories to identify the EQA schemes that they should participate in based on the evolution of their activities and type of EQA schemes considered. They might also serve as basis for the Belgian Accreditation Body for accreditation

assessments and for the Belgian health care authorities to estimate the necessary budget that should be foreseen and attributed by the National Institute for Health and Disability Insurance to the BMCHG to cover participation fees.

### Comparison With Prior Work

Similar recommendations have already been developed by other countries, for instance, Dutch, Slovenian, and Estonian laboratories have to participate in a minimum of one EQA scheme for each accredited analysis of their scope during an accreditation cycle (during 3 years, till the suspension of the accreditation, and during 5 years, respectively), while other (eg, Lithuanian) laboratories are requested to participate twice during this period of time or every year for specific fields [43]. It is unfortunate that no European consensus exists at this time [19]. However, we hope that the development of guidelines on this topic in different European countries should be a catalyst to the initiation of a general reflection on the harmonization of the quality assessment of genetic testing at a European level.

Our guidelines reflect the opinion that the scope of quality controls should be broad enough to cover all methods, technologies, and tests included in the scope of the centers. It is not acceptable that a laboratory would only be accredited for a (small) fraction of its testing offers and thus avoid EQA participation.

### Limitations

Regarding the limits of this study, we have to mention that these guidelines only concern EQA schemes from accredited providers. Ring tests [44] to which BMCHG may also participate in with a small number of other Belgian or foreign genetic centers were excluded. Nonetheless, the preliminary phase of the study revealed that approximately 30% of the quality controls to which the BMCHG participate in are ring tests. They were not considered in this study because we wanted to give priority to EQAs offered by accredited providers. Ring tests are often highly specific and involve a limited number of participants. The difficulty to get enough test material for all participants make the standardization of their organization difficult. However, this opens the door to future improvements in the harmonization process of the quality management of human genetic analyses when no formal EQA scheme is available.

Another limitation is that the majority of the EQAs considered are specific for hereditary rare diseases and not for all diseases.

Finally, the guidelines have been developed at the Belgian level, without asking the opinions of foreign experts. However, several members of the working group act as assessors in international schemes and have good insights into practice, evaluation, and (poor) performance management.

### Conclusion

These first Belgian guidelines will help the BMCHG to improve their quality management system with recommendation on the frequency of participation in EQA schemes and on dealing with poor performance and change management. Moreover, they help the Belgian health care authorities to estimate the budget required to cover the participation of the BMCHG in EQAs.

We are convinced that these Belgian guidelines could be used by foreign human genetics medical centers and can serve as a starting point for discussion about the harmonization of quality processes at a broader level.

## Acknowledgments

The authors thank the BMCHG, the Belgian College of Human Genetics and Rare Diseases, and the Belgian Accreditation body for their contribution to this study and pertinent suggestions during the redaction of the guidelines. JL and NMV are scientific collaborators from Sciensano and supported by the Belgian National Institute for Health and Disability Insurance (grants W4043.0100.6 and W4043.0100.8).

All authors confirmed they have contributed to the intellectual content of this paper, and they have met the following four requirements: (1) substantial contributions to the conception or design of the study, or the acquisition, analysis, or interpretation of data for the study; (2) drafting the study or revising it critically for important intellectual content; (3) final approval of the version to be published; and (4) agreement to be accountable for all aspects of the study in ensuring that questions related to the accuracy or integrity of any part of the study are appropriately investigated and resolved.

## Conflicts of Interest

None declared.

### Multimedia Appendix 1

Statistical modeling of the maximal error rates when performing genetic analyses for rare diseases.

[\[DOCX File , 62 KB - medinform\\_v9i7e27980\\_app1.docx \]](#)

### Multimedia Appendix 2

List of the inventoried external quality assessment schemes considered for the establishment of the guidelines. EQA schemes offered by several providers are marked with an asterisk.

[\[DOCX File , 66 KB - medinform\\_v9i7e27980\\_app2.docx \]](#)

## References

1. Evangelista T, Hedley V, Atalaia A, Johnson M, Lynn S, Le Cam Y, et al. The context for the thematic grouping of rare diseases to facilitate the establishment of European Reference Networks. *Orphanet J Rare Dis* 2016 Feb 24;11:17 [FREE Full text] [doi: [10.1186/s13023-016-0398-y](https://doi.org/10.1186/s13023-016-0398-y)] [Medline: [26911987](https://pubmed.ncbi.nlm.nih.gov/26911987/)]
2. Bavisetty S, Grody WW, Yazdani S. Emergence of pediatric rare diseases: review of present policies and opportunities for improvement. *Rare Dis* 2013;1:e23579 [FREE Full text] [doi: [10.4161/rdis.23579](https://doi.org/10.4161/rdis.23579)] [Medline: [25002987](https://pubmed.ncbi.nlm.nih.gov/25002987/)]
3. Mazzucato M, Visonà Dalla Pozza L, Minichiello C, Manea S, Barbieri S, Toto E, et al. The epidemiology of transition into adulthood of rare diseases patients: results from a population-based registry. *Int J Environ Res Public Health* 2018 Oct 10;15(10):2212 [FREE Full text] [doi: [10.3390/ijerph15102212](https://doi.org/10.3390/ijerph15102212)] [Medline: [30309015](https://pubmed.ncbi.nlm.nih.gov/30309015/)]
4. Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, et al. International cooperation to enable the diagnosis of all rare genetic diseases. *Am J Hum Genet* 2017 May 04;100(5):695-705 [FREE Full text] [doi: [10.1016/j.ajhg.2017.04.003](https://doi.org/10.1016/j.ajhg.2017.04.003)] [Medline: [28475856](https://pubmed.ncbi.nlm.nih.gov/28475856/)]
5. Gainotti S, Mascalzoni D, Bros-Facer V, Petrini C, Florida G, Roos M, et al. Meeting patients' right to the correct diagnosis: ongoing international initiatives on undiagnosed rare diseases and ethical and social issues. *Int J Environ Res Public Health* 2018 Sep 21;15(10):2072 [FREE Full text] [doi: [10.3390/ijerph15102072](https://doi.org/10.3390/ijerph15102072)] [Medline: [30248891](https://pubmed.ncbi.nlm.nih.gov/30248891/)]
6. Dharssi S, Wong-Rieger D, Harold M, Terry S. Review of 11 national policies for rare diseases in the context of key patient needs. *Orphanet J Rare Dis* 2017 Mar 31;12(1):63 [FREE Full text] [doi: [10.1186/s13023-017-0618-0](https://doi.org/10.1186/s13023-017-0618-0)] [Medline: [28359278](https://pubmed.ncbi.nlm.nih.gov/28359278/)]
7. Berwouts S, Fanning K, Morris MA, Barton DE, Dequeker E. Quality assurance practices in Europe: a survey of molecular genetic testing laboratories. *Eur J Hum Genet* 2012 Nov;20(11):1118-1126 [FREE Full text] [doi: [10.1038/ejhg.2012.125](https://doi.org/10.1038/ejhg.2012.125)] [Medline: [22739339](https://pubmed.ncbi.nlm.nih.gov/22739339/)]
8. Content sheet 10-1: overview of External Quality Assessment (EQA). World Health Organization. 2011 Jan 01. URL: [https://terrance.who.int/mediacentre/data/ebola/training-packages/LQMS/10\\_b\\_eqa\\_contents.pdf](https://terrance.who.int/mediacentre/data/ebola/training-packages/LQMS/10_b_eqa_contents.pdf) [accessed 2021-06-16]
9. Alkhenizan A, Shaw C. Impact of accreditation on the quality of healthcare services: a systematic review of the literature. *Ann Saudi Med* 2011;31(4):407-416 [FREE Full text] [doi: [10.4103/0256-4947.83204](https://doi.org/10.4103/0256-4947.83204)] [Medline: [21808119](https://pubmed.ncbi.nlm.nih.gov/21808119/)]
10. Cassiman J. Can (should) molecular diagnostic labs improve the quality of their services? *Eur J Hum Genet* 2012 Nov;20(11):1103-1104 [FREE Full text] [doi: [10.1038/ejhg.2012.126](https://doi.org/10.1038/ejhg.2012.126)] [Medline: [22739345](https://pubmed.ncbi.nlm.nih.gov/22739345/)]
11. Tembuysen L, Tack V, Zwaenepoel K, Pauwels P, Miller K, Bubendorf L, et al. The relevance of external quality assessment for molecular testing for ALK positive non-small cell lung cancer: results from two pilot rounds show room for optimization. *PLoS One* 2014;9(11):e112159 [FREE Full text] [doi: [10.1371/journal.pone.0112159](https://doi.org/10.1371/journal.pone.0112159)] [Medline: [25386659](https://pubmed.ncbi.nlm.nih.gov/25386659/)]

12. Claustres M, Kožich V, Dequeker E, Fowler B, Hehir-Kwa JY, Miller K, European Society of Human Genetics. Recommendations for reporting results of diagnostic genetic testing (biochemical, cytogenetic and molecular genetic). *Eur J Hum Genet* 2014 Feb;22(2):160-170 [FREE Full text] [doi: [10.1038/ejhg.2013.125](https://doi.org/10.1038/ejhg.2013.125)] [Medline: [23942201](https://pubmed.ncbi.nlm.nih.gov/23942201/)]
13. ISO 15189:2012 Laboratoires de biologie médicale — Exigences concernant la qualité et la compétence. International Organization for Standardization. 2012 Nov. URL: [http://www.iso.org/iso/fr/catalogue\\_detail?csnumber=56115](http://www.iso.org/iso/fr/catalogue_detail?csnumber=56115) [accessed 2020-10-22]
14. ILAC policy for participation in proficiency testing activities. International Laboratory Accreditation Cooperation. 2014 Jun. URL: [https://ilac.org/latest\\_ilac\\_news/ilac-p9062014-published/](https://ilac.org/latest_ilac_news/ilac-p9062014-published/) [accessed 2020-10-22]
15. Dequeker E. Quality assurance in genetic laboratories. In: Patrinos GP, editor. *Molecular Diagnostics*. Amsterdam, Netherlands: Elsevier Ltd; 2017:493-500.
16. Tack V, Ligtenberg MJL, Siebers AG, Rombout PDM, Dabir PD, Weren RDA, et al. RAS testing for colorectal cancer patients is reliable in European laboratories that pass external quality assessment. *Virchows Arch* 2018 May;472(5):717-725. [doi: [10.1007/s00428-017-2291-z](https://doi.org/10.1007/s00428-017-2291-z)] [Medline: [29333594](https://pubmed.ncbi.nlm.nih.gov/29333594/)]
17. Keppens C, Dufrainig K, van Krieken HJ, Siebers AG, Kafatos G, Lowe K, et al. European follow-up of incorrect biomarker results for colorectal cancer demonstrates the importance of quality improvement projects. *Virchows Arch* 2019 Jul;475(1):25-37. [doi: [10.1007/s00428-019-02525-9](https://doi.org/10.1007/s00428-019-02525-9)] [Medline: [30719547](https://pubmed.ncbi.nlm.nih.gov/30719547/)]
18. Keppens C, Tack V, Hart N, Tembuysen L, Ryska A, Pauwels P, EQA assessors expert group. A stitch in time saves nine: external quality assessment rounds demonstrate improved quality of biomarker analysis in lung cancer. *Oncotarget* 2018 Apr 17;9(29):20524-20538 [FREE Full text] [doi: [10.18632/oncotarget.24980](https://doi.org/10.18632/oncotarget.24980)] [Medline: [29755669](https://pubmed.ncbi.nlm.nih.gov/29755669/)]
19. Cassiman J. EuroGentest NoE, the ESHG, and genetic services. *Eur J Hum Genet* 2017 Dec;25(s2):S47-S49 [FREE Full text] [doi: [10.1038/ejhg.2017.155](https://doi.org/10.1038/ejhg.2017.155)] [Medline: [29297885](https://pubmed.ncbi.nlm.nih.gov/29297885/)]
20. Dufrainig K, Lierman E, Vankeerberghen A, Franke S, Dequeker E. External quality assessment for molecular diagnostic laboratories in Belgium: can we improve it? *Accred Qual Assur* 2019 Nov 25;25(1):39-49. [doi: [10.1007/s00769-019-01410-x](https://doi.org/10.1007/s00769-019-01410-x)]
21. Matthijs G, Limaye N. Belgian Medical Centers for Human Genetics. Belgian Society for Human Genetics. 2021. URL: <https://www.beshg.be/partners/centers> [accessed 2021-06-16]
22. Sciensano. 2021. URL: <https://www.sciensano.be/> [accessed 2021-06-16]
23. Collège belge de Génétique. 2021. URL: <https://www.college-genetics.be/> [accessed 2021-06-16]
24. European Molecular Genetics Quality Network. URL: <https://www.emqn.org> [accessed 2021-06-16]
25. Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat Rev Genet* 2018 May;19(5):269-285 [FREE Full text] [doi: [10.1038/nrg.2017.117](https://doi.org/10.1038/nrg.2017.117)] [Medline: [29576615](https://pubmed.ncbi.nlm.nih.gov/29576615/)]
26. Petrackova A, Vasinek M, Sedlarikova L, Dyskova T, Schneiderova P, Novosad T, et al. Standardization of sequencing coverage depth in NGS: recommendation for detection of clonal and subclonal mutations in cancer diagnostics. *Front Oncol* 2019;9:851. [doi: [10.3389/fonc.2019.00851](https://doi.org/10.3389/fonc.2019.00851)] [Medline: [31552176](https://pubmed.ncbi.nlm.nih.gov/31552176/)]
27. Hofgärtner WT, Tait JF. Frequency of problems during clinical molecular-genetic testing. *Am J Clin Pathol* 1999 Jul;112(1):14-21. [doi: [10.1093/ajcp/112.1.14](https://doi.org/10.1093/ajcp/112.1.14)] [Medline: [10396281](https://pubmed.ncbi.nlm.nih.gov/10396281/)]
28. Ewen KR, Bahlo M, Treloar SA, Levinson DF, Mowry B, Barlow JW, et al. Identification and analysis of error types in high-throughput genotyping. *Am J Hum Genet* 2000 Sep;67(3):727-736 [FREE Full text] [doi: [10.1086/303048](https://doi.org/10.1086/303048)] [Medline: [10924406](https://pubmed.ncbi.nlm.nih.gov/10924406/)]
29. Bonin A, Bellemain E, Bronken Eidesen P, Pompanon F, Brochmann C, Taberlet P. How to track and assess genotyping errors in population genetics studies. *Mol Ecol* 2004 Nov;13(11):3261-3273. [doi: [10.1111/j.1365-294X.2004.02346.x](https://doi.org/10.1111/j.1365-294X.2004.02346.x)] [Medline: [15487987](https://pubmed.ncbi.nlm.nih.gov/15487987/)]
30. Hoffman JI, Amos W. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Mol Ecol* 2005 Feb;14(2):599-612. [doi: [10.1111/j.1365-294X.2004.02419.x](https://doi.org/10.1111/j.1365-294X.2004.02419.x)] [Medline: [15660949](https://pubmed.ncbi.nlm.nih.gov/15660949/)]
31. Gilles A, Megléc E, Pech N, Ferreira S, Malausa T, Martin J. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* 2011 May 19;12:245 [FREE Full text] [doi: [10.1186/1471-2164-12-245](https://doi.org/10.1186/1471-2164-12-245)] [Medline: [21592414](https://pubmed.ncbi.nlm.nih.gov/21592414/)]
32. Roden DM, Wilke RA, Kroemer HK, Stein CM. Pharmacogenomics: the genetics of variable drug responses. *Circulation* 2011 Apr 19;123(15):1661-1670 [FREE Full text] [doi: [10.1161/CIRCULATIONAHA.109.914820](https://doi.org/10.1161/CIRCULATIONAHA.109.914820)] [Medline: [21502584](https://pubmed.ncbi.nlm.nih.gov/21502584/)]
33. Our EQAs. GenQA. 2021. URL: <https://www.genqa.org/eqa> [accessed 2021-06-16]
34. Biomedical Quality Assurance KU Leuven. Leuven: KU Leuven Biomedical Quality Assurance; 2020. URL: <https://eqascheme.org/> [accessed 2021-06-16]
35. 2020 surveys and anatomic pathology education programs. College of American Pathologists. 2020. URL: <https://www.cap.org/laboratory-improvement/catalogs-ordering-and-shipping> [accessed 2021-06-16]
36. United Kingdom National External Quality Assessment Service. URL: <https://www.ukneqash.org/> [accessed 2021-06-16]
37. Schellenberg I, Spannagl M, Hunfeld KP. EQAS Program 2021. INSTAND. 2021 Mar 01. URL: <https://www.instand-ev.de/en/instand-eqas/eqas-program-2021/> [accessed 2021-06-16]
38. Ringversuche. Referenzinstitut für Bioanalytik. 2021. URL: <https://www.rfb.bio/cgi/surveys> [accessed 2021-06-16]



39. Brookman B, Butler O, Koch M, Noblett T, Örnemark U, Patriarca M, et al. Proficiency testing in analytical chemistry, microbiology and laboratory medicine: discussions on current practice and future directions. *Accred Qual Assur* 2015 Apr 8;20(4):339-344. [doi: [10.1007/s00769-015-1120-9](https://doi.org/10.1007/s00769-015-1120-9)]
40. Essais d'aptitude (proficiency testing): lignes directrices en matière de participation et évaluation des performances dans le cadre des adultes d'accréditation. FPS Economy. Brussels: BELAC; 2011 Feb 01. URL: <https://economie.fgov.be/sites/default/files/Files/Publications/files/Belac-FR/2-106-FR.pdf> [accessed 2020-10-22]
41. OECD guidelines for quality assurance in molecular genetic testing. Organisation for Economic Co-operation and Development. Paris: OECD; 2007. URL: <http://www.oecd.org/science/emerging-tech/38839788.pdf> [accessed 2021-06-16]
42. EA-4/18 rev.00 – Guidance on the level and frequency of proficiency testing participation. Accredia. 2010 Jun. URL: <https://www.accredia.it/en/documento/ea-418-rev-00-guidance-on-the-level-and-frequency-of-proficiency-testing-participation/> [accessed 2021-06-16]
43. Örnemark U, Fostel H, Straub R, van de Kreeke J. Policies, requirements and surveys concerning frequency for participation in proficiency testing schemes. *Accred Qual Assur* 2004 Sep 4;9(11-12):729-732. [doi: [10.1007/s00769-004-0858-2](https://doi.org/10.1007/s00769-004-0858-2)]
44. Vandeveldel NM. Strategies to improve the quality of reference networks for rare diseases. *Translational Sci Rare Dis* 2020 Aug 03;5(1-2):59-79. [doi: [10.3233/trd-190032](https://doi.org/10.3233/trd-190032)]

## Abbreviations

**BMCHG:** Belgian Medical Centers of Human Genetics

**CE:** European Conformity

**EQA:** external quality assessment

*Edited by C Lovis; submitted 09.03.21; peer-reviewed by K Cato, A Mavragani; accepted 25.04.21; published 12.07.21.*

*Please cite as:*

*Lantoine J, Brysse A, Dideberg V, Claes K, Symoens S, Coucke W, Benoit V, Rombout S, De Rycke M, Seneca S, Van Laer L, Wuyts W, Corveleyn A, Van Den Bogaert K, Rydlewski C, Wilkin F, Ravoet M, Fastré E, Capron A, Vandeveldel NM*

*Frequency of Participation in External Quality Assessment Programs Focused on Rare Diseases: Belgian Guidelines for Human Genetics Centers*

*JMIR Med Inform* 2021;9(7):e27980

URL: <https://medinform.jmir.org/2021/7/e27980>

doi: [10.2196/27980](https://doi.org/10.2196/27980)

PMID: [34255700](https://pubmed.ncbi.nlm.nih.gov/34255700/)

©Joséphine Lantoine, Anne Brysse, Vinciane Dideberg, Kathleen Claes, Sofie Symoens, Wim Coucke, Valérie Benoit, Sonia Rombout, Martine De Rycke, Sara Seneca, Lut Van Laer, Wim Wuyts, Anniek Corveleyn, Kris Van Den Bogaert, Catherine Rydlewski, Françoise Wilkin, Marie Ravoet, Elodie Fastré, Arnaud Capron, Nathalie Monique Vandeveldel. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org>), 12.07.2021. 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

# Preferences of the Public for Sharing Health Data: Discrete Choice Experiment

Jennifer Viberg Johansson<sup>1</sup>, PhD; Heidi Beate Bentzen<sup>2</sup>, LLM; Nisha Shah<sup>3</sup>, MSc; Eik Haraldsdóttir<sup>4</sup>, MSc; Guðbjörg Andrea Jónsdóttir<sup>4</sup>, PhD; Jane Kaye<sup>3,5</sup>, PhD; Deborah Mascalonzi<sup>1,6</sup>, PhD; Jorien Veldwijk<sup>1,7</sup>, PhD

<sup>1</sup>Centre for Research Ethics & Bioethics, Department of Public Health and Caring Sciences, Uppsala Universitet, Uppsala, Sweden

<sup>2</sup>Norwegian Research Center for Computers and Law, Faculty of Law, University of Oslo, Oslo, Norway

<sup>3</sup>Centre for Health, Law, and Emerging Technologies, Faculty of Law, University of Oxford, Oxford, United Kingdom

<sup>4</sup>Social Science Research Institute, University of Iceland, Reykjavik, Iceland

<sup>5</sup>Centre for Health, Law and Emerging Technologies, Melbourne Law School, University of Melbourne, Melbourne, Australia

<sup>6</sup>Institute for Biomedicine, Bolzano, Italy

<sup>7</sup>Erasmus School of Health Policy and Management, Erasmus University, Rotterdam, Netherlands

**Corresponding Author:**

Jennifer Viberg Johansson, PhD

Centre for Research Ethics & Bioethics

Department of Public Health and Caring Sciences

Uppsala Universitet

Box 564

Uppsala, SE-751 22

Sweden

Phone: 46 184716288

Email: [jennifer.viberg@crb.uu.se](mailto:jennifer.viberg@crb.uu.se)

## Abstract

**Background:** Digital technological development in the last 20 years has led to significant growth in digital collection, use, and sharing of health data. To maintain public trust in the digital society and to enable acceptable policy-making in the future, it is important to investigate people's preferences for sharing digital health data.

**Objective:** The aim of this study is to elicit the preferences of the public in different Northern European countries (the United Kingdom, Norway, Iceland, and Sweden) for sharing health information in different contexts.

**Methods:** Respondents in this discrete choice experiment completed several *choice tasks*, in which they were asked if data sharing in the described hypothetical situation was acceptable to them. Latent class logistic regression models were used to determine attribute-level estimates and heterogeneity in preferences. We calculated the relative importance of the attributes and the predicted acceptability for different contexts in which the data were shared from the estimates.

**Results:** In the final analysis, we used 37.83% (1967/5199) questionnaires. All attributes influenced the respondents' willingness to share health information ( $P < .001$ ). The most important attribute was whether the respondents were informed about their data being shared. The possibility of opting out from sharing data was preferred over the opportunity to consent (opt-in). Four classes were identified in the latent class model, and the average probabilities of belonging were 27% for class 1, 32% for class 2, 23% for class 3, and 18% for class 4. The uptake probability varied between 14% and 85%, depending on the least to most preferred combination of levels.

**Conclusions:** Respondents from different countries have different preferences for sharing their health data regarding the value of a review process and the reason for their new use. Offering respondents information about the use of their data and the possibility to opt out is the most preferred governance mechanism.

(*JMIR Med Inform* 2021;9(7):e29614) doi:[10.2196/29614](https://doi.org/10.2196/29614)

**KEYWORDS**

preferences; discrete choice experiment; health data; secondary use; willingness to share

## Introduction

### Background

Digital technological development in the last 20 years has led to significant growth in digitally collecting, using, and sharing health data. This is partly due to the development and adoption of electronic medical records, genotyping, biobanking, and self-tracking applications via mobile devices. Different domains, such as health care, medical research, and technological and pharmaceutical companies, have become increasingly dependent on collecting and sharing data digitally to develop health care and new medical and technological products [1-3]. It has also led individuals to take a more active role in seeking out health information, thus managing and promoting their own health by having access to new health websites and mobile apps [4,5].

As different domains are dependent on public data, it is important to maintain public trust in the digital world. There is growing literature about preferences of the public, research participants, and patients for data sharing. People's willingness to share data for secondary use is dependent on contextual factors such as the type of data being linked, level of identification, and the new purpose for the data being shared [6-9]. A study that investigated the public's preferences regarding data linkage for health research showed that the type of information shared is the most important factor for people deciding whether they are willing to consent to the new use of their data [10]. Other studies show that people are interested in sharing their health information to improve health but are less willing to make data available to companies and insurance companies whose purpose for using the data may be unclear or not align with the public's expectations [7,11,12].

Previous research on people's willingness to share health data digitally has focused on one particular factor, such as the purpose of data sharing [6,7]. In addition, these studies in this area have been constrained by a specific context, such as looking at data movement within a health care setting [8,13,14]. To our knowledge, no studies have investigated how individuals' preferences change depending on the context in which health data are used, what type of information is involved, which different control mechanisms are considered appropriate for different contexts, and how an individual's acceptance of sharing data might change in response to changing contexts. There is a lack of knowledge on individuals' trade-off behavior in the current situation where data are linked across fields. Such studies are needed to provide a comprehensive understanding of the trade-offs between different factors, thereby informing policymaking and legal development therein [15].

We would like to evoke the need for a stated choice method that investigates behavioral intention to share health information, which is captured through trade-offs between varying levels, such as who the new user of the data is and for what reason the data will be shared. It is necessary to move away from the problematic single (fixed) scenario that captures people's behavioral intentions using Likert scales [15,16]. A stated choice method such as discrete choice experiment (DCE) require respondents to make a decision when the circumstances change. This method provides a deep understanding of context-specific

factors that people value. Moreover, using DCE as a method aligns with the theory by philosopher Helen Nissenbaum, which emphasizes that privacy is perceived and expected differently, depending on the norms and values surrounding the context [17].

### Objective

The aim of this study is to elicit the public's preferences and the heterogeneity in preferences in different Northern European countries (Sweden, Norway, Iceland, and the United Kingdom) to share health information in different contexts in order to determine what governance structures should be in place in the health sphere.

## Methods

### Discrete Choice Experiment

The DCE method is increasingly used in health care fields to quantify the preferences of specific target populations concerning any health-related product or service [18-20]. In a DCE, respondents are asked to complete several *choice tasks*. Each choice task describes the situation at hand. The description of the situation is based on its characteristics or *attributes* with systematically varying levels. In our case, respondents were asked to choose, to accept, or reject a situation several times. By monitoring their decisions in each choice task, their preferences were elicited. DCEs draw upon *random utility theory*, according to which an individual derives a certain *utility* for what the individual is confronted with in a choice task [21-24]. By comparing the attribute-level estimates, conclusions can be drawn about the importance of the attributes relative to each other. Moreover, the utility and acceptability of different data sharing situations can be calculated based on the attribute-level estimates from the experiment.

### DCE Development

The salient factors of digital health data sharing were identified through a three-step approach [25]. First, a literature review was performed to identify the possible factors that influence respondents' willingness to share their health data. Second, based on the output of the literature review, 14 focus group discussions were conducted with members of the public in the United Kingdom, Iceland, and Sweden. We carried out a comparative investigation of the respondents' attitudes, expectations, and beliefs about sharing health data. Focus group participants from all three countries mentioned the following factors as important when allowing data to be shared: level of identification, the reason for the new use, type of information being shared, the data subject being informed, and the monitoring of sharing. After the focus group discussion, a nominal group technique was used to ask participants to rank the importance of the different aspects or factors discussed in the focus groups, in addition to an a priori list of factors identified from the prior literature review. During the nominal group technique session, participants were asked to rank the potential factors from most to least important and then discuss them in the group. The authors and content experts thoroughly discussed the attributes and levels to confirm their relevance. On the basis of these steps, seven attributes were selected

(Textbox 1). During a two-hour webinar, content experts were asked to comment on the attributes as well as framing of the levels. Eight think-aloud interviews [26] were conducted (four in Sweden, two in Iceland, and two in the United Kingdom) to evaluate whether correct wording was used and whether the target population understood the attributes, levels, educational

information, and choice tasks. Finally, a two-day workshop was held where both method and content experts were invited to reach a consensus on chosen attributes and levels. Areas of expertise included law, philosophy, ethics, social science, and stated preference research.

**Textbox 1.** List of all attributes and levels included in the final discrete choice experiment.

<b>Attributes and Levels</b>	
1.	<p>Health information collector: different collectors can collect health information. The different collectors are as follows:</p> <ul style="list-style-type: none"> <li>• A technological company with which you have used a service, program, or application for your phone or computer. You may have used a service through the company's website, where you have entered information about yourself. Alternatively, you have downloaded an app to your phone, and it has collected information about your health.</li> <li>• An academic research project where you have participated and they have collected health information about you.</li> <li>• Your health care provider (hospital or general practitioner) who has collected health information about you regarding your care.</li> </ul>
2.	<p>Data user: your health information will be shared to a new data user. This new recipient may be:</p> <ul style="list-style-type: none"> <li>• A technological company that develops health app which can be used to predict diagnoses.</li> <li>• A pharmaceutical company that develops and manufactures new medicines.</li> <li>• An academic research project that produces new knowledge by testing hypotheses and theories about human health.</li> <li>• A national authority, for example, the public health authority or information and commissioner's office, which is responsible for the health of the population. They can track peoples' health through population registers to prevent disease.</li> </ul>
3.	<p>The reason of data use: this aspect describes the reason why the data user wants to have access to your health information. The different reasons may be:</p> <ul style="list-style-type: none"> <li>• Develop a new product or service. It can be a medical device, a drug, or app for your phone, or a new health service or program.</li> <li>• Promote, advertise, or market their product or service to personalize communication. For direct advertising to a specific target group for a new service or product.</li> <li>• Investigate a policy initiative. Your health information can provide a basis for a new policy initiative at a national level. It may be to improve services for a specific part of the population or to identify new preventive measures to improve public health.</li> <li>• Evaluate the quality of the data user's product or service, and planning resource distribution in the future.</li> </ul>
4.	<p>Information and consent: this aspect is about whether you will be informed if your health information is being shared.</p> <ul style="list-style-type: none"> <li>• You will not be informed that health information about you is being shared and used in a new context.</li> <li>• You will be informed that health information about you is being shared and used in a new context.</li> <li>• You will be informed that health information about you is being shared and used in a new context as well as be told that you can opt out.</li> <li>• You will be informed and asked to consent that health information about you is being shared and used in a new context.</li> </ul>
5.	<p>Review of data sharing: before your data are shared, there might be a review of the reason and how the data user will store and use your health information. The data user needs to apply for access to the health information. The reviewer makes a decision based on national law.</p> <ul style="list-style-type: none"> <li>• There will be no review of the data sharing.</li> <li>• A committee will review the transfer of your health information to the new context.</li> <li>• A committee will review the transfer and the use of your health information in the new context.</li> </ul>

A Bayesian D-efficient design was used for this DCE to strive for reliable parameter estimates [21,27,28]. The design was developed using NGene (version 1.2.1; ChoiceMetrics 2012). This is the most commonly used design strategy and is congruent with the guidelines of the International Society of Pharmacoeconomics and Outcome Research on good research practice [27]. Pilot testing priors based on best guesses were used to inform the design using 500 Halton draws and 1000 repetitions. For this design, we assumed that there would be no

interaction between attributes. The level balance (ie, all levels appearing an equal number of times) was optimized. The pilot design had a D-error of 0.31. A total of 28 unique choice tasks were generated and divided into four blocks. Respondents were randomly assigned to either block and answered seven choice tasks.

We pilot tested the draft questionnaire among our target population (n=50) in each of the four countries. The



attribute-level estimates that significantly contributed to the choice from the pilot study served as direct prior input for the design of the final DCE questionnaire.

### Questionnaire

The questionnaire consisted of three parts. The first part contained questions regarding demographic characteristics (eg, age, gender, educational level, self-reported health status, and long-term health conditions). The eHealth Literacy Scale is designed to assess people’s perceived skills at using information technology for health [29], and it comprises eight items assessing different aspects of eHealth literacy (eHL). Each item had five response categories: strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree.

The second part was the DCE. Each participant was given an alternative choice that they were asked to accept or reject. An additional level was added to the attribute *Information and*

*consent* in the final design. Therefore, the final DCE consisted of 32 unique choice tasks divided into four blocks, and each participant answered eight choice tasks for two types of health information (16 choice tasks). Before respondents were asked to complete the choice tasks, they received detailed information on the meaning of all attributes and levels, as well as an example of how to complete a choice task. This particular DCE topic describes a situation. It is not tradable in the ordinary sense (one product or service over another). Given that this topic is not tradable like regular DCE, each participant was given a choice alternative where the participant was asked to accept or reject. **Figure 1** shows a choice task with one situation. The remaining attributes changed in a systematic manner between the different levels.

The third part of the questionnaire related to trust in different domains and other people, attitudes toward new technology, and self-assessed eHL.

**Figure 1.** An example of a discrete choice experiment with one choice situation.

Imagine that lifestyle information about you has been collected. The organisation that collected your data is now proposing to share it. Please read the situation description below and let us know if you think that sharing this information in this situation is acceptable or not.

	Situation 1 of 8	
The organisation collecting my information is	an <b>academic research project.</b>	
They will share it with	a <b>technological company.</b>	
The reason they want to use my information is to	develop a <b>new product or service.</b>	
When they share my information, I will	be offered an <b>opt-out and information.</b>	
There will be	a committee that <b>reviews the sharing of information.</b>	
Do you think this situation is acceptable?	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No

### Study Population

Ethics approval was obtained before the start of the study from the Ethical Review Boards in the countries where this was required (University of Oxford Central University Ethics Committee REF: R63378/RE002; Swedish Ethical Review Authority Dnr 2020-00623; University of Iceland Science Ethics Committee VSH-2019-019).

The DCE was web-based, and respondents were invited to participate via the recruitment service SurveyEngine [30]. The recruitment company performed opt-in survey panels in Sweden, Norway, and the UK. The respondents decided on incentive models that worked best for their specific membership, such as cash, vouchers, virtual currencies, points for gift cards, or participation in raffles. The respondents received €1.50 (US \$1.80) for answering our survey. The Icelandic respondents were randomly selected from the Social Science Research Institute’s Online Panel at the University of Iceland. Respondents in the Social Science Research Institute Online

Panel were recruited through random samples drawn from the National Population Register in Iceland. A lottery to win one of the two gift vouchers of €65 (US \$77.40) was used as an incentive.

We aim to obtain a representative sample of the general population of each country in terms of gender and age. Data were collected from August to November 2020.

### Statistical Analysis

#### Descriptive Statistics

Descriptive statistics (means and frequencies) were used to summarize all the variables of interest. The overall level of eHL was calculated for each respondent. Individuals responding strongly disagree or disagree to one of the items were categorized as having inadequate eHL. Individuals responding with neither agree nor disagree with one of the items were categorized as having problematic eHL. Individuals responding

agree or strongly agree to all the items were categorized as having sufficient eHL.

One-way analysis of variance and nonparametric measures were used to test the differences between the personal characteristics of each country.

### Preferences for Sharing Health Information

The discrete choice data collected in the survey were first analyzed separately for each country using a binary logit model. The Swait-Louviere test was performed to investigate whether there were significant scale differences across samples from different countries [31]. Latent class models were then used to determine the attribute-level estimates and importance weights of the attributes. The latent class model identifies classes of respondents based on unobserved (latent) heterogeneity in preferences [32]. Akaike information criteria and log-likelihood were used to determine the best-fitting model [33]. All attributes were effect-coded [34], meaning that the reference category was coded as  $-1$ , and the sum of all the coded levels for each attribute was zero. A constant term was also estimated to quantify the utility associated with rejecting information sharing under the presented situation (Intercept). All results were considered statistically significant at  $P < .05$ .

The final utility function was as follows:

$$\begin{aligned}
 U = V_{rta\&b|c} + \varepsilon = & \beta_0|c * \text{reject}_{rta\&b|c} + \beta_1 * \\
 & \text{collector\_technological}_{rta\&b|c} + \beta_2 * \\
 & \text{collector\_research}_{rta\&b|c} + \beta_3 * \\
 & \text{user\_technological}_{rta\&b|c} + \beta_4 * \\
 & \text{user\_pharmaceutical}_{rta\&b|c} + \beta_5 * \text{user\_research}_{rta\&b|c} \\
 & + \beta_6 * \text{reason\_develop}_{rta\&b|c} + \beta_7 * \\
 & \text{reason\_promoting}_{rta\&b|c} + \beta_8 * \text{reason\_policy}_{rta\&b|c} + \\
 & \beta_9 * \text{information\_not\_informed}_{rta\&b|c} + \beta_{10} * \\
 & \text{information\_informed}_{rta\&b|c} + \beta_{11} * \\
 & \text{information\_opt-out}_{rta\&b|c} + \beta_{12} * \text{review\_no} \\
 & \text{review}_{rta\&b|c} + \beta_{13} * \text{review\_sharing}_{rta\&b|c} + \varepsilon
 \end{aligned}
 \tag{1}$$

$V$  is the observed utility of accepting to share health data with a second user based on what respondents  $r$  belonging to class  $c$  reported for the alternative  $a$  in choice task  $t$ . The  $\beta_0$  represents the alternative specific constant, and  $\beta_1$ - $\beta_{13}$  are attribute-level estimates that indicate the relative importance of each attribute level. Data cleaning and descriptive statistics were performed using R (version 4.0.2; R Core Team). The latent class logistic regression was performed with the econometric software NLogit 5.0 (Econometric Software, Inc), using 100 random draws. In latent class analysis, unobserved preference heterogeneity among respondents is modeled as discrete classes with similar preferences or choice patterns but with different variances across classes [35,36]. As the probability of a participant belonging to any specific class cannot be directly observed, the model searches for groups of respondents sharing similar choice patterns. Once choice patterns have been stratified into classes, the model could determine the probability of a participant with certain characteristics being assigned to each class (class assignment model). This separate logit model was fitted to

determine the associations between individual class membership and country. We also explored potential associations with other variables, such as age, sex, and E-HL. When individually added to the model, they all significantly contributed to latent class assignment. However, when adding multiple covariates into a one-class assignment model, we observed multicollinearity between the variables. As *country* was the most important variable for this overall study (and a necessity to include as we pooled data from multiple countries into one data set), we focused on that variable in this study. We will explore the impact of other variables separately in the analysis conducted on data from separate countries to avoid this collinearity caused by country differences.

### Relative Importance of the Attributes

Using the relative preference weights, that is, the attribute-level estimates from the DCE, we calculated the relative importance of the attributes. For each attribute, the total impact on utility was determined by subtracting the lowest from the highest estimate within each attribute. All attributes were divided according to the highest difference value. This provided a relative distance between the most important attributes and all other attributes.

### Acceptance Uptake

The acceptance uptake (also referred to as predicted probability [37], participation probability [38], predicted uptake [24,39], or subsequent uptake [40]) was calculated for different scenarios for sharing health data. This was determined for different potential scenarios and could inform future implementation strategies. Acceptance uptake can be understood as the probability that a participant would choose the described scenarios; alternatively, the number of respondents out of 100 that would accept the scenarios described. These scenarios represent existing or hypothetical scenarios. Using the attribute levels, scenarios based on specific data sharing and governance features were assembled. The utility for a specific scenario is calculated by using the following equation:

$$V_{\text{Scenario } 1} = \beta_A + \beta_B + \beta_C
 \tag{2}$$

The acceptance uptake, the probability of accepting, was then calculated by using the following equation:

$$\text{Acceptance uptake} = 1/(1 + \exp^{-V_{\text{Scenario } 1}})
 \tag{3}$$

## Results

### Respondents' Characteristics

In total, 5199 respondents answered the questionnaire (Sweden,  $n=1208$ ; Norway,  $n=928$ ; Iceland,  $n=2187$ ; United Kingdom,  $n=876$ ). Respondents who completed the survey in less than 5 minutes ( $n=97$ ) or did not complete the entire survey ( $n=3135$ ) were excluded. In the final analysis, we used 37.83% (1967/5199) of the questionnaires. The mean ages of the respondents were 50.4 years (SD 16.9) in Sweden, 48.3 years (SD 17.2) in Norway, 49.9 years (SD 15.9) in the United Kingdom, and 48.2 years (SD 17.2) in Iceland. Respondents

with university education included 36.2% (162/447) in Sweden, 39.3% (167/425) in Norway, 52.1% (232/445) in the United Kingdom, and 57.3% (287/501) in Iceland. Respondents with sufficient eHL included 30.6% (137/447) in Sweden, 22.8% (97/425) in Norway, 36.6% (163/445) in the United Kingdom, and 20.6% (103/501) in Iceland. The respondents' characteristics are presented in Tables 1-3.

**Table 1.** Descriptive statistics of the respondents presented as percentages, mean, or median with statistical testing between the different countries.

Variates	Sweden (n=481)	Norway (n=465)	United Kingdom (n=477)	Iceland (n=544)	P value (ANOVA <sup>a</sup> )
<b>Age (years)</b>					.15
Mean (SD)	50.3 (16.9)	48.1 (17.2)	49.6 (15.9)	48.3 (17.2)	
Median (range)	53 (18-88)	50 (18-84)	49 (18-90)	47 (19-88)	
Survey duration, mean (SD)	15.5 (11.5)	15.4 (10.2)	12.8 (7.36)	20.7 (14.9)	<.001

<sup>a</sup>ANOVA: analysis of variance.

**Table 2.** Descriptive statistics of the respondents presented as percentages with Chi-square testing between the different countries.

Variates	Sweden (n=438)	Norway (n=424)	United Kingdom (n=450)	Iceland (n=538)	P value (Chi-square test)
<b>Gender, n (%)</b>					.79
Female	219 (50)	226 (53.3)	236 (52.4)	268 (49.8)	
Male	218 (49.8)	197 (46.5)	214 (47.6)	268 (49.8)	
Other	1 (0.2)	1 (0.2)	0 (0)	2 (0.4)	
<b>General health status, n (%)</b>					<.001
Good	296 (67.6)	285 (67.2)	335 (74.4)	445 (82.7)	
<b>Chronic health condition, n (%)</b>					<.001
No	203 (46.3)	179 (42.2)	262 (58.2)	304 (56.5)	

**Table 3.** Descriptive statistics of the respondents presented as percentages with Kruskal-Wallis testing between the different countries.

Variates	Sweden (n=447)	Norway (n=425)	United Kingdom (n=445)	Iceland (n=501)	P value (Kruskal-Wallis test)
<b>Highest educational level, n (%)</b>					<.001
High school	251 (56.2)	234 (55.1)	129 (29)	181 (36.1)	
Primary school	34 (7.6)	24 (5.6)	84 (18.9)	33 (6.6)	
University	162 (36.2)	167 (39.3)	232 (52.1)	287 (57.3)	
<b>eHealth literacy, n (%)</b>					<.001
Insufficient	115 (25.7)	126 (29.6)	116 (26.1)	183 (36.5)	
Problematic	195 (43.6)	202 (47.5)	166 (37.3)	215 (42.9)	
Sufficient	137 (30.6)	97 (22.8)	163 (36.6)	103 (20.6)	
<b>How often they are using apps related to health, n (%)</b>					<.001
Daily	64 (14.3)	71 (16.7)	107 (24)	87 (17.4)	
Weekly	52 (11.6)	44 (10.4)	51 (11.5)	69 (13.8)	
Monthly or more seldom	121 (27.1)	152 (35.8)	57 (12.8)	145 (28.9)	
Never	176 (39.4)	109 (25.6)	212 (47.6)	144 (28.7)	
I don't know	34 (7.6)	49 (11.5)	18 (4)	56 (11.2)	
<b>Internet is useful, n (%)</b>					.04
Yes	312 (69.8)	261 (61.4)	306 (68.8)	341 (68.1)	
<b>Internet is an important source for health information, n (%)</b>					.05
Yes	365 (81.7)	328 (77.2)	330 (74.2)	395 (78.8)	

## Preferences for Sharing Health Information

The coefficients for all attributes were statistically significant and had signs consistent with our expectations in the binary logit model (Table 4). The respondents found situations such as when the collector was their health care provider, if the new

user was a national authority, and the reason was to evaluate the quality of the care as being more acceptable. Moreover, respondents thought it was important to be informed and preferred situations that offered the opportunity to opt out, and that there was a review of the sharing and use of the health information in place.

**Table 4.** Estimates for the multinomial logit model with all countries together.

Attribute and level	Logit		
	Estimate (SE)	P value	95% CI
<b>Collector</b>			
A technological company	-0.19 (0.02)	<.001	-0.22 to -0.16
A research project	0.08 (0.02)	<.001	0.05 to 0.11
Your health care provider (Ref <sup>a</sup> )	0.11 (N/A) <sup>b</sup>	N/A	N/A
<b>New user</b>			
A technological company	-0.26 (0.02)	<.001	-0.31 to -0.22
A pharmaceutical company	-0.03 (0.02)	.15	-0.08 to 0.012
A research project	0.12 (0.02)	<.001	0.08 to 0.16
A national authority (Ref)	0.17 (N/A)	N/A	N/A
<b>Reason</b>			
Develop a new product or service	0.15 (0.02)	<.001	0.11 to 0.20
Promoting, advertising, or marketing	-0.47 (0.02)	<.001	-0.52 to -0.42
Investigate a policy initiative	0.11 (0.02)	N/A	0.07 to 0.15
Evaluate the quality (Ref)	0.21 (N/A)	N/A	N/A
<b>Information</b>			
Not informed	-0.90 (0.02)	<.001	-0.95 to -0.85
Informed	-0.08 (0.02)	.001	-0.12 to -0.03
Informed and ability to opt out	0.51 (0.02)	<.001	0.46 to 0.55
Informed and consent (Ref)	0.47 (N/A)	N/A	N/A
<b>Reviewing</b>			
No specific review	-0.52 (0.02)	<.001	-0.56 to -0.49
Review of sharing	0.25 (0.02)	<.001	0.22 to 0.30
Review of sharing and use (Ref)	0.27 (N/A)	N/A	N/A
Intercept	0.51 (0.01)	<.001	0.49 to 0.54

<sup>a</sup>Reference category.

<sup>b</sup>N/A: not applicable.

Four classes were identified as providing the best fit in the latent class model (Table 5). The information criteria suggested a significant improvement in the fit for the latent class specification over the binary model. All attributes were statistically significant in all classes (besides the new user for class 3), which means that all attributes influenced the decision to accept or reject health information being shared.

The average probability of belonging to class 1 was 27%, class 2 was 32%, class 3 was 23%, and class 4 was 18%. The four

classes displayed some important differences. The intercept term, which reflects the average utility associated with the rejection option, was positive and significant in the binary model (0.51; Table 4). This finding suggests that, on average, respondents in this study preferred *not to share* their health data. The intercept term in the latent class model was negative in class 4, which suggests that class 4 was positive for sharing data. Classes 1, 3, and 4 found a review process regarding the use to be insufficient, whereas class 2 did not (Figure 2).

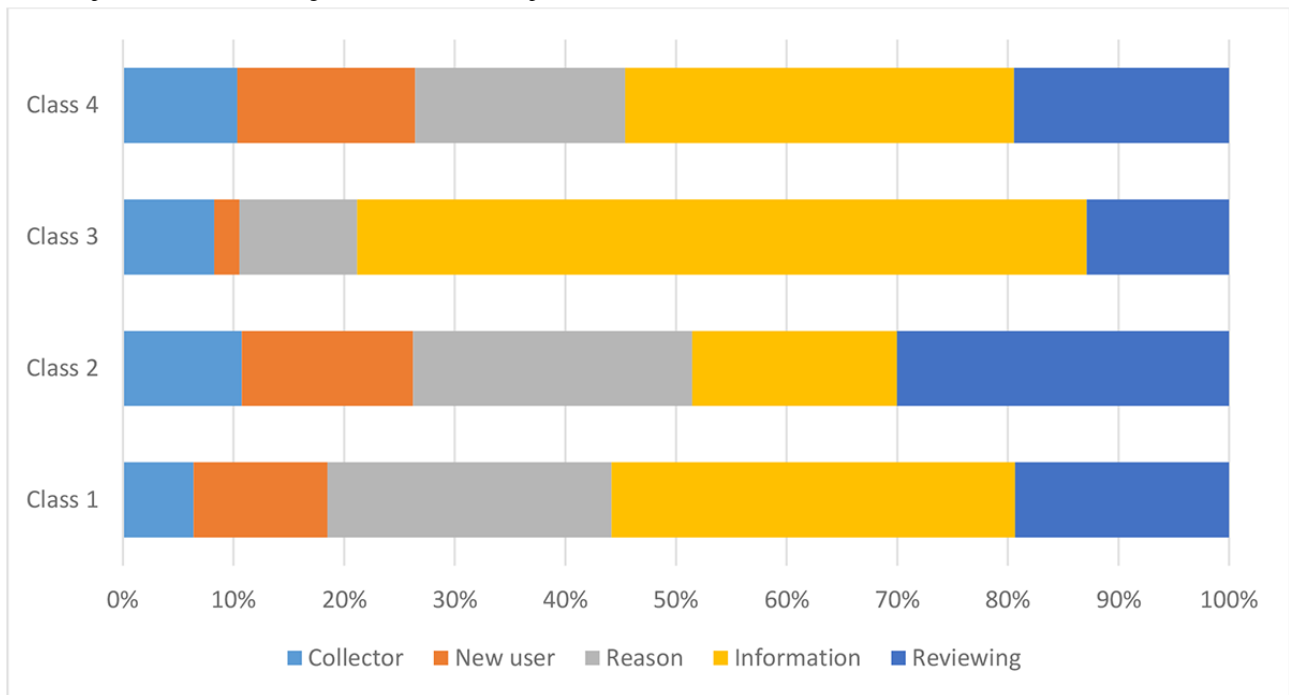


**Table 5.** Estimates for the latent class model, four classes with country as class membership.

Attribute and level	Latent class			
	Class 1	Class 2	Class 3	Class 4
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
<b>Collector</b>				
A technological company	-0.26 <sup>a</sup> (0.10)	-0.37 <sup>b</sup> (0.04)	-0.36 <sup>b</sup> (0.06)	-0.40 <sup>b</sup> (0.08)
A research project	0.14 (0.11)	0.15 <sup>b</sup> (0.03)	0.21 <sup>b</sup> (0.05)	0.16 <sup>a</sup> (0.07)
Your health care provider (Ref <sup>c</sup> )	0.13 (N/A <sup>d</sup> )	0.22 (N/A)	0.15 (N/A)	0.25 (N/A)
<b>New user</b>				
A technological company	-0.41 <sup>a</sup> (0.16)	-0.53 <sup>b</sup> (0.04)	-0.07 (0.07)	-0.59 <sup>b</sup> (0.09)
A pharmaceutical company	-0.12 (0.14)	-0.08 (0.04)	-0.08 (0.07)	-0.05 (0.08)
A research project	0.17 (0.13)	0.29 <sup>b</sup> (0.05)	0.07 (0.07)	0.23 (0.09)
A national authority (Ref)	0.36 (N/A)	0.32 (N/A)	0.08 (N/A)	0.41 (N/A)
<b>Reason</b>				
Develop a new product or service	0.56 <sup>b</sup> (0.16)	0.33 <sup>b</sup> (0.04)	0.04 (0.07)	0.43 <sup>b</sup> (0.09)
Promoting, advertising, or marketing	-1.06 <sup>b</sup> (0.19)	-0.98 <sup>b</sup> (0.06)	-0.41 <sup>b</sup> (0.09)	-0.76 <sup>b</sup> (0.10)
Investigate a policy initiative	0.10 (0.14)	0.25 <sup>b</sup> (0.04)	0.05 (0.07)	-0.05 <sup>b</sup> (0.09)
Evaluate the quality (Ref)	0.40 (N/A)	0.40 (N/A)	0.32 (N/A)	0.38 (N/A)
<b>Information</b>				
Not informed	-1.47 <sup>b</sup> (0.19)	-0.66 <sup>b</sup> (0.06)	-2.95 <sup>b</sup> (0.12)	-1.36 <sup>b</sup> (0.12)
Informed	0.02 (0.18)	-0.04 (0.05)	-0.41 <sup>b</sup> (0.07)	-0.11 (0.09)
Informed and ability to opt out	0.82 <sup>b</sup> (0.19)	0.35 <sup>b</sup> (0.05)	1.61 <sup>b</sup> (0.10)	0.84 <sup>b</sup> (0.08)
Informed and consent (Ref)	0.65 (N/A)	0.31 (N/A)	1.34 (N/A)	0.51 (N/A)
<b>Reviewing</b>				
No specific review	-0.78 <sup>b</sup> (0.13)	-1.07 <sup>b</sup> (0.05)	-0.58 <sup>b</sup> (0.07)	-0.75 <sup>b</sup> (0.09)
Review of sharing	0.43 <sup>b</sup> (0.12)	0.50 <sup>b</sup> (0.04)	0.28 <sup>b</sup> (0.06)	0.46 <sup>b</sup> (0.07)
Review of sharing and use (Ref)	0.35 (N/A)	0.57 (N/A)	0.30 (N/A)	0.29 (N/A)
Intercept	3.60 <sup>b</sup> (0.14)	0.31 <sup>b</sup> (0.04)	0.84 <sup>b</sup> (0.06)	-2.01 <sup>b</sup> (0.09)
AIC <sup>e</sup>	29,730 (N/A)	N/A	N/A	N/A
Log-likelihood	-14,797 (N/A)	N/A	N/A	N/A
Average class probability (%)	27 (N/A)	32 (N/A)	23 (N/A)	18 (N/A)
<b>Class membership</b>				
Constant	0.96 <sup>b</sup> (0.17)	0.73 <sup>b</sup> (0.19)	0.87 <sup>b</sup> (0.18)	Ref
Sweden	-0.29 (0.22)	0.09 (0.23)	-0.91 <sup>b</sup> (0.25)	Ref
Norway	-1.06 <sup>b</sup> (0.22)	-0.78 <sup>b</sup> (0.24)	-0.50 <sup>a</sup> (0.22)	Ref
Iceland	-0.76 <sup>b</sup> (0.22)	0.15 (0.23)	-0.98 <sup>b</sup> (0.24)	Ref

<sup>a</sup>Significance at 5% level.<sup>b</sup>Significance at 1% level.<sup>c</sup>Ref: Reference category.<sup>d</sup>N/A: not applicable.<sup>e</sup>AIC: Akaike information criteria.

**Figure 2.** Relative importance score for respondents' preferences stratified using the four-class model. The reason and having a review process in place were most important for class 2. Being informed was most important for classes 1, 3, and 4.

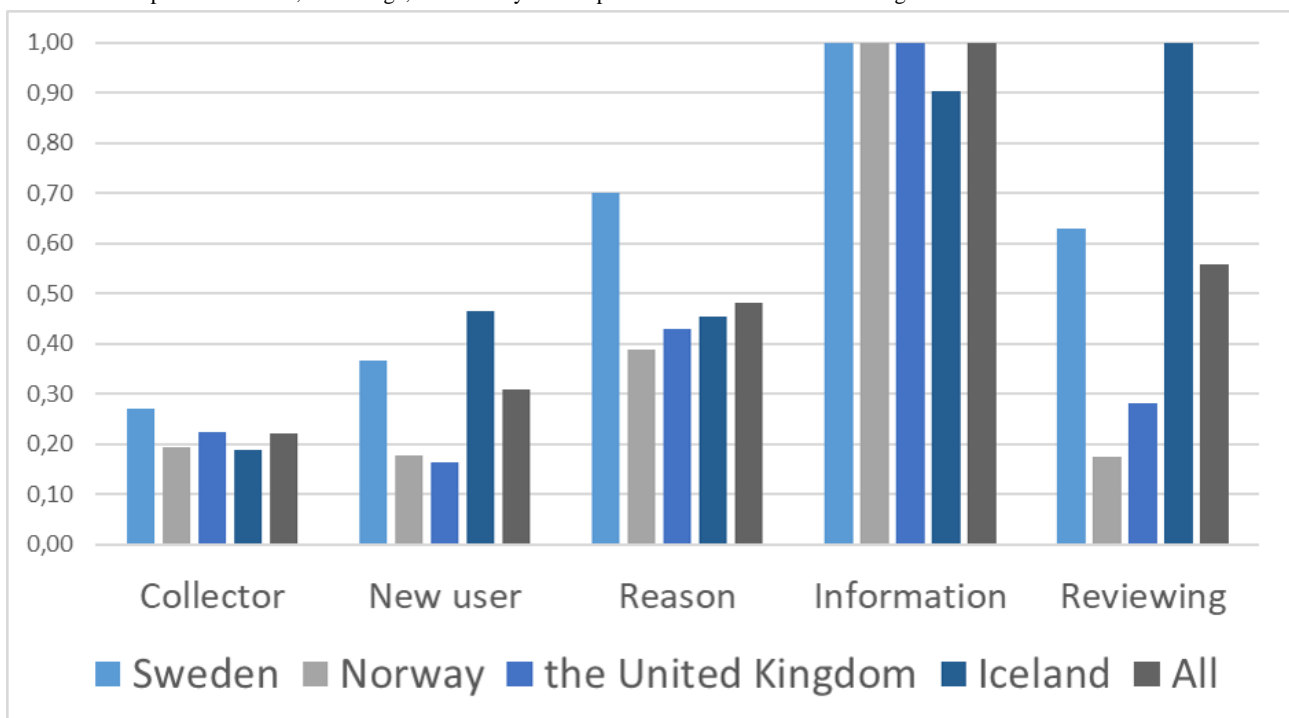


The country significantly predicted class membership as follows: Norwegians were more likely to think that being informed and knowing the purpose would be important when deciding on data sharing (class 1). Icelanders were more likely to think that a review of the sharing and knowing the purpose would be important when deciding on data sharing (class 2). Swedes were more likely to belong to classes 1 and 2. Respondents from the United Kingdom were divided evenly into all classes.

**Relative Importance of the Attributes**

In a situation where health data were about to be transferred to new users, respondents reported the importance of being informed. Having a review process in place was the second most important attribute (Figure 3). Swedish respondents placed more importance on the reason that their health information would be shared compared with respondents from Norway, the United Kingdom, and Iceland. Having a review process in place was the most important attribute for respondents in Iceland.

**Figure 3.** Relative importance score for all respondents' preferences, stratified by country. Receiving information and having the opportunity to opt out was the most important attribute, on average, followed by review process and the reason for sharing the information.



The type of shared information does not change the relative order of the attributes. However, the reason for sharing information was more important when genetic information was shared, as opposed to lifestyle information ([Multimedia Appendix 1](#)).

### Acceptance Uptake

The combination of levels that was most preferred gave an acceptance uptake of 85%, that is, health information collected by a health care provider, evaluating the quality of the national authority's service, planning how resources should be distributed in the future, being informed, having the ability to opt out, with a review process of sharing and use of information. The least

preferred combination gave 14% acceptance uptake, that is, health information collected by a technological company; promoting, advertising, marketing for a new technological company; not being informed of the sharing; and with no review process in place.

Depending on the attribute levels combined into new hypothetical scenarios in which data will be shared from a health care setting to a technological company, the uptake probability varies between 18% and 77% ([Table 6](#)). In situations where people were not informed about their data being transferred, the uptake probability increased if a review process was in place. For further scenarios, view [Multimedia Appendix 2](#).

**Table 6.** The acceptance uptake (class adjusted probability) when health information is shared from a health care setting to a technological company.

Scenarios	Not informed (%)	Informed (%)	Opt-out (%)	Consent (%)
<b>No review</b>				
Develop a new product or service	32	49	63	63
Promoting, advertising, or marketing	18	37	53	60
Investigate a policy initiative	29	47	61	61
Evaluate the quality	32	51	64	64
<b>Review of use</b>				
Develop a new product or service	45	65	77	76
Promoting, advertising, or marketing	32	52	65	65
Investigate a policy initiative	43	63	75	74
Evaluate the quality	46	66	77	76
<b>Review of use and transfer</b>				
Develop a new product or service	45	65	77	76
Promoting, advertising, or marketing	31	53	65	64
Investigate a policy initiative	43	63	74	73
Evaluate the quality	46	66	77	76

## Discussion

### Principal Findings

This DCE study elicited preferences of citizens in Sweden, Norway, Iceland, and the United Kingdom for sharing health data digitally. Respondents in this study indicated that they preferred to share their data when a national authority was going to be the new user of the data. The second preferred new user was an academic research project. On average, and in almost all classes, the respondents preferred a pharmaceutical company as a new user, instead of a technological company. This might be because pharmaceutical companies are well regulated.

The findings show that, on average, respondents from these countries find it more acceptable if they are at least informed about the fact that data will be shared. In addition, having a review process in place to oversee the sharing and use of data was important to people, including the reason the new user had to request data to be shared. These findings provide evidence that supports the European Union General Data Protection Regulation (GDPR) 2016/679 [41], where transparency is one of the foundational principles and serves as a cornerstone of the

Regulation. The GDPR advocates informational self-determination by increasing transparency requirements for data collection practices. It also strengthens individuals' rights regarding their personal data. However, consent is only one of several lawful bases to process personal data listed in Articles 6 and 9 of the GDPR, and cannot be used for the sake of appearances; it is only a lawful base if the data subject is offered control and a genuine choice [42].

Even though all participating countries are required to adhere to the same regulation, the results of this study show that respondents in different countries value different factors when health information is shared. Our results indicate that having a review process in place can be more important for respondents in Sweden and Iceland. However, it can be practically and economically challenging to implement a review process, especially among all private companies. Moreover, having different governance mechanisms in each country can be problematic for cross-border sharing. Therefore, we emphasize the purpose limitation principle, Article 5(1)(b) GDPR [41], that the collection purposes shall be specified, explicit, and legitimate, and that the personal data shall not be further

processed in a manner incompatible with those purposes. Respect for purpose limitation can meet peoples' concerns and requests for contextual control and respect for expectations. This can make the difference between success and failure for the population's acceptance of sharing [43].

It was hypothesized that respondents would prefer to share data if they were offered the opportunity to consent. However, the results show that respondents preferred an opportunity to opt out of the opportunity to consent. It might be enough to have the ability to opt out where other governance mechanisms are in place. From a learning health system perspective in countries with government-financed health care, such as the ones studied here, it might be easier to argue in favor of extensive data sharing between health care providers and medical researchers to account for the shared interest in improved health care for patients [44]. Other rights such as access to health care and quality of health care are also vital concerns [45].

Health care service providers need collaboration with technological companies to improve health via new technological products [46]. Therefore, it could be valuable to evaluate a *thick* opt-out procedure that will provide more participation, and simultaneously acknowledge people's rights to decide over their own private sphere [47]. A *thick* opt-out, or an informed opt-out, in this context means that people become well informed that their data will be shared, but they will not actively agree to the sharing. The default position is that data will be shared and people who do not want to share their health information can actively disagree by opting out. Hence, we identified a need to investigate whether a *thick* opt-out procedure can be sufficient in some contexts. When calculating the acceptance uptake (Table 6), we found an even proportion of people accepting and rejecting their data being shared when offered an opportunity to consent rather than opting out. To account for the range of governance preferences of individuals, dynamic, and meta-consent models [48,49], which allow individuals to first choose their preferred governance model, ought to be studied in the context of not only medical research but also, more generally, for sharing health data in society.

Collecting, storing, and sharing health data is now part of our society. A study by Xafis [50] found that many respondents were of the opinion that data that cannot be traced back to an individual holds a different status compared with identifiable data and could be used without consent. Most respondents in our study preferred to be informed and had the opportunity to opt out or consent to the transfer, even when the data in question would be processed anonymously. This is not required by GDPR. O'Doherty et al [51] advocate the need for a broader consideration, which does not only rely on governance mechanisms such as informed consent and anonymization, as it tends to focus merely on the individual. We also advocate that further aspects need to be considered when sharing information, such as the provision of information, opportunity to opt out, and a review mechanism. Governance mechanisms that also need to be considered are cyber security technologies (eg, access controls and encryption) to safeguard data, along with fostering greater public involvement, transparency, and democratic discourse about this issue [51]. Information security focusing on consent and anonymization as a legal basis is too

narrowly focused; wider societal concerns are not addressed. Hence, they are, as O'Doherty et al [51] state, "insufficient to protect against subversion of health databases for nonsanctioned secondary uses, or to provide guidance for reasonable but controversial secondary uses." Our results support the finding that if the purposes are of great societal value and not only advertising or marketing, then people find it more acceptable to share their health information. Adding a review process increased the probability of people accepting their health information being shared even further.

Aitken et al [10] investigated public preferences for sharing health information in the context of research. In their study, similar attribute selection was made regarding the identity of the new user, what type of information was shared, the purpose of data sharing, and oversight of the process. In contrast to the study by Aitken et al [10], our study results emphasize the importance of respondents being informed of the new user and further use of health information. The reason that our respondents valued the opportunity to be informed might be due to the scope of our study. Our study includes data sharing between different contexts, whereas the previous study only examined data sharing within the academic research context.

An earlier study indicated that people were moderately happy to share most types of information, with least support for sharing personal information such as marital status, age, and income status [52]. Other studies also showed support for data sharing in medical research, as long as the data are pseudonymized [53,54]. In our study, we asked respondents to assume that all shared health information would be pseudonymized. The reason for sharing became more important when genetic information was shared. If the health information collectors and the new users can ensure that there is a guarantee for people remaining anonymous, and successfully manage to communicate this, it will facilitate data sharing in the future. However, this is not applicable for genetic data, as such data are uniquely identifiable and can generally not be anonymized. However, it might be possible to compensate for this if the reason for using the data can be well communicated.

The results provided a deeper understanding of context-specific factors that people value and provide a robust evidence base, which both confirm and challenge the current policy [55]. We should understand that the willingness to share health information varies depending on contextual properties. In particular, the situation in which information is gathered, who the data is being shared with, for what purpose and whether consent is provided, and the extent to which these preferences change depending on which Northern European country the respondents live in. We hypothesized that respondents would make a different choice depending on the context, and all attributes significantly contributed to the choice. This is in accordance with the theory of contextual integrity [17], which finds that privacy is perceived and expected differently depending on the norms and values surrounding the context. People do not request complete control over information about themselves, or that no information about them should be shared. It is important to note that this is shared appropriately. In Table 6, we can see the different probabilities of respondents accepting to share their data in different scenarios. The reason for sharing



plays a major role for respondents, as does the opportunity to opt out or consent. Adding a second governance mechanism would increase the number of people accepting to share their health information.

This is one of the first DCE studies on this topic and is very valuable for ongoing cyber security discussions. However, there might be a hypothetical bias, as in all DCEs. This risk is due to respondents not being bound by their hypothetical choices: they might, in reality, choose something different from what they stated.

This is the first DCE study to compare the preferences of people in Nordic countries for sharing health information. Moderating and mediating factors such as level of education, gender, health status, and e-literacy need further investigation, as they may affect the differences between the countries in preference choices.

In this study, any attribute referring to personal benefits to individuals concerned when health information was shared was excluded. Aitken et al [10] included the attribute *profit-making*, which we considered included in our DCE because of its relative ranking. However, following discussions with the research team and the cognitive interviews, it was excluded because benefit is already incorporated in the nature of some combinations. For example, when a new user is a technological company and the new reason is to develop a new product or service; then, it is

understood that the company will benefit financially from the shared health information. Similarly, if health information is shared with health care to evaluate care, it is implicit that both society and individuals benefit. However, whether the actual transfer involves a monetary exchange could still be a relevant attribute in some contexts.

## Conclusions

Taking the public's acceptance of sharing data into account becomes more important in policy making in the digital world. This study provides insights into the cyber security and privacy research areas on how important specific elements of data sharing are for the public when they consider sharing their data. This is useful for further policy making on the governance of health data in the digital world. At the same time, this provides crucial insights into how to approach people about sharing their data with health care, research projects, national authorities, or different companies. On average, respondents were hesitant to share health information. Respondents' willingness to share their data was most impacted by giving them information about what would happen with their data and the possibility of opting out. To have a review system in place is important for the respondents. Respondents from the studied countries differed in their preferences for sharing health data. This choice of consent or opt out should be further investigated to meet the challenges of the extensive need to share health data digitally and the heterogeneity in people's preferences.

---

## Conflicts of Interest

None declared.

---

### Multimedia Appendix 1

The relative importance stratified by the type of health information.

[[PDF File \(Adobe PDF File\), 299 KB - medinform\\_v9i7e29614\\_app1.pdf](#)]

---

### Multimedia Appendix 2

Acceptance uptake (%) when health information is shared from a technological company to a new technological company and the acceptance uptake when health information is shared to develop a new product or service and with no review process.

[[PDF File \(Adobe PDF File\), 34 KB - medinform\\_v9i7e29614\\_app2.pdf](#)]

---

## References

1. Gadde S, Kalli V. Descriptive analysis of machine learning and its application in healthcare. *Int J Comp Sci Trends Technol* 2020 Mar;8(2):189-196. [doi: [10.33144/23478578](#)]
2. Badawi O, Brennan T, Celi LA, Feng M, Ghassemi M, Ippolito A, MIT Critical Data Conference 2014 Organizing Committee. Making big data useful for health care: a summary of the inaugural mit critical data conference. *JMIR Med Inform* 2014 Aug 22;2(2):e22 [FREE Full text] [doi: [10.2196/medinform.3447](#)] [Medline: [25600172](#)]
3. Raguseo E. Big data technologies: an empirical investigation on their adoption, benefits and risks for companies. *Int J Inf Manage* 2018 Feb;38(1):187-195. [doi: [10.1016/j.ijinfomgt.2017.07.008](#)]
4. Lupton D. The digitally engaged patient: Self-monitoring and self-care in the digital health era. *Soc Theory Health* 2013 Jun 19;11(3):256-270. [doi: [10.1057/sth.2013.10](#)]
5. Vaid S, Harari G. Smartphones in personal informatics: a framework for self-tracking research with mobile sensing. In: Baumeister H, Montag G, editors. *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics*. New York, USA: Springer; Nov 2019:65-92.
6. Clayton EW, Halverson CM, Sathe NA, Malin BA. A systematic literature review of individuals' perspectives on privacy and genetic information in the United States. *PLoS One* 2018;13(10):e0204417 [FREE Full text] [doi: [10.1371/journal.pone.0204417](#)] [Medline: [30379944](#)]

7. Shabani M, Bezuidenhout L, Borry P. Attitudes of research participants and the general public towards genomic data sharing: a systematic literature review. *Expert Rev Mol Diagn* 2014 Nov;14(8):1053-1065. [doi: [10.1586/14737159.2014.961917](https://doi.org/10.1586/14737159.2014.961917)] [Medline: [25260013](https://pubmed.ncbi.nlm.nih.gov/25260013/)]
8. Aitken M, de St Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. *BMC Med Ethics* 2016 Nov 10;17(1):73 [FREE Full text] [doi: [10.1186/s12910-016-0153-x](https://doi.org/10.1186/s12910-016-0153-x)] [Medline: [27832780](https://pubmed.ncbi.nlm.nih.gov/27832780/)]
9. Tully MP, Bernsten C, Aitken M, Vass C. Public preferences regarding data linkage for research: a discrete choice experiment comparing Scotland and Sweden. *BMC Med Inform Decis Mak* 2020 Jun 16;20(1):109 [FREE Full text] [doi: [10.1186/s12911-020-01139-5](https://doi.org/10.1186/s12911-020-01139-5)] [Medline: [32546147](https://pubmed.ncbi.nlm.nih.gov/32546147/)]
10. Aitken M, McAtteer G, Davidson S, Frostick C, Cunningham-Burley S. Public preferences regarding data linkage for health research: a discrete choice experiment. *Int J Popul Data Sci* 2018 Jun 26;3(1):429 [FREE Full text] [doi: [10.23889/ijpds.v3i1.429](https://doi.org/10.23889/ijpds.v3i1.429)] [Medline: [32935004](https://pubmed.ncbi.nlm.nih.gov/32935004/)]
11. Goodman D, Johnson CO, Bowen D, Smith M, Wenzel L, Edwards K. De-identified genomic data sharing: the research participant perspective. *J Community Genet* 2017 Jul;8(3):173-181 [FREE Full text] [doi: [10.1007/s12687-017-0300-1](https://doi.org/10.1007/s12687-017-0300-1)] [Medline: [28382417](https://pubmed.ncbi.nlm.nih.gov/28382417/)]
12. Jones RD, Sabolch AN, Aakhus E, Spence RA, Bradbury AR, Jaggi R. Patient perspectives on the ethical implementation of a rapid learning system for oncology care. *J Oncol Pract* 2017 Mar;13(3):e163-e175. [doi: [10.1200/JOP.2016.016782](https://doi.org/10.1200/JOP.2016.016782)] [Medline: [28118107](https://pubmed.ncbi.nlm.nih.gov/28118107/)]
13. Hassan L, Dalton A, Hammond C, Tully MP. A deliberative study of public attitudes towards sharing genomic data within NHS genomic medicine services in England. *Public Underst Sci* 2020 Oct;29(7):702-717 [FREE Full text] [doi: [10.1177/0963662520942132](https://doi.org/10.1177/0963662520942132)] [Medline: [32664786](https://pubmed.ncbi.nlm.nih.gov/32664786/)]
14. Karampela M, Ouhbi S, Isomursu M. Connected health user willingness to share personal health data: questionnaire study. *J Med Internet Res* 2019 Nov 27;21(11):e14537 [FREE Full text] [doi: [10.2196/14537](https://doi.org/10.2196/14537)] [Medline: [31774410](https://pubmed.ncbi.nlm.nih.gov/31774410/)]
15. Potoglou D, Palacios J, Feijóo C. An integrated latent variable and choice model to explore the role of privacy concern on stated behavioural intentions in e-commerce. *J Choice Model* 2015 Dec;17:10-27. [doi: [10.1016/j.jocm.2015.12.002](https://doi.org/10.1016/j.jocm.2015.12.002)]
16. Denman DC, Baldwin AS, Betts AC, McQueen A, Tiro JA. Reducing 'I don't know' responses and missing survey data: implications for measurement. *Med Decis Making* 2018 Aug;38(6):673-682 [FREE Full text] [doi: [10.1177/0272989X18785159](https://doi.org/10.1177/0272989X18785159)] [Medline: [29962272](https://pubmed.ncbi.nlm.nih.gov/29962272/)]
17. Nissenbaum H. Privacy as contextual integrity. *Wash L Rev* 2004 Feb;79:119-158.
18. Clark MD, Determann D, Petrou S, Moro D, de Bekker-Grob EW. Discrete choice experiments in health economics: a review of the literature. *Pharmacoeconomics* 2014 Sep;32(9):883-902. [doi: [10.1007/s40273-014-0170-x](https://doi.org/10.1007/s40273-014-0170-x)] [Medline: [25005924](https://pubmed.ncbi.nlm.nih.gov/25005924/)]
19. de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. *Health Econ* 2012 Feb;21(2):145-172. [doi: [10.1002/hec.1697](https://doi.org/10.1002/hec.1697)] [Medline: [22223558](https://pubmed.ncbi.nlm.nih.gov/22223558/)]
20. Viberg Johansson J, Langenskiöld S, Segerdahl P, Hansson MG, Hösterey UU, Gummesson A, et al. Research participants' preferences for receiving genetic risk information: a discrete choice experiment. *Genet Med* 2019 Oct;21(10):2381-2389. [doi: [10.1038/s41436-019-0511-4](https://doi.org/10.1038/s41436-019-0511-4)] [Medline: [30992550](https://pubmed.ncbi.nlm.nih.gov/30992550/)]
21. Hensher D, Rose J, Greene W. *Applied Choice Analysis* 2nd Edition. Cambridge, UK: Cambridge University Press; 2015.
22. Louviere J, Hensher D, Swait J. *Stated Choice Methods: Analysis and Applications*. Cambridge, UK: Cambridge university press; 2000.
23. Ryan M, Farrar S. Using conjoint analysis to elicit preferences for health care. *Br Med J* 2000 Jun 3;320(7248):1530-1533 [FREE Full text] [doi: [10.1136/bmj.320.7248.1530](https://doi.org/10.1136/bmj.320.7248.1530)] [Medline: [10834905](https://pubmed.ncbi.nlm.nih.gov/10834905/)]
24. Ryan M, Gerard K, Amaya-Amaya M. *Using Discrete Choice Experiments to Value Health and Health Care*. Berlin, Germany: Springer Science & Business Media; 2007.
25. Viberg Johansson J, Shah N, Haraldsdóttir E, Bentzen HB, Coy S, Kaye J, et al. Governance mechanisms for sharing of health data: an approach towards selecting attributes for complex discrete choice experiment studies. *Technol Soc* 2021 Aug;66:101625. [doi: [10.1016/j.techsoc.2021.101625](https://doi.org/10.1016/j.techsoc.2021.101625)]
26. Cheraghi-Sohi S, Bower P, Mead N, McDonald R, Whalley D, Roland M. Making sense of patient priorities: applying discrete choice methods in primary care using 'think aloud' technique. *Fam Pract* 2007 Jun;24(3):276-282. [doi: [10.1093/fampra/cmm007](https://doi.org/10.1093/fampra/cmm007)] [Medline: [17478438](https://pubmed.ncbi.nlm.nih.gov/17478438/)]
27. Bridges JF, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA, et al. Conjoint analysis applications in health-a checklist: a report of the ISPOR good research practices for conjoint analysis task force. *Value Health* 2011 Jun;14(4):403-413 [FREE Full text] [doi: [10.1016/j.jval.2010.11.013](https://doi.org/10.1016/j.jval.2010.11.013)] [Medline: [21669364](https://pubmed.ncbi.nlm.nih.gov/21669364/)]
28. Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making: a user's guide. *Pharmacoeconomics* 2008;26(8):661-677. [doi: [10.2165/00019053-200826080-00004](https://doi.org/10.2165/00019053-200826080-00004)] [Medline: [18620460](https://pubmed.ncbi.nlm.nih.gov/18620460/)]
29. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. *J Med Internet Res* 2006 Nov 14;8(4):e27 [FREE Full text] [doi: [10.2196/jmir.8.4.e27](https://doi.org/10.2196/jmir.8.4.e27)] [Medline: [17213046](https://pubmed.ncbi.nlm.nih.gov/17213046/)]
30. SurveyEngine. URL: <https://surveyengine.com/> [accessed 2021-06-22]

31. Swait J, Louviere J. The role of the scale parameter in the estimation and comparison of multinomial logit models. *J Mark Res* 1993 Aug;30(3):305. [doi: [10.2307/3172883](https://doi.org/10.2307/3172883)]
32. Greene WH, Hensher DA. A latent class model for discrete choice analysis: contrasts with mixed logit. *Transport Res Part B: Method* 2003 Sep;37(8):681-698. [doi: [10.1016/s0191-2615\(02\)00046-2](https://doi.org/10.1016/s0191-2615(02)00046-2)]
33. Dziak J, Coffman D, Lanza S, Li R, Jermiin LS. Sensitivity and specificity of information criteria. *Brief Bioinform* 2020 Mar 23;21(2):553-565 [FREE Full text] [doi: [10.1093/bib/bbz016](https://doi.org/10.1093/bib/bbz016)] [Medline: [30895308](https://pubmed.ncbi.nlm.nih.gov/30895308/)]
34. Bech M, Gyrd-Hansen D. Effects coding in discrete choice experiments. *Health Econ* 2005 Oct;14(10):1079-1083. [doi: [10.1002/hec.984](https://doi.org/10.1002/hec.984)] [Medline: [15852455](https://pubmed.ncbi.nlm.nih.gov/15852455/)]
35. Burke PF, Burton C, Huybers T, Islam T, Louviere JJ, Wise C. The scale-adjusted latent class model: application to museum visitation. *Tour Anal* 2010 Jul 1;15(2):147-165. [doi: [10.3727/108354210x12724863327605](https://doi.org/10.3727/108354210x12724863327605)]
36. Magidson J, Vermunt J. Removing the Scale Factor Confound in Multinomial Logit Choice Models to Obtain Better Estimates of Preference. In: *Sawtooth Software Conference. 2007 Presented at: SSC'07; October 17-19, 2007; Santa Rosa, California* URL: <https://sawtoothsoftware.com/resources/technical-papers/conferences/sawtooth-software-conference-2007>
37. Hall J, Kenny P, King M, Louviere J, Viney R, Yeoh A. Using stated preference discrete choice modelling to evaluate the introduction of varicella vaccination. *Health Econ* 2002 Jul;11(5):457-465. [doi: [10.1002/hec.694](https://doi.org/10.1002/hec.694)] [Medline: [12112494](https://pubmed.ncbi.nlm.nih.gov/12112494/)]
38. Hensher DA, Rose J, Greene WH. The implications on willingness to pay of respondents ignoring specific attributes. *Transportation* 2005 May 21;32(3):203-222. [doi: [10.1007/s11116-004-7613-8](https://doi.org/10.1007/s11116-004-7613-8)]
39. Aaron FD, Abramowicz H, Abt I, Adamczyk L, Adamus M, Aggarwal R, et al. . [doi: [10.1140/epic/s10052-012-2175-y](https://doi.org/10.1140/epic/s10052-012-2175-y)]
40. Jonker M, de Bekker-Grob E, Veldwijk J, Goossens L, Bour S, Rutten-Van MM. COVID-19 contact tracing apps: predicted uptake in the Netherlands based on a discrete choice experiment. *JMIR Mhealth Uhealth* 2020 Oct 9;8(10):e20741 [FREE Full text] [doi: [10.2196/20741](https://doi.org/10.2196/20741)] [Medline: [32795998](https://pubmed.ncbi.nlm.nih.gov/32795998/)]
41. General Data Protection Regulation GDPR. General Data Protection Regulation. 2016. URL: <https://gdpr-info.eu> [accessed 2021-06-22]
42. Guidelines 05/2020 on Consent Under Regulation 2016/679. European Data Protection Board. 2020. URL: [https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-052020-consent-under-regulation-2016679\\_en](https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-052020-consent-under-regulation-2016679_en) [accessed 2021-06-22]
43. Bentzen H. Context as key: the protection of personal integrity by means of the purpose limitation principle. In: *EaL RK, editor. Handbook on European Data Protection Law. Luxembourg: Publications Office of the EU; 2021.*
44. Faden RR, Kass NE, Goodman SN, Pronovost P, Tunis S, Beauchamp TL. An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. *Hastings Cent Rep* 2013;Spec No:S16-S27. [doi: [10.1002/hast.134](https://doi.org/10.1002/hast.134)] [Medline: [23315888](https://pubmed.ncbi.nlm.nih.gov/23315888/)]
45. Mascalonzi D, Paradiso A, Hansson M. Rare disease research: breaking the privacy barrier. *Appl Transl Genom* 2014 Jun 1;3(2):23-29 [FREE Full text] [doi: [10.1016/j.atg.2014.04.003](https://doi.org/10.1016/j.atg.2014.04.003)] [Medline: [27275410](https://pubmed.ncbi.nlm.nih.gov/27275410/)]
46. Taneja H. How Tech Companies Can Help Fix U.S. Health Care. *Harvard Business Review*. 2020 Apr 28. URL: <https://hbr.org/2020/04/how-big-tech-can-help-fix-u-s-health-care?ab=hero-subleft-2> [accessed 2021-06-22]
47. Giesbertz NA, Bredenoord AL, van Delden JJ. A thick opt-out is often sufficient. *Am J Bioeth* 2013;13(4):44-46. [doi: [10.1080/15265161.2013.767962](https://doi.org/10.1080/15265161.2013.767962)] [Medline: [23514397](https://pubmed.ncbi.nlm.nih.gov/23514397/)]
48. Ploug T, Holm S. Meta consent - a flexible solution to the problem of secondary use of health data. *Bioethics* 2016 Nov;30(9):721-732 [FREE Full text] [doi: [10.1111/bioe.12286](https://doi.org/10.1111/bioe.12286)] [Medline: [27628305](https://pubmed.ncbi.nlm.nih.gov/27628305/)]
49. Budin-Ljøsne I, Teare HJ, Kaye J, Beck S, Bentzen HB, Caenazzo L, et al. Dynamic consent: a potential solution to some of the challenges of modern biomedical research. *BMC Med Ethics* 2017 Jan 25;18(1):4 [FREE Full text] [doi: [10.1186/s12910-016-0162-9](https://doi.org/10.1186/s12910-016-0162-9)] [Medline: [28122615](https://pubmed.ncbi.nlm.nih.gov/28122615/)]
50. Xafis V. The acceptability of conducting data linkage research without obtaining consent: lay people's views and justifications. *BMC Med Ethics* 2015 Nov 17;16(1):79 [FREE Full text] [doi: [10.1186/s12910-015-0070-4](https://doi.org/10.1186/s12910-015-0070-4)] [Medline: [26577591](https://pubmed.ncbi.nlm.nih.gov/26577591/)]
51. O'Doherty KC, Christofides E, Yen J, Bentzen HB, Burke W, Hallowell N, et al. If you build it, they will come: unintended future uses of organised health data collections. *BMC Med Ethics* 2016 Sep 6;17(1):54 [FREE Full text] [doi: [10.1186/s12910-016-0137-x](https://doi.org/10.1186/s12910-016-0137-x)] [Medline: [27600117](https://pubmed.ncbi.nlm.nih.gov/27600117/)]
52. Shah N, Coathup V, Teare H, Forgie I, Giordano GN, Hansen TH, et al. Sharing data for future research-engaging participants' views about data governance beyond the original project: a DIRECT Study. *Genet Med* 2019 May;21(5):1131-1138. [doi: [10.1038/s41436-018-0299-7](https://doi.org/10.1038/s41436-018-0299-7)] [Medline: [30262927](https://pubmed.ncbi.nlm.nih.gov/30262927/)]
53. Kaufman D, Murphy J, Scott J, Hudson K. Subjects matter: a survey of public opinions about a large genetic cohort study. *Genet Med* 2008 Nov;10(11):831-839. [doi: [10.1097/GIM.0b013e31818bb3ab](https://doi.org/10.1097/GIM.0b013e31818bb3ab)] [Medline: [19011407](https://pubmed.ncbi.nlm.nih.gov/19011407/)]
54. Spencer K, Sanders C, Whitley EA, Lund D, Kaye J, Dixon WG. Patient perspectives on sharing anonymized personal health data using a digital system for dynamic consent and research feedback: a qualitative study. *J Med Internet Res* 2016 Apr 15;18(4):e66 [FREE Full text] [doi: [10.2196/jmir.5011](https://doi.org/10.2196/jmir.5011)] [Medline: [27083521](https://pubmed.ncbi.nlm.nih.gov/27083521/)]
55. Potoglou D, Dunkerley F, Patil S, Robinson N. Public preferences for internet surveillance, data retention and privacy enhancing services: evidence from a pan-European study. *Comput Hum Behav* 2017 Oct;75:811-825 [FREE Full text] [doi: [10.1016/j.chb.2017.06.007](https://doi.org/10.1016/j.chb.2017.06.007)]

---

**Abbreviations****DCE:** discrete choice experiment**eHL:** eHealth literacy**GDPR:** General Data Protection Regulation

---

*Edited by C Lovis; submitted 14.04.21; peer-reviewed by X Cheng, R Krukowski; comments to author 02.05.21; revised version received 11.05.21; accepted 17.05.21; published 05.07.21.*

*Please cite as:*

*Viberg Johansson J, Bentzen HB, Shah N, Haraldsdóttir E, Jónsdóttir GA, Kaye J, Mascalzoni D, Veldwijk J*

*Preferences of the Public for Sharing Health Data: Discrete Choice Experiment*

*JMIR Med Inform 2021;9(7):e29614*

*URL: <https://medinform.jmir.org/2021/7/e29614>*

*doi: [10.2196/29614](https://doi.org/10.2196/29614)*

*PMID: [36260402](https://pubmed.ncbi.nlm.nih.gov/36260402/)*

©Jennifer Viberg Johansson, Heidi Beate Bentzen, Nisha Shah, Eik Haraldsdóttir, Guðbjörg Andrea Jónsdóttir, Jane Kaye, Deborah Mascalzoni, Jorien Veldwijk. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 05.07.2021. 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

# Predicting Biologic Therapy Outcome of Patients With Spondyloarthritis: Joint Models for Longitudinal and Survival Analysis

Carolina Barata<sup>1,2</sup>, MSc; Ana Maria Rodrigues<sup>3,4</sup>, MD, PhD; Helena Canhão<sup>3,4</sup>, MD, PhD; Susana Vinga<sup>1,5,6</sup>, PhD; Alexandra M Carvalho<sup>1,2,6</sup>, PhD

<sup>1</sup>Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

<sup>2</sup>Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

<sup>3</sup>Comprehensive Health Research Center, NOVA Medical School, NOVA University of Lisbon, Lisbon, Portugal

<sup>4</sup>EpiDoC Unit, The Chronic Diseases Research Centre, NOVA Medical School, NOVA University of Lisbon, Lisbon, Portugal

<sup>5</sup>Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento em Lisboa (INESC-ID), Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

<sup>6</sup>Lisbon Unit for Learning and Intelligent Systems, Lisbon, Portugal

**Corresponding Author:**

Alexandra M Carvalho, PhD  
Instituto de Telecomunicações  
Instituto Superior Técnico  
Universidade de Lisboa  
Lisbon  
Portugal  
Phone: 351 218 418 454  
Email: [alexandra.carvalho@tecnico.ulisboa.pt](mailto:alexandra.carvalho@tecnico.ulisboa.pt)

## Abstract

**Background:** Rheumatic diseases are one of the most common chronic diseases worldwide. Among them, spondyloarthritis (SpA) is a group of highly debilitating diseases, with an early onset age, which significantly impacts patients' quality of life, health care systems, and society in general. Recent treatment options consist of using biologic therapies, and establishing the most beneficial option according to the patients' characteristics is a challenge that needs to be overcome. Meanwhile, the emerging availability of electronic medical records has made necessary the development of methods that can extract insightful information while handling all the challenges of dealing with complex, real-world data.

**Objective:** The aim of this study was to achieve a better understanding of SpA patients' therapy responses and identify the predictors that affect them, thereby enabling the prognosis of therapy success or failure.

**Methods:** A data mining approach based on joint models for the survival analysis of the biologic therapy failure is proposed, which considers the information of both baseline and time-varying variables extracted from the electronic medical records of SpA patients from the database, Reuma.pt.

**Results:** Our results show that being a male, starting biologic therapy at an older age, having a larger time interval between disease start and initiation of the first biologic drug, and being human leukocyte antigen (HLA)-B27 positive are indicators of a good prognosis for the biological drug survival; meanwhile, having disease onset or biologic therapy initiation occur in more recent years, a larger number of education years, and higher values of C-reactive protein or Bath Ankylosing Spondylitis Functional Index (BASFI) at baseline are all predictors of a greater risk of failure of the first biologic therapy.

**Conclusions:** Among this Portuguese subpopulation of SpA patients, those who were male, HLA-B27 positive, and with a later biologic therapy starting date or a larger time interval between disease start and initiation of the first biologic therapy showed longer therapy adherence. Joint models proved to be a valuable tool for the analysis of electronic medical records in the field of rheumatic diseases and may allow for the identification of potential predictors of biologic therapy failure.

(*JMIR Med Inform* 2021;9(7):e26823) doi:[10.2196/26823](https://doi.org/10.2196/26823)

**KEYWORDS**

data mining; survival analysis; joint models; spondyloarthritis; drug survival; rheumatic disease; electronic medical records; medical records

## Introduction

### Motivation

Rheumatic diseases are chronic diseases that, being the leading cause of disability in developed countries, consume many health and social resources. Among these diseases, spondyloarthritis (SpA) is a group of several related disorders that can be highly debilitating and significantly impact patients' quality of life, health care systems, and society [1].

As there is no cure, treatment focuses on the relief of symptoms and the delay of the disease's progression. Biologic therapies are the most recent approach for treating these disorders, and their use is recommended when all other methods have failed. However, the therapy selection follows no specific criteria, and trying to establish which patients benefit the most from each drug is still a problem that needs to be solved [2].

A better understanding of therapy responses for these patients and identifying the predictors that affect these responses would allow for a prognosis of therapy success or failure and thus be highly valuable in conserving the resources and time of both patients and medical doctors. Moreover, this understanding could be used to aid medical experts in tailoring the treatment to the patient by using a more personalized approach.

Meanwhile, the emerging availability of electronic medical records has enabled the storage of great amounts of information that can be used to extract insightful knowledge. Data mining is a rapidly growing field that focuses on developing the techniques necessary for insightfully using this information.

The analysis of an outcome of interest is usually performed using survival analysis methods, such as the Kaplan-Meier estimator [3] and the Cox model [4]. Nevertheless, these methods are only able to deal with time-static variables. For dealing with time-varying variables, methods such as the extended Cox model [5] have been introduced. However, they are not appropriate for dealing with biomarkers [6,7].

Joint models have been presented in the literature as a useful approach for handling these types of analysis, having been used in a wide range of medical studies, including the most common disease areas of cancer and HIV and AIDS [8].

In the field of rheumatic diseases, these joint modeling and machine learning approaches were studied to evaluate the clinical impact on flare occurrence in patients undergoing biologic treatments for rheumatoid arthritis. Both models were proven to assist in decisions on biologic dose reduction with the potential to reduce the occurrence of flares significantly [9]. The development of juvenile dermatomyositis was also studied using longitudinal approaches, allowing for a better perception of longitudinal outcomes and a more accurate comprehension of predictors' effects [10].

However, the use of these methods has been less explored for other diseases in this field, such as SpA.

Our main goal was to propose a data mining approach based on joint models to infer relationships between time-to-event and longitudinal electronic medical record data, retrieved from the Rheumatic Disease Portuguese Register (Reuma.pt) [11]. We further aimed to study the predictors of failure of the first biologic therapy for patients with SpA and verify the applicability of joint models for the study of therapeutic response in rheumatic diseases.

### Background

#### *Spondyloarthritis*

Spondyloarthritis is the name given to a family of inflammatory rheumatic diseases that share distinctive pathophysiologic, clinical, radiographic, and genetic features. This includes ankylosing spondylitis (AS)—the characteristic type of this group—psoriatic arthritis, reactive arthritis, enteropathic arthritis, and so called undifferentiated SpA.

AS is characterized by chronic inflammation predominantly affecting the axial skeleton. Although its pathogenesis is poorly understood, there is a strong association between AS and the human leukocyte antigen B27 (HLA-B27), and the typical age at onset of this condition is at the second or third decade of life [12].

The first population-based study on rheumatic diseases in Portugal, EpiReumaPt, reported the national health survey results in 2015, revealing a general SpA prevalence of 1.6% and a prevalence of 2.0% and 1.2% for women and men, respectively [1].

The socioeconomic impact can be rather high for these conditions. A recent study [13] revealed that AS has a total annual economic impact of €639 million (US \$773 million) in Portugal. This value includes the disease-related costs for the patient and the national health system and the economic impact of the lost workdays.

Clinical monitoring of a disease is of extreme importance to understanding disease progression, better assessing patient response to treatment, and guiding therapeutic decisions.

Laboratory exams include erythrocyte sedimentation rate (ESR) and levels of C-reactive protein (CRP), which are markers of inflammation, and other laboratory data that are considered to show relevant alterations.

Functional ability can be evaluated using the Bath Ankylosing Spondylitis Functional Index (BASFI) [14] score, and activity disease can be evaluated using the Bath Ankylosing Disease Status in Ankylosing Spondylitis (BASDAI) score [15] or the more recently developed Ankylosing Spondylitis Disease Activity Score (ASDAS) [16]. The ASDAS has two different formulas, ASDAS-CRP (which uses the C-reactive protein) and ASDAS-ESR (which uses ESR), with ASDAS-CRP usually being the preferred system.

Treatment of SpA should be tailored to the patient, and patient signs, symptoms, and characteristics should be taken into account, with the most common goal being the attainment of a state of inactive disease.

Treatment options can include physical therapy, nonsteroidal anti-inflammatory drugs, disease-modifying antirheumatic drugs, and, if patients remain in a high disease activity state when trying the referred options, treatment with biologic agents, namely tumor necrosis factor inhibitors and interleukin-17 or interleukin-23 inhibitors.

### Time-to-Event Analysis

Survival analysis, or time-to-event analysis, is the collection of statistical procedures for the analysis in which the outcome variable of interest is the time until an event occurs.

Let  $T$  denote a random, nonnegative, continuous variable representing the patient's survival time and let  $t$  be an observed value of  $T$ .

The hazard function can be interpreted as the instantaneous potential per unit time for the event to occur, given that the individual has survived to time  $t$ . It is calculated as follows:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

The main feature that distinguishes survival data from other types of data is the possible presence of censored survival times.

Considering a specific individual  $i$ , let  $T_i$  be the random variable representing its true survival time and  $C_i$  the potential censoring time. With consideration to right censoring, the censoring indicator variable,  $\delta_i$ , is defined as  $\delta_i = I(T_i \leq C_i)$ , where  $I(\cdot)$  is the indicator function.

The Cox proportional hazards model [4] allows us to estimate the hazard function and explore how the survival of a group of patients depends on the values of one or more explanatory variables.

Let  $\mathbf{x} = (x_1, x_2, \dots, x_k)$  be the values of the  $k$  explanatory variables of an individual and  $\beta = (\beta_1, \beta_2, \dots, \beta_k)$  the vector of its correspondent unknown regression coefficients. The hazard function is given by the following:

$$h(t; \mathbf{x}) = h_0(t) \exp(\beta^T \mathbf{x})$$

where  $h_0(t)$  is the baseline hazard function, representing the hazard for a patient when its vector of explanatory variables is equal to zero ( $\mathbf{x} = 0$ ).

It is possible to extend the previously presented Cox model for handling time-dependent variables [5]. This model is referred to as the extended Cox model. However, it is not theoretically appropriate to deal with biomarkers since it assumes that the time-dependent variables are predictable processes, measured without error and with a full path completely known.

### Longitudinal Analysis

Longitudinal data can be defined as the data obtained from multiple measurements of individuals throughout time.

Linear mixed effects (LME) models are a common way of modeling this data. In an LME model, the individual's response is assumed to follow a linear regression model where some of the regression parameters are population specific and others are patient specific. These are referred to as fixed effects and random effects, respectively [17].

Let  $\mathbf{Y}_i$  be the  $n_i$ -dimensional response vector for subject  $i$ . In general, a linear mixed-effects model satisfies the following:

$$\mathbf{Y}_i = \mathbf{X}_i \beta + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i$$

where  $\beta$  is a  $p$ -dimensional vector that contains the fixed effects;  $\mathbf{b}_i$  is the  $q$ -dimensional vector containing the random effects, and  $\boldsymbol{\varepsilon}_i$  is a  $n_i$ -dimensional vector of random errors;  $\mathbf{X}_i$  and  $\mathbf{Z}_i$  are the  $(n_i \times p)$  and  $(n_i \times q)$  fixed-effects and random effects design matrices, respectively;  $D$  is a  $(q \times q)$  positive-definite covariance matrix; and  $\Sigma_i$  is a  $(n_i \times n_i)$  positive-definite covariance matrix that depends on  $i$  through its dimension  $n_i$ . The  $\boldsymbol{\varepsilon}_i$  is normally distributed with mean zero and covariance matrix  $D$ , and  $\mathbf{b}_i$  is normally distributed with mean zero and covariance matrix  $\Sigma_i$ . Both  $\mathbf{b}_i$  and  $\boldsymbol{\varepsilon}_i$  are assumed to be independent of each other and of groups [17].

### Joint Models for Longitudinal and Time-to-Event Data

The basic idea of joint models is to perform combined analysis, in which a relative risk model is estimated for the time-to-event outcome, taking into account the effect of the longitudinal data measurements—this is usually done by combining a survival model with a mixed-effects model [6].

The first step is modeling the continuous longitudinal outcomes with LME models. Let  $\mathbf{y}_k$  denote the  $(n_{ki} \times 1)$  longitudinal response vector for the  $k$ -th outcome ( $k = 1, \dots, K$ ) and the  $i$ -th subject that is composed by elements  $y_{kil}$ , which represent the value of the  $k$ -th longitudinal outcome taken at time point  $t_{kil}$ . Let  $\mathbf{b}_{ki}$  be a vector of random effects and  $\beta_k$  a vector of fixed effects. We have that the conditional expectation of  $\mathbf{y}_k$  given  $\mathbf{b}_{ki}$ ;  $\eta_{ki}(t)$  is modeled through the LME model as follows:

$$E(\mathbf{y}_k | \mathbf{b}_{ki}) = \mathbf{X}_{ki}(t) \beta_k + \mathbf{Z}_{ki}(t) \mathbf{b}_{ki}$$

where  $\mathbf{x}_{ki}(t)$  and  $\mathbf{z}_{ki}(t)$  are the design vectors for the random and fixed effects, respectively.

Let  $\tau_i$  be the true event time for the  $i$ -th subject. We can now postulate the relative risk model for the survival process as follows:

$$h_i(t) = h_0(t) \exp(\boldsymbol{\alpha}_{ki}^T \mathbf{w}_i(t))$$

where  $M_i(t) = \{M_{1i}(t), \dots, M_{ki}(t)\}$  and  $M_{ki}(t) = \{\eta_{ki}(s), 0 \leq s < t\}$  denotes the history of the true unobserved longitudinal process up to time point  $t$ ,  $h_0(\cdot)$  denotes the baseline risk function, and  $\mathbf{w}_i(t)$  is a vector of exogenous covariates with a corresponding vector of regression coefficients  $\gamma$ . The  $f_{kl}$  functions, parametrized by vector  $\alpha_{kl}$ , specify which components of each longitudinal outcome will be present in the relative risk

model, allowing up to  $L_k$  functional forms for each of  $K$  longitudinal outcomes. The parameters contained in  $\alpha_{ki}$  quantify the effect of the correspondent underlying longitudinal outcome to the risk for an event.

One of the basic approaches for the functional form is to model the event's hazard as having an association only with the current value of the longitudinal outcome at the same time point. Considering a single outcome, this is given by  $f\{\alpha, \mathbf{w}_i(t), \mathbf{b}_i, M_i(t)\} = \alpha\eta_i(t)$ , where  $\alpha$  is the strength of association parameter that indicates the change in the log hazard when there is a unit change in the patient's longitudinal outcome value.

Estimation of joint models is performed by exploiting the full joint likelihood that is derived from the joint distribution of the longitudinal and survival outcomes. Methods for this estimation can follow, among others, a frequentist or a Bayesian paradigm [18,19].

## Methods

### SpA Patients on Biologic Therapies: Data Description and Preprocessing

The data used in this study were retrieved from Reuma.pt [11] on July 22, 2019. This register was developed by the Portuguese Society of Rheumatology, has been active since June 2008, and contains information retrieved on a routine basis of rheumatic patients in Portugal receiving biological therapies. Although Reuma.pt also contains patients with several rheumatic diseases, the focus of this work was on patients with SpA.

With the data extracted from this database, 4 different data sets, A, B, C and D, were obtained according to different strategies for handling the missing values. A set of different joint models was then fitted to all data sets, as well as the equivalent extended Cox models using R software (The R Foundation for Statistical Computing) [20], namely packages “JM” [18] and “Jmbayes” [19]. This resulted in a total of 49 joint models and 29 extended Cox models. All the steps of the data processing and modeling framework are described in detail in this section.

The Reuma.pt database contains information regarding patients and patient visits, including identification data, demographic data, previous medical history, comorbidities, laboratory results, past and current therapies, adverse events, and disease activity scores, among others.

The follow-up of patients through this registry enables the monitoring of treatment efficacy, safety, and comorbidities.

The goal of this work was to perform a survival analysis that takes into account both time-independent and time-dependent variables and understand how these impact the outcome of interest. Therefore, 3 types of variables were needed: time-independent (baseline) variables, time-dependent variables, and time-to-event variables. The latter type is not directly found in the database and therefore needed to be processed from the existing data.

Our event of interest was the failure of the first biological therapy for each patient, where failure was defined as the discontinuation of the biological therapy due to inefficacy

(evaluated by ASDAS) or adverse events (such as infection or hypersensitivity).

In this context, the time-to-event variables indicated if the biologic therapy failed or if the patient was censored—henceforth referred to as the failure index. The time until the occurrence of either failure index is referred to as time to failure.

The data extracted from the database had to go through several preprocessing steps in order to reach a format compatible with the models to be fitted.

First, a set of variables to be considered was selected, with 3 main aspects being taken into account: level of missing values, relevance to the study, and variable equivalence. Variables with more than 60% of values missing were not considered nor were variables that were considered to be irrelevant for the goal of our study as determined according to the feedback given by medical specialists. Furthermore, some variables available in the raw data set were equivalent in the sense that they represented the same information.

After this set of assumptions and processing steps, we obtained our initial data set, which consisted of the following 3 variables:

1. Time-to-event variables—time to failure and failure index;
2. Time-independent variables—sex, marital status, year of diagnosis, age at diagnosis, year of disease beginning, age at disease beginning, year of start of the first biologic therapy, age at start of the first biologic therapy, years from diagnosis to start of the first biologic therapy, disease years until start of the first biologic therapy, HLA-B27, employment status before disease, employment status at baseline, years of education, smoking habits, alcohol consumption habits, weight, height, BMI, number of pathologies, biologic therapy, concomitant disease-modifying antirheumatic drug at baseline, concomitant corticoid at baseline, baseline CRP, baseline ESR, baseline BASDAI, baseline BASFI, and baseline ASDAS;
3. Time-dependent variables—CRP, ESR, BASDAI, BASFI, and ASDAS.

The ASDAS we refer to here and henceforth is the one that incorporates the CRP value into its calculation and corresponds to the ASDAS-CRP.

A thorough inspection of all variables was made to identify any incomplete or incorrect values. Some examples of issues that arose were values with incorrect formats or incoherent with these variables' possible range. If possible, by crossing information and with medical professionals' help and consultation, the values were corrected, but whenever it was impossible to draw conclusions, the observations were eliminated.

In the presented initial data set, not all patients have every baseline variable available. This poses an issue and is a challenge that needs to be dealt with in most studies that use data from clinical settings. The issue arises because most methods of variable selection and statistical models cannot handle missing values. Therefore, to further proceed with our



analysis, we needed to understand the different approaches that can be used to handle this problem according to our needs. Common approaches include performing complete-case analysis, removing individuals with incomplete data for a subset of covariates, and multiple imputation techniques.

We decided not to perform any imputation techniques, as the imputation of baseline variables could introduce a high bias in our models' estimates.

On the other hand, there is always value in keeping the most amount of data as possible to avoid wasting relevant information.

As there was no obvious choice regarding which approach would be the most appropriate and to enable the drawing of valuable insights from the data, the decision to create 4 different data sets, according to different approaches, was made. Furthermore, this allowed us to study how the strategy for handling missing data and the resulting data differences can influence the modeling process and the subsequent results. The overall process for the creation of these data sets is depicted in Figure 1.

The first approach consisted of keeping only the patients for whom all baseline variables were available (ie, keeping only the complete cases).

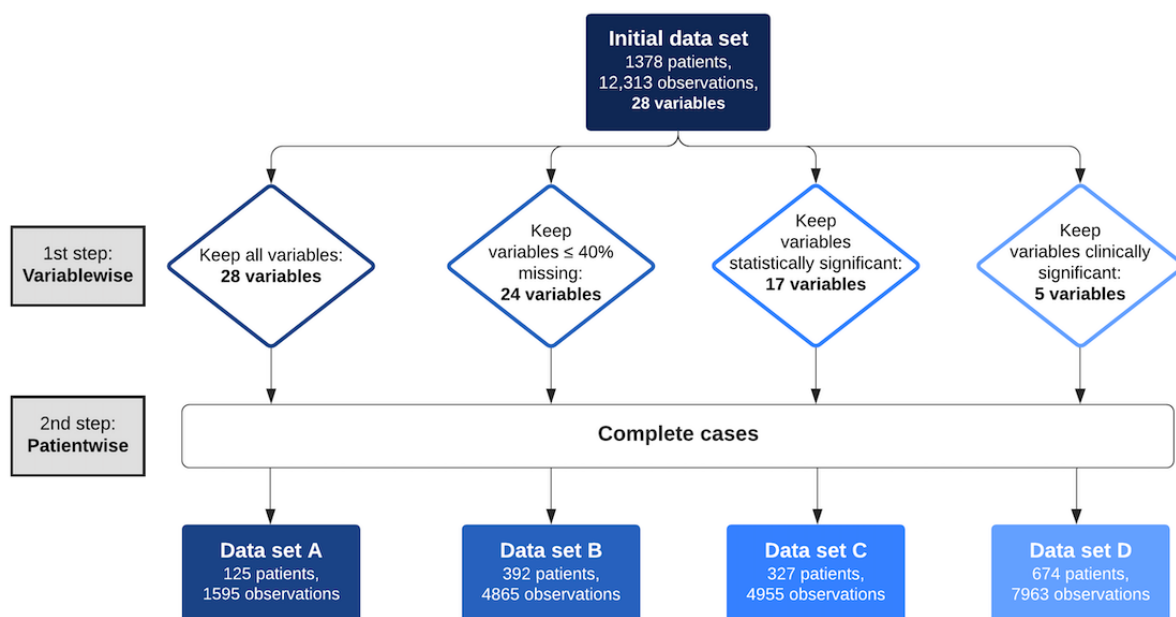
The second approach was to consider only the variables with fewer than 40% of missing values and then keeping the complete cases of those variables. The percentage of 40% was chosen since it seemed to provide a good balance between the number of eliminated variables and the number of eliminated patients.

The third approach consisted of fitting a univariate Cox model for each initial baseline variable and then keeping only the statistically significant ones in those models, according to a 5% level. After that, the complete cases of those variables were once again kept.

The last approach was to keep only those variables that were considered clinically relevant by expert medical doctors' knowledge and according to insight from literature research where predictors of biologic drug survival in SpA were studied [21-23].

The variables selected for consideration were sex, disease years to first biologic therapy, age at start of the first biologic therapy, education years, baseline CRP, baseline BASDAI, and baseline BASFI. Variables age at start of the first biologic therapy and baseline BASDAI were later dropped due to violation of the proportional hazards assumption.

**Figure 1.** Flowchart representing the overall approach for the preprocessing of the initial data into 4 new data sets: A, B, C, and D.



### Statistical Model Implementation

For the initial data set, both an overall survival curve and curves for survival according to the biologic therapy were fitted with the Kaplan-Meier estimator [3].

Regarding the 4 processed data sets, the same approach was used for all data sets, which proceeded as follows.

The first step in the analysis was to perform variable selection for the baseline covariates. This removed any unnecessary predictors that could have added noise to the estimations.

Five different methods were used to compare and study the variability of the obtained results. These were backward stepwise selection using Akaike information criterion (AIC), forward stepwise selection using AIC, best subset selection using a primal-dual active set approach, lasso regression, and the stepwise likelihood ratio variable selection strategy presented by Collett [24].

Despite this, the variables obtained from the stepwise likelihood ratio variable selection were the ones ultimately selected for the next steps of the analysis, namely for building the survival submodel.

This variable selection was not performed for data set D, as the medical experts selected the variables of interest for this specific case.

A Cox model for the survival submodel was then fitted using the selected baseline covariates, constituting the survival submodel.

For each of the time-dependent variables, 7 different LME models were fitted. The one with the better fit according to AIC and Bayesian information criterion was chosen as the time-dependent submodel.

The formulae of the different LME models fitted and the corresponding names are presented in Table 1.

Having both submodels, the same joint models were fitted with R packages “JM” [18] and “Jmbayes” [19]. The former estimates the model under a maximum likelihood approach and the latter under a Bayesian approach, more specifically, using Markov chain Monte Carlo algorithms.

The R package “Jmbayes” also enables the fitting of multivariate joint models. Considering that variables CPR and ESR are both measurements of inflammation and that ASDAS uses CRP and BASDAI elements in its composition, we chose not to fit together CRP with ESR or ASDAS with CRP and BASDAI. Thus, 2 different combinations of time-dependent variables were considered for the multivariate joint models: CRP, BASDAI, and BASFI; and BASFI and ASDAS.

The equivalent models, both univariate and multivariate, were also fitted with an extended Cox model, which enabled the comparison of both methods.

Should a case arise where the survival submodel contained any of the variables of baseline CRP, baseline ESR, baseline BASDAI, baseline BASFI, or baseline ASDAS, these baseline variables would be dropped when the correspondent time-dependent variable was present in the univariate joint model or extended Cox model. This would enable us to compare the effect of the variable in its baseline form with its time-dependent form. Similarly, if more than one of the 5 baseline variables were present in the survival submodel, a multivariate joint model or extended Cox model would also be fitted with those variables in the time-dependent form, and the baseline form would be dropped from the survival submodel.

The overall process for fitting the joint models and extended Cox models is schematically presented in Figure 2 and Figure 3, respectively, where it is also possible to observe the numbers given to the models that were fitted for every data set.

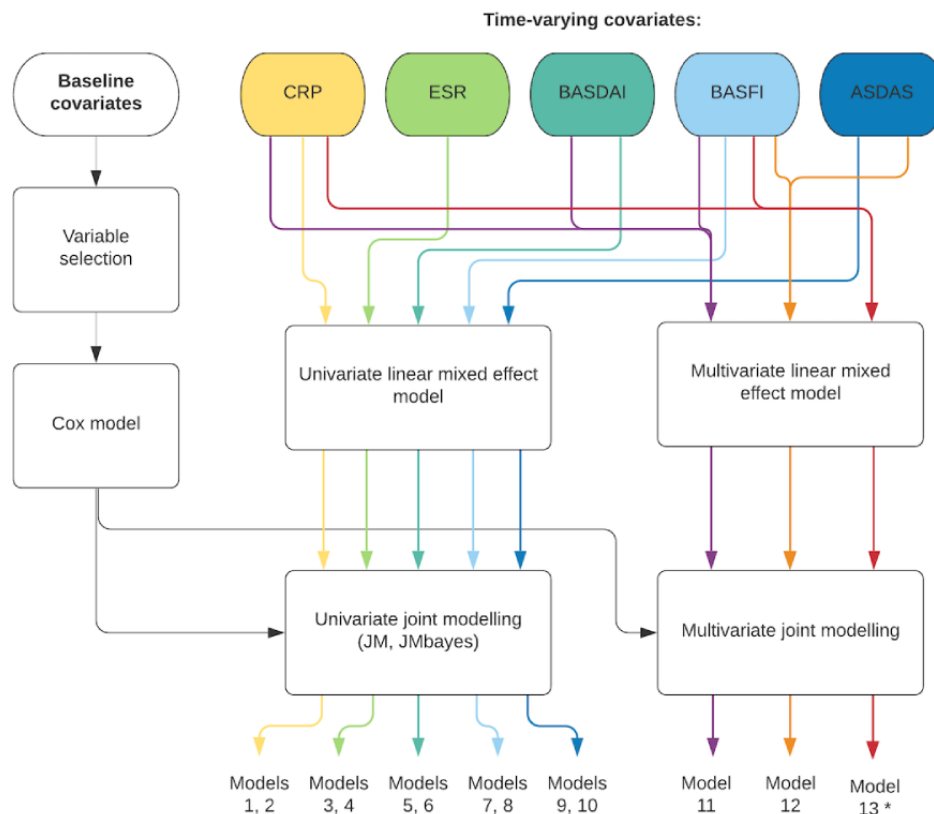
The exhaustive tests performed attempted to cover several types of strategies and models, and were aimed at identifying key covariates involved in the prognosis of the disease, particularly in the response to treatment. Indeed, the rationale for this approach was to comprehensively span the described methods due to the fact that this specific Reuma.pt data set did not contain any prior studies that focused on the identification of specific markers for the prognosis of the patient’s therapy response.

All the analysis was performed using R software [20], particularly, the “MASS” [25], “BeSS” [26], and “glmnet” [27] packages for the forward and backward stepwise variable selection, best subset selection, and lasso regression, respectively. Furthermore “car” [28] was used for multicollinearity testing; “survival” [29] for the Kaplan-Meier curves, Cox model, extended Cox model, and proportional hazards testing; “survminer” [30] for the plotting of survival curves; “nlme” [31] for fitting the linear mixed-effects models; and “JM” [18] and “Jmbayes” [19] for fitting the joint models.

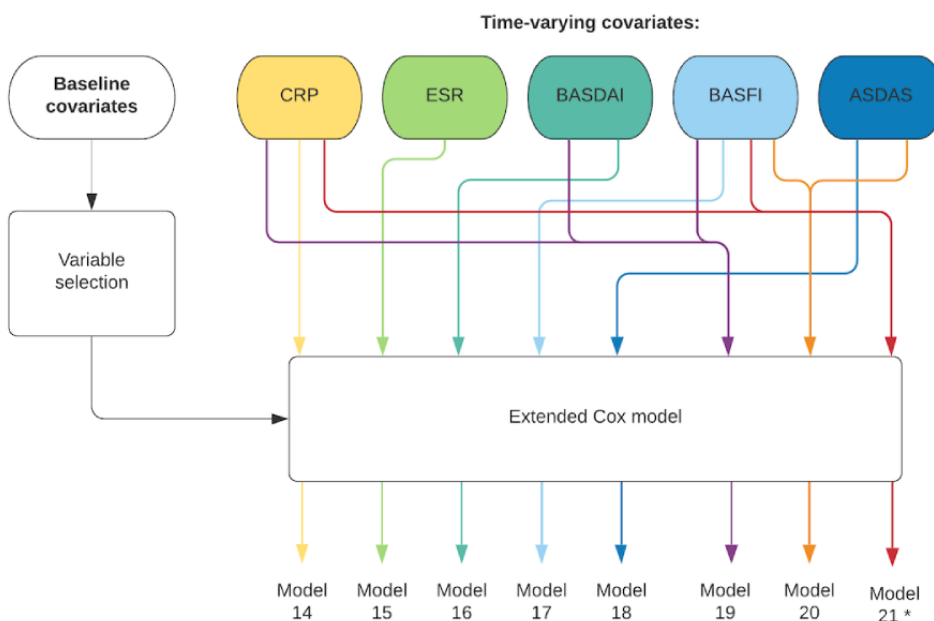
**Table 1.** Time-dependent functions used given a time-varying variable *y*. NC represents natural cubic spline function;  $\beta$ , fixed effects, *b*, random effects; and *t*, time.

Model	Time-dependent functions
Linear and random intercept	$\beta_0 + \beta_1 t_{ij} + b_{i0} + \epsilon_{ij}$
Linear and random slope	$\beta_0 + \beta_1 t_{ij} + b_{i0} + b_{i0} t_{ij} + \epsilon_{ij}$
Cubic and random intercept	$\beta_0 + \beta_1 t_{ij} + \text{NC}(t_{ij}) + b_{i0} + \epsilon_{ij}$
Cubic and random slope	$\beta_0 + \beta_1 t_{ij} + \text{NC}(t_{ij}) + b_{i0} + b_{i1} t_{ij} + \epsilon_{ij}$
Spline and random intercept	$\text{NC}(t_{ij}, 2, (\beta_0, \beta_1, \beta_2, \beta_3)^T, (b_{i0}) + \epsilon_{ij})$
Spline and random slope	$\text{NC}(t_{ij}, 2, (\beta_0, \beta_1, \beta_2, \beta_3)^T, (b_{i0}, b_{i1})^T + \epsilon_{ij})$
Spline and random spline	$\text{NC}(t_{ij}, 2, (\beta_0, \beta_1, \beta_2, \beta_3)^T, (b_{i0}, b_{i1}, b_{i2}, b_{i3})^T + \epsilon_{ij})$

**Figure 2.** Flowchart representing the overall approach for the data analysis using univariate and multivariate joint modeling. The variable selection step is not performed for data set D. ASDAS: Ankylosing Spondylitis Disease Activity Score; BASDAI: Bath Ankylosing Spondylitis Disease Activity Index; BASFI: Bath Ankylosing Spondylitis Functional Index; CRP: C-reactive protein; ESR: erythrocyte sedimentation rate. \*This model is only fitted for data set D.



**Figure 3.** Flowchart representing the overall approach for the data analysis using univariate and multivariate extended Cox modeling. The variable selection step is not performed for data set D. ASDAS: Ankylosing Spondylitis Disease Activity Score; BASDAI: Bath Ankylosing Spondylitis Disease Activity Index; BASFI: Bath Ankylosing Spondylitis Functional Index; CRP: C-reactive protein; ESR: erythrocyte sedimentation rate. \*This model is only fitted for data set D.



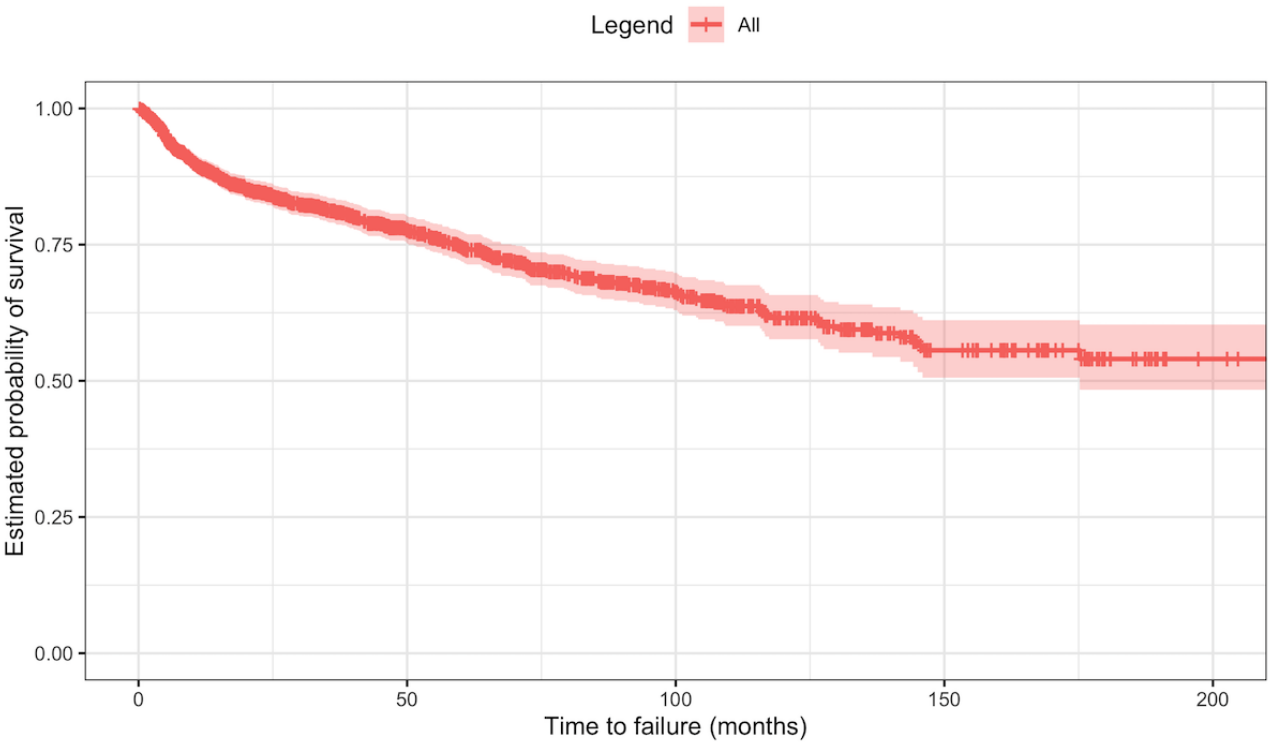
**Ethics Approval and Consent To Participate**

Reuma.pt was approved by the National Data Protection Board (Comissão Nacional de Protecção de Dados, Portugal) and by the Ethics Committee of Centro Hospitalar Lisboa Norte,

Hospital de Santa Maria, Lisbon, Portugal. Patients signed Reuma.pt's informed and written consent.

## Results

### Initial SpA Data Set

The survival probability curve, , of the first biologic therapy for the overall population from the initial data set obtained with the Kaplan-Meier estimator is presented in Figure 4, where vertical ticks along the curve indicate censored patients. We can observe that the slope of the curve is higher at the initial months, indicating that there are more failures closer to the beginning of the therapy.

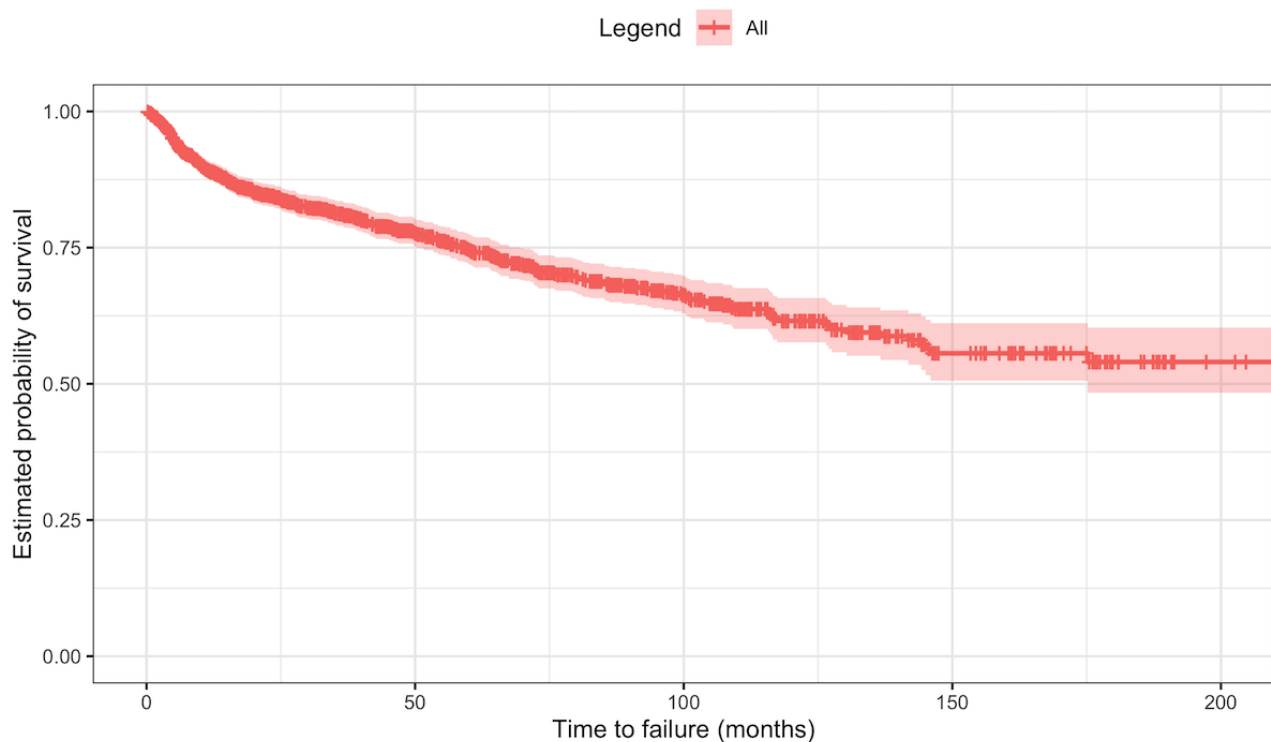
A comparison of the survival probabilities between the different biologic therapies can be seen in Figure 5, where it is possible to observe a clear distinction between the different biologic drugs' curves. The *P* value of the log-rank test is also presented in the figure, indicating that the biologic drug curves differ significantly in survival at a 5% level.

A comparison of the survival probabilities between the different biologic therapies can be seen in Figure 5, where the survival curves for the different biologics were estimated using the Kaplan-Meier method. It is also possible to observe the *P* value of the log-rank test, whose null hypothesis is that all the groups have identical hazard functions. As this value is equal to .03, we can reject this hypothesis at a 5% level of significance.

The pairwise log-rank tests with corrections for multiple testing were also performed for all pairs of biologics to better compare the survival of the therapy between biologics. According to the tests, only 1 pair, etanercept and golimumab, had significantly different survival curves. An analysis of the curve indicates that golimumab conferred better survival than did etanercept at a 5% level of significance.

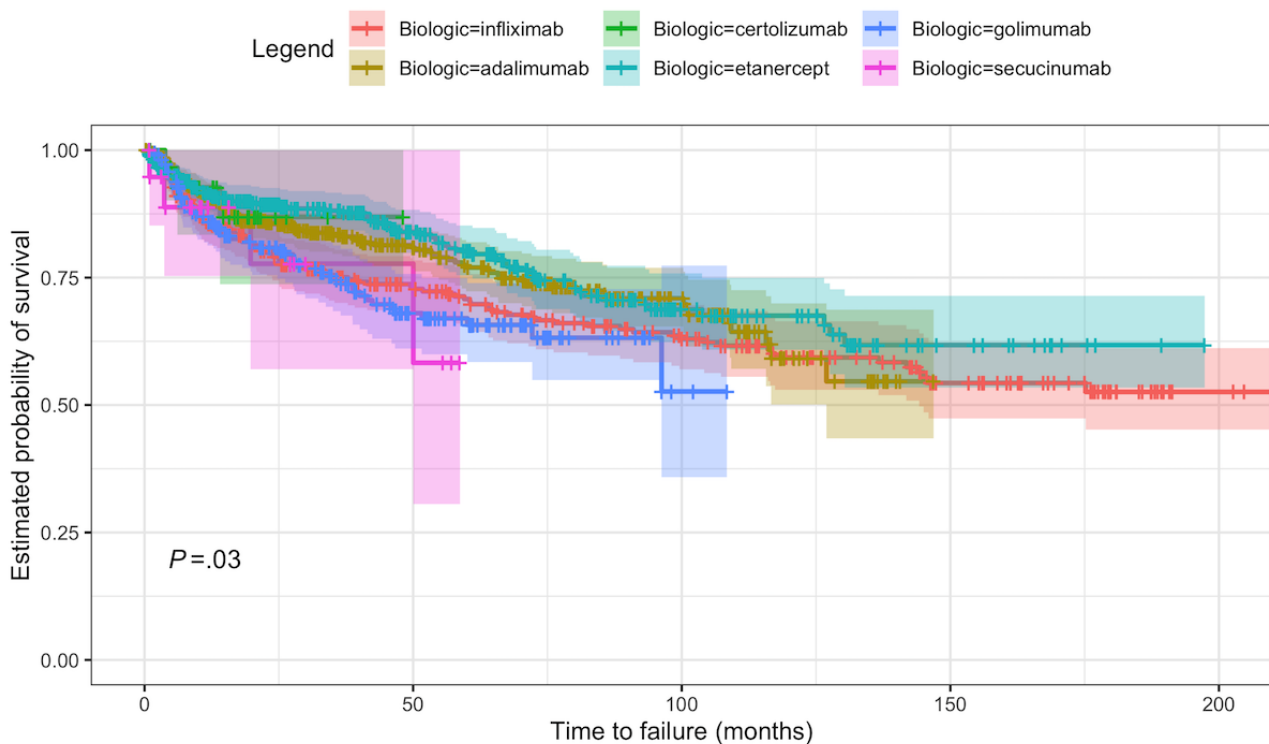
It should also be noted that in the Portuguese SpA subpopulation studied, there were some biologic drugs that had a very small number of observations. This difference in number of observations between different drugs could have also increased the difficulty in properly comparing the survival between them.

**Figure 4.** Kaplan-Meier estimation of the biologic therapy survival curve for the overall population of the initial data set with the 95% CI [5].





**Figure 5.** Kaplan-Meier estimation of the biologic therapy survival curve by biologic drug for the initial data set and *P* value of the correspondent log-rank test with the 95% CI [5].



**Figure 6.** Coefficient signs of the covariates present in the survival submodels fitted with a Cox regression for data sets A, B, C, and D. BASFI: Bath Ankylosing Spondylitis Functional Index; CRP, C-reactive protein; HLA-B27, human leukocyte antigen B27.

Variable	Data set			
	A	B	C	D
<b>Sex (ref: Female)</b>				
Male				-
Year of disease beginning			+	
Year of start of the first biologic	+	+		
Age at start of the first biologic	-			
Disease years until start of the first biologic				-
HLA-B27			-	
Years of education	+	+		+
Baseline CRP				+
Baseline BASFI	+	+	+	+

**Comparison of Results: Data Sets A, B, C, and D**

The first main step of the modeling process consisted of selecting the baseline variables of interest.

The comparison of the selected variables only took data sets A, B, and C into consideration. The percentage of times a variable was selected for each data set, considering the 5 variable selection approaches tested, is presented in Table 2, along with the average number of times it was selected overall.

Although not all variables were present in every data set, the variability in the covariates selected for each set of data was still noticeable. This indicates that the initial process of handling the missing values and the initial selection of variables to be considered at this stage have a somewhat elevated influence on

the results that are obtained later; in short, the results are sensitive to the parameter choice. This difference may be justified by the existence of different variables and even different patients even if many are common between data sets.

Only 1 variable was selected by all methods and for all data sets where it was considered: years of education. On average, the variables that were selected in at least more than 50% of the methods used for variable selection were year of disease onset, age at start of the first biologic therapy, baseline BASFI, and baseline ASDAS.

Table 3 shows the sign of the coefficient for each of the covariates that were included in the survival submodel for each data set. The sign of the respective coefficient indicates the effect of this covariate on the outcome of interest, which we

took as the failure of the first biologic therapy. A positive sign indicates that the variable increases the risk of failure for higher values of that variable (for a continuous variable) or for that value in comparison to the reference level (for a categorical variable); a negative coefficient indicates the opposite: a decrease in the risk of failure.

**Table 2.** Percentage of times a variable was selected for each data set (A, B, C, D) and the average of those percentages across all data sets.

Variable	Data set (%)			Average
	A	B	C	
Sex	0	20	0	7
Marital status	20	N/A <sup>a</sup>	N/A	20
Year of diagnosis	20	40	20	27
Age at diagnosis	40	40	N/A	40
Year of disease onset	60	20	100	60
Age at disease onset	20	20	N/A	20
Year of start of the first biologic therapy	40	60	20	40
Age at start of the first biologic therapy	60	40	N/A	50
Years from diagnosis to start of the first biologic therapy	20	20	N/A	20
Disease years until start of the first biologic therapy	20	20	N/A	20
HLA-B27 <sup>b</sup>	0	40	100	47
Employment status before disease	40	20	N/A	30
Employment status at baseline	20	20	0	13
Years of education	100	100	N/A	100
Smoking	40	40	20	33
Alcohol	20	40	20	27
Weight	40	N/A	N/A	40
Height	20	N/A	20	20
BMI	40	N/A	N/A	40
Number of pathologies	20	0	N/A	10
Biologic	20	20	20	20
Concomitant DMARD <sup>c</sup>	0	20	0	7
Concomitant corticoid	60	20	20	33
Baseline CRP <sup>d</sup>	20	0	20	13
Baseline ESR <sup>e</sup>	40	20	40	33
Baseline BASDAI <sup>f</sup>	40	0	60	33
Baseline BASFI <sup>g</sup>	80	100	40	73
Baseline ASDAS <sup>h</sup>	60	40	60	53

<sup>a</sup>N/A: not applicable.

<sup>b</sup>HLA-B27: human leukocyte antigen B27.

<sup>c</sup>DMARD: disease-modifying antirheumatic drug.

<sup>d</sup>CRP: C-reactive protein.

<sup>e</sup>ESR: erythrocyte sedimentation rate.

<sup>f</sup>BASDAI: Bath Ankylosing Spondylitis Disease Activity Index.

<sup>g</sup>BASFI: Bath Ankylosing Spondylitis Functional Index.

<sup>h</sup>ASDAS: Ankylosing Spondylitis Disease Activity Score.

**Table 3.** Coefficient signs of the covariates present in the survival submodels fitted with a Cox regression for data sets A, B, C, and D.

Variable	Data set			
	A	B	C	D
<b>Sex</b>				
Female (ref)				
Male				-
Year of disease beginning			+	
Year of start of the first biologic therapy	+	+		
Age at start of the first biologic therapy	-			
Disease years until start of the first biologic therapy				-
HLA-B27 <sup>a</sup>			-	
Years of education	+	+		+
Baseline CRP <sup>b</sup>				+
Baseline BASFI <sup>c</sup>	+	+	+	+

<sup>a</sup>HLA-B27: human leukocyte antigen B27.

<sup>b</sup>CRP: C-reactive protein.

<sup>c</sup>BASFI: disease-modifying antirheumatic drug.

It can be noticed that the sign of the coefficient is coherent between the data sets for all variables even if the number of data sets where that variable is present differs.

According to the results obtained in the Cox regression, the factors that indicate a good prognosis for the biologic drug survival were being a male, starting the biologic therapy at an older age, having a larger time interval between disease start and initiation of the first biologic therapy, and being HLA-B27 positive. On the contrary, a disease onset or initiation of biologic therapy in more recent years, a higher number of years of education, and higher values of CRP or BASFI at baseline were all predictors of a greater risk of failure of the first biologic therapy.

Given the elevated number of models fitted and to aid in the drawing of comprehensive conclusions, [Table 4](#) and [Table 5](#) were created to depict the joint and extended Cox models, respectively. For each data set and for every variable present in the model, the tables show the percentage of models (relative to the total number of fitted models for each data set) in which the covariate was statistically significant, the percentage of models in which the variable had a positive regression coefficient, and the percentage of models in which the variable had a negative regression coefficient. Furthermore, the average of these percentages was calculated to obtain an overall view of the most common behavior of each variable as determined by information gathered from all data sets.

**Table 4.** Percentage of models in which a variable was statistically significant; percentage of models in which a variable had a positive coefficient sign; the percentage of models in which a variable had a negative coefficient sign for the covariates present in the joint models fitted for data sets A, B, C, and D; and the average of those percentages across all data sets.

Variable	Data set (%)												Average (%)			
	A			B			C			D			ss	pos	neg	
	ss <sup>a</sup>	pos <sup>b</sup>	neg <sup>c</sup>	ss	pos	neg	ss	pos	neg	ss	pos	neg				
Male (ref: female)	N/A <sup>d</sup>	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0	46	54	0	46	54
Year of disease onset	N/A	N/A	N/A	N/A	N/A	N/A	92	75	25	N/A	N/A	N/A	92	75	25	
Year biologic therapy initiation	82	36	64	83	67	25	N/A	N/A	N/A	N/A	N/A	N/A	83	52	45	
Age at start of the first biologic therapy	82	0	100	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	82	0	100	
Disease years until start of the first biologic therapy	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	15	0	100	15	0	100
HLA-B27 <sup>e</sup>	N/A	N/A	N/A	N/A	N/A	N/A	83	0	100	N/A	N/A	N/A	83	0	100	
Years of education	82	100	0	100	100	0	N/A	N/A	N/A	85	100	0	91	100	0	
Baseline CRP <sup>f</sup>	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	44	100	0	44	100	0	
Baseline BASFI <sup>g</sup>	50	100	0	75	100	0	25	63	37	50	100	0	50	91	9	
CRP	67	100	0	100	100	0	100	100	0	100	100	0	92	100	0	
ESR <sup>h</sup>	50	100	0	100	100	0	100	100	0	100	100	0	88	100	0	
BASDAI <sup>i</sup>	100	100	0	100	100	0	100	100	0	100	100	0	100	100	0	
BASFI	33	33	66	50	50	50	50	50	50	60	60	40	48	48	52	
ASDAS <sup>j</sup>	100	100	0	100	100	0	100	100	0	100	100	0	100	100	0	

<sup>a</sup>ss: percentage of models in which variable is statistically significant.

<sup>b</sup>pos: percentage of models in which variable has positive coefficient.

<sup>c</sup>neg: percentage of models in which variable has negative coefficient.

<sup>d</sup>N/A: not applicable.

<sup>e</sup>HLA-B27: human leukocyte antigen B27.

<sup>f</sup>CRP: C-reactive protein.

<sup>g</sup>BASFI: Bath Ankylosing Spondylitis Functional Index.

<sup>h</sup>ESR: erythrocyte sedimentation rate.

<sup>i</sup>BASDAI: Bath Ankylosing Spondylitis Disease Activity Index.

<sup>j</sup>ASDAS: Ankylosing Spondylitis Disease Activity Score.



**Table 5.** Percentage of models in which a variable was statistically significant; percentage of models in which a variable had a positive coefficient sign; and percentage of models in which a variable had a negative coefficient sign for the covariates present in the extended Cox models fitted for data sets A, B, C, and D; and average of those percentages across all data sets.

Variable	Data set												Average			
	A			B			C			D			ss	pos	neg	
	ss <sup>a</sup>	pos <sup>b</sup>	neg <sup>c</sup>	ss	pos	neg	ss	pos	neg	ss	pos	neg				
Male (ref: female)	N/A <sup>d</sup>	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0	25	75	0	25	75
Year of disease onset	N/A	N/A	N/A	N/A	N/A	N/A	100	100	0	N/A	N/A	N/A	100	100	0	0
Year of biologic therapy initiation	100	100	0	100	100	0	N/A	N/A	N/A	N/A	N/A	N/A	100	100	0	0
Age at start of the first biologic therapy	86	0	100	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	86	0	100	0
Disease years until start of the first biologic therapy	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0	0	100	0	0	100
HLA-B27 <sup>e</sup>	N/A	N/A	N/A	N/A	N/A	N/A	86	0	100	N/A	N/A	N/A	86	0	100	0
Years of education	100	100	0	100	100	0	N/A	N/A	N/A	75	100	0	92	100	0	0
Baseline CRP <sup>f</sup>	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	40	80	20	40	80	20	0
Baseline BASFI <sup>g</sup>	50	100	0	50	100	0	0	50	50	50	100	0	38	88	13	0
CRP	0	100	0	50	100	0	0	100	0	100	100	0	38	100	0	0
ESR <sup>h</sup>	0	100	0	100	100	0	100	100	0	100	100	0	75	100	0	0
BASDAI <sup>i</sup>	100	100	0	100	100	0	100	100	0	100	100	0	100	100	0	0
BASFI	33	100	0	67	100	0	33	100	0	100	100	0	58	100	0	0
ASDAS <sup>j</sup>	100	100	0	100	100	0	100	100	0	100	100	0	100	100	0	0

<sup>a</sup>ss: percentage of models in which variable is statistically significant.

<sup>b</sup>pos: percentage of models in which variable has positive coefficient.

<sup>c</sup>neg: percentage of models in which variable has negative coefficient.

<sup>d</sup>N/A: not applicable.

<sup>e</sup>HLA-B27: human leukocyte antigen B27.

<sup>f</sup>CRP: C-reactive protein.

<sup>g</sup>BASFI: Bath Ankylosing Spondylitis Functional Index.

<sup>h</sup>ESR: erythrocyte sedimentation rate.

<sup>i</sup>BASDAI: Bath Ankylosing Spondylitis Disease Activity Index.

<sup>j</sup>ASDAS: Ankylosing Spondylitis Disease Activity Score.

This consensus or ensemble approach was conducted to facilitate the identification of the most significant variables. The rationale is that if a feature always appears as significant, independently of the specific chosen model, then there is evidence that this feature is associated with the outcome.

Similar reasoning is applicable for identifying the covariates' effect on the event of interest; that is, its positive or negative contribution for the risk of the therapy failure.

Focusing on the time-independent variables and starting with the covariate that represents male sex, we can see that this variable is not statistically significant in any joint or extended Cox model. This is coherent with what was observed in the Cox model, where sex was not a statistically significant predictor for our outcome. Regarding the effect of the variable on the event of interest, being male was more frequently a good predictor of biologic therapy survival than a bad predictor although this ratio was very small for the joint models.

The year of disease beginning was statistically significant for most models it was present in even if only 1 data set analyzed this variable. Its associated coefficient was positive for all the extended Cox models and for an average of 92% (12/13) of the joint models, which is consistent with the result obtained in the Cox models, indicating that patients with a more recent onset of disease have a higher risk of treatment failure.

The year of start of the first biologic therapy appears as a statistically significant predictor in most joint models and in all extended Cox models. For the majority of models, biologic therapy initiated in more recent years appeared to increase the risk of its failure.

The age at the start of the first biologic was statistically significant in most joint and extended Cox models, and older age at the time of therapy initiation was consistently a predictor of decreased risk of failure.

The year interval between disease beginning and start of the first biologic therapy was only statistically significant in a small percentage of joint models and showed no significance in any of the extended Cox models, echoing the results for the Cox model. The coefficient sign for this covariate was coherent among all models, indicating that a larger year interval reduces the chance of biologic therapy failure.

Being HLA-B27 positive was statistically significant as a predictor for biologic therapy failure in approximately 80% of all models and was consistently associated with a decreased risk of failure when in comparison with HLA-B27-negative patients.

The number of years of education showed statistical significance in roughly 90% of all joint and extended Cox models, and a higher number of education years increased the hazard of biologic therapy failure for all Cox, extended Cox, and joint models.

The value of CRP at baseline had an associated positive regression coefficient in all joint models and in 80% (4/5) of extended Cox models, indicating an increased risk of failure for higher CRP values, which was also verified in the Cox model. This covariate was not statistically significant in the Cox regression, but was significant in approximately 40% of the joint and extended Cox models.

The baseline BASFI was statistically significant in fewer than half of the joint and extended Cox models, even though it was always statistically significant in the survival submodels fitted with a Cox model. This variable appeared to be a predictor for increased risk of biologic therapy failure in most joint and extended Cox models, which is concordant with the Cox models' results.

Regarding the time-dependent variables, we noticed that, for the joint models, variables CRP, ESR, BASDAI, and ASDAS appeared to be statistically significant in most models. Furthermore, all were predictors of increased therapy failure for all the joint models that were fitted. Variable BASFI was only statistically significant in approximately half of the joint models, and the sign of its coefficient also varied considerably, not showing any clear tendency regarding the effect of this variable on the outcome.

In the extended Cox models, only variables BASDAI and ASDAS were statistically significant for all models. Variable ESR, BASFI, and CRP showed statistical significance in 75%, 58%, and 38% of the models, respectively. In all the models, all 5 time-varying covariates were predictors of increased risk of biologic failure.

## Discussion

### Principal Results

Overall, the results obtained from the Cox models, extended Cox models, and joint models all indicated similar effects of the covariates on the treatment outcome.

The biomarkers that indicated a good prognosis for the biologic drug survival were being male, starting biologic therapy at an

older age, having a larger time interval between disease onset and initiation of the first biologic drug, and being HLA-B27 positive.

Conversely, disease onset or initiation of biologic therapy in more recent years, a greater number of education years, and higher values of CRP or BASFI at baseline all appeared to be predictors of a greater risk of failure of the first biologic therapy.

### Comparison With Prior Work

Male sex [22], HLA-B27-positive status [32], and longer disease duration [21] have been reported in the literature as being good predictors of biologic drug survival, which concurs with the results obtained in the Cox models of our study. On the other hand, older age at the start of the biologic therapy [32] has been reported to increase the risk of failure of the therapy, which is contrary to what was found in our data. We could interpret our result by speculating that older patients are more complacent due to the perceived efficacy of the therapy or because their symptoms are more intense than those of younger people and thus the relative improvement of symptoms is more noticeable, therefore increasing their satisfaction levels and decreasing the chances of therapy switch.

Regarding the predictors that were found to increase the risk of failure in our study, starting treatment in more recent years [33], and higher values of BASFI at baseline [21] were likewise found to be predictors of biologic drug discontinuation in research publications. A higher number of education years [23] was reported in one study as decreasing the risk of therapy failure, which differed from our results. Again, we could speculate and say that patients with a higher academic level are more comfortable with expressing their discontent with the lack of therapy response or that they are more aware of new therapeutic options and for that reason, request a switch of the therapy more often.

Higher values of CRP at baseline were found to increase [34] the hazard of biologic therapy failure in some publications but to have the opposite effect [23,33] in others.

### Limitations

Some limitations of our study include the suboptimal fitting of the longitudinal variables. Therefore, the choice of the LME function for describing the biomarker trajectories and the different functional forms available that specify the association between the longitudinal biomarker and the hazard function of the event should be further explored.

### Conclusions

Joint models are statistical models that can analyze both time-static and time-varying variables and therefore enable the inference of relationships between time-to-event and longitudinal data that are widely present in electronic medical records.

In this work, this modeling approach was selected to investigate biologic drug survival and its predictors for SpA patients in Portugal.

Furthermore, the insights obtained throughout the process that culminated in the fitting of these models are also highly valuable.

This study was the first to use the data of SpA patients from Reuma.pt in this capacity. The entire preprocessing work performed for enabling the use of Reuma.pt produced a data set that can be used by researchers who wish to investigate this group of diseases.

The variable selection process appears to be sensitive to this data preprocessing step depending on which variables and patients are described in the data set.

The tested methods for variable selection yielded quite different results for the same set of data. The process of selection of covariates should be analyzed carefully, as fully automated methods may not be the most appropriate ones for establishing which variables should be included in the statistical model. A wise approach consists of a balance between statistical significance and clinical significance, with the study's goal always being kept in mind.

We demonstrated that joint models, particularly the functions implemented in the R software packages “JM” and “Jmbayes,” can be successfully used for the simultaneous analysis of time-to-event and longitudinal data.

Health care providers use rheumatic disease progression measures computed from disease activity scores to shape treatment strategies and improve the quality of life of their patients [35]. However, targeted treatments that save patients from the potential side effects of high-cost, unsatisfactory treatments require the identification of biomarkers that can determine which patients can profit from a given therapy [36]. Computational methods using statistical and machine learning methods hold promise for the overall understanding of rheumatic diseases and can aid in formulating therapeutic strategies and predicting prognosis and outcome [37,38].

With this study, it was possible to identify the potential predictors of biologic therapy failure for this Portuguese population of SpA patients. This can aid the prognosis of these rheumatic diseases and potentially predict the most adequate treatment option according to the patient's characteristics.

---

## Acknowledgments

This study was supported by the Portuguese Foundation for Science and Technology (Fundação para a Ciência e Tecnologia) through the Instituto de Telecomunicações (UIDB/50008/2020), INESC-ID (UIDB/50021/2020), and projects MATISSE (DSAIPA/DS/0026/2019) and PREDICT (PTDC/CCI-CIF/29877/2017). This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 951970 (OLISSIPO project). We also acknowledge Sociedade Portuguesa de Reumatologia and all Reuma.pt contributors.

---

## Authors' Contributions

CB processed the data, performed the fitting of the models and computational experiments, and wrote the first draft of the manuscript (all authors made the required updates). AMR and HC provided the data, clinical insights, and interpretation. AMC and SV conceived the study, supervised the research, generated the final results, and wrote the manuscript. All authors contributed to the final draft, and read and approved the final version of the manuscript.

---

## Conflicts of Interest

None declared.

---

## References

1. Branco JC, Rodrigues AM, Gouveia N, Eusébio M, Ramiro S, Machado PM, EpiReumaPt study group. Prevalence of rheumatic and musculoskeletal diseases and their impact on health-related quality of life, physical function and mental health in Portugal: results from EpiReumaPt- a national health survey. *RMD Open* 2016;2(1):e000166 [FREE Full text] [doi: [10.1136/rmdopen-2015-000166](https://doi.org/10.1136/rmdopen-2015-000166)] [Medline: [26848402](https://pubmed.ncbi.nlm.nih.gov/26848402/)]
2. Jones A, Ciurtin C, Ismajli M, Leandro M, Sengupta R, Machado PM. Biologics for treating axial spondyloarthritis. *Expert Opin Biol Ther* 2018 Jun;18(6):641-652. [doi: [10.1080/14712598.2018.1468884](https://doi.org/10.1080/14712598.2018.1468884)] [Medline: [29681195](https://pubmed.ncbi.nlm.nih.gov/29681195/)]
3. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958 Jun;53(282):457-481. [doi: [10.1080/01621459.1958.10501452](https://doi.org/10.1080/01621459.1958.10501452)]
4. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 2018 Dec 05;34(2):187-202. [doi: [10.1111/j.2517-6161.1972.tb00899.x](https://doi.org/10.1111/j.2517-6161.1972.tb00899.x)]
5. Kalbfleisch J, Prentice, RL. *The Statistical Analysis of Failure Time Data*. In: 2nd edition. John Wiley & Sons. New York: John Wiley & Sons; 2011:978.
6. Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Boca Raton: CRC press; 2012.
7. Rizopoulos D, Takkenberg JJM. *EuroIntervention* 2014 Jun;10(2):285-288 [FREE Full text] [doi: [10.4244/EIJV10I2A47](https://doi.org/10.4244/EIJV10I2A47)] [Medline: [24952063](https://pubmed.ncbi.nlm.nih.gov/24952063/)]

8. Sudell M, Kolamunnage-Dona R, Tudur-Smith C. Joint models for longitudinal and time-to-event data: a review of reporting quality with a view to meta-analysis. *BMC Med Res Methodol* 2016 Dec 05;16(1):168 [FREE Full text] [doi: [10.1186/s12874-016-0272-6](https://doi.org/10.1186/s12874-016-0272-6)] [Medline: [27919221](https://pubmed.ncbi.nlm.nih.gov/27919221/)]
9. Welsing P, Broeder A, Tekstra J. SAT0116 Dynamic prediction of flares in rheumatoid arthritis using joint modeling and machine learning: simulation of clinical impact when used as decision aid in a disease activity guided dose reduction strategy. *Annals of the Rheumatic Diseases* 2019;78:1125-1126. [doi: [10.1136/annrheumdis-2019-eular.2881](https://doi.org/10.1136/annrheumdis-2019-eular.2881)]
10. Lim LSH, Pullenayegum E, Moineddin R, Gladman DD, Silverman ED, Feldman BM. *Pediatr Rheumatol Online J* 2017 Mar 29;15(1):18 [FREE Full text] [doi: [10.1186/s12969-017-0148-2](https://doi.org/10.1186/s12969-017-0148-2)] [Medline: [28356102](https://pubmed.ncbi.nlm.nih.gov/28356102/)]
11. Canhão H, Faustino A, Martins F, Fonseca JE, Rheumatic Diseases Portuguese Register Board Coordination, Portuguese Society of Rheumatology. Reuma.pt - the rheumatic diseases Portuguese register. *Acta Reumatol Port* 2011;36(1):45-56 [FREE Full text] [Medline: [21483280](https://pubmed.ncbi.nlm.nih.gov/21483280/)]
12. Khan, MA. *Ankylosing spondylitis*. Oxford: Oxford University Press; 2009:9780195368079.
13. Espondilite Anquilosante tem um impacto económico em Portugal de 639 milhões de euros por ano. Novartis. URL: <https://www.novartis.pt/news/media-releases/espondilite-anquilosante-tem-um-impacto-economico-em-portugal-de-639-milhoes-de> [accessed 2020-11-26]
14. Calin A, Garrett S, Whitelock H, Kennedy LG, O'Hea J, Mallorie P, et al. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. *J Rheumatol* 1994 Dec;21(12):2281-2285. [Medline: [7699629](https://pubmed.ncbi.nlm.nih.gov/7699629/)]
15. Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. *J Rheumatol* 1994 Dec;21(12):2286-2291. [Medline: [7699630](https://pubmed.ncbi.nlm.nih.gov/7699630/)]
16. Lukas C, Landewé R, Sieper J, Dougados M, Davis J, Braun J, Assessment of SpondyloArthritis International Society. Development of an ASAS-endorsed disease activity score (ASDAS) in patients with ankylosing spondylitis. *Ann Rheum Dis* 2009 Jan;68(1):18-24. [doi: [10.1136/ard.2008.094870](https://doi.org/10.1136/ard.2008.094870)] [Medline: [18625618](https://pubmed.ncbi.nlm.nih.gov/18625618/)]
17. Molenberghs G, Verbeke G. *Linear Mixed Models for Longitudinal Data*. New York: Springer; 2000.
18. Rizopoulos D. JM: an R package for the joint modelling of longitudinal and time-to-event data. *J. Stat. Soft* 2010;35(9):1-33. [doi: [10.18637/jss.v035.i09](https://doi.org/10.18637/jss.v035.i09)]
19. Rizopoulos D. The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *J. Stat. Soft* 2016;72(7):1-46. [doi: [10.18637/jss.v072.i07](https://doi.org/10.18637/jss.v072.i07)]
20. R: a language and environment for statistical computing. R Core Team. URL: <http://softlibre.unizar.es/manuales/aplicaciones/r/fullrefman.pdf> [accessed 2020-11-26]
21. Glinborg B, Østergaard M, Krogh NS, Tarp U, Manilo N, Loft AGR, et al. Clinical response, drug survival and predictors thereof in 432 ankylosing spondylitis patients after switching tumour necrosis factor  $\alpha$  inhibitor therapy: results from the Danish nationwide DANBIO registry. *Ann Rheum Dis* 2013 Jul;72(7):1149-1155. [doi: [10.1136/annrheumdis-2012-201933](https://doi.org/10.1136/annrheumdis-2012-201933)] [Medline: [22941767](https://pubmed.ncbi.nlm.nih.gov/22941767/)]
22. Flouri ID, Markatseli TE, Boki KA, Papadopoulos I, Skopouli FN, Voulgari PV, et al. Comparative analysis and predictors of 10-year tumor necrosis factor inhibitors drug survival in patients with spondyloarthritis: first-year response predicts longterm drug persistence. *J Rheumatol* 2018 Jun;45(6):785-794 [FREE Full text] [doi: [10.3899/jrheum.170477](https://doi.org/10.3899/jrheum.170477)] [Medline: [29606666](https://pubmed.ncbi.nlm.nih.gov/29606666/)]
23. Hebeisen M, Scherer A, Micheroli R, Nissen MJ, Tamborrini G, Möller B, et al. Comparison of drug survival on adalimumab, etanercept, golimumab and infliximab in patients with axial spondyloarthritis. *PLoS One* 2019;14(5):e0216746 [FREE Full text] [doi: [10.1371/journal.pone.0216746](https://doi.org/10.1371/journal.pone.0216746)] [Medline: [31145730](https://pubmed.ncbi.nlm.nih.gov/31145730/)]
24. Collett D. Variable selection procedures. In: Collett D, editor. *Modeling Survival Data in Medical Research*. Boca Raton: CRC press; 2015:252-272.
25. Venables W, Ripley B. *Modern Applied Statistics With S-PLUS*. New York: Springer Science & Business Media; 2013.
26. Wen C, Zhang A, Quan S, Wang X. BeSS: an R Package for best subset selection in linear, logistic and Cox proportional hazards models. *J. Stat. Soft* 2020;94(4):1-24. [doi: [10.18637/jss.v094.i04](https://doi.org/10.18637/jss.v094.i04)]
27. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Soft* 2010;33(1):1-22. [doi: [10.18637/jss.v033.i01](https://doi.org/10.18637/jss.v033.i01)]
28. Fox J, Weisberg S. *An R Companion to Applied Regression*. eBook: Sage publications; 2018.
29. Therneau T, Grambsch P. *Modeling Survival Data: extending the Cox model*. New York: Springer Science & Business Media; 2013.
30. Kassambara A, Kosinski M, Biecek P. survminer: drawing survival curves using 'ggplot2'. *Datanovia*. 2017. URL: <https://rpkgs.datanovia.com/survminer/> [accessed 2020-11-24]
31. Pinheiro J, Bates D, DebRoy S. Package nlme: linear and nonlinear mixed effects models., The R Project. 2017. URL: <https://cran.r-project.org/web/packages/nlme/nlme.pdf> [accessed 2020-11-24]
32. Maneiro JR, Souto A, Salgado E, Mera A, Gomez-Reino JJ. Predictors of response to TNF antagonists in patients with ankylosing spondylitis and psoriatic arthritis: systematic review and meta-analysis. *RMD Open* 2015;1(1):e000017 [FREE Full text] [doi: [10.1136/rmdopen-2014-000017](https://doi.org/10.1136/rmdopen-2014-000017)] [Medline: [26509050](https://pubmed.ncbi.nlm.nih.gov/26509050/)]



33. Sepriano A, Ramiro S, van der Heijde D, Ávila-Ribeiro P, Fonseca R, Borges J, et al. eEffect of comedication with conventional synthetic disease-modifying antirheumatic drugs on retention of tumor necrosis factor inhibitors in patients with spondyloarthritis: a prospective cohort study. *Arthritis Rheumatol* 2016 Nov;68(11):2671-2679 [FREE Full text] [doi: [10.1002/art.39772](https://doi.org/10.1002/art.39772)] [Medline: [27273894](https://pubmed.ncbi.nlm.nih.gov/27273894/)]
34. Glinthorg B, Ostergaard M, Krogh NS, Dreyer L, Kristensen HL, Hetland ML. Predictors of treatment response and drug continuation in 842 patients with ankylosing spondylitis treated with anti-tumour necrosis factor: results from 8 years' surveillance in the Danish nationwide DANBIO registry. *Ann Rheum Dis* 2010 Nov;69(11):2002-2008. [doi: [10.1136/ard.2009.124446](https://doi.org/10.1136/ard.2009.124446)] [Medline: [20511613](https://pubmed.ncbi.nlm.nih.gov/20511613/)]
35. Farheen K, Agarwal SK. Assessment of disease activity and treatment outcomes in rheumatoid arthritis. *J Manag Care Pharm* 2011;17(9 Suppl B):S09-S13. [doi: [10.18553/jmcp.2011.17.s9-b.s09](https://doi.org/10.18553/jmcp.2011.17.s9-b.s09)] [Medline: [22073934](https://pubmed.ncbi.nlm.nih.gov/22073934/)]
36. Gavrilă BI, Ciofu C, Stoica V. Biomarkers in rheumatoid arthritis, what is new? *J Med Life* 2016;9(2):144-148 [FREE Full text] [Medline: [27453744](https://pubmed.ncbi.nlm.nih.gov/27453744/)]
37. Pandit A, Radstake TRDJ. Machine learning in rheumatology approaches the clinic. *Nat Rev Rheumatol* 2020 Feb;16(2):69-70. [doi: [10.1038/s41584-019-0361-0](https://doi.org/10.1038/s41584-019-0361-0)] [Medline: [31908355](https://pubmed.ncbi.nlm.nih.gov/31908355/)]
38. Jiang M, Li Y, Jiang C, Zhao L, Zhang X, Lipsky PE. Machine learning in rheumatic diseases. *Clin Rev Allergy Immunol* 2021 Feb;60(1):96-110. [doi: [10.1007/s12016-020-08805-6](https://doi.org/10.1007/s12016-020-08805-6)] [Medline: [32681407](https://pubmed.ncbi.nlm.nih.gov/32681407/)]

## Abbreviations

**AIC:** Akaike information criterion

**AS:** ankylosing spondylitis

**ASDAS:** ankylosing spondylitis disease activity score

**BASDAI:** Bath Ankylosing Disease Status in Ankylosing Spondylitis

**BASFI:** Bath Ankylosing Spondylitis Functional Index

**CRP:** C-reactive protein

**ESR:** erythrocyte sedimentation rate

**HLA:** human leukocyte antigen

**LME:** linear mixed effects

**SpA:** spondyloarthritis

*Edited by C Lovis; submitted 29.12.20; peer-reviewed by C Amado, S Kardes; comments to author 20.02.21; revised version received 13.04.21; accepted 23.04.21; published 30.07.21.*

*Please cite as:*

Barata C, Rodrigues AM, Canhão H, Vinga S, Carvalho AM

*Predicting Biologic Therapy Outcome of Patients With Spondyloarthritis: Joint Models for Longitudinal and Survival Analysis*

*JMIR Med Inform* 2021;9(7):e26823

URL: <https://medinform.jmir.org/2021/7/e26823>

doi: [10.2196/26823](https://doi.org/10.2196/26823)

PMID: [34328435](https://pubmed.ncbi.nlm.nih.gov/34328435/)

©Carolina Barata, Ana Maria Rodrigues, Helena Canhão, Susana Vinga, Alexandra M Carvalho. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 30.07.2021. 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

# Relation Classification for Bleeding Events From Electronic Health Records Using Deep Learning Systems: An Empirical Study

Avijit Mitra<sup>1</sup>, MSc; Bhanu Pratap Singh Rawat<sup>1</sup>, MSc; David D McManus<sup>2</sup>, MSc, MD; Hong Yu<sup>1,2,3,4</sup>, PhD

<sup>1</sup>College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA, United States

<sup>2</sup>Department of Medicine, University of Massachusetts Medical School, Worcester, MA, United States

<sup>3</sup>Department of Computer Science, University of Massachusetts Lowell, Lowell, MA, United States

<sup>4</sup>Center for Healthcare Organization and Implementation Research, Bedford Veterans Affairs Medical Center, Bedford, MA, United States

**Corresponding Author:**

Hong Yu, PhD

Department of Computer Science

University of Massachusetts Lowell

1 University Avenue

Lowell, MA,

United States

Phone: 1 508 612 7292

Email: [Hong\\_Yu@uml.edu](mailto:Hong_Yu@uml.edu)

## Abstract

**Background:** Accurate detection of bleeding events from electronic health records (EHRs) is crucial for identifying and characterizing different common and serious medical problems. To extract such information from EHRs, it is essential to identify the relations between bleeding events and related clinical entities (eg, bleeding anatomic sites and lab tests). With the advent of natural language processing (NLP) and deep learning (DL)-based techniques, many studies have focused on their applicability for various clinical applications. However, no prior work has utilized DL to extract relations between bleeding events and relevant entities.

**Objective:** In this study, we aimed to evaluate multiple DL systems on a novel EHR data set for bleeding event-related relation classification.

**Methods:** We first expert annotated a new data set of 1046 deidentified EHR notes for bleeding events and their attributes. On this data set, we evaluated three state-of-the-art DL architectures for the bleeding event relation classification task, namely, convolutional neural network (CNN), attention-guided graph convolutional network (AGGCN), and Bidirectional Encoder Representations from Transformers (BERT). We used three BERT-based models, namely, BERT pretrained on biomedical data (BioBERT), BioBERT pretrained on clinical text (Bio+Clinical BERT), and BioBERT pretrained on EHR notes (EhrBERT).

**Results:** Our experiments showed that the BERT-based models significantly outperformed the CNN and AGGCN models. Specifically, BioBERT achieved a macro F1 score of 0.842, outperforming both the AGGCN (macro F1 score, 0.828) and CNN models (macro F1 score, 0.763) by 1.4% ( $P < .001$ ) and 7.9% ( $P < .001$ ), respectively.

**Conclusions:** In this comprehensive study, we explored and compared different DL systems to classify relations between bleeding events and other medical concepts. On our corpus, BERT-based models outperformed other DL models for identifying the relations of bleeding-related entities. In addition to pretrained contextualized word representation, BERT-based models benefited from the use of target entity representation over traditional sequence representation

(*JMIR Med Inform* 2021;9(7):e27527) doi:[10.2196/27527](https://doi.org/10.2196/27527)

**KEYWORDS**

bleeding; relation classification; electronic health records; CNN; GCN; BERT

## Introduction

### Background

Bleeding refers to the escape of blood from the circulatory system either internally or externally. Bleeding events are common and frequently have a major impact on patient quality of life and survival. Bleeding events are common adverse drug events, particularly among patients with cardiovascular conditions who are prescribed anticoagulant medications [1].

We are seeing a marked increase in the use of anticoagulants, driven predominantly by the increased prevalence of atrial fibrillation (AF), a prothrombotic condition for which anticoagulants are frequently indicated. In the United States, the number of AF patients is increasing rapidly, mostly in the elderly population, with a projection of 12 million by 2050 [2,3]. The chance of having a stroke from AF can be as high as 10% within 5 years of AF diagnosis [4]. Clinicians must weigh stroke risk against the risk of bleeding from anticoagulants [5,6]. Most published data on the risks of anticoagulants come from clinical trials, where major bleeding outcomes are rigorously adjudicated by trained abstractors. However, there are limitations to this approach, as there are many important groups that are underrepresented in clinical trials. Real-world data are lacking, in part owing to the significant time and cost associated with manual chart review, which is the current gold standard for bleeding classification. With a lack of available risk calculators for a situation like this, it is challenging to advise anticoagulants to older AF patients as they are at high risk for both stroke and anticoagulant complications, for example, bleeding [7-9]. Clinicians and researchers would benefit from new ways to classify the relations between bleeding events and related medical entities to provide more accurate risk and benefit assessments of commonly used medications, particularly anticoagulants.

Clinical notes, such as electronic health records (EHRs), contain rich information for various studies including but not limited to epidemiological research, pharmacovigilance, and drug safety surveillance [10,11]. However, bleeding and its attributes are mostly documented in the unstructured EHR narratives instead of the structured fields [10]. With the availability and success of different deep learning (DL) techniques, building accurate and effective DL-based natural language processing (NLP) systems can alleviate this problem and prove viable against more expensive and time-consuming manual annotations. Therefore, in this work, we evaluated different DL models for relation classification between bleeding events and related medical concepts. Relation classification is the task of classifying relations for a pair of target entities from a text span. For example, given the text span “clotted blood was found in the entire colon,” the task is to detect the relation between the bleeding event “clotted blood” and anatomic site “colon.”

A majority of previous studies on clinical text have primarily focused on the relations between medications and other factors such as adverse drug effects (ADEs) [12-15]. However, to our knowledge, there has been no prior work that aims at identifying bleeding event-related relations from EHRs using DL-based

NLP systems. The advantages of such systems make them the right group of candidates to investigate for this task.

### Relevant Literature

Realizing the importance of relation classification tasks for clinical narratives, different research groups released several publicly available data sets and launched shared tasks with a focus on relation classification in the clinical domain [15-19]. These include detecting relation types among medical problems, tests, and treatments [16], as well as relations between medications and their various attributes, such as dosage and ADEs [15,17-19]. Our task can be closely compared to any of these tasks.

In general, the relation classification problem can be solved by different systems or models, including rule-based systems, non-DL-based machine learning models, and DL models, depending on the domain and context. For example, Kang et al [20] used the Unified Medical Language System (UMLS) [21] to build a knowledge base where relations between medications and ADEs can be detected based on the shortest path between them. Xu et al [22] applied support vector machines (SVMs) to determine the relation between drugs and diseases, while Henriksson et al [11] used random forest.

Studies have compared non-DL-based machine learning models with DL models for relation classification, and the results are mixed. Munkhdalai et al [12] used a recurrent neural network (RNN) on clinical notes for relation identification and found that an SVM with a rich feature set outperformed the RNN on their data set. In contrast, Luo et al [23] showed that a convolutional neural network (CNN) with pretrained medical word embeddings is superior to traditional machine learning methods. A similar observation was made by He et al [24] for their CNN model with a multipooling operation.

Beyond traditional RNN and CNN models, Li and Yu [13] evaluated a capsule network and multilayer perceptron (MLP) for single domain and multidomain relation classification tasks on EHR data sets and found that although there was a slight improvement, the capsule network model was not superior to the MLP model. Christopoulou et al [14] developed intrasentence models based on bidirectional long short-term memory (bi-LSTM) and attention mechanism. The authors also employed a transformer network [25] for building an intersentence model. For clinical conversations, Du et al [26] proposed a relation span attribute tagging (R-SAT) model that utilizes bi-LSTM and has been shown to outperform the baseline by a large margin for two relation classification tasks.

Recent DL architectures, such as Bidirectional Encoder Representations from Transformers (BERT) [27] and graph convolutional network (GCN), have shown promising results for relation classification across different domains. Wu and He [28] used BERT with entity information for relation classification on the SemEval-2010 Task 8 data set [29] and obtained better results than other state-of-the-art methods. Soares et al [30] introduced a new training scheme for BERT, matching the blank (MTB), which gave superior performance on three different data sets. Lin et al [31] used BERT to solve the sentence-agnostic temporal relation extraction problem for

clinical text. Guo et al [32] proposed a novel GCN model with attention and densely connected layers, named the attention-guided graph convolutional network (AGGCN), which utilizes the full dependency tree information of the input sequences. In their experiments, the AGGCN achieved significant performance gain over the other GCN-based systems on multiple relation classification data sets. A GCN has also been employed on different biomedical tasks successfully, including biomedical event extraction [33] and measurement of semantic relatedness between UMLS concepts [34], among others.

Among different DL models, CNN, BERT, and AGGCN are currently the most representative architectures. However, despite being state-of-the-art models, few studies have evaluated the three models parallelly for clinical relation classification, which is the focus of this study.

## Objective

In this study, we focused on the evaluation of three different state-of-the-art DL systems for the relation classification task on a new curated EHR data set. These systems included a CNN, a GCN with attention (AGGCN), and models based on BERT. In particular, a GCN has not yet been explored in any clinical setting for relation classification. The contributions of this work can be summarized as follows: (1) this is the first study to identify the relations between bleeding events and other relevant medical concepts; (2) we provide comparative analyses of three different DL architectures for the relation classification task on a new EHR data set; and (3) we explored the effects of additional domain knowledge on the AGGCN model, as well as how entity position representations influence BERT models' predictions.

## Methods

### Data Set

With approval from the Institutional Review Board at the University of Massachusetts Medical School and a memorandum

**Table 1.** Data statistics.

Relation type	Occurrences	Relation length, mean (SD)
Event-Site	3495	4.81 (10.20)
Event-Lab	3314	93.69 (137.99)
Event-AltCause <sup>a</sup>	4947	48.08 (94.02)
Lab-Severity	3470	3.26 (4.82)

<sup>a</sup>AltCause: suspected alternative cause.

We used the NLTK package [35] to tokenize EHR text. For all experiments, we maintained a train, validation, and test split of 60:20:20 on the note level. We also generated negative relation instances by taking permutations of all possible entity pairs that did not have any relationship between them. For all three splits, this resulted in a set of negative relations that was two to three times the other relations combined. For the training and development sets, we down-sampled the negative relations such that their frequency was similar to the other four relation types

of understanding between the University of Massachusetts Medical School and Northwestern University, we annotated 1046 deidentified discharge summaries from patients with cardiovascular diseases who received anticoagulants during their stays at hospitals affiliated with Northwestern University. The notes were annotated by five medical experts under the supervision of two senior physicians. From the comprehensive list of 13 entity types, we chose five relevant to bleeding and the relations among them. This resulted in four relation types for our relation classification study as follows: (1) bleeding event-bleeding anatomic site (Event-Site), (2) bleeding event-bleeding lab evaluation (Event-Lab), (3) bleeding event-suspected alternative cause (Event-AltCause), and (4) bleeding lab evaluation-severity (Lab-Severity).

A *bleeding event* indicates the escape of blood from the circulatory system. Examples of bleeding events from our cohort include mentions such as “hemorrhage,” “black tarry stools,” and “clotted blood.” *Bleeding anatomic site* is the corresponding anatomic site for a bleeding event, for example, “esophagus” in the phrase “blood oozing in esophagus.” *Bleeding lab evaluation* is any relevant laboratory test, and *severity* is the test value when in an abnormal range. *Suspected alternative cause* indicates possible alternative causes for bleeding other than anticoagulants.

Our cohort of 1046 notes included 15,363 relation instances. There was a large variation in token length, ranging from 3 to 985. For our task, we chose a subset that had instances with token length no more than 1000. Since most DL models do not handle long input sequences and 99.11% (15,226) of the 15,363 relation instances had a token length less than 1000, we used these 15,226 instances to build the final data set. This included both intrasentence and intersentence relations. All the relation types and their frequencies for this cohort are provided in Table 1. We also list relation lengths for each relation type, which is the number of tokens between the two target entities. It can be noticed that out of the four relation types on average, Event-Lab and Event-AltCause had significantly longer relation lengths with wider spreads.

combined. We did not perform down-sampling for the test set, so it would be representative of the real EHR note distribution.

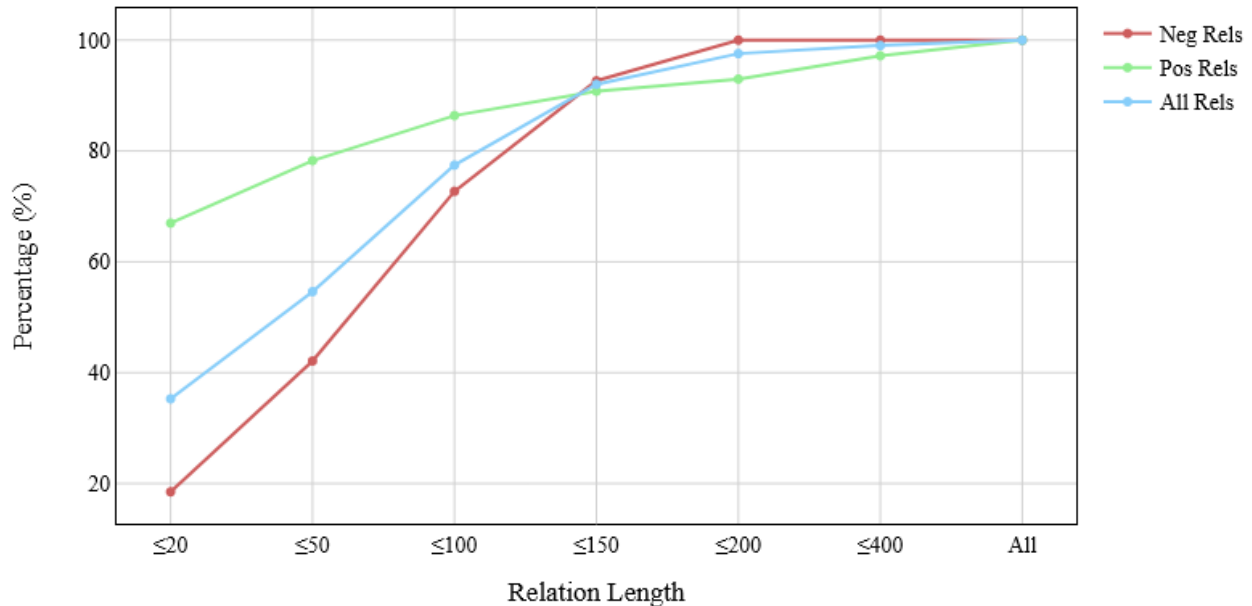
Figure 1 shows the relation distribution in our data set for different relation lengths. The x-axis indicates the range (eg,  $\leq 20$  indicates all instances that have a relation length of 20 or lower), and the y-axis indicates the percentage of instances at that range. Positive relations are all relation instances that belong to the four relation types described above. Here, we can see a steep increase for the negative relations compared to the positive



relations. This shows that, on average, negative relations had longer relation lengths. For example, as we increased the relation length upper bound from 50 to 100, there was almost a 30% increase in negative relations, whereas for positive relations, it

was less than 10%. In particular, negative relations had a mean relation length of 74.01 (SD 49.40). We discuss the implications of relation length in the Results section.

**Figure 1.** Relation distribution for different relation lengths. Neg: negative; Pos: positive; Rels: relations.



## Models

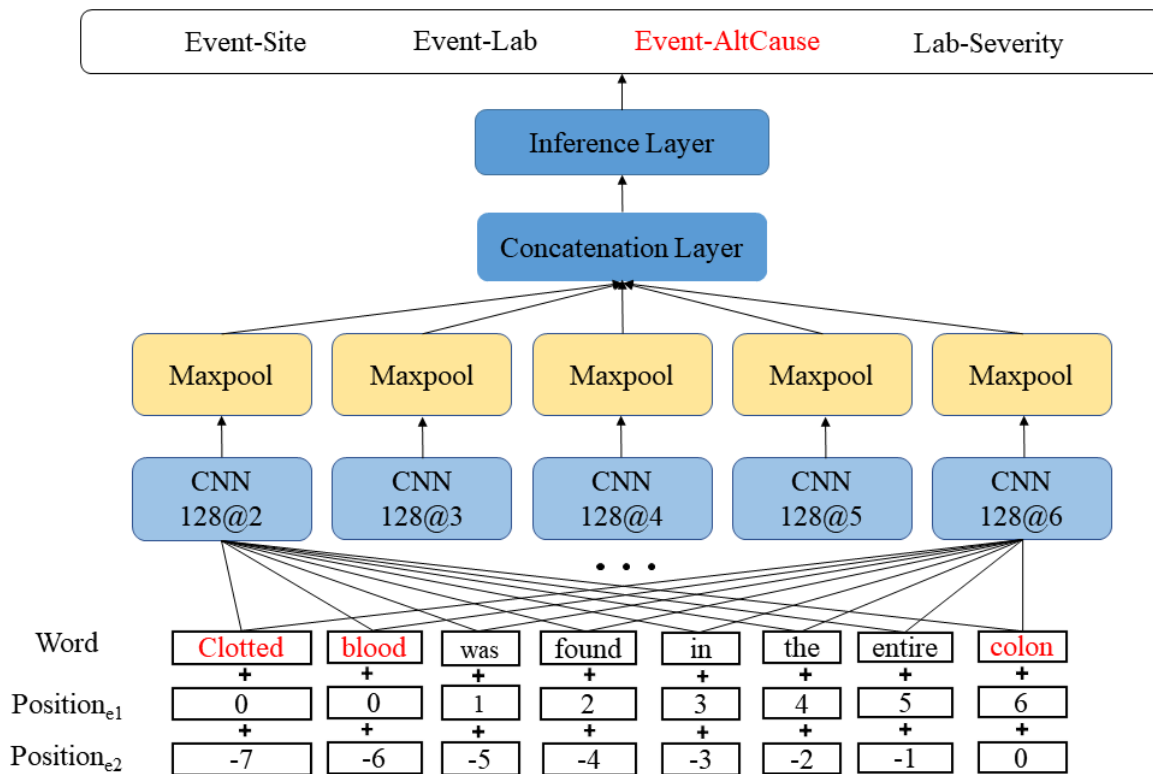
For this work, we evaluated three different state-of-the-art DL architectures (CNN, GCN, and BERT), which we describe briefly below.

### CNN

CNN is a class of deep feed-forward neural networks that is specialized for data with a high degree of temporal or spatial correlation such as image data. CNNs have also been widely used for various NLP tasks with success, including relation classification [36-39]. Our CNN relation classification model was built upon the work of Nguyen and Grishman [37], which is a state-of-the-art CNN architecture for relation classification

in the open domain. As shown in Figure 2, the model utilizes five separate convolutional layers with filters of different window sizes to capture rich local n-gram features. For example, “128@2” in the first CNN block indicates 128 filters with a window size of 2. Each layer is followed by a tanh nonlinearity. Finally, we used a maxpool layer, concatenated the output, applied dropout, and added a fully connected layer, followed by a softmax layer for the final classification. As input, we used pretrained word embeddings concatenated with randomly initialized positional embeddings. We used positional embeddings to embed the relative positions of the target entities and other words in a relation instance, as it has been shown to improve various NLP tasks including relation classification [24,40].

**Figure 2.** The high-level view of our convolutional neural network (CNN) model. It has five different CNN modules with filters of different window sizes, followed by maxpooling and concatenation. The inference layer includes a dropout, a fully connected layer, and a softmax layer. Positione1 and Positione2 refer to the relative positions of each word from entity1 and entity2, respectively. AltCause: suspected alternative cause.



**GCN**

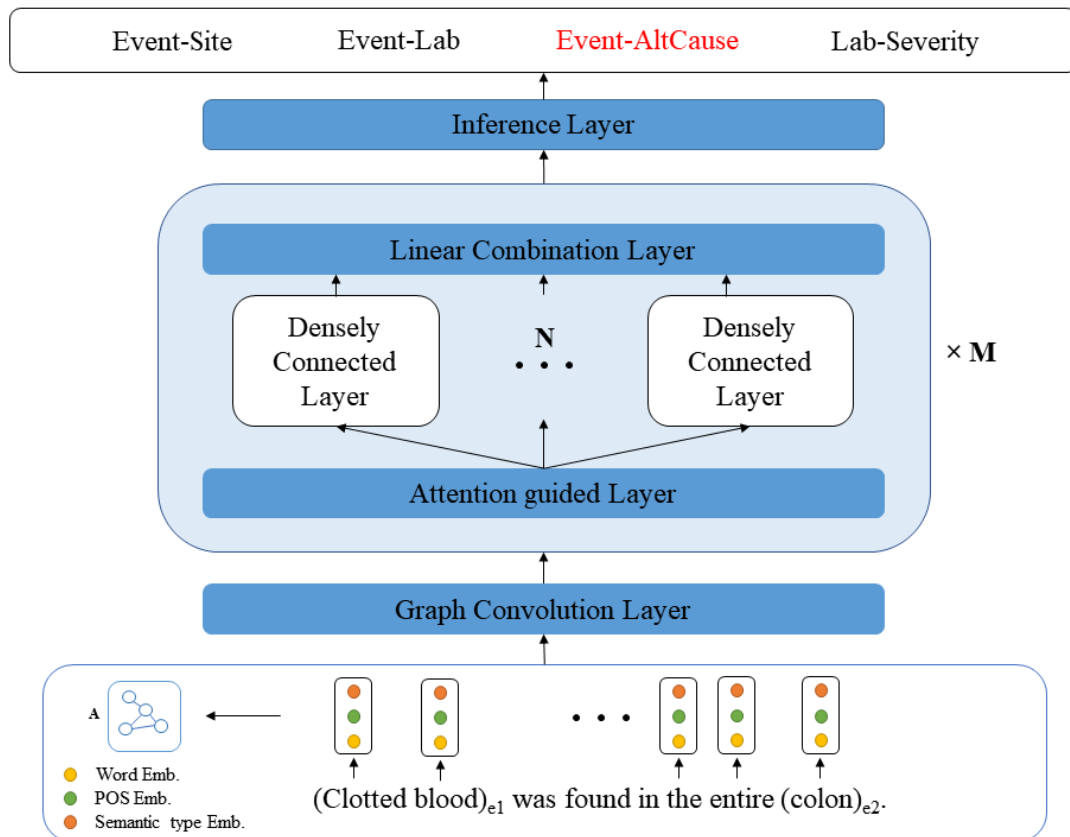
Since semantic coding has enjoyed success in clinical NLP [41], GCNs [42] may be effective and powerful as they represent the semantic or syntactic dependency of input sequences as graphs, which have shown superior performance for the relation classification task in the open domain [32,43]. We implemented the AGGCN [32], which incorporates dense connections for rich dependency information and multihead attention [25] for soft pruning the trees (Figure 3). Here, each sentence corresponds to a graph, represented in the form of an adjacency matrix A, where  $A_{ij}=1$  if node  $i$  and node  $j$  have an edge between them and  $A_{ij}=0$  otherwise. Additional model details are available in Multimedia Appendix 1.

Unlike the previous work [32], we built semantic graphs instead of syntactic graphs. This was motivated by decades of NLP work in the clinical domain that highlights the advantages of semantic parsers [41,44]. To construct the graph, we used the

UMLS Metathesaurus [21]. First, we mapped an input sentence to the UMLS concepts using MetaMap [44]. We considered all words in an input sequence as the nodes in a graph, each with a self-loop. Then, for every two nodes, we connected them if they had a semantic relation (eg, child-of) and were identified as at least one of the 26 preselected semantic types. These semantic types were chosen to prioritize bleeding events and relevant entities (Multimedia Appendix 2). However, owing to data sparsity, this resulted in disconnected graphs where most of the nodes had no incoming or outgoing edge. As an alternative, we relaxed the criteria by connecting nodes to each other (belonged to any of the 26 semantic types). In a separate experiment, we repeated the same process with all 127 semantic types from the UMLS Metathesaurus.

In addition, we investigated two different methods, namely, initializing A from a uniform distribution and initializing A with all 1s (all nodes are connected to each other). Finally, we explored semantic-type embeddings (STEs). A comparison of these methods is available in the Results section.

**Figure 3.** The high-level view of our attention-guided graph convolutional network (AGGCN) model. A is the adjacency matrix used to represent the graph data. The core of the model is comprised of M identical blocks (AGGCN blocks), each with three types of layers as follows: one attention-guided layer, N densely connected layers, and one linear combination layer. Details are available in [Multimedia Appendix 1](#). AltCause: suspected alternative cause; Emb: embedding; POS: parts of speech.

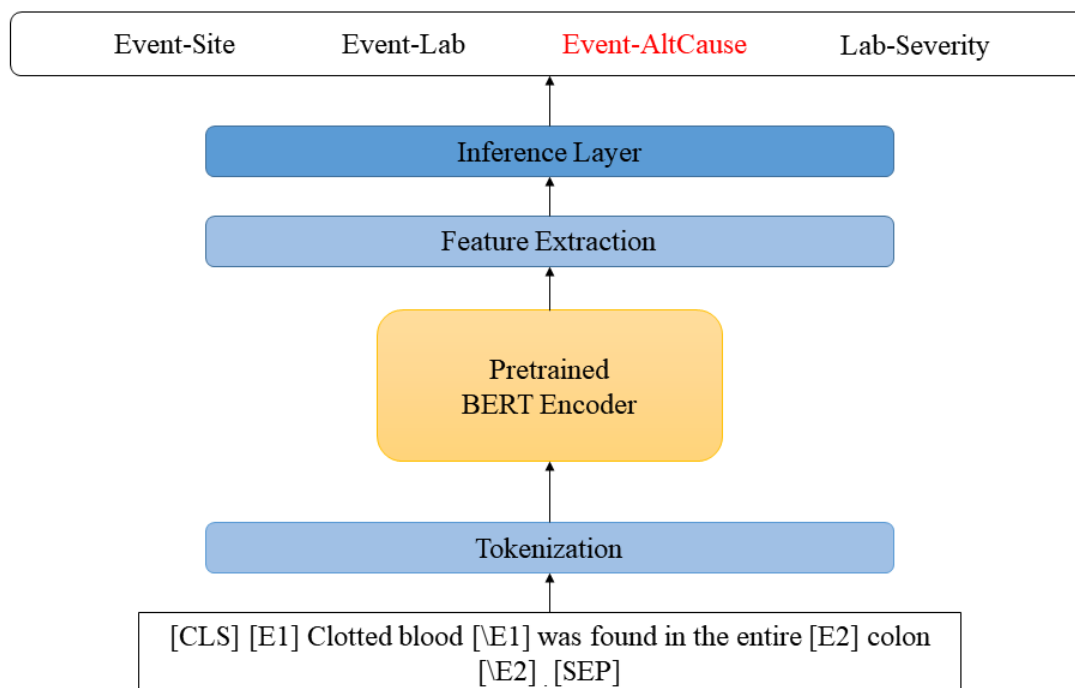


### BERT

BERT [27] is a language representation model that was pretrained on a large text corpus using unsupervised objectives. BERT has been shown to outperform most of the DL models in various NLP tasks, including clinical applications [45]. At its core, BERT employs bidirectional transformers [25] with multihead attention mechanisms. Paired with an effective pretraining scheme for unsupervised tasks, namely, masked language modeling and next sentence prediction, BERT can provide a rich contextual representation for any text sequence. BERT's contextualized word representations can be fine-tuned for any downstream NLP task. In this work, we used three variants of BERT (BERT pretrained on biomedical data [BioBERT] [46], BioBERT pretrained on clinical text [Bio+Clinical BERT] [47], and BioBERT pretrained on EHR notes [EhrBERT] [45]), all of which have been shown to improve clinical NLP applications. They all share the same architecture with a difference in their pretraining corpora.

In our implementation (Figure 4), for a target entity pair, we used four reserved tokens ([E1], [E2], [\E1], and [\E2]) to mark the start and end of the entities. For our task, to handle an input sequence larger than 512 word pieces, we modified the BERT encoder so that it could slide over any input sequence with a stride, essentially splitting the sequence into multiple 512 word piece-long subsequences. It later merges the fine-tuned hidden representations of the subsequences depending on the maximum context window. A maxpool operation is performed over the subsequences' [CLS] tokens to create the final [CLS] representation. Later the feature extraction module constructs features from the final hidden representations. It can be from either the [CLS] token or a fusion of entity start or end tokens. In particular, we experimented with approaches, such as the maxpool of entity-start tokens ([E1] and [E2]), concatenation of entity-start tokens, and max-pool of entity-end tokens ([\E1] and [\E2]). Details about these are provided in the Results section (Experiments With BioBERT subsection). Finally, we added a fully connected layer on top for the relation classification.

**Figure 4.** The high-level view of a Bidirectional Encoder Representations from Transformers (BERT)-based model. AltCause: suspected alternative cause.



### Evaluation Metrics

All the models were evaluated using precision, recall, and F1 score. We report both micro- and macro-averaged scores. Averaged over all the instances, micro-averaged scores give an overall evaluation and therefore are biased toward the class with the highest instances. On the contrary, macro-averaged scores help obtain a better understanding of the models' performance across different classes as it is averaged over all the classes.

### Experimental Setup

All model hyperparameters were fine-tuned on the development set. For the CNN model, we included five convolutional layers, each with 128 filters and different window sizes (2, 3, 4, 5, and 6). We chose Adam as the optimizer with a learning rate of 0.01, and the dropout rate was 0.5. The model was trained for 300 epochs. We found 300 and 10 to work the best as the dimensions for word and position embeddings, respectively. For the AGGCN model, we used part-of-speech (POS) embeddings in addition to pretrained word embeddings. Here, the dimensions were 30 and 300, respectively. We ran the AGGCN model for 100 epochs with a learning rate of 0.5 and stochastic gradient descent optimizer. Other hyperparameters included three heads for the attention layer, three AGGCN blocks, two and five sublayers in the first and second dense layers, etc. For both the CNN and AGGCN models, we used global vectors for word representation (GLOVE) [48] as pretrained word embeddings.

We used the popular library Transformers [49] for implementing our BERT models. As mentioned in the Models subsection, we

modified the existing implementation so that it could cover sequences of all lengths. We used a stride of 128 with a maximum sequence length of 512. The learning rate was  $5 \times 10^{-5}$  and the dropout rate was 0.1. We initialized each BERT model's encoder with corresponding pretrained weights. All models were fine-tuned for 15 epochs.

Cross-entropy loss was used for training all the models. In each experiment, we used an early stopping criterion based on the model's performance on the development set. All models were evaluated on the same hold out test set, and the reported results were averaged over three independent runs. All model trainings and evaluations were performed on Tesla V100 GPUs (Nvidia).

## Results

### Comparison of the Models

We report our results for the relation classification task in Table 2. All BERT-based models did comparatively better than the CNN and AGGCN models. The BioBERT model achieved a 1.3% absolute improvement ( $P < .001$ ) over the AGGCN model in both micro and macro F1 scores, while the difference with the CNN model was even more significant at almost 8% ( $P < .001$ ). A similar performance improvement was observed for the Bio+Clinical BERT model but with a lower recall. The CNN model performed the worst for all relation types. For each model, we also report the macro scores of two ensemble methods (last two rows) where both improved the model performance.  $P$  values were calculated following the work by Berg-Kirkpatrick et al [50]



**Table 2.** Performance comparison of convolutional neural network (CNN), attention-guided graph convolutional network (AGGCN), and Bidirectional Encoder Representations from Transformers–based models (BERT).

Relation type and performance	Model				
	CNN <sup>a</sup>	AGGCN <sup>b</sup>	BioBERT <sup>c</sup>	Bio+Clinical BERT <sup>d</sup>	EhrBERT <sup>e</sup>
<b>Event-Site</b>					
Precision, mean (SD)	0.910 (0.003)	0.941 (0.009)	0.916 (0.058)	0.929 (0.020)	0.942 (0.024)
Recall, mean (SD)	0.817 (0.003)	0.947 (0.006)	0.942 (0.009)	0.930 (0.016)	0.920 (0.024)
F1 score, mean (SD)	0.861 (0.003)	0.944 (0.002)	0.928 (0.027)	0.929 (0.003)	0.977 (0.003)
<b>Event-Lab</b>					
Precision, mean (SD)	0.653 (0.014)	0.619 (0.014)	0.616 (0.029)	0.618 (0.023)	0.587 (0.031)
Recall, mean (SD)	0.629 (0.011)	0.737 (0.022)	0.793 (0.027)	0.785 (0.010)	0.802 (0.012)
F1 score, mean (SD)	0.641 (0.003)	0.672 (0.002)	0.692 (0.009)	0.691 (0.011)	0.677 (0.022)
<b>Event-AltCause<sup>f</sup></b>					
Precision, mean (SD)	0.640 (0.006)	0.718 (0.017)	0.708 (0.048)	0.718 (0.026)	0.721 (0.014)
Recall, mean (SD)	0.596 (0.012)	0.723 (0.030)	0.828 (0.029)	0.792 (0.015)	0.803 (0.009)
F1 score, mean (SD)	0.617 (0.004)	0.720 (0.007)	0.761 (0.017)	0.753 (0.008)	0.760 (0.006)
<b>Lab-Severity</b>					
Precision, mean (SD)	0.907 (0.004)	0.967 (0.003)	0.977 (0.005)	0.974 (0.007)	0.963 (0.011)
Recall, mean (SD)	0.963 (0.001)	0.986 (0.004)	0.993 (0.001)	0.991 (0.001)	0.991 (0.004)
F1 score, mean (SD)	0.934 (0.002)	0.976 (0.002)	0.985 (0.003)	0.982 (0.004)	0.977 (0.003)
<b>Micro</b>					
Precision, mean (SD)	0.768 (0.006)	0.800 (0.014)	0.786 (0.038)	0.793 (0.020)	0.783 (0.013)
Recall, mean (SD)	0.739 (0.006)	0.838 (0.015)	0.885 (0.017)	0.868 (0.009)	0.873 (0.005)
F1 score, mean (SD)	0.753 (0.002)	0.818 (0.001)	0.832 (0.015)	0.829 (0.007)	0.826 (0.009)
<b>Macro</b>					
Precision, mean (SD)	0.777 (0.005)	0.811 (0.010)	0.804 (0.032)	0.810 (0.017)	0.803 (0.007)
Recall, mean (SD)	0.751 (0.005)	0.848 (0.014)	0.889 (0.016)	0.874 (0.009)	0.879 (0.006)
F1 score, mean (SD)	0.763 (0.003)	0.828 (0.001)	0.842 (0.012)	0.839 (0.005)	0.836 (0.007)
<b>Macro (majority voting)</b>					
Precision	0.778	0.813	0.822	0.824	0.823
Recall	0.752	0.849	0.895	0.882	0.887
F1 score	0.764	0.829	0.855	0.851	0.851
<b>Macro (averaging predictions)</b>					
Precision	0.779	0.813	0.824	0.826	0.828
Recall	0.753	0.855	0.879	0.879	0.886
F1 score	0.765	0.833	0.850	0.850	0.854

<sup>a</sup>CNN: convolutional neural network.

<sup>b</sup>AGGCN: attention-guided graph convolutional network.

<sup>c</sup>BioBERT: BERT pretrained on biomedical data.

<sup>d</sup>Bio+Clinical BERT: BioBERT pretrained on clinical text.

<sup>e</sup>EhrBERT: BioBERT pretrained on electronic health record notes.

<sup>f</sup>AltCause: suspected alternative cause.

## Domain Knowledge for the AGGCN

For the AGGCN, we first experimented with different approaches to encode information from graph inputs. The AGGCN uses an  $n \times n$  adjacency matrix  $A$  to represent a graph with  $n$  nodes. For our inputs, we built the graph based on MetaMap [44], as explained in the Models subsection. To understand the importance of domain-specific knowledge (UMLS), we also removed the UMLS knowledge by connecting all the nodes (tokens) of a graph (input sequence) to each other (all connected). This is equivalent to setting all the elements in

$A$  to 1. In addition, we also explored a weighted graph (Uniform). For this, we built  $A$  using a uniform distribution with the half-open interval  $[0,1)$ .

As shown in Table 3, predefining the graph using the domain knowledge did not improve the overall performance. Several factors may have contributed to this result, including the noise introduced by MetaMap for mapping text to the UMLS concepts and the incompleteness of concept relations in the UMLS. Our results showed that the weighted graph (Uniform) achieved the best performance.

**Table 3.** AGGCN (Attention-guided graph convolutional network) performance with different methods.

Metric and performance	Method <sup>a</sup>				
	MetaMap (26) <sup>b</sup>	MetaMap (All) <sup>c</sup>	All Connected	Uniform	Uniform + STE <sup>d</sup>
<b>Micro</b>					
Precision, mean (SD)	0.774 (0.008)	0.757 (0.026)	0.783 (0.025)	0.800 (0.014)	0.796 (0.011)
Recall, mean (SD)	0.829 (0.007)	0.852 (0.019)	0.845 (0.018)	0.838 (0.015)	0.836 (0.007)
F1 score, mean (SD)	0.800 (0.003)	0.801 (0.006)	0.812 (0.005)	0.818 (0.001)	0.816 (0.007)
<b>Macro</b>					
Precision, mean (SD)	0.787 (0.008)	0.781 (0.018)	0.798 (0.019)	0.811 (0.010)	0.805 (0.011)
Recall, mean (SD)	0.844 (0.007)	0.865 (0.018)	0.855 (0.017)	0.848 (0.014)	0.848 (0.008)
F1 score, mean (SD)	0.813 (0.003)	0.816 (0.003)	0.824 (0.003)	0.828 (0.001)	0.825 (0.008)

<sup>a</sup>All methods used global vectors for word representation (GLOVE) and part-of-speech (POS) embeddings.

<sup>b</sup>MetaMap (26) used 26 specific semantic types.

<sup>c</sup>MetaMap (All) used all 127 semantic types from the Unified Medical Language System Metathesaurus.

<sup>d</sup>STE: semantic-type embedding.

We also evaluated the effects of STEs. The UMLS had a total of 127 semantic types, from which we identified 26 semantic types relevant to our work (Uniform + STE). For a word with multiple semantic types, we used the semantic type with the highest MetaMap Indexing (MMI) score. Our results with STEs, however, did not improve the performance. We also evaluated POS embeddings and entity-type embeddings. Results from our experiments suggested that only POS embeddings improved performance, while entity-type embeddings slightly degraded performance. Other experiments included the use of different pretrained word embeddings. Surprisingly, we found that the biomedical word embeddings [51] did not perform well compared with the GLOVE embeddings on our data set. In summary, the best combination for AGGCN includes adjacency matrix initialization from uniform distribution and the use of GLOVE and POS embeddings.

## Experiments With BERT

For classification, there are various ways to extract the contextualized sequence representations from BERT. The most

common approach is to use [CLS] token embedding. In this work, since entity positions were already encoded in the input sequence, we explored different alternatives [30]. For example, we considered fusing the entity start tokens' embeddings ([E1] and [E2]) and the entity end tokens' embeddings ([\E1] and [\E2]). The fusion function was either maxpooling or concatenation. To our knowledge, this is the first study to evaluate different approaches for extracting BERT representation for clinical relation classification.

We used BioBERT as a representative of the BERT-based models, and the results are shown in Table 4. Although [CLS] token embedding is the most common approach, our results suggested that its performance is close to taking the concatenation of the entity start or end tokens' embeddings. In fact, the best performing method was the maxpool of the entity start tokens' embeddings, resulting in 1% improvement in the macro F1 score over [CLS]-only representation.

**Table 4.** Effect of different sequence representation methods on the BioBERT (BERT pretrained on biomedical data) model.

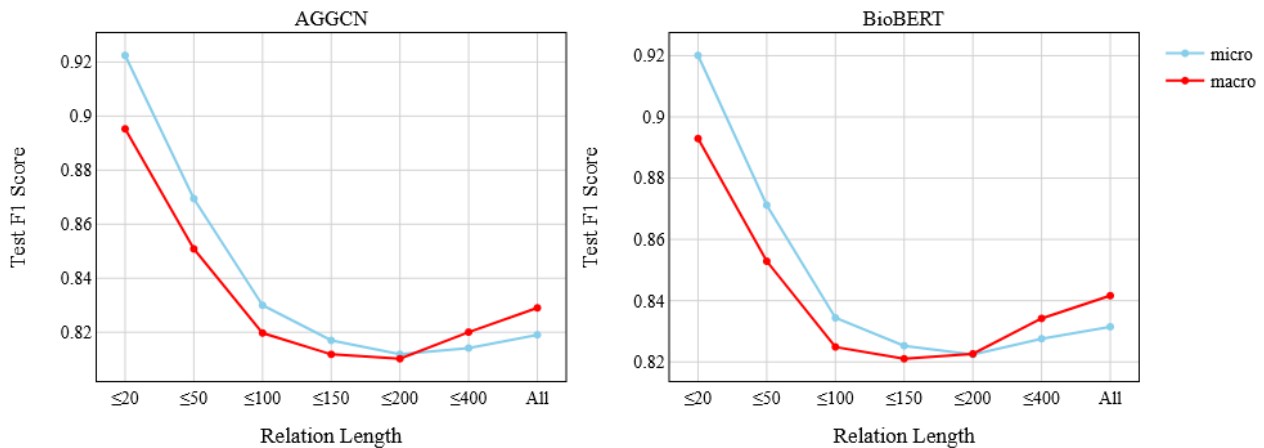
Method and performance	Micro	Macro
<b>[CLS] only</b>		
Precision, mean (SD)	0.779 (0.040)	0.803 (0.024)
Recall, mean (SD)	0.866 (0.015)	0.873 (0.011)
F1 score, mean (SD)	0.819 (0.015)	0.832 (0.010)
<b>Maxpool-start tokens</b>		
Precision, mean (SD)	0.786 (0.038)	0.804 (0.032)
Recall, mean (SD)	0.885 (0.017)	0.889 (0.016)
F1 score, mean (SD)	0.832 (0.015)	0.842 (0.012)
<b>Maxpool-end tokens</b>		
Precision, mean (SD)	0.780 (0.036)	0.800 (0.027)
Recall, mean (SD)	0.878 (0.015)	0.885 (0.014)
F1 score, mean (SD)	0.825 (0.014)	0.837 (0.010)
<b>Maxpool-start tokens + [CLS]</b>		
Precision, mean (SD)	0.775 (0.034)	0.794 (0.028)
Recall, mean (SD)	0.882 (0.014)	0.887 (0.014)
F1 score, mean (SD)	0.824 (0.014)	0.835 (0.012)
<b>Concatenate-start tokens</b>		
Precision, mean (SD)	0.762 (0.021)	0.787 (0.015)
Recall, mean (SD)	0.886 (0.011)	0.891 (0.008)
F1 score, mean (SD)	0.819 (0.008)	0.832 (0.006)
<b>Concatenate-end tokens</b>		
Precision, mean (SD)	0.768 (0.007)	0.793 (0.005)
Recall, mean (SD)	0.880 (0.009)	0.885 (0.009)
F1 score, mean (SD)	0.820 (0.005)	0.833 (0.005)
<b>Concatenate-start tokens + [CLS]</b>		
Precision, mean (SD)	0.743 (0.034)	0.777 (0.021)
Recall, mean (SD)	0.895 (0.008)	0.898 (0.006)
F1 score, mean (SD)	0.811 (0.017)	0.827 (0.013)

### Effect of Relation Length

As pointed out in [Table 1](#), the four relation types have a wide range of relation lengths. Relation length (ie, the number of words between the target entities) acts as context and hence can influence the training process. To demonstrate how it affected

our trained models, we created multiple subsets of our test set, each with a different range for relation length. Each subset contained only those test instances that had a relation length within the subset range. We chose the AGGCN and BioBERT models and ran inference on all the test subsets. The results are shown in [Figure 5](#).

**Figure 5.** Effect of relation length on model performance. The x-axis indicates the subset range, for example, " $\leq 20$ " indicates the test subset that consists of all the instances with a relation length of 20 or lower. AGGCN: attention-guided graph convolutional network; BERT: Bidirectional Encoder Representation from Transformers; BioBERT: BERT pretrained on biomedical data.



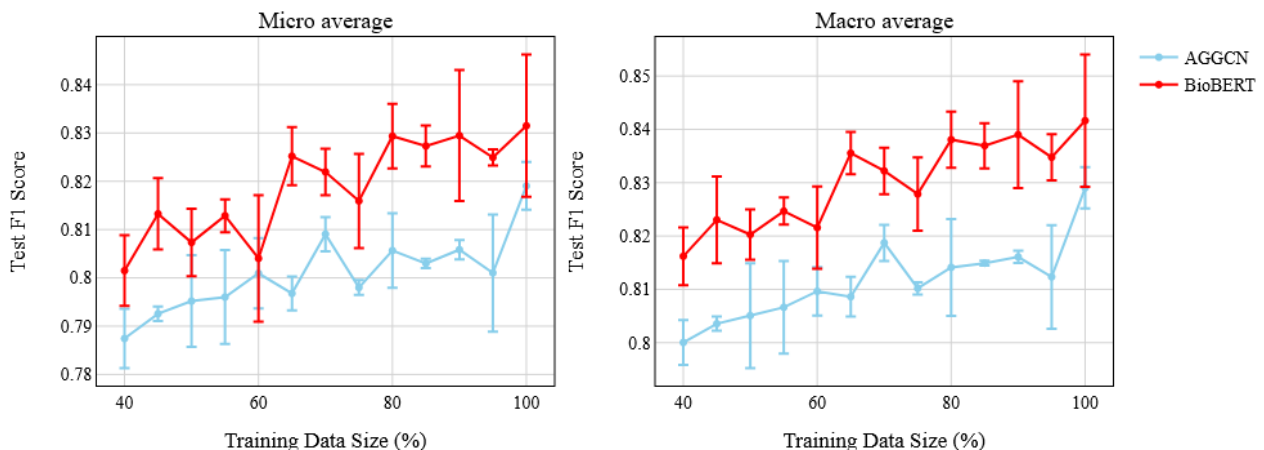
For both models, the test F1 scores kept decreasing until the relation length range reached 200, with an exception for the BioBERT macro score that had the lowest F1 score at 150. After this point, the macro F1 scores surpassed their respective micro scores, and surprisingly, both models' F1 scores improved despite the increase in relation length. This is slightly counterintuitive, as a larger relation length should have been difficult for the models to understand. To understand this behavior, we manually reviewed the gold labels and model predictions for all the test instances that had a relation length of 200 or higher. As expected, we found that all these instances were from either the relation types (Event-Lab and Event-AltCause) or a negative relation. Interestingly, all the model predictions were also within these three types. This shows that the models learned the correlation between relation length and relation type as a shortcut [52] and consequently did not consider Event-Site and Lab-Severity as possible relation types for longer relation lengths, resulting in improved overall performance. Our analyses showed the limitations of machine learning models in that they might learn from correlations, not causality, and this might lead to model overfitting.

However, a drawback of learning this shortcut is labeling many negative relations as Event-Lab or Event-AltCause, as negative relations have long relation lengths on average (refer to the Dataset subsection). For both models, this generated many false positives, resulting in low precision. This also explains the huge difference between precision and recall for these two relation types (Table 2).

**Model Performance With Data Size**

For any supervised DL method, the amount of available labeled data almost always plays a key role in the overall model performance. In our task, we wanted to evaluate how this affects the models, namely AGGCN and BioBERT. To this end, we trained both models with different portions of the training data separately and measured their performances. We observed an upward trend (Figure 6) for both, indicating that more training data would be better for our clinical relation classification task irrespective of the model type and metric averaging criterion. However, the AGGCN appeared to have less deviation (low standard deviation) with more data (a high slope), as opposed to BioBERT, for which the deviations were higher, although the performance differences remained statistically significant between the two models.

**Figure 6.** Effect of training data size on model performance. Each error bar indicates the standard deviation range at the corresponding point. AGGCN: attention-guided graph convolutional network; BERT: Bidirectional Encoder Representation from Transformers; BioBERT: BERT pretrained on biomedical data.





## Discussion

### Principal Findings

The results of our experiments demonstrated that fine-tuned BERT-based models outperformed both the CNN and AGGCN models by a significant margin. This can be attributed to the richer and contextualized representation of the pretrained BERT models compared to pretrained word embeddings, such as GLOVE, even when paired with POS embeddings and domain knowledge (for AGGCN). In our experiment, we found that the CNN significantly underperformed the AGGCN and BERT-based models by a large margin, primarily because of its inability to capture the global context of the input sequences. On the other hand, although all BERT-based models outperformed the AGGCN model by relatively small margins, they were statistically significant ( $P < .001$ ).

Despite model architectural differences, all models had better performance on the Event-Site and Lab-Severity relation types (eg, F1 scores of 0.928 and 0.985, respectively, for BioBERT). However, their performances for Event-Lab and Event-AltCause were relatively poor (eg, F1 scores of 0.692 and 0.761, respectively, for BioBERT). As shown in Table 1, these two relation types had comparatively larger relation lengths. This phenomenon would result in difficulty in annotation, thereby negatively impacting performance. Moreover, the lengthy context could pose challenges for the DL models as well. Both could have contributed to the overall poor performance for these two categories. In addition, except for the CNN model, we observed significant differences between precision and recall.

Our results showed that incorporating the concept relations from the UMLS did not improve AGGCN's performance. One possible reason might be the data sparsity, that is, only few concepts were connected in the graph input for the AGGCN. When a token is not identified by MetaMap as relevant but is important for classifying the instance, putting a 0 in its corresponding node position in the adjacency matrix  $A$  sends an erroneous signal to the model. This is a possible area for improvement, and we will work on this as part of our future work. On the other hand,  $A$  initialized with a uniform distribution gave the best recall and a better F1 score. This approach might seem counterintuitive as it does not necessarily pass any useful information unlike a dependency tree. However, this can be reasoned as the input dependency tree serves as an initialization, helping the attention-guided layers to build multiple edge-weighted graphs. This acts as a soft-pruning strategy where the model learns how the nodes should be connected to each other and on which connections to focus.

A quick look at the standard deviations reveals that Bio+Clinical BERT and EhrBERT were more stable than BioBERT, as both had utilized large scale EHR notes for the pretraining process. BioBERT had the highest F1 score, but different instantiations of the network gave widely different results, contributing to the higher standard deviation. The AGGCN was also better than BioBERT in this regard. Thus, we suggest using Bio+Clinical BERT or EhrBERT when stability is the primary concern. BioBERT on the other hand had the highest recall, which may be an important criterion for clinical applications. For the

AGGCN, the key advantage was the model being lightweight and consequently having a faster inference (Multimedia Appendix 3).

### Error Analysis

We conducted error analysis for the two relations (Event-Lab and Event-AltCause) where models performed poorly for both recall and precision scores. We analyzed the BioBERT model and made the following observations:

1. Most incorrect predictions were false positives, driven by the target entity types. For example, the model incorrectly predicted an Event-Lab relation in "Irrigation catheter was placed in ED and [hematuria]<sub>e1</sub> has improved. Repeat [H&H]<sub>e2</sub> is >8 and bleeding has stopped."
2. Another common source of error was the model incorrectly labeling a negative relation sequence that described a patient's medical history that was not directly related to the present diagnosis. For example, "Likely source thought to be upper GIB given hx of bleeding [ulcer]<sub>e1</sub> in past + [hematemesis]<sub>e2</sub>." Here, the model predicts the relation Event-AltCause between the target entities. Though the entity *GIB* can be a suspected alternative cause, both target entities are from the patient's previous history.
3. Another reason for error was the existence of the relation in the instance but between different entities. For example, take the negative relation instance "Daily CBC show anemia ([Hbg]<sub>e1</sub> 8.7 - 8.8, current at 8.7), with low Fe, transferrin+TIBC wnl, high ferritin. Labs support hemolytic anemia with low haptoglobin, high LDH, high tbili and indirect bili. Per inpatient attending read, blood smear showed no schistocytes, bite cells or heinz bodies, with few reticulocytes visualized per hpf, final report pending. CT kidney/pelvis showed no gross GU abnormalities and left gluteal [hematoma]<sub>e2</sub>." Here, the model predicted an Event-Lab relation though *Hbg* and *hematoma* do not have any such relation. However, there is an Event-Lab relation here between *Hbg* and *anemia*.
4. Limited corpus size and no additional domain knowledge made it difficult for the model to make predictions on relation instances with never-observed words or medical acronyms. In some cases, it was worsened due to the lack of grammatical consistency and coherent patterns.

### Conclusions

In this work, we studied three state-of-the-art DL architectures for a relation classification task on a novel EHR data set. Our work is the first to identify the relations between a bleeding event and related clinical concepts. Our results showed that BERT-based models performed better than attention-guided GCN and CNN models. Further experiments suggested that semantic graphs built using the UMLS semantic types and relations between them did not help the GCN model. On the other hand, incorporating entity token information improved the performance of BERT-based models. We also demonstrated the impacts of relation length and training data size. In our future work, we plan to explore richer domain knowledge and distant supervision. Additionally, leveraging our earlier work on named entity recognition (NER) [53], we aim to build a joint learning

pipeline that integrates both NER and relation classification for bleeding events and relevant medical concepts.

## Acknowledgments

This work was supported in part by grants HL125089 and R01HL137794 from the National Institutes of Health (NIH). HY is supported by R01MH125027, R01DA045816, and R01LM012817, all from the NIH. This work was also supported in part by the Center for Intelligent Information Retrieval (CIIR). We thank our annotators Raelene Goodwin, Edgard Granillo, Nadiya Frid, Heather Keating, and Brian Corner for annotating the discharge summaries. The contents of this paper do not represent the views of the CIIR or NIH.

## Conflicts of Interest

DDM received grants and personal fees from Bristol Myers Squibb, grants and personal fees from Pfizer, grants and personal fees from Heart Rhythm Society, grants and personal fees from Fitbit, personal fees from Samsung, grants from Boehringer Ingelheim, grants and personal fees from Flexcon, nonfinancial support from Apple, personal fees from Rose Consulting, and personal fees from Boston Biomedical during this study.

### Multimedia Appendix 1

Attention-guided graph convolutional network (AGGCN).

[DOCX File , 22 KB - [medinform\\_v9i7e27527\\_app1.docx](#) ]

### Multimedia Appendix 2

Relevant semantic types from the Unified Medical Language System.

[DOCX File , 19 KB - [medinform\\_v9i7e27527\\_app2.docx](#) ]

### Multimedia Appendix 3

Training time and parameter size.

[DOCX File , 19 KB - [medinform\\_v9i7e27527\\_app3.docx](#) ]

## References

1. Reynolds MR, Shah J, Essebag V, Olshansky B, Friedman PA, Hadjis T, et al. Patterns and predictors of warfarin use in patients with new-onset atrial fibrillation from the FRACTAL Registry. *Am J Cardiol* 2006 Feb 15;97(4):538-543. [doi: [10.1016/j.amjcard.2005.09.086](#)] [Medline: [16461052](#)]
2. Colilla S, Crow A, Petkun W, Singer DE, Simon T, Liu X. Estimates of current and future incidence and prevalence of atrial fibrillation in the U.S. adult population. *Am J Cardiol* 2013 Oct 15;112(8):1142-1147. [doi: [10.1016/j.amjcard.2013.05.063](#)] [Medline: [23831166](#)]
3. Miyasaka Y, Barnes ME, Bailey KR, Cha SS, Gersh BJ, Seward JB, et al. Mortality trends in patients diagnosed with first atrial fibrillation: a 21-year community-based study. *J Am Coll Cardiol* 2007 Mar 06;49(9):986-992 [FREE Full text] [doi: [10.1016/j.jacc.2006.10.062](#)] [Medline: [17336723](#)]
4. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. *Stroke* 1991 Aug;22(8):983-988. [doi: [10.1161/01.str.22.8.983](#)] [Medline: [1866765](#)]
5. Lip G, Lane D, Buller H, Apostolakis S. Development of a novel composite stroke and bleeding risk score in patients with atrial fibrillation: the AMADEUS Study. *Chest* 2013 Dec;144(6):1839-1847. [doi: [10.1378/chest.13-1635](#)] [Medline: [24009027](#)]
6. Di Biase L, Burkhardt JD, Santangeli P, Mohanty P, Sanchez JE, Horton R, et al. Periprocedural stroke and bleeding complications in patients undergoing catheter ablation of atrial fibrillation with different anticoagulation management: results from the Role of Coumadin in Preventing Thromboembolism in Atrial Fibrillation (AF) Patients Undergoing Catheter Ablation (COMPARE) randomized trial. *Circulation* 2014 Jun 24;129(25):2638-2644. [doi: [10.1161/CIRCULATIONAHA.113.006426](#)] [Medline: [24744272](#)]
7. Hobbs FDR, Roalfe AK, Lip GYH, Fletcher K, Fitzmaurice DA, Mant J. Performance of stroke risk scores in older people with atrial fibrillation not taking warfarin: comparative cohort study from BAFTA trial. *BMJ* 2011 Jun 23;342:d3653. [doi: [10.1136/bmj.d3653](#)] [Medline: [21700651](#)]
8. Hylek EM, Evans-Molina C, Shea C, Henault LE, Regan S. Major Hemorrhage and Tolerability of Warfarin in the First Year of Therapy Among Elderly Patients With Atrial Fibrillation. *Circulation* 2007 May 29;115(21):2689-2696. [doi: [10.1161/circulationaha.106.653048](#)]
9. Mant J, Hobbs FR, Fletcher K, Roalfe A, Fitzmaurice D, Lip GY, et al. Warfarin versus aspirin for stroke prevention in an elderly community population with atrial fibrillation (the Birmingham Atrial Fibrillation Treatment of the Aged Study, BAFTA): a randomised controlled trial. *The Lancet* 2007 Aug;370(9586):493-503. [doi: [10.1016/s0140-6736\(07\)61233-1](#)]

10. Turchin A, Shubina M, Breydo E, Pendergrass ML, Einbinder JS. Comparison of Information Content of Structured and Narrative Text Data Sources on the Example of Medication Intensification. *Journal of the American Medical Informatics Association* 2009 May 01;16(3):362-370. [doi: [10.1197/jamia.m2777](https://doi.org/10.1197/jamia.m2777)]
11. Henriksson A, Kvist M, Dalianis H, Duneld M. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *J Biomed Inform* 2015 Oct;57:333-349 [FREE Full text] [doi: [10.1016/j.jbi.2015.08.013](https://doi.org/10.1016/j.jbi.2015.08.013)] [Medline: [26291578](https://pubmed.ncbi.nlm.nih.gov/26291578/)]
12. Munkhdalai T, Liu F, Yu H. Clinical Relation Extraction Toward Drug Safety Surveillance Using Electronic Health Record Narratives: Classical Learning Versus Deep Learning. *JMIR Public Health Surveill* 2018 Apr 25;4(2):e29 [FREE Full text] [doi: [10.2196/publichealth.9361](https://doi.org/10.2196/publichealth.9361)] [Medline: [29695376](https://pubmed.ncbi.nlm.nih.gov/29695376/)]
13. Li F, Yu H. An investigation of single-domain and multidomain medication and adverse drug event relation extraction from electronic health record notes using advanced deep learning models. *J Am Med Inform Assoc* 2019 Jul 01;26(7):646-654 [FREE Full text] [doi: [10.1093/jamia/ocz018](https://doi.org/10.1093/jamia/ocz018)] [Medline: [30938761](https://pubmed.ncbi.nlm.nih.gov/30938761/)]
14. Christopoulou F, Tran T, Sahu S, Miwa M, Ananiadou S. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *J Am Med Inform Assoc* 2020 Jan 01;27(1):39-46 [FREE Full text] [doi: [10.1093/jamia/ocz101](https://doi.org/10.1093/jamia/ocz101)] [Medline: [31390003](https://pubmed.ncbi.nlm.nih.gov/31390003/)]
15. Jagannatha A, Liu F, Liu W, Yu H. Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0). *Drug Saf* 2019 Jan 16;42(1):99-111 [FREE Full text] [doi: [10.1007/s40264-018-0762-z](https://doi.org/10.1007/s40264-018-0762-z)] [Medline: [30649735](https://pubmed.ncbi.nlm.nih.gov/30649735/)]
16. Uzuner, South B, Shen S, DuVall S. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;18(5):552-556 [FREE Full text] [doi: [10.1136/amiajnl-2011-000203](https://doi.org/10.1136/amiajnl-2011-000203)] [Medline: [21685143](https://pubmed.ncbi.nlm.nih.gov/21685143/)]
17. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J Biomed Inform* 2012 Oct;45(5):885-892 [FREE Full text] [doi: [10.1016/j.jbi.2012.04.008](https://doi.org/10.1016/j.jbi.2012.04.008)] [Medline: [22554702](https://pubmed.ncbi.nlm.nih.gov/22554702/)]
18. Roberts K, Demner-Fushman D, Topping J. Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track. In: *Proceedings of the 10th Text Analysis Conference*. 2017 Presented at: 10th Text Analysis Conference; 2017; Gaithersburg, MD URL: [https://tac.nist.gov/publications/2017/additional.papers/TAC2017.ADR\\_overview.proceedings.pdf](https://tac.nist.gov/publications/2017/additional.papers/TAC2017.ADR_overview.proceedings.pdf)
19. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J Am Med Inform Assoc* 2020 Jan 01;27(1):3-12 [FREE Full text] [doi: [10.1093/jamia/ocz166](https://doi.org/10.1093/jamia/ocz166)] [Medline: [31584655](https://pubmed.ncbi.nlm.nih.gov/31584655/)]
20. Kang N, Singh B, Bui C, Afzal Z, van Mulligen EM, Kors JA. Knowledge-based extraction of adverse drug events from biomedical text. *BMC Bioinformatics* 2014 Mar 04;15:64 [FREE Full text] [doi: [10.1186/1471-2105-15-64](https://doi.org/10.1186/1471-2105-15-64)] [Medline: [24593054](https://pubmed.ncbi.nlm.nih.gov/24593054/)]
21. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: [10.1093/nar/gkh061](https://doi.org/10.1093/nar/gkh061)] [Medline: [14681409](https://pubmed.ncbi.nlm.nih.gov/14681409/)]
22. Xu J, Wu Y, Zhang Y, Wang J, Lee H, Xu H. CD-REST: a system for extracting chemical-induced disease relation in literature. *Database (Oxford)* 2016;2016:baw036 [FREE Full text] [doi: [10.1093/database/baw036](https://doi.org/10.1093/database/baw036)] [Medline: [27016700](https://pubmed.ncbi.nlm.nih.gov/27016700/)]
23. Luo Y, Cheng Y, Uzuner Ö, Szolovits P, Starren J. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. *J Am Med Inform Assoc* 2018 Jan 01;25(1):93-98 [FREE Full text] [doi: [10.1093/jamia/ocx090](https://doi.org/10.1093/jamia/ocx090)] [Medline: [29025149](https://pubmed.ncbi.nlm.nih.gov/29025149/)]
24. He B, Guan Y, Dai R. Classifying medical relations in clinical text via convolutional neural networks. *Artif Intell Med* 2019 Jan;93:43-49. [doi: [10.1016/j.artmed.2018.05.001](https://doi.org/10.1016/j.artmed.2018.05.001)] [Medline: [29778673](https://pubmed.ncbi.nlm.nih.gov/29778673/)]
25. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. 2017 Presented at: 31st Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA URL: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
26. Du N, Wang M, Tran L, Li G, Shafran I. Learning to infer entities, properties and their relations from clinical conversations. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019 Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; November 3-7, 2019; Hong Kong, China p. 4979-4990. [doi: [10.18653/v1/d19-1503](https://doi.org/10.18653/v1/d19-1503)]
27. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2-7, 2019; Minneapolis, MN p. 4171-4186. [doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423)]
28. Wu S, He Y. Enriching pre-trained language model with entity information for relation classification. In: *CIKM '19: Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2019 Presented at: 28th ACM International Conference on Information and Knowledge Management; November 3-7, 2019; Beijing, China p. 2361-2364. [doi: [10.1145/3357384.3358119](https://doi.org/10.1145/3357384.3358119)]

29. Hendrickx I, Kim S, Kozareva Z, Nakov P, Séaghdha D, Padó S, et al. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In: SEW '09: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. 2009 Presented at: Workshop on Semantic Evaluations: Recent Achievements and Future Directions; June 4, 2009; Boulder, CO p. 94-99. [doi: [10.3115/1621969.1621986](https://doi.org/10.3115/1621969.1621986)]
30. Soares L, FitzGerald N, Ling J, Kwiatkowski T. Matching the blanks: Distributional similarity for relation learning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 28-August 2, 2019; Florence, Italy p. 2895-2905. [doi: [10.18653/v1/p19-1279](https://doi.org/10.18653/v1/p19-1279)]
31. Lin C, Miller T, Dligach D, Bethard S, Savova G. A BERT-based Universal Model for Both Within- and Cross-sentence Clinical Temporal Relation Extraction. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June 7, 2019; Minneapolis, MN p. 65-71. [doi: [10.18653/v1/W19-1908](https://doi.org/10.18653/v1/W19-1908)]
32. Guo Z, Zhang Y, Lu W. Attention guided graph convolutional networks for relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 28-August 2, 2019; Florence, Italy p. 241-251. [doi: [10.18653/v1/p19-1024](https://doi.org/10.18653/v1/p19-1024)]
33. Zhao W, Zhang J, Yang J, He T, Ma H, Li Z. A novel joint biomedical event extraction framework via two-level modeling of documents. *Information Sciences* 2021 Mar;550:27-40. [doi: [10.1016/j.ins.2020.10.047](https://doi.org/10.1016/j.ins.2020.10.047)]
34. Mao Y, Fung K. Use of word and graph embedding to measure semantic relatedness between Unified Medical Language System concepts. *J Am Med Inform Assoc* 2020 Oct 01;27(10):1538-1546 [FREE Full text] [doi: [10.1093/jamia/ocaa136](https://doi.org/10.1093/jamia/ocaa136)] [Medline: [33029614](https://pubmed.ncbi.nlm.nih.gov/33029614/)]
35. Loper E, Bird S. NLTK: the Natural Language Toolkit. In: ETMTNLP '02: Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1. 2002 Presented at: ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics; July 7, 2002; Philadelphia, PA p. 63-70. [doi: [10.3115/1118108.1118117](https://doi.org/10.3115/1118108.1118117)]
36. Xu K, Feng Y, Huang S, Zhao D. Semantic relation classification via convolutional neural networks with simple negative sampling. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 Presented at: 2015 Conference on Empirical Methods in Natural Language Processing; September 17-21, 2015; Lisbon, Portugal p. 536-540. [doi: [10.18653/v1/d15-1062](https://doi.org/10.18653/v1/d15-1062)]
37. Nguyen T, Grishman R. Relation Extraction: Perspective from Convolutional Neural Networks. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015 Presented at: 1st Workshop on Vector Space Modeling for Natural Language Processing; June 5, 2015; Denver, CO p. 39-48. [doi: [10.3115/v1/w15-1506](https://doi.org/10.3115/v1/w15-1506)]
38. Wang L, Cao Z, De MG, Liu Z. Relation classification via multi-level attention CNNs. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016 Presented at: 54th Annual Meeting of the Association for Computational Linguistics; August 7-12, 2016; Berlin, Germany p. 1298-1307. [doi: [10.18653/v1/p16-1123](https://doi.org/10.18653/v1/p16-1123)]
39. Liu C, Sun W, Chao W, Che W. Convolution Neural Network for Relation Extraction. In: Motoda H, Wu Z, Cao L, Zaiane O, Yao M, Wang W, editors. *Advanced Data Mining and Applications. ADMA 2013. Lecture Notes in Computer Science*, vol 8347. Berlin, Heidelberg: Springer; 2013:231-242.
40. Li F, Liu W, Yu H. Extraction of Information Related to Adverse Drug Events from Electronic Health Record Notes: Design of an End-to-End Model Based on Deep Learning. *JMIR Med Inform* 2018 Nov 26;6(4):e12159 [FREE Full text] [doi: [10.2196/12159](https://doi.org/10.2196/12159)] [Medline: [30478023](https://pubmed.ncbi.nlm.nih.gov/30478023/)]
41. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc* 2004;11(5):392-402 [FREE Full text] [doi: [10.1197/jamia.M1552](https://doi.org/10.1197/jamia.M1552)] [Medline: [15187068](https://pubmed.ncbi.nlm.nih.gov/15187068/)]
42. Kipf T, Welling M. Semi-supervised classification with graph convolutional networks. 2017 Presented at: 5th International Conference on Learning Representations, ICLR; April 24-26, 2017; Toulon, France URL: <https://openreview.net/pdf?id=SJU4ayYgl>
43. Zhang Y, Qi P, Manning C. Graph convolution over pruned dependency trees improves relation extraction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: 2018 Conference on Empirical Methods in Natural Language Processing; October 31-November 4, 2018; Brussels, Belgium p. 2205-2215. [doi: [10.18653/v1/d18-1244](https://doi.org/10.18653/v1/d18-1244)]
44. Aronson A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001:17-21 [FREE Full text] [Medline: [11825149](https://pubmed.ncbi.nlm.nih.gov/11825149/)]
45. Li F, Jin Y, Liu W, Rawat B, Cai P, Yu H. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)-Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. *JMIR Med Inform* 2019 Sep 12;7(3):e14830 [FREE Full text] [doi: [10.2196/14830](https://doi.org/10.2196/14830)] [Medline: [31516126](https://pubmed.ncbi.nlm.nih.gov/31516126/)]
46. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682)] [Medline: [31501885](https://pubmed.ncbi.nlm.nih.gov/31501885/)]



47. Alsentzer E, Murphy J, Boag W, Weng W, Jindi D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June 7, 2019; Minneapolis, MN p. 72-78. [doi: [10.18653/v1/w19-1909](https://doi.org/10.18653/v1/w19-1909)]
48. Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25-29, 2014; Doha, Qatar p. 2014-2014. [doi: [10.3115/v1/d14-1162](https://doi.org/10.3115/v1/d14-1162)]
49. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020 Presented at: 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; November 16-20, 2020; Online p. 38-45. [doi: [10.18653/v1/2020.emnlp-demos.6](https://doi.org/10.18653/v1/2020.emnlp-demos.6)]
50. Berg-Kirkpatrick T, Burkett D, Klein D. An empirical investigation of statistical significance in NLP. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012 Presented at: 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning; July 12-14, 2012; Jeju Island, Korea p. 995-1005 URL: <https://www.aclweb.org/anthology/D12-1091>
51. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional Semantics Resources for Biomedical Text Processing. In: Proceedings of LBM. 2013 Presented at: LBM; December 12-13, 2013; Tokyo, Japan p. 39-44 URL: <http://bio.nplab.org/pdf/pyysalo13literature.pdf>
52. Geirhos R, Jacobsen J, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nat Mach Intell 2020 Nov 10;2(11):665-673. [doi: [10.1038/s42256-020-00257-z](https://doi.org/10.1038/s42256-020-00257-z)]
53. Mitra A, Rawat B, McManus D, Kapoor A, Yu H. Bleeding Entity Recognition in Electronic Health Records: A Comprehensive Analysis of End-to-End Systems. AMIA Annu Symp Proc 2020;2020:860-869 [FREE Full text] [Medline: [33936461](https://pubmed.ncbi.nlm.nih.gov/33936461/)]

---

## Abbreviations

**ADE:** adverse drug effect  
**AF:** atrial fibrillation  
**AGGCN:** attention-guided graph convolutional network  
**AltCause:** suspected alternative cause  
**BERT:** Bidirectional Encoder Representations from Transformers  
**bi-LSTM:** bidirectional long short-term memory  
**Bio+Clinical BERT:** BioBERT pretrained on clinical text  
**BioBERT:** BERT pretrained on biomedical data  
**CNN:** convolutional neural network  
**DL:** deep learning  
**EHR:** electronic health record  
**EhrBERT:** BioBERT pretrained on EHR notes  
**GCN:** graph convolutional network  
**GLOVE:** global vectors for word representation  
**MLP:** multilayer perceptron  
**NER:** named entity recognition  
**NLP:** natural language processing  
**POS:** part-of-speech  
**RNN:** recurrent neural network  
**STE:** semantic-type embedding  
**SVM:** support vector machine  
**UMLS:** Unified Medical Language System

*Edited by G Eysenbach; submitted 27.01.21; peer-reviewed by Y Fan, H Park, Y Mao; comments to author 19.02.21; revised version received 19.03.21; accepted 30.05.21; published 02.07.21.*

*Please cite as:*

*Mitra A, Rawat BPS, McManus DD, Yu H*

*Relation Classification for Bleeding Events From Electronic Health Records Using Deep Learning Systems: An Empirical Study*  
*JMIR Med Inform 2021;9(7):e27527*

*URL: <https://medinform.jmir.org/2021/7/e27527>*

*doi: [10.2196/27527](https://doi.org/10.2196/27527)*

*PMID: [34255697](https://pubmed.ncbi.nlm.nih.gov/34255697/)*

©Avijit Mitra, Bhanu Pratap Singh Rawat, David D McManus, Hong Yu. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 02.07.2021. 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

# Machine Learning Methods for the Diagnosis of Chronic Obstructive Pulmonary Disease in Healthy Subjects: Retrospective Observational Cohort Study

Shigeo Muro<sup>1</sup>, MD, PhD; Masato Ishida<sup>2</sup>, MSc; Yoshiharu Horie<sup>3</sup>, PhD; Wataru Takeuchi<sup>4</sup>, MEng; Shunki Nakagawa<sup>4</sup>, MSc; Hideyuki Ban<sup>4</sup>, PhD; Tohru Nakagawa<sup>5</sup>, MD, PhD; Tetsuhisa Kitamura<sup>6</sup>, MD, MSc, DPH

<sup>1</sup>Department of Respiratory Medicine, Nara Medical University, Nara, Japan

<sup>2</sup>Department of Respiratory and Immunology, Medical, AstraZeneca KK, Osaka, Japan

<sup>3</sup>Department of Data Science, Medical, AstraZeneca KK, Osaka, Japan

<sup>4</sup>Center for Technology Innovation–Artificial Intelligence, Research & Development Group, Hitachi, Ltd, Tokyo, Japan

<sup>5</sup>Hitachi Health Care Center, Hitachi, Ltd, Ibaraki, Japan

<sup>6</sup>Division of Environmental Medicine and Population Sciences, Department of Social and Environmental Medicine, Graduate School of Medicine, Osaka University, Osaka, Japan

**Corresponding Author:**

Yoshiharu Horie, PhD

Department of Data Science, Medical

AstraZeneca KK

3-1, Ofuka-cho, Kita-ku

Osaka, 5300011

Japan

Phone: 81 81 6 4802 3600

Email: [yoshiharu.horie@astrazeneca.com](mailto:yoshiharu.horie@astrazeneca.com)

## Abstract

**Background:** Airflow limitation is a critical physiological feature in chronic obstructive pulmonary disease (COPD), for which long-term exposure to noxious substances, including tobacco smoke, is an established risk. However, not all long-term smokers develop COPD, meaning that other risk factors exist.

**Objective:** This study aimed to predict the risk factors for COPD diagnosis using machine learning in an annual medical check-up database.

**Methods:** In this retrospective observational cohort study (ARTDECO [Analysis of Risk Factors to Detect COPD]), annual medical check-up records for all Hitachi Ltd employees in Japan collected from April 1998 to March 2019 were analyzed. Employees who provided informed consent via an opt-out model were screened and those aged 30 to 75 years without a prior diagnosis of COPD/asthma or a history of cancer were included. The database included clinical measurements (eg, pulmonary function tests) and questionnaire responses. To predict the risk factors for COPD diagnosis within a 3-year period, the Gradient Boosting Decision Tree machine learning (XGBoost) method was applied as a primary approach, with logistic regression as a secondary method. A diagnosis of COPD was made when the ratio of the prebronchodilator forced expiratory volume in 1 second (FEV<sub>1</sub>) to prebronchodilator forced vital capacity (FVC) was <0.7 during two consecutive examinations.

**Results:** Of the 26,101 individuals screened, 1213 met the exclusion criteria, and thus, 24,815 individuals were included in the analysis. The top 10 predictors for COPD diagnosis were FEV<sub>1</sub>/FVC, smoking status, allergic symptoms, cough, pack years, hemoglobin A<sub>1c</sub>, serum albumin, mean corpuscular volume, percent predicted vital capacity, and percent predicted value of FEV<sub>1</sub>. The areas under the receiver operating characteristic curves of the XGBoost model and the logistic regression model were 0.956 and 0.943, respectively.

**Conclusions:** Using a machine learning model in this longitudinal database, we identified a number of parameters as risk factors other than smoking exposure or lung function to support general practitioners and occupational health physicians to predict the development of COPD. Further research to confirm our results is warranted, as our analysis involved a database used only in Japan.

**KEYWORDS**

chronic obstructive pulmonary disease; airflow limitation; medical check-up; Gradient Boosting Decision Tree; logistic regression

## Introduction

Chronic obstructive pulmonary disease (COPD) is characterized by airflow limitation associated with persistent respiratory symptoms. Most patients with COPD experience exacerbation of symptoms and are at high risk of developing comorbidities such as cardiovascular disease [1].

Long-term exposure to tobacco smoke, vapor, gas, dust, and fumes is an established major risk factor for COPD [2]. However, only a small percentage of smokers develop airflow limitation, while nonsmokers can develop COPD [3]. These inconsistencies indicate that risk factors other than long-term smoking are associated with COPD [4].

The prevalence of COPD has been reported to be 12% to 13% among smokers [5]. However, only 9.4% of patients with airflow limitation have a previous diagnosis of COPD, and European data indicate that up to 80% of COPD cases are undiagnosed [6], suggesting delays in the diagnosis of COPD. The ARCTIC observational cohort study showed that late COPD diagnosis was associated with a higher exacerbation rate and increased comorbidities and costs compared with early diagnosis [7].

To address the issue of undiagnosed COPD, significant risk factors for airflow limitation other than smoking should be identified and evaluated in routine clinical practice. In a cohort study of 9040 individuals from the Japanese general population, concomitant *Chlamydia pneumoniae* and *Mycoplasma pneumoniae* seropositivity was found to be an independent risk factor for airflow limitation [8]. Additionally, Sato et al employed an annual health examination with pulmonary function tests measuring airflow limitation to identify undiagnosed patients with COPD among the Japanese population and found that iron deficiency might be associated with COPD development [9]. However, the follow-up duration of these cohorts was short (<3 years), limiting their ability to identify risk factors for COPD in the general population.

A large questionnaire-based surveillance demonstrated some improvement in diagnostic rates for COPD; however, approximately 60% of eligible participants failed to respond to the questionnaire [10]. While these results suggest that identifying robust and relevant risk factors is likely to improve early diagnosis, the slow progression and heterogeneity of the disease have hindered the identification of such risk factors for COPD development.

The recently reported “Subtype and Stage Inference” machine learning computational model identified subtypes of patients with COPD [11]. Compared with traditional approaches, the advantages of machine learning include the ability to process complex nonlinear relationships between predictors and to provide novel outputs. Therefore, the aim of this study was to apply machine learning methods to predict possible risk factors for the development of airflow limitation, an essential feature

of COPD diagnosis, using a Japanese medical check-up database comprising data from a number of healthy subjects to support the early diagnosis of COPD by general practitioners and occupational health physicians.

## Methods

### Study Design and Population

This was a retrospective observational cohort study to predict the risk factors for COPD diagnosis in healthy individuals. The analysis data set comprised individuals aged  $\geq 30$  years who had undertaken more than two medical check-ups, had no history of lung cancer or asthma at the first medical check-up, and could be classified as either having a diagnosis of COPD or as not having COPD. This study was designed according to the *Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Diagnosis* guidelines for prognostic studies [12].

The study protocol was reviewed and approved by the ethics committee of MINS (a nonprofit organization in Tokyo, Japan) and the Research & Development Group and Corporate Hospital Group of Hitachi, Ltd (Tokyo, Japan) prior to the start of data analysis. Individual informed consent was obtained using an opt-out model in agreement with the Institutional Review Board at Hitachi, Ltd. This study was conducted in accordance with the ethical principles of the Declaration of Helsinki.

### Data Source

The data source was annual medical check-up data for all Hitachi employees from April 1998 to March 2019. Data were archived in a high-security server that was managed with limited access rights by Hitachi. The annual medical check-up includes clinical measurements and questionnaires to examine the health of employees (Multimedia Appendix 1). Such questionnaires are utilized by Japanese organizations to evaluate their employees' health and give advice about health promotion, such as giving up smoking and exercising regularly based on the second term of the National Health Promotion Movement in the 21st century (Health Japan 21) issued by the Ministry of Health, Labour, and Welfare in Japan [13].

### Definition of COPD

COPD was considered according to the lung function status at two consecutive measurements during an annual lung function test when the prebronchodilator (pre-BD) forced expiratory volume in 1 second/forced vital capacity (FEV<sub>1</sub>/FVC) was  $< 0.7$ , as previously employed in a large population-based cohort study [14]. Individuals having a pre-BD FEV<sub>1</sub>/FVC  $\geq 0.7$  in at least three consecutive annual lung function test measurements were classified as non-COPD. For individuals with more than three records in the non-COPD group, the most recent three records were analyzed. Individuals having less than two lung function tests were excluded from all analyses. Spirometry was calibrated and performed by trained paramedical personnel according to



the American Thoracic Society/European Respiratory Society guidelines [15,16].

## Statistical Analysis

### *Age at COPD Diagnosis*

The age distribution for disease diagnosis was evaluated and stratified by smoking status (current smoker, exsmoker, or nonsmoker). The age at COPD diagnosis was defined as the age at the first of two consecutive measurements in which the pre-BD FEV<sub>1</sub>/FVC was <0.7.

### *Risk Factor Prediction Using Machine Learning*

Two types of models were constructed for predicting the risk factors for COPD diagnosis within 3 years as follows: a machine learning method (Gradient Boosting Decision Tree machine learning [XGBoost] [17]) and an established statistical method (logistic regression [18]). Individuals who did not meet the study inclusion criteria and/or had lung cancer/asthma were excluded from the analyses. Any individuals with missing data during the 3 years prior to the diagnosis year in the COPD group or during the most recent 3 years in the non-COPD group were excluded from the analyses. Propensity scores were calculated based on age, sex, smoking status, BMI, eosinophil count (EOS), and FEV<sub>1</sub>.

Data were randomly divided into a training data set and a test data set at a ratio of 7:3, with the same ratio of COPD to non-COPD individuals. Propensity scoring was used to balance the characteristics of COPD and non-COPD individuals (caliper: 0.2) in the training and test data sets. Next, the training data set was randomly divided 8:2 for model construction (XGBoost and logistic) and evaluation of model performance, respectively. The data split, model construction, and evaluation processes were repeated five times for cross-validation (5-CV approach) [19]. Model parameters, including the depth of the tree and regularization factor, were refined during performance evaluation by the 5-CV approach. Finally, the most optimal model was generated by applying the best parameters confirmed by the 5-CV approach. To evaluate model performance in the

unlearned data, the most optimized model was used to evaluate the test data set.

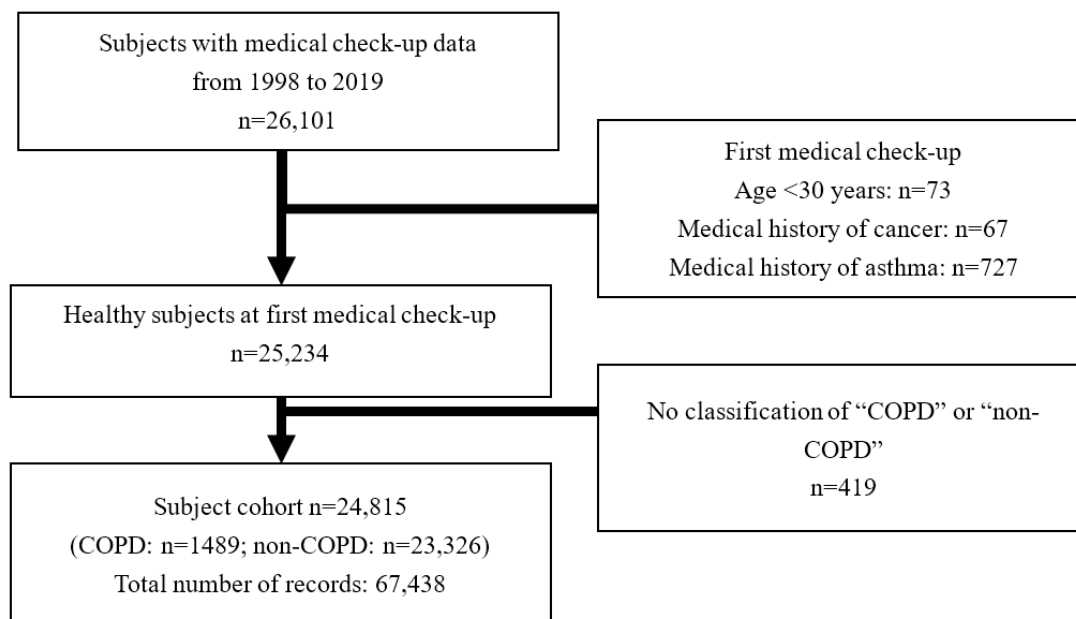
Model construction by logistic regression was performed in a similar way to the XGBoost method. Models were constructed in the training data set (randomly sampled data from the entire data set) and subsequently validated in the test data set after model evaluation.

Following model construction, the area under the receiver operating characteristic curve (AUC), positive predictive value, sensitivity, specificity, and F1-measure were calculated for each model to evaluate the performance under both 5-CV and test conditions [20]. The feature importance of the machine learning model was calculated to examine the contribution of each predictor to the model constructed using the Gini impurity method [19]. The feature weight of the logistic regression model was also calculated. All analyses were performed using Python 3.6 software (Python Software Foundation).

## Results

### Individuals

Data from 26,101 individuals (employees and their families) aged 30 to 75 years, who underwent annual check-ups between April 1998 and March 2019 were included in our analysis. The total number of medical check-up records was 318,568. All 26,101 individuals had lung function test measurements for 3 consecutive years. The medical records for 73 individuals aged <30 years at the first medical check-up, 67 individuals with a history of cancer, and 727 individuals with a history of asthma were excluded, as were data from 419 individuals who had already been diagnosed with COPD (subjects for whom all data points of pre-BD FEV<sub>1</sub>/FVC were <0.7 during the observational period) or had not been classified as either COPD or non-COPD (subjects with pre-BD FEV<sub>1</sub>/FVC <0.7 without two consecutive measurements). Accordingly, data for 24,815 individuals (corresponding to 67,438 records) were included in the analyses (Figure 1).

**Figure 1.** Flow diagram of the study. COPD: chronic obstructive pulmonary disease.

### Baseline Characteristics

Table 1 shows the baseline characteristics of the COPD and non-COPD groups. Overall, 1489 individuals were considered as having COPD (pre-BD FEV<sub>1</sub>/FVC <0.7 at two consecutive measurements during annual lung function tests). In comparison with the non-COPD group, the COPD group had a lower BMI, worse lung function (pre-BD FEV<sub>1</sub>, pre-BD percent predicted value of FEV<sub>1</sub> [%FEV<sub>1</sub>], and pre-BD FEV<sub>1</sub>/FVC), and greater emphysematous change and chronic inflammation as determined

by computed tomography. Furthermore, comorbidities, such as arrhythmia, duodenal ulcer, colorectal polyp, angina, stomach ulcer, and kidney disease, were more prevalent in the COPD group. Statistically significant differences in hematological parameters (mean corpuscular volume [MCV], mean corpuscular hemoglobin concentration [MCHC], mean corpuscular hemoglobin [MCH], hemoglobin [Hb], and hematocrit [HT] [15]) between the COPD and non-COPD groups were also observed. Inflammatory markers, particularly white blood cell (WBC) count and EOS, were also significantly higher in the COPD group.

**Table 1.** Subject characteristics stratified by chronic obstructive pulmonary disease status.

Characteristic	Non-COPD <sup>a</sup> (n=23,326)	COPD (n=1489)	P value
Age (years), mean (SD)	42 (9.1)	48 (9.3)	<.001
Female, n (%)	3841 (16.5%)	58 (3.9%)	<.001
<b>Smoking status, n (%)</b>			<.001
Current smoker	10,632 (45.6%)	1,021 (68.6%)	
Exsmoker	3534 (15.2%)	202 (13.6%)	
Nonsmoker	9153 (39.3%)	266 (17.9%)	
Unknown/missing	7 (0.0%)	0 (0.0%)	
BMI (kg/m <sup>2</sup> ), mean (SD)	23 (3.2)	22 (2.7)	<.001
<b>Lung function test, mean (SD)</b>			
Prebronchodilator FEV <sub>1</sub> <sup>b</sup>	3.4 (0.7)	3.1 (0.6)	<.001
Prebronchodilator FVC <sup>c</sup>	4.1 (0.8)	4.2 (0.8)	<.001
Prebronchodilator FEV <sub>1</sub> /FVC	83.7 (5.4)	74.9 (5.1)	<.001
<b>Comorbidity, n (%)</b>			
Arrhythmia	107 (0.5%)	16 (1.1%)	.003
Duodenal ulcer	158 (0.7%)	19 (1.3%)	.02
Colorectal polyp	43 (0.2%)	13 (0.9%)	<.001
Angina	56 (0.2%)	10 (0.7%)	.006
Stomach ulcer	180 (0.8%)	29 (1.9%)	<.001
Kidney disease	77 (0.3%)	12 (0.8%)	.01
<b>Computed tomography finding, n (%)</b>			
Bulla, bleb	108 (0.5%)	31 (2.1%)	<.001
Moderate emphysema	18 (0.1%)	13 (0.9%)	<.001
Mild emphysema	96 (0.4%)	27 (1.8%)	<.001
Calcification of left anterior descending coronary artery	128 (0.5%)	16 (1.1%)	.02
Chronic inflammation	342 (1.5%)	43 (2.9%)	<.001
<b>Laboratory parameters, mean (SD)</b>			
Albumin (U/L)	4.4 (0.2)	4.3 (0.2)	<.001
Alanine aminotransferase (U/L)	209.5 (53.7)	215.2 (54.6)	<.001
Aspartate aminotransferase (U/L)	26.4 (14.8)	24.3 (12.8)	<.001
Blood urea nitrogen (mg/dL)	14.1 (3.2)	14.7 (3.3)	<.001
Cholinesterase (U/L)	320.8 (60.1)	307.8 (58.8)	<.001
Estimated glomerular filtration rate (mL/min/1.73 m <sup>2</sup> )	83.6 (14.6)	80.2 (14.1)	<.001
Eosinophil count (cells/mm <sup>3</sup> )	183.2 (124.5)	195.4 (125.9)	<.001
Gamma-glutamyl transferase (U/L)	42.7 (34.4)	45.8 (34.5)	<.001
Hemoglobin (g/dL)	14.7 (1.4)	14.9 (1.1)	<.001
Hemoglobin A <sub>1c</sub> (%)	5.3 (0.7)	5.4 (0.7)	<.001
Hematocrit (%)	44.0 (3.6)	44.6 (3.1)	<.001
MCH <sup>d</sup> (pg)	30.5 (1.8)	31.1 (1.7)	<.001
MCHC <sup>e</sup> (g/L)	33.4 (1.0)	33.3 (0.8)	<.001
MCV <sup>f</sup> (fL)	91.3 (4.6)	93.4 (4.4)	<.001

Characteristic	Non-COPD <sup>a</sup> (n=23,326)	COPD (n=1489)	P value
WBC <sup>g</sup> count ( $\times 10^2$ cells/ $\mu$ L)	58.8 (15.0)	62.8 (15.5)	<.001

<sup>a</sup>COPD: chronic obstructive pulmonary disease.

<sup>b</sup>FEV<sub>1</sub>: forced expiratory volume in 1 second.

<sup>c</sup>FVC: forced vital capacity.

<sup>d</sup>MCH: mean corpuscular hemoglobin.

<sup>e</sup>MCHC: mean corpuscular hemoglobin concentration.

<sup>f</sup>MCV: mean corpuscular volume.

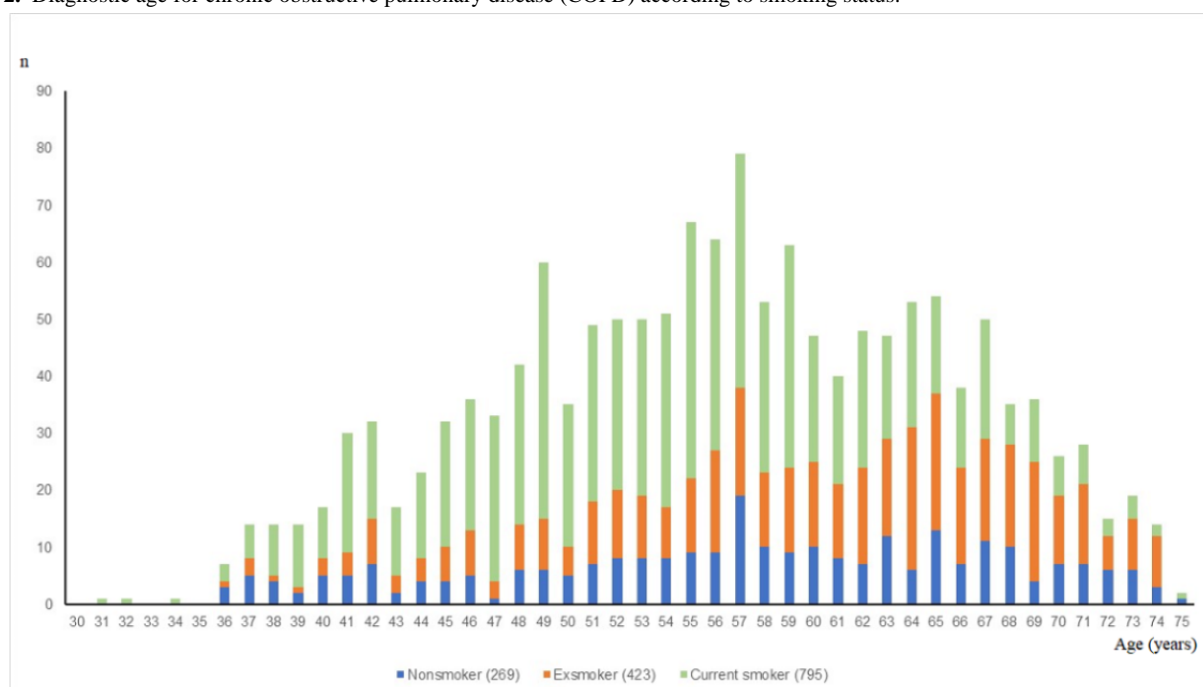
<sup>g</sup>WBC: white blood cell.

## Percentage of Individuals With COPD

The overall percentage of individuals with COPD was 6.0% (1489/24,815). According to smoking status, the percentage of individuals with COPD was 8.8% (1021/11,653) among current

smokers and 5.4% (202/3736) among exsmokers. Notably, 2.8% (266/9419) of nonsmokers had developed COPD. The peak age at diagnosis of COPD among current smokers and exsmokers was 55 years and 65 years, respectively (Figure 2).

**Figure 2.** Diagnostic age for chronic obstructive pulmonary disease (COPD) according to smoking status.



## Risk Factors for COPD Diagnosis

Overall, 20,265 individuals (COPD: n=954; non-COPD: n=19,311) with 51,432 records (COPD: n=2435; non-COPD: n=48,997) out of 24,815 individuals who met the criteria (Multimedia Appendix 2) were included in the machine learning analysis. Table 2 shows the model performance of the XGBoost and logistic regression models. For both models, the AUC, accuracy, sensitivity, specificity, and F-measure were generally similar between the training and test data sets. The XGBoost model had a higher positive predictive value (0.505) than the logistic regression model (0.441). The AUC was high in the training and test sets for both models (range: 0.892-0.956). Additionally, the accuracy and specificity exceeded 0.883 and 0.879, respectively, for both models.

The most important predictive factors for COPD diagnosis were lung function tests (ie, FEV<sub>1</sub>/FVC, percent vital capacity [%VC], and %FEV<sub>1</sub>) and smoking status, followed by cough, hematological indices (ie, MCV, MCHC, MCH, Hb, and HT), treatment with antidiabetic drugs, hemoglobin A<sub>1c</sub>, serum albumin, total protein, and BMI. Other predictive risk factors were EOS, serum alanine aminotransferase, WBC count, and urinary WBC count (Table 3). Logistic regression analysis showed that low FEV<sub>1</sub>/FVC and %FEV<sub>1</sub>; high %VC; high MCV, MCHC, and Hb; and low HT and MCH were related factors, and that individuals treated with antidiabetic drugs had a higher number of associated risk factors for COPD. Low serum albumin, low total protein, and low BMI were also confirmed as risk factors (Multimedia Appendix 3).



**Table 2.** Comparison of performance of the Gradient Boosting Decision Tree machine learning (XGBoost) and logistic regression models.

Variable	XGBoost <sup>a</sup> model		Logistic regression model	
	Training, mean (SE)	Test, mean	Training, mean (SE)	Test, mean
Positive predictive value	0.505 (0.099)	0.362	0.441 (0.110)	0.285
AUC <sup>b</sup>	0.956 (0.015)	0.898	0.943 (0.022)	0.892
Accuracy	0.917 (0.032)	0.918	0.884 (0.049)	0.883
Sensitivity	0.845 (0.021)	0.877	0.874 (0.039)	0.901
Specificity	0.960 (0.016)	0.919	0.946 (0.025)	0.882
F-measure	0.370 (0.107)	0.513	0.306 (0.110)	0.434

<sup>a</sup>XGBoost: Gradient Boosting Decision Tree machine learning.

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

**Table 3.** Importance of each predictor in the XGBoost model.

Variable	Importance value
Forced expiratory volume in 1 second/forced vital capacity	0.2824
Smoking status	0.0329
Allergic symptoms (yes/no)	0.0303
Symptom-cough (yes/no)	0.0294
Smoking-pack year	0.0222
Hemoglobin A <sub>1c</sub>	0.0197
Albumin	0.0195
Mean corpuscular volume	0.0177
%Vital capacity	0.0165
%Forced expiratory volume in 1 second	0.0164
Treatment with an antidiabetic drug (yes/no)	0.0162
Allergic disease (yes/no)	0.0146
Hematocrit	0.0144
Urinary red blood cells	0.0143
Hemoglobin	0.0138
Age	0.0128
Smoking duration	0.0127
High density lipoprotein cholesterol	0.0123
Mean corpuscular hemoglobin concentration	0.0122
Total protein	0.0118
BMI	0.0118
Number of eosinophils	0.0115
Mean corpuscular hemoglobin	0.0114
Serum white blood cells	0.0111
Fasting blood sugar	0.0110
Serum alanine aminotransferase	0.0108
Pulse rate	0.0108
Forced expiratory volume in 1 second	0.0107
Urinary white blood cells	0.0104
Diastolic blood pressure	0.0103

For future utilization of risk factors for disease assessment in daily clinical practice, the machine learning process was validated using a questionnaire to predict risk factors for COPD development (Multimedia Appendix 1). Of 30 variables, 25 were clinical parameters that overlapped between the two methods. The top 30 risk factors also included the following five questions: "I am regularly doing exercise," "I have chest compression and pain," "Average sleeping time in the past 1 month," "I have breakfast every day," and "Body fat ratio" (Multimedia Appendix 4). Among these, logistic regression analysis showed that insufficient sleeping time and not having breakfast every day were risk factors for COPD (Multimedia Appendix 3).

## Discussion

This study applied a machine learning method, a powerful tool to analyze large quantities of complex data, to predict risk factors for COPD. This is the first study to investigate more than 300,000 records from working-age adults in Japan utilizing an annual medical check-up database. This system allows healthy employees to track their health conditions over time by clinical measurements and questionnaires. We found that the most significant predictor of COPD diagnosis was the absolute value of  $FEV_1/FVC$ , indicating that low  $FEV_1$  in early adulthood is an important factor in the development of COPD. Childhood asthma is associated with impaired lung function, lower lung function in adulthood, and higher risk of COPD even for nonsmoker participants, as previously reported by Martinez et al [21]. In our speculation, some part of the nonsmoker COPD population might have had a history of childhood asthma, increasing susceptibility to passive smoke exposure or airway pollution and resulting in the early diagnosis of COPD in nonsmokers compared with exsmokers in the study. Smoking status had the second highest impact on disease diagnosis. Among individuals with a smoking history, the peak age of COPD diagnosis was older in exsmokers than in current smokers. This finding suggests that smoking cessation delays the diagnosis of COPD, consistent with a previous study in which smoking cessation was reported to affect the natural history of COPD [22].

Erythrocyte indices (MCV and MCHC) might also be available as potential predictors of COPD diagnosis in addition to lung function measurements. These data are supported by a previous report in which continuous smoking had a significant effect on hematological parameters compared with nonsmoking, and it may be associated with an increased risk of COPD [23]. The increased levels of MCV and MCHC in individuals with COPD support a previous finding that impaired lung function has a strong association with ischemic heart disease [24]. Conversely, the presence of an allergic disease appeared to have a preventive effect on airflow limitation, which is in contrast with observations from the Tasmanian Longitudinal Health Study in which the presence of allergic diseases was an early predictor of lung trajectories toward COPD [25]. However, the Hokkaido cohort study showed that subjects with multiple asthma-like features had slower lung function decline [26]. From the findings of our observational study in Japan, we can speculate that early diagnosis and intervention for allergic diseases may have less

impact on lung function and that regular and frequent medical intervention could lead to an overall increase in life expectancy among patients who can readily access appropriate treatment by respiratory specialists.

Furthermore, individuals with decreased levels of serum albumin and total protein, as well as lower hemoglobin  $A_{1c}$  and BMI may be at risk of developing cachexia, a common condition among patients with COPD [27]. With respect to other identified risk factors, a retrospective cross-sectional study showed an association between EOS and airflow limitation in patients with COPD [28]. Given that increased alanine aminotransferase levels have been observed in patients with obstructive sleep apnea [29], individuals at risk of developing COPD might be exposed to intermittent hypoxia, indicating that a reduced sleeping time, as determined in the study questionnaire, might also represent a risk factor for COPD. Even minor changes in hematological parameters might be attributable to hypoxic conditions, leading to sleep disruption. Additionally, frequently missing breakfast might accelerate malnutrition in the COPD group. Furthermore, significantly higher prevalence rates of chronic neck and lower back pain in patients with COPD compared with healthy individuals were observed in a population-based study, although the findings were not confirmed by logistic regression analysis [30], and the link between COPD and back pain remains unknown. The observation of increased WBC counts in patients with COPD compared with healthy controls [31] suggests that systemic inflammation may be involved in the pathogenesis of COPD [32].

Our results also indicate that smoking cessation should be prioritized for the prevention of COPD and that smokers with sleep disturbances, back pain, and/or low BMI and malnutrition may be at increased risk of developing COPD and should be considered as candidates for lifestyle intervention therapy. Furthermore, the five key questions included in our questionnaire should be validated in future investigations and potentially implemented in daily practice as part of an annual medical check-up to prevent COPD.

The positive predictive value of the XGBoost model was comparable to that of a self-scored persistent airflow obstruction screening questionnaire in the Japanese population previously reported by Samukawa et al [33]. However, our models showed more accuracy because the sensitivity and specificity of our models achieved higher figures, and the AUC reached over 0.9 compared with that of the questionnaire, which ranged from 0.595 to 0.612. The AUCs of the XGBoost and logistic regression models were similar, while the most important factor related to COPD diagnosis was  $FEV_1$  in both models. However, some variables differed in importance in each model. Kuhn et al reported that machine learning approaches can incorporate high-order nonlinear interactions among predictors that cannot be addressed by traditional modeling approaches (eg, logistic regression models) [34]. However, machine learning methods cannot elucidate whether a causal relationship exists between the identified variable and the disease. Thus, the association between risk factors detected using a machine learning model and COPD requires validation in future prospective studies.

A strength of this study was the use of longitudinal lung function test data from healthy individuals from April 1998 to March 2019. In general, medical checkup data are not linked to medical records, meaning that profiles of lung function tests over time could not be investigated. However, it was possible to evaluate longitudinal lung function tests because the database included data from individuals from a point in time when they were healthy until they had developed COPD. Additionally, data from healthy individuals were included, allowing lung function test results from when they were diagnosed with COPD to be investigated. Finally, both clinical measurements and questionnaire variables were included in the database, thereby increasing the potential to identify several different risk factors for COPD.

The limitations of this study include the definition of COPD diagnosis by airflow limitation with pulmonary function tests. Instead of post-BD spirometry data as suggested by the ATS/European respiratory guidelines, we employed pre-BD spirometry data for the diagnosis of COPD since no post-BD spirometry was performed in the annual medical check-up. The precise diagnosis of COPD cannot always be demonstrated by airflow limitation alone; however, we believe that the diagnostic approach was reasonable from a clinical perspective as airflow limitation has been reported to be a poor prognostic factor in the general population [35]. Low lung function values

( $FEV_1/FVC < 0.7$ ) might be observed at a single time point in some individuals for no discernable reason. Therefore, we considered COPD as lung function of  $FEV_1/FVC < 0.7$  on two consecutive occasions. In terms of differentiation between asthma and COPD, we cannot exclude the possibility of misclassification of asthma as COPD in some patients since reversibility tests were not performed in the annual medical check-up, but participants with a medical history of asthma were excluded. Additionally, a database from a single organization was analyzed in this study; thus, the results might include bias based on the type of industry or the organizational structure of the company, limiting the generalizability of the findings. To obtain more generalizable findings, studies using other databases are necessary. Finally, some unknown confounders may have remained; therefore, we plan to perform model validation by analyzing other databases. Well-controlled prospective studies should be conducted to confirm the predictive factors for COPD diagnosis.

In conclusion, our machine learning method applied to longitudinal medical check-up data, including general questionnaires and laboratory parameters, identified hematological, nutritional, and inflammatory parameters as potential risk factors for COPD. These parameters, along with lung function and smoking status, may be useful in identifying at-risk individuals and may lead to an earlier diagnosis.

---

## Acknowledgments

The authors thank ND Smith, Y Baba, and K Dohi of EMC Japan (Osaka, Japan) for critical reading and native checking of the manuscript. We also thank Tricia Newell and Clare Cox of Edanz Evidence Generation for providing editing support. This study was funded by AstraZeneca KK.

---

## Authors' Contributions

SM and TK contributed to data interpretation and reviewed the manuscript. YH planned the analyses, contributed to data interpretation, and drafted the manuscript. MI designed the study and drafted the manuscript. SN, WT, and HB conducted the analyses. TN corrected and provided the data, and reviewed the analysis including data cleansing and preprocessing. All authors take full responsibility for the content and editorial decisions, and approved the final version.

---

## Conflicts of Interest

SM has received honoraria from AstraZeneca KK, Boehringer Ingelheim Japan, GlaxoSmithKline KK, Novartis Pharma KK, Meiji Seika Pharma Co, Ltd, Kyorin Pharmaceutical Co, Ltd, Otsuka Pharmaceutical Co, Ltd, Teijin Pharma Ltd, CHEST MI, Inc, Daiichi Sankyo Co, Ltd, Chugai Pharmaceutical Co, Ltd, Sanofi KK, Actelion Pharmaceuticals Japan Ltd, and Olympus Corporation. MI and YH are employees of AstraZeneca KK. WT, SN, HB, and TN are employees of Hitachi, Ltd.

---

### Multimedia Appendix 1

Clinical assessments and questions used in the analysis, and list of variables.

[\[DOCX File, 56 KB - medinform\\_v9i7e24796\\_app1.docx\]](#)

---

### Multimedia Appendix 2

Numbers of records and individuals included in the machine learning model.

[\[DOCX File, 42 KB - medinform\\_v9i7e24796\\_app2.docx\]](#)

---

### Multimedia Appendix 3

Association between chronic obstructive pulmonary disease and the top 30 variables of importance based on logistic regression.

[\[DOCX File, 45 KB - medinform\\_v9i7e24796\\_app3.docx\]](#)

## Multimedia Appendix 4

Importance of each predictor in the XGBoost model (including questionnaire items).

[\[DOCX File, 44 KB - medinform\\_v9i7e24796\\_app4.docx\]](#)

## References

1. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. *Am J Respir Crit Care Med* 2017 Mar 01;195(5):557-582. [doi: [10.1164/rccm.201701-0218PP](https://doi.org/10.1164/rccm.201701-0218PP)] [Medline: [28128970](https://pubmed.ncbi.nlm.nih.gov/28128970/)]
2. de Marco R, Accordini S, Marcon A, Cerveri I, Antó JM, Gislason T, European Community Respiratory Health Survey (ECRHS). Risk factors for chronic obstructive pulmonary disease in a European cohort of young adults. *Am J Respir Crit Care Med* 2011 Apr 01;183(7):891-897. [doi: [10.1164/rccm.201007-1125OC](https://doi.org/10.1164/rccm.201007-1125OC)] [Medline: [20935112](https://pubmed.ncbi.nlm.nih.gov/20935112/)]
3. Fletcher C, Peto R. The natural history of chronic airflow obstruction. *Br Med J* 1977 Jun 25;1(6077):1645-1648 [[FREE Full text](#)] [doi: [10.1136/bmj.1.6077.1645](https://doi.org/10.1136/bmj.1.6077.1645)] [Medline: [871704](https://pubmed.ncbi.nlm.nih.gov/871704/)]
4. Stang P, Lydick E, Silberman C, Kempel A, Keating ET. The prevalence of COPD: using smoking rates to estimate disease frequency in the general population. *Chest* 2000 May;117(5 Suppl 2):354S-359S. [doi: [10.1378/chest.117.5\\_suppl\\_2.354S](https://doi.org/10.1378/chest.117.5_suppl_2.354S)] [Medline: [10843976](https://pubmed.ncbi.nlm.nih.gov/10843976/)]
5. Fukuchi Y, Nishimura M, Ichinose M, Adachi M, Nagai A, Kuriyama T, et al. COPD in Japan: the Nippon COPD Epidemiology study. *Respirology* 2004 Nov;9(4):458-465. [doi: [10.1111/j.1440-1843.2004.00637.x](https://doi.org/10.1111/j.1440-1843.2004.00637.x)] [Medline: [15612956](https://pubmed.ncbi.nlm.nih.gov/15612956/)]
6. Soriano JB, Zielinski J, Price D. Screening for and early detection of chronic obstructive pulmonary disease. *The Lancet* 2009 Aug 29;374(9691):721-732. [doi: [10.1016/S0140-6736\(09\)61290-3](https://doi.org/10.1016/S0140-6736(09)61290-3)] [Medline: [19716965](https://pubmed.ncbi.nlm.nih.gov/19716965/)]
7. Larsson K, Janson C, Ställberg B, Lisspers K, Olsson P, Kostikas K, et al. Impact of COPD diagnosis timing on clinical and economic outcomes: the ARCTIC observational cohort study. *Int J Chron Obstruct Pulmon Dis* 2019;14:995-1008 [[FREE Full text](#)] [doi: [10.2147/COPD.S195382](https://doi.org/10.2147/COPD.S195382)] [Medline: [31190785](https://pubmed.ncbi.nlm.nih.gov/31190785/)]
8. Muro S, Tabara Y, Matsumoto H, Setoh K, Kawaguchi T, Takahashi M, Nagahama Study Group. Relationship Among Chlamydia and Mycoplasma Pneumoniae Seropositivity, IKZF1 Genotype and Chronic Obstructive Pulmonary Disease in A General Japanese Population: The Nagahama Study. *Medicine (Baltimore)* 2016 Apr;95(15):e3371 [[FREE Full text](#)] [doi: [10.1097/MD.0000000000003371](https://doi.org/10.1097/MD.0000000000003371)] [Medline: [27082601](https://pubmed.ncbi.nlm.nih.gov/27082601/)]
9. Sato K, Shibata Y, Inoue S, Igarashi A, Tokairin Y, Yamauchi K, et al. Impact of cigarette smoking on decline in forced expiratory volume in 1s relative to severity of airflow obstruction in a Japanese general population: The Yamagata-Takahata study. *Respir Investig* 2018 Mar;56(2):120-127. [doi: [10.1016/j.resinv.2017.11.011](https://doi.org/10.1016/j.resinv.2017.11.011)] [Medline: [29548649](https://pubmed.ncbi.nlm.nih.gov/29548649/)]
10. Jordan RE, Adab P, Sitch A, Enocson A, Blissett D, Jowett S, et al. Targeted case finding for chronic obstructive pulmonary disease versus routine practice in primary care (TargetCOPD): a cluster-randomised controlled trial. *The Lancet Respiratory Medicine* 2016 Sep;4(9):720-730. [doi: [10.1016/S2213-2600\(16\)30149-7](https://doi.org/10.1016/S2213-2600(16)30149-7)] [Medline: [27444687](https://pubmed.ncbi.nlm.nih.gov/27444687/)]
11. Young AL, Bragman FJS, Rangelov B, Han MK, Galbán CJ, Lynch DA, COPDGene Investigators. Disease Progression Modeling in Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* 2020 Feb 01;201(3):294-302 [[FREE Full text](#)] [doi: [10.1164/rccm.201908-1600OC](https://doi.org/10.1164/rccm.201908-1600OC)] [Medline: [31657634](https://pubmed.ncbi.nlm.nih.gov/31657634/)]
12. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015 Jan 06;162(1):55-63. [doi: [10.7326/M14-0697](https://doi.org/10.7326/M14-0697)] [Medline: [25560714](https://pubmed.ncbi.nlm.nih.gov/25560714/)]
13. Sugiyama K, Tomata Y, Takemi Y, Tsushita K, Nakamura M, Hashimoto S, et al. Awareness and health consciousness regarding the national health plan "Health Japan 21" (2nd edition) among the Japanese population in 2013 and 2014. *Nihon Koshu Eisei Zasshi* 2016;63(8):424-431 [[FREE Full text](#)] [doi: [10.11236/jph.63.8\\_424](https://doi.org/10.11236/jph.63.8_424)] [Medline: [27681283](https://pubmed.ncbi.nlm.nih.gov/27681283/)]
14. Terzikhan N, Verhamme KMC, Hofman A, Stricker BH, Brusselle GG, Lahousse L. Prevalence and incidence of COPD in smokers and non-smokers: the Rotterdam Study. *Eur J Epidemiol* 2016 Aug;31(8):785-792 [[FREE Full text](#)] [doi: [10.1007/s10654-016-0132-z](https://doi.org/10.1007/s10654-016-0132-z)] [Medline: [26946425](https://pubmed.ncbi.nlm.nih.gov/26946425/)]
15. Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, ATS/ERS Task Force. Standardisation of spirometry. *Eur Respir J* 2005 Aug;26(2):319-338 [[FREE Full text](#)] [doi: [10.1183/09031936.05.00034805](https://doi.org/10.1183/09031936.05.00034805)] [Medline: [16055882](https://pubmed.ncbi.nlm.nih.gov/16055882/)]
16. Celli BR, MacNee W, ATS/ERS Task Force. Standards for the diagnosis and treatment of patients with COPD: a summary of the ATS/ERS position paper. *Eur Respir J* 2004 Jun;23(6):932-946 [[FREE Full text](#)] [doi: [10.1183/09031936.04.00014304](https://doi.org/10.1183/09031936.04.00014304)] [Medline: [15219010](https://pubmed.ncbi.nlm.nih.gov/15219010/)]
17. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016 Presented at: 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)]
18. Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, et al. Strong rules for discarding predictors in lasso-type problems. *J R Stat Soc Series B Stat Methodol* 2012 Mar;74(2):245-266 [[FREE Full text](#)] [doi: [10.1111/j.1467-9868.2011.01004.x](https://doi.org/10.1111/j.1467-9868.2011.01004.x)] [Medline: [25506256](https://pubmed.ncbi.nlm.nih.gov/25506256/)]
19. James G, Witten D, Hastie T, Tibshirani R. Introduction. In: *An Introduction to Statistical Learning*. Springer Texts in Statistics, vol 103. New York, USA: Springer; 2013.



20. 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. [Medline: [3203132](#)]
21. Martinez FJ, Han MK, Allinson JP, Barr RG, Boucher RC, Calverley PMA, et al. At the Root: Defining and Halting Progression of Early Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* 2018 Jun 15;197(12):1540-1551 [FREE Full text] [doi: [10.1164/rccm.201710-2028PP](#)] [Medline: [29406779](#)]
22. Bai J, Chen X, Liu S, Yu L, Xu J. Smoking cessation affects the natural history of COPD. *Int J Chron Obstruct Pulmon Dis* 2017;12:3323-3328 [FREE Full text] [doi: [10.2147/COPD.S150243](#)] [Medline: [29180862](#)]
23. Malenica M, Prnjavorac B, Bego T, Dujic T, Semiz S, Skrbo S, et al. Effect of Cigarette Smoking on Haematological Parameters in Healthy Population. *Med Arch* 2017 Apr;71(2):132-136 [FREE Full text] [doi: [10.5455/medarh.2017.71.132-136](#)] [Medline: [28790546](#)]
24. Eriksson B, Lindberg A, Müllerova H, Rönmark E, Lundbäck B. Association of heart diseases with COPD and restrictive lung function--results from a population survey. *Respir Med* 2013 Jan;107(1):98-106 [FREE Full text] [doi: [10.1016/j.rmed.2012.09.011](#)] [Medline: [23127573](#)]
25. Bui DS, Lodge CJ, Burgess JA, Lowe AJ, Perret J, Bui MQ, et al. Childhood predictors of lung function trajectories and future COPD risk: a prospective cohort study from the first to the sixth decade of life. *Lancet Respir Med* 2018 Jul;6(7):535-544. [doi: [10.1016/S2213-2600\(18\)30100-0](#)] [Medline: [29628376](#)]
26. Suzuki M, Makita H, Konno S, Shimizu K, Kimura H, Kimura H, Hokkaido COPD Cohort Study Investigators. Asthma-like Features and Clinical Course of Chronic Obstructive Pulmonary Disease. An Analysis from the Hokkaido COPD Cohort Study. *Am J Respir Crit Care Med* 2016 Dec 01;194(11):1358-1365. [doi: [10.1164/rccm.201602-0353OC](#)] [Medline: [27224255](#)]
27. von Haehling S, Anker MS, Anker SD. Prevalence and clinical impact of cachexia in chronic illness in Europe, USA, and Japan: facts and numbers update 2016. *J Cachexia Sarcopenia Muscle* 2016 Dec;7(5):507-509 [FREE Full text] [doi: [10.1002/jcsm.12167](#)] [Medline: [27891294](#)]
28. Huang W, Huang C, Wu P, Chen C, Cheng Y, Chen H, et al. The association between airflow limitation and blood eosinophil levels with treatment outcomes in patients with chronic obstructive pulmonary disease and prolonged mechanical ventilation. *Sci Rep* 2019 Sep 17;9(1):13420 [FREE Full text] [doi: [10.1038/s41598-019-49918-z](#)] [Medline: [31530874](#)]
29. Chin K, Nakamura T, Takahashi K, Sumi K, Ogawa Y, Masuzaki H, et al. Effects of obstructive sleep apnea syndrome on serum aminotransferase levels in obese patients. *Am J Med* 2003 Apr 01;114(5):370-376. [doi: [10.1016/s0002-9343\(02\)01570-x](#)] [Medline: [12714126](#)]
30. de Miguel-Díez J, López-de-Andrés A, Hernandez-Barrera V, Jimenez-Trujillo I, Del Barrio JL, Puente-Maestu L, et al. Prevalence of Pain in COPD Patients and Associated Factors: Report From a Population-based Study. *Clin J Pain* 2018 Sep;34(9):787-794. [doi: [10.1097/AJP.0000000000000598](#)] [Medline: [29485534](#)]
31. Biljak VR, Pancirov D, Cepelak I, Popović-Grle S, Stjepanović G, Grubišić T. Platelet count, mean platelet volume and smoking status in stable chronic obstructive pulmonary disease. *Platelets* 2011;22(6):466-470. [doi: [10.3109/09537104.2011.573887](#)] [Medline: [21506665](#)]
32. Agustí A, Edwards LD, Rennard SI, MacNee W, Tal-Singer R, Miller BE, Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) Investigators. Persistent systemic inflammation is associated with poor clinical outcomes in COPD: a novel phenotype. *PLoS One* 2012;7(5):e37483 [FREE Full text] [doi: [10.1371/journal.pone.0037483](#)] [Medline: [22624038](#)]
33. Samukawa T, Matsumoto K, Tsukuya G, Koriyama C, Fukuyama S, Uchida A, et al. Development of a self-scored persistent airflow obstruction screening questionnaire in a general Japanese population: the Hisayama study. *Int J Chron Obstruct Pulmon Dis* 2017;12:1469-1481 [FREE Full text] [doi: [10.2147/COPD.S130453](#)] [Medline: [28553099](#)]
34. Kuhn S, Egert B, Neumann S, Steinbeck C. Building blocks for automated elucidation of metabolites: machine learning methods for NMR prediction. *BMC Bioinformatics* 2008 Sep 25;9:400 [FREE Full text] [doi: [10.1186/1471-2105-9-400](#)] [Medline: [18817546](#)]
35. Akkermans RP, Biermans M, Robberts B, ter Riet G, Jacobs A, van Weel C, et al. COPD prognosis in relation to diagnostic criteria for airflow obstruction in smokers. *Eur Respir J* 2014 Jan;43(1):54-63 [FREE Full text] [doi: [10.1183/09031936.00158212](#)] [Medline: [23563262](#)]

## Abbreviations

- 5-CV:** five times for cross-validation
- AUC:** area under the receiver operating characteristic curve
- BD:** bronchodilator
- COPD:** chronic obstructive pulmonary disease
- EOS:** eosinophil count
- FEV1:** forced expiratory volume in 1 second
- FVC:** forced vital capacity
- Hb:** hemoglobin

**HT:** hematocrit

**MCHC:** mean corpuscular hemoglobin concentration

**MCV:** mean corpuscular volume

**WBC:** white blood cell

**XGBoost:** Gradient Boosting Decision Tree machine learning method

*Edited by G Eysenbach; submitted 05.10.20; peer-reviewed by C Gandhi; comments to author 28.10.20; revised version received 17.11.20; accepted 11.04.21; published 06.07.21.*

*Please cite as:*

*Muro S, Ishida M, Horie Y, Takeuchi W, Nakagawa S, Ban H, Nakagawa T, Kitamura T*

*Machine Learning Methods for the Diagnosis of Chronic Obstructive Pulmonary Disease in Healthy Subjects: Retrospective Observational Cohort Study*

*JMIR Med Inform 2021;9(7):e24796*

*URL: <https://medinform.jmir.org/2021/7/e24796>*

*doi: [10.2196/24796](https://doi.org/10.2196/24796)*

*PMID: [34255684](https://pubmed.ncbi.nlm.nih.gov/34255684/)*

©Shigeo Muro, Masato Ishida, Yoshiharu Horie, Wataru Takeuchi, Shunki Nakagawa, Hideyuki Ban, Tohru Nakagawa, Tetsuhisa Kitamura. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 06.07.2021. 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

# Ambulatory Risk Models for the Long-Term Prevention of Sepsis: Retrospective Study

Jewel Y Lee<sup>1</sup>, MSc; Sevda Molani<sup>1</sup>, PhD; Chen Fang<sup>1</sup>, PhD; Kathleen Jade<sup>1</sup>, ND; D Shane O'Mahony<sup>2</sup>, MD; Sergey A Kornilov<sup>1</sup>, PhD; Lindsay T Mico<sup>3</sup>, MSc; Jennifer J Hadlock<sup>1</sup>, MD

<sup>1</sup>Institute for Systems Biology, Seattle, WA, United States

<sup>2</sup>Swedish Center for Research and Innovation, Swedish Medical Center, Seattle, WA, United States

<sup>3</sup>Providence St Joseph Health, Renton, WA, United States

**Corresponding Author:**

Jennifer J Hadlock, MD

Institute for Systems Biology

401 Terry Ave N

Seattle, WA, 98109

United States

Email: [jhadlock@isbscience.org](mailto:jhadlock@isbscience.org)

## Abstract

**Background:** Sepsis is a life-threatening condition that can rapidly lead to organ damage and death. Existing risk scores predict outcomes for patients who have already become acutely ill.

**Objective:** We aimed to develop a model for identifying patients at risk of getting sepsis within 2 years in order to support the reduction of sepsis morbidity and mortality.

**Methods:** Machine learning was applied to 2,683,049 electronic health records (EHRs) with over 64 million encounters across five states to develop models for predicting a patient's risk of getting sepsis within 2 years. Features were selected to be easily obtainable from a patient's chart in real time during ambulatory encounters.

**Results:** The models showed consistent prediction scores, with the highest area under the receiver operating characteristic curve of 0.82 and a positive likelihood ratio of 2.9 achieved with gradient boosting on all features combined. Predictive features included age, sex, ethnicity, average ambulatory heart rate, standard deviation of BMI, and the number of prior medical conditions and procedures. The findings identified both known and potential new risk factors for long-term sepsis. Model variations also illustrated trade-offs between incrementally higher accuracy, implementability, and interpretability.

**Conclusions:** Accurate implementable models were developed to predict the 2-year risk of sepsis, using EHR data that is easy to obtain from ambulatory encounters. These results help advance the understanding of sepsis and provide a foundation for future trials of risk-informed preventive care.

(*JMIR Med Inform* 2021;9(7):e29986) doi:[10.2196/29986](https://doi.org/10.2196/29986)

**KEYWORDS**

sepsis; machine learning; electronic health records; risk prediction; clinical decision making; prevention; risk factors

## Introduction

Sepsis is a life-threatening condition characterized by a systemic immunological response to infection. Each year, more than 1.7 million adults in the United States develop sepsis, and nearly 16% of them die [1]. It is the leading cause of death in hospitals worldwide and puts a huge burden on health care systems [2-4]. Research to date has primarily focused on the inpatient setting, where timely treatment can improve sepsis-associated mortality and morbidity [5-9]. Commonly used risk scores, such as the systemic inflammatory response syndrome (SIRS) score [10],

quick sequential organ failure assessment (qSOFA) score [11], and modified early warning score (MEWS) [12], offer benefit once patients are acutely ill, but are less useful for early detection [13-16]. Advanced machine learning has led to more efficient models based on data from larger populations and a greater number of risk factors [17-21], but these are designed for emergency and inpatient settings [21-27].

Better risk models are needed to support community-acquired sepsis prevention. In 2016, Wang et al were the first to develop a risk score for long-term sepsis [28]. Using the REGARDS cohort (n=30,239), they predicted an individual's 10-year risk

of sepsis (REGARD SRS), with a bootstrapped C index of 0.703. The REGARD SRS and SSRS rely on demographic and medical history features that could be obtained by patient self-report, but they also depend on clinical laboratory results from blood and urine, including laboratory tests, such as cystatin-C and high-sensitivity C-reactive protein, which are not routinely measured in community-dwelling patients. Thus, there is a pressing need for a noninvasive solution to guide interventions for preventing sepsis, including immunization, education on infection prevention, and early symptom recognition [29,30]. Published guidelines currently recommend these interventions for some patients, such as those who will be experiencing neutropenia secondary to chemotherapy or posttransplant immunosuppression [31,32], but many other patients at high risk are overlooked. An implementable model that works on real-world patient data could support risk stratification for population health outreach or at the point of care.

Given the increased adoption of electronic health records (EHRs) in ambulatory care [33], a wealth of longitudinal phenotype and exposure data is now accessible to support predictive analytics. Sepsis risk research can move beyond inpatient encounters toward investigation of long-term patient trajectories. Historical data can support more accurate models for clinical decision support and improved resource stewardship. Yet, accuracy is only one dimension of model quality. Two other considerations are implementability in real-world settings and biomedical relevance for discovery of new hypotheses about the mechanisms of disease, prevention, and treatment.

In this study, we developed EHR-based models using supervised machine learning methods to predict the long-term risk of sepsis, investigating both time-invariant and temporal synopsis features. For each model, we reported results for both performance and feature importance, and discussed trade-offs between accuracy, interpretability, implementability, and biomedical relevance. This research investigated the potential to predict long-term sepsis risk in ways that can inform clinical decisions and lead to a better understanding of the disease.

## Methods

### Data and Study Setting

Providence St. Joseph Health (PSJH) is a community health system that includes over 51 hospitals and 1085 clinics. This retrospective study used clinical data from PSJH EHRs for patients who presented for health care at Providence, Swedish, or Kadlec sites in Alaska, California, Montana, Oregon, and Washington. Research was conducted within a Health Insurance Portability and Accountability Act (HIPAA)-secure data

platform, after date shifting had been applied to reduce the risk of reidentification. Dates were shifted using a randomly selected offset per patient of up to  $\pm 365$  days. All time windows below were defined on postshifted dates. Procedures were approved by the Institutional Review Board (IRB) at PSJH (IRB Study Number STUDY2019000389). Records were included for patients who presented for health care at least one time between 2017 and 2019. Our prediction model used records from patients over 18 years of age during a 3-year observation window starting in 2014 to predict sepsis in a 2-year window, starting in 2017. Patient age was calculated for the prediction window start date. Patients with no valid birth date or no encounters prior to 2014 were excluded. Our final study cohort consisted of 2,683,049 patients, including 1,558,851 (58.1%) women and 1,124,198 (41.9%) men, and the median age was 51.36 years. Over 64,000,000 encounters were collected from the cohort patients for feature extraction.

### Feature and Label Extraction

Features represent information about the data used as model inputs, and the label is the outcome that the model is trained to predict. In this study, we selected features that can be easily obtained from EHRs, including previously reported long-term risk factors for sepsis [34] and potential risk factors for investigation. Binary outcome variables were used in labeling for classification (1 for sepsis and 0 for no sepsis). Sepsis was defined using the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [35] hierarchical terminology system. The label was set to 1 if the parent concept for sepsis, SNOMED CT identifier (SCTID = 91302008), or any of its descendants was found in the problem list during the prediction window.

The following features were extracted from the observation window: sex, age, ethnicity, race, height, weight, BMI, ambulatory vital signs, history of medical conditions, hospital length of stay, encounters, problem list entries, medical history entries, medication orders, and procedures. Medical conditions were considered present if the SNOMED CT parent concept or any of its descendant concepts were found in the problem list during the observation window. The sepsis feature was included to investigate whether having a history of sepsis is a risk factor for developing sepsis in the future. Ratio features with repeated observations (eg, BMI, vital signs, and hospital length of stay) were transformed through statistical aggregation (minimum, maximum, mean, and standard deviation). All features are defined in Table 1 and categorized into four feature sets as follows: basic, vital signs, medical history, and health care delivery data. In total, 49 features were entered into the supervised machine learning process.



**Table 1.** Definitions of features used for models in the study for the observation window.

Category	Definition
<b>Basic features</b>	
Sex	Male (1), female (0), missing (-1)
Age	Age calculated at the start of the prediction window
Race	Native Hawaiian/Pacific Islander, American Indian/Alaska Native, Asian, Black/African American (1); White (0); other/missing (-1)
Ethnicity	Hispanic/Latino (1), not Hispanic/Latino (0), missing (-1)
Height	Last observed height
Weight	Last observed weight
Std_BMI	Standard deviation of BMI
<b>Vital sign features</b>	
BP_sys	Average and standard deviation of systolic blood pressure
BP_dia	Average and standard deviation of diastolic blood pressure
BT	Average and standard deviation of body temperature
HR	Average and standard deviation of heart rate
RR	Average and standard deviation of respiratory rate
<b>Medical history features</b>	
Sepsis	Sepsis (SCTID <sup>a</sup> 91302008)
Pneumonia	Pneumonia (SCTID 233604007)
Bacterial infection	Bacterial infectious disease (SCTID 87628006)
Fungal infection	Mycosis (SCTID 3218000)
Protein-energy malnutrition	Deficiency of macronutrients (SCTID 238107002)
Cancer	Malignant neoplastic disease (SCTID 363346000)
COPD <sup>b</sup>	Chronic obstructive lung disease (SCTID 13645005)
Diabetes	Diabetes mellitus (SCTID 73211009)
Chronic kidney disease	Chronic kidney disease (SCTID 709044004)
Hypertension	Hypertensive disorder, systemic arterial (SCTID 38341003)
Deep vein thrombosis	Deep venous thrombosis (SCTID 128053003)
Arteriosclerosis	Arteriosclerotic vascular disease (SCTID 72092001)
Peripheral artery disease	Peripheral arterial occlusive disease (SCTID 399957001)
Coronary artery disease	Coronary arteriosclerosis (SCTID 53741008)
Heart attack	Myocardial infarction (SCTID 22298006)
Atrial fibrillation	Atrial fibrillation (SCTID 49436004)
Stroke	Cerebrovascular accident (SCTID 230690007)
Heart failure	Heart failure (SCTID 84114007)
<b>Health care delivery features</b>	
n_encounter	Total count of clinical encounters
n_hospitalization	Total count of hospitalizations
LOS	Average, minimum, maximum, and standard deviation of length of hospital stay
n_problem	Total count of problem list entries
u_problem	Number of unique problem list entries
n_medical_hx	Total count of medical history entries
u_medical_hx	Number of unique medical history entries

Category	Definition
n_medication	Total count of prescription medication orders
u_medication	Number of unique prescription medication orders
n_procedure	Total count of ordered medical procedures
u_procedure	Number of unique ordered medical procedures

<sup>a</sup>SCTID: Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) identifier.

<sup>b</sup>COPD: chronic obstructive pulmonary disease.

### Machine Learning

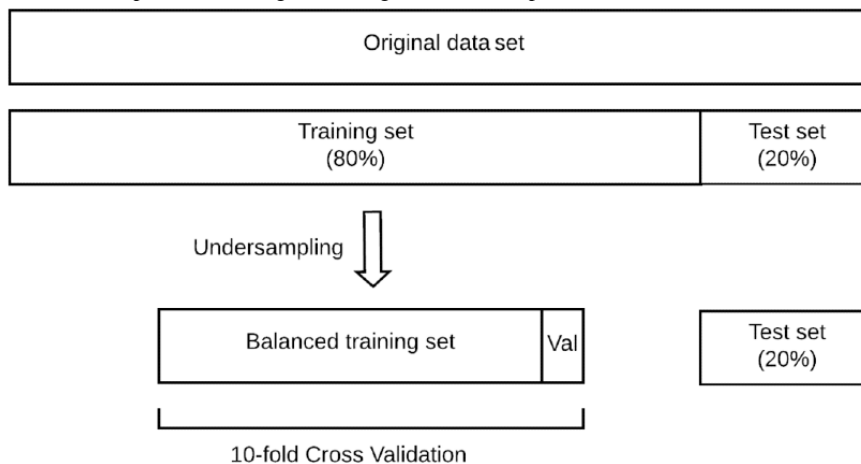
Data preprocessing and cleaning were conducted as follows. Missing data in categorical features (sex, race, and ethnicity) were assigned to be -1. Missing data in height, weight, and vital signs were imputed using the carry-forward method if previous observations were available; otherwise, median imputation was used. Outliers in height and weight were detected by calculating the modified z-score based on median absolute deviation (MAD) [36] in equation 1 with a threshold of 3.5. Both outliers and missing data were imputed with the median. Equation 1 is as follows:

$$M_i = 0.6745 (x_i - \tilde{x}) / MAD \quad (1)$$

where MAD is the median absolute deviation and  $\tilde{x}$  is the median of  $x$ .

Patients diagnosed with sepsis accounted for only about 0.8% of the cohort, leading to extremely imbalanced data. To ensure the validity of the model but, at the same time, overcome the class imbalance in the medical data set, we reserved 20% of the original data as a test set and undersampled the other 80% of the data by randomly selecting the same number of patients from the majority class (no sepsis) as the minority class (sepsis) to construct a balanced training set. The train/test split process is shown in Figure 1. This training set was then trained with several machine learning methods, including gradient boosting (GB), support vector machine (SVM), and logistic regression (LR), and validated with 10-fold cross validation. Four models were constructed with different combinations of feature sets. Model 1 used only the basic features. Sequentially, we added vital sign features to model 2, medical history features to model 3, and health care delivery data features to model 4.

**Figure 1.** Training, validation, and test split for modeling of the long-term risk of sepsis.



### Model Performance Evaluation

All classification models were built using scikit-learn [37], an open-source Python machine learning library. Widely adopted performance measures, such as area under the receiver operating characteristic curve (AUROC), precision, sensitivity (or recall), specificity, and likelihood ratio, were used to evaluate the discrimination ability of our prediction models. Appropriate measures were selected based on the class distribution in the models. We also analyzed relative feature importance using the following three methods: (1) Shapley Additive exPlanations (SHAP) algorithm, (2) permutation testing, and (3) model coefficients from L1-regularized logistic regression (L1-LR). SHAP, an algorithm developed from coalition game theory, calculates the average marginal contribution of a feature across all possible coalitions [38]. Permutation testing estimates feature

importance by calculating the drop in the performance after permuting the feature. A feature is considered important if shuffling its values increases the model prediction error. Shapley values and permutation feature importance computed on test data avoid the systematic bias in feature selection found with mean decrease impurity-based measures [39]. We also retrieved coefficients from L1-LR to investigate the relevance and directionality of features. LR with L1 regularization is a sparse linear model in which coefficients for unimportant features are reduced to zero [40], and the sign of the coefficient suggests positive or negative association with the model outcome (sepsis) [41].

## Results

Table 2 shows the results of 10-fold cross-validation based on training data using GB, SVM, and LR. The results show a consistent trend of model performance, increasing as more features were added. GB slightly outperformed linear classifiers (SVM and LR) in all four models. The best AUROC of 0.8216 was achieved by model 4. The trained GB models were then used to make predictions on the 20% test data set, and they were

evaluated with precision, sensitivity (or recall), specificity, positive and negative likelihood ratios, and diagnostic odds ratios because of the highly imbalanced class distribution (Table 3). The test set prevalence was 0.0079 with the population size of 536610. The results showed that the positive likelihood ratio ranged from 2.1135 to 2.8897, and the negative likelihood ratio ranged from 0.3192 to 0.4997. Sensitivity and specificity in each model had similar results in the training set and test set for predicting the sepsis outcome.

**Table 2.** Ten-fold cross-validation results on the training set.

Model and classifier	Precision	Sensitivity	Specificity	AUROC <sup>a</sup>	Ten-fold error (%)
<b>Model 1 (basic)</b>					
GB <sup>b</sup>	0.6727	0.6725	0.6725	0.7349	0.29%
SVM <sup>c</sup>	0.6607	0.6606	0.6606	0.7167	0.27%
LR <sup>d</sup>	0.6569	0.6565	0.6565	0.7134	0.29%
<b>Model 2 (basic + VS<sup>e</sup>)</b>					
GB	0.6947	0.6946	0.6946	0.7595	0.28%
SVM	0.6812	0.6811	0.6811	0.7425	0.29%
LR	0.6776	0.6775	0.6775	0.7399	0.26%
<b>Model 3 (basic + VS + MHX<sup>f</sup>)</b>					
GB	0.7008	0.7006	0.7006	0.7671	0.20%
SVM	0.6897	0.6868	0.6868	0.7502	0.17%
LR	0.6893	0.6891	0.6891	0.7523	0.18%
<b>Model 4 (basic + VS + MHX + HCD<sup>g</sup>)</b>					
GB	0.7483	0.7481	0.7481	0.8216	0.27%
SVM	0.7191	0.7169	0.7169	0.7910	0.26%
LR	0.7185	0.7175	0.7175	0.7835	0.19%

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

<sup>b</sup>GB: gradient boosting.

<sup>c</sup>SVM: support vector machine.

<sup>d</sup>LR: logistic regression.

<sup>e</sup>VS: vital signs.

<sup>f</sup>MHX: medical history.

<sup>g</sup>HCD: health care delivery data.

**Table 3.** Prediction results and 95% confidence intervals for the test set using the trained gradient boosting model.

Model	Precision, value (95% CI)	Sensitivity, value (95% CI)	Specificity, value (95% CI)	LR <sup>a</sup> , value (95% CI)	LR <sup>b</sup> , value (95% CI)	DOR <sup>c</sup>
Model 1 (basic)	0.0165 (0.0159-0.0171)	0.6552 (0.6407-0.6694)	0.6900 (0.6887-0.6912)	2.1135 (2.0670-2.1611)	0.4997 (0.4793-0.5209)	4
Model 2 (basic + VS <sup>d</sup> )	0.0177 (0.0171-0.0184)	0.6862 (0.6721-0.7001)	0.6980 (0.6968-0.6993)	2.2724 (2.2256-2.3202)	0.4495 (0.4299-0.4701)	5
Model 3 (basic + VS + MHX <sup>e</sup> )	0.0184 (0.0177-0.0190)	0.6874 (0.6733-0.7012)	0.7084 (0.7071-0.7096)	2.3570 (2.3086-2.4065)	0.4413 (0.4220-0.4615)	5
Model 4 (basic + VS + MHX + HCD <sup>f</sup> )	0.0224 (0.0217-0.0231)	0.7653 (0.7523-0.7779)	0.7352 (0.7340-0.7363)	2.8897 (2.8401-2.9401)	0.3192 (0.3023-0.3371)	9

<sup>a</sup>LR+: positive likelihood ratio.

<sup>b</sup>LR-: negative likelihood ratio.

<sup>c</sup>DOR: diagnostic odds ratio.

<sup>d</sup>VS: vital signs.

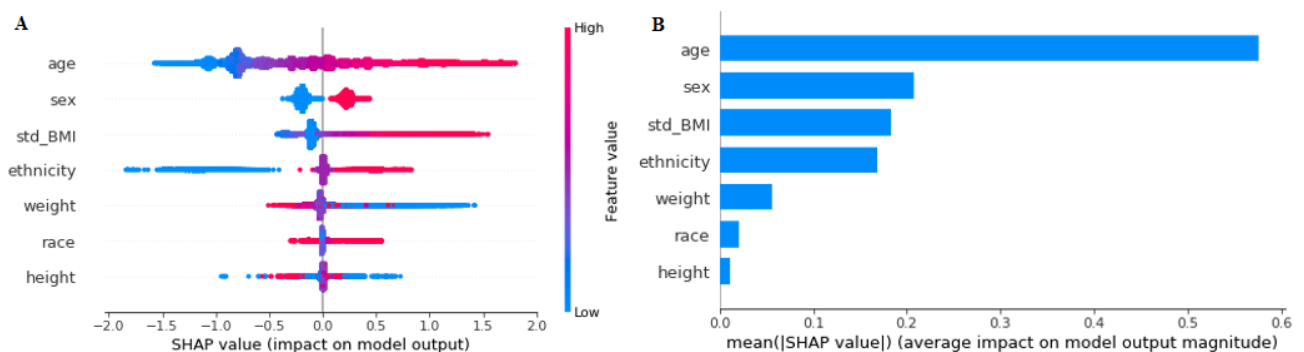
<sup>e</sup>MHX: medical history.

<sup>f</sup>HCD: health care delivery data.

To ensure the stability and reliability of the model, SHAP and permutation testing methods were implemented on the GB model. These methods improve the interpretability of the black box model and give a reasonable explanation for the prediction of each outcome. The results for the SHAP algorithm are shown in Figures 2-5. In addition, L1-LR and permutation results for model 4 are presented in Figure 6. In models 1-3, where health care delivery data features were not used, SHAP showed age as the dominant feature for predicting sepsis. Other important features included sex, ethnicity, respiratory rate, heart rate, standard deviation of BMI, history of sepsis, diabetes, and chronic kidney disease. In model 4, where health care delivery data features were added, the most predictive features were the

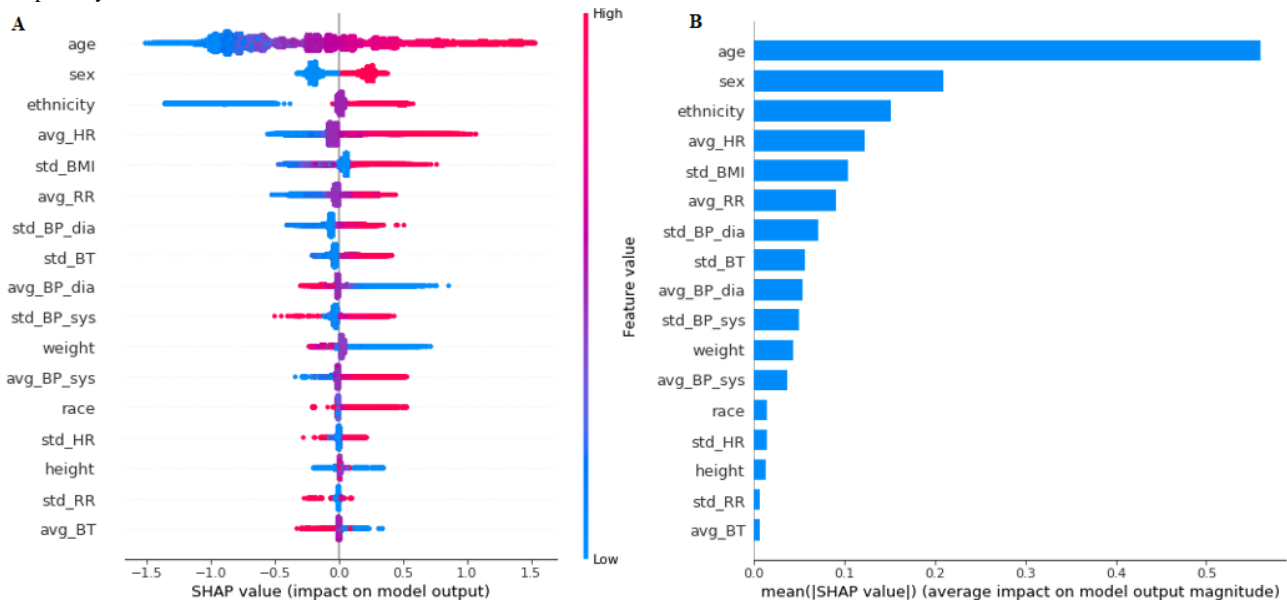
number of unique entries (u\_medical\_hx), followed by age, the total count of medical history entries (n\_medical\_hx), the total count of encounters (n\_encounter), sex, and the total count of ordered medical procedures (n\_procedure). The important features identified in the SHAP algorithm have high permutation importance and high absolute values of coefficients learned by L1-LR models. The sign of the coefficients showed the directionality of those features. Moreover, the average diastolic blood pressure (avg\_BP\_dia) and the total count of encounters (n\_encounter) were assigned with a negative coefficient in all three models, which implied the effect of high values for these features in decreasing the risk of developing sepsis.

**Figure 2.** The Shapley Additive exPlanations (SHAP) algorithm results for long-term sepsis risk in model 1. (A) The influence of higher and lower values of the feature on the patient's outcome. The left side of this graph represents reduced risk of developing sepsis, and the right side of the graph represents increased risk of developing sepsis. Red dots represent higher values of the feature, and blue dots represent lower values of the feature. Nominal classes are binary (0,1). (B) The ranking of feature importance indicated by SHAP.

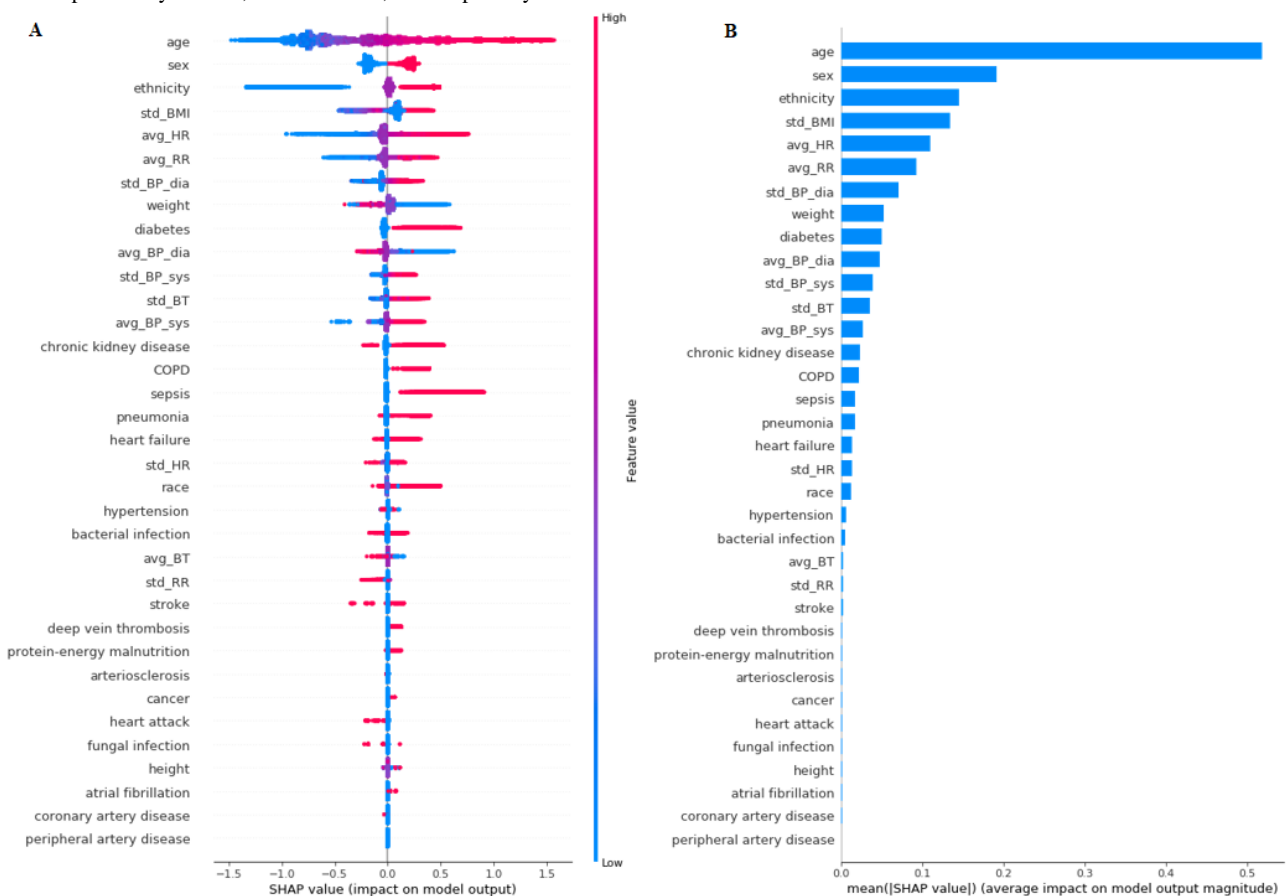




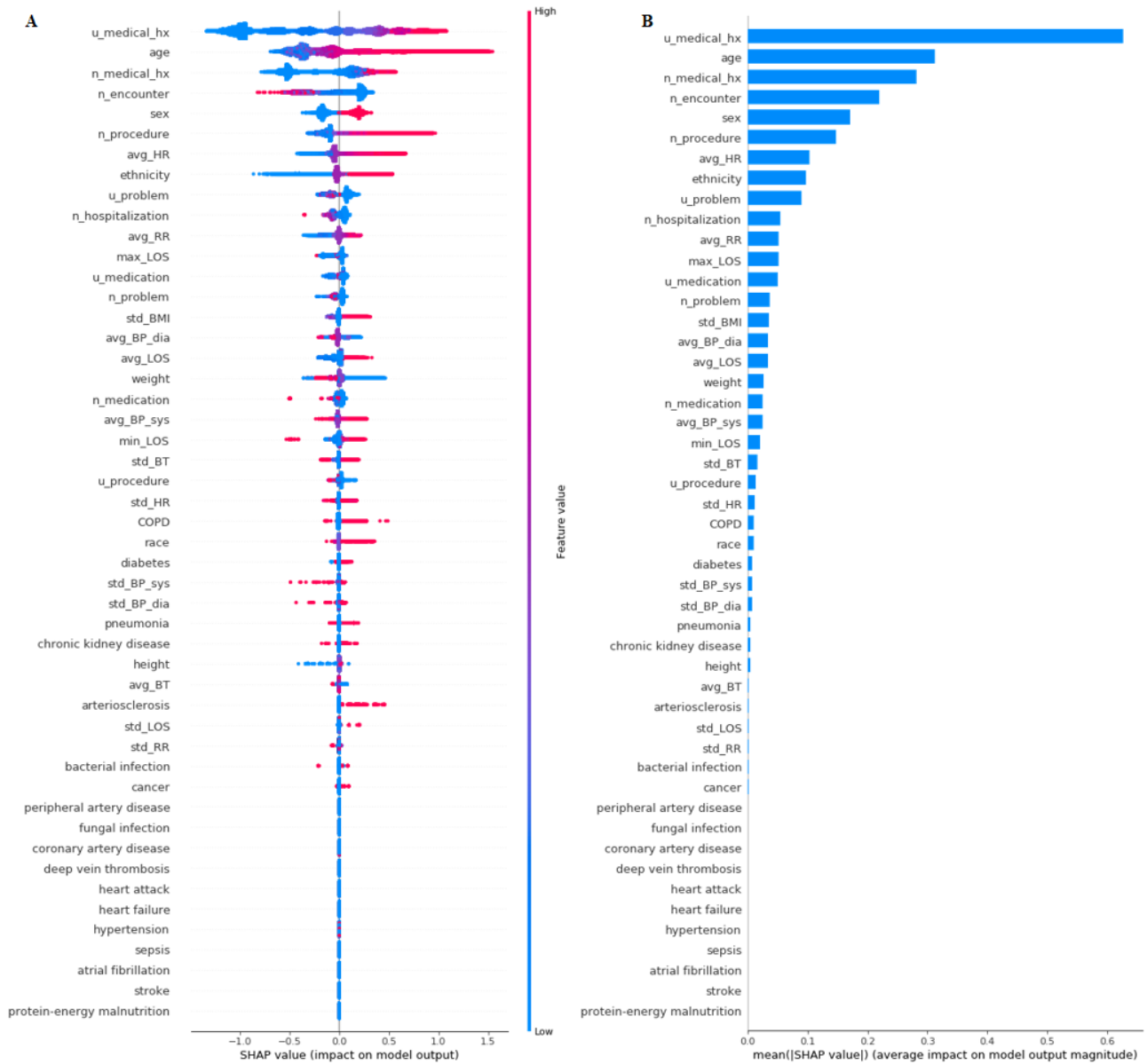
**Figure 3.** The Shapley Additive exPlanations (SHAP) algorithm results for long-term sepsis risk in model 2. (A) The influence of higher and lower values of the feature on the patient's outcome. The left side of this graph represents reduced risk of developing sepsis, and the right side of the graph represents increased risk of developing sepsis. Red dots represent higher values of the feature, and blue dots represent lower values of the feature. Nominal classes are binary (0,1). (B) The ranking of feature importance indicated by SHAP. BP: blood pressure; BT: body temperature; HR: heart rate; RR: respiratory rate.



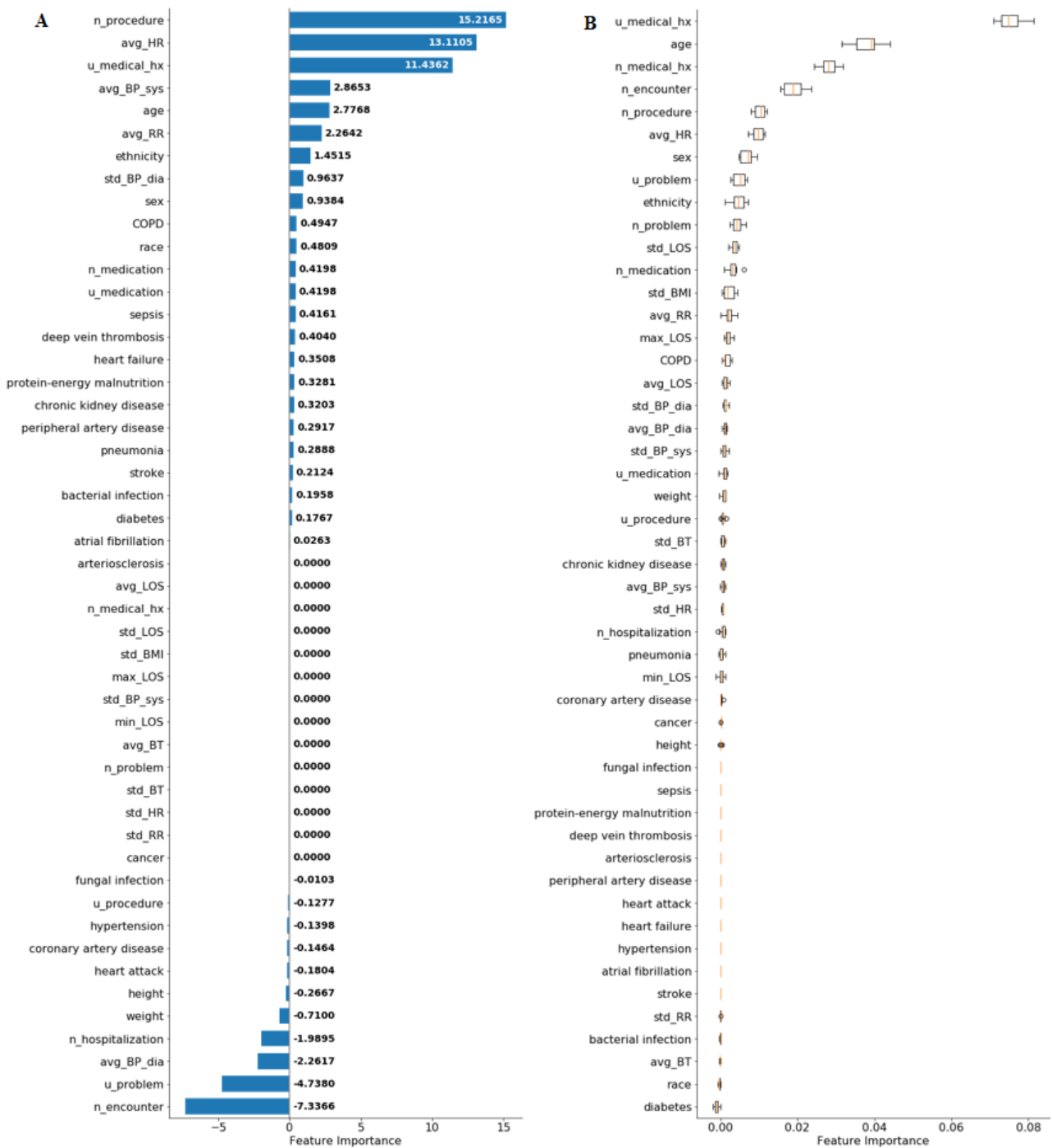
**Figure 4.** The Shapley Additive exPlanations (SHAP) algorithm results for long-term sepsis risk in model 3. (A) The influence of higher and lower values of the feature on the patient's outcome. The left side of this graph represents reduced risk of developing sepsis, and the right side of the graph represents increased risk of developing sepsis. Red dots represent higher values of the feature, and blue dots represent lower values of the feature. Nominal classes are binary (0,1). (B) The ranking of feature importance indicated by SHAP. BP: blood pressure; BT: body temperature; COPD: chronic obstructive pulmonary disease; HR: heart rate; RR: respiratory rate.



**Figure 5.** The Shapley Additive exPlanations (SHAP) algorithm results for long-term sepsis risk in model 4. (A) The influence of higher and lower values of the feature on the patient's outcome. The left side of this graph represents reduced risk of developing sepsis, and the right side of the graph represents increased risk of developing sepsis. Red dots represent higher values of the feature, and blue dots represent lower values of the feature. Nominal classes are binary (0,1). (B) The ranking of feature importance indicated by SHAP. BP: blood pressure; BT: body temperature; COPD: chronic obstructive pulmonary disease; HR: heart rate; LOS: length of hospital stay; RR: respiratory rate.



**Figure 6.** The L1-regularized logistic regression (L1-LR) algorithm results (A) and permutation testing results for long-term sepsis risk in model 4 (B). BP: blood pressure; BT: body temperature; COPD: chronic obstructive pulmonary disease; HR: heart rate; LOS: length of hospital stay; RR: respiratory rate.



## Discussion

### Principal Findings

In this study, we constructed four interpretable implementable EHR-based models to predict the 2-year risk of sepsis in adults. Each model performed well, considering the complexity of the features included. As expected, model 4 with all 49 features outperformed the others, with an AUROC of 0.8216 achieved by the GB algorithm in the training set. Due to the low prevalence of sepsis outcomes in the 20% test set, the precision was low in all models. However, the positive likelihood ratio

of 2.8897 and negative likelihood ratio of 0.3192 achieved by model 4 showed that our model has the ability to identify patients with higher risk of sepsis. The dominant features in this model, accounting for more than half of the feature importance, were the numbers of unique and total medical history entries (u\_medical\_hx and n\_medical\_hx), and age. Medical history features suggest an increased burden of underlying health conditions, and aging is the most substantial risk factor for multimorbidity [42]. Comorbidities are known to be significantly higher in patients with sepsis compared to those without sepsis [1,43], but previous models have not included multimorbidity as a distinct feature. Another strong

predictor in model 4 was the total number of ordered medical procedures ( $n_{\text{procedure}}$ ). Procedures, particularly those that are invasive, increase the risk of hospital-acquired infections, and may also be indicative of health status and multimorbidity. The total number of encounters ( $n_{\text{encounter}}$ ), which was assigned a negative coefficient in L1-LR, was also a strong predictor in model 4. Although it requires further investigation, one possible reason could be that a greater number of health care visits is associated with better access to preventative health care.

Age, ethnicity, sex, average heart rate ( $\text{avg\_HR}$ ), and standard deviation of BMI ( $\text{std\_BMI}$ ) were the most important features in models 2 and 3. In addition to increasing the risk of multimorbidity, age is a known independent risk factor for sepsis incidence, severity, and outcomes [44]. Whether ethnicity represents a sepsis risk factor is not yet established. Results from epidemiological studies are contrasting [45-47]. Ethnicity may also be associated with socioeconomic status, a health determinant recently found to be associated with a higher rate of hospital admissions for infection [48]. Future tracking of health-related social needs in structured EHR data [49] will support deeper investigation. A higher resting heart rate, which is common in infection, is also a risk factor for all-cause mortality [50] and may suggest a poorer health status. Patients with higher average heart rates may have had infections during previous encounters. Obesity and malnourishment are known risk factors for sepsis [51], but the standard deviation of BMI (change over time) is a new potential risk factor and merits investigation. In models 3 and 4, basic features and vital signs (age, ethnicity, sex, BMI, and heart rate) appeared to be more stronger predictors than well-established medical conditions known to be sepsis risk factors, including heart failure [52], chronic kidney disease [53], chronic obstructive pulmonary disease (COPD) [54], and diabetes [55,56]. Taken together, these findings suggest the possibility that sepsis risk is associated with not only age and medical conditions, but also vital signs and features related to health care delivery.

Although the highest performance was achieved with the health care delivery data features set, it has limited usefulness for discovering potential risk factors given its reliance on aggregated features, such as the number of medical history entries. Inclusion of these aggregated features weakens other predictors that are potentially more biomedically informative, including medical conditions and biomarkers, such as vital signs. The second-best performing model (model 3) identified a subset of biomarkers as strong predictors, including the standard deviation of BMI and average resting heart rate.

In models 3 and 4 that incorporated medical history, the conditions with greater importance for long-term sepsis risk were history of sepsis, heart failure, chronic kidney disease, pneumonia, COPD, and diabetes. In contrast, the most impactful chronic diseases in the REGARDS 10-year prediction score were chronic lung disease, followed by diabetes and peripheral artery disease [28,34]. The difference in risk factors between REGARDS and our models may reflect a different population sample and prediction window, but could also reflect differing definitions for conditions. For example, Wang et al used laboratory markers (estimated glomerular filtration rate, urinary

albumin-to-creatinine ratio, and cystatin-C) for chronic kidney disease [28]. We selected diagnostic codes, which are less precise, but more likely to be consistently implementable on EHR data. SNOMED CT was selected because it is a medically curated semantic ontology, which is structured as a directed acyclic graph and used in EHRs across many countries. These codes can be mapped to ICD-10 codes, but different health care systems would likely benefit from retraining and retesting the model for their specific population.

The primary goal of this study was to investigate whether readily available EHR data can predict the long-term risk of developing sepsis during ambulatory visits in real time. Performance could also be useful for assessing population health. Interpretability was a secondary concern, and the feature importance estimates discussed above should be taken as exploratory. Relationships identified in the models reflected shared information content, but not necessarily biomedical relevance or causality. However, feature importance models suggested new insights on potential risk factors for sepsis that merit further investigation.

### Limitations

The studied population may have sample bias toward patients with continuous care within one health care system. There are also many common issues with structured EHR data that hamper the extraction of accurate information, including missing data, erroneous data, differences in EHR conventions among providers, and changes in how data are stored in EHRs over time [57]. These were only partially offset by terminology mapping, data removal, or imputation.

Using EHR diagnostic codes to identify sepsis patients also has limitations. First, it may miss cases where patients had sepsis at a different health care system. Second, because there is no confirmatory diagnostic test for sepsis, this model included patients who were treated empirically for sepsis but might not have had it. Third, variations in sepsis diagnosis, documentation, and coding practices could lead to missing sepsis labels [58]. Fourth, it does not differentiate between severe and milder forms of sepsis, or between hospital-acquired and community-acquired sepsis [43].

Future models can take advantage of the Adult Sepsis Event surveillance definition optimized for EHRs, which was recently released by the CDC [1,59]. This criterion uses objective clinical data to identify severe sepsis in hospitalized patients and displays superior sensitivity to diagnostic codes [1]. Lastly, our definition of ambulatory vital signs may include those that were taken in urgent or emergency situations. This is valid for prediction on real-world EHR data, but future models could better distinguish urgent encounters from those that are more likely to represent outpatient baseline.

### Conclusions

Strategies for long-term sepsis risk prediction are needed to advance the understanding of the disease and guide efforts for prevention. We used retrospective EHR data from 2,683,049 adults across five US states to develop models for predicting adult patients' long-term risk of sepsis. Our models achieved a high AUROC and suggested new insights into potential long-term risk factors, including changes in BMI and a higher



mean heart rate in ambulatory settings. These models could be implemented at a low cost, requiring only information that is easy to obtain from EHRs in real time. Ambulatory patients at the highest risk for sepsis could benefit from personalized preventative approaches, including increased emphasis on

immunization, and education on reducing the risk of infection and recognizing early symptoms of sepsis. This implementable model provides a path toward clinical trials of risk-informed interventions for long-term sepsis prevention.

## Acknowledgments

This work was funded in part by the Washington Research Foundation. We thank Ryan T Roper and Venkata R Duvvuri for their design and implementation assistance for biomedical concept extraction. We are grateful to Providence St. Joseph Health for sharing their data, data engineering expertise, and computational resources. We appreciate the technical assistance of Mark Premo, Jennifer Jones, and Andrey Dubovoy. We would also like to acknowledge SNOMED International for developing and maintaining SNOMED CT.

## Conflicts of Interest

None declared.

## References

1. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, CDC Prevention Epicenter Program. Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014. *JAMA* 2017 Oct 03;318(13):1241-1249 [FREE Full text] [doi: [10.1001/jama.2017.13836](https://doi.org/10.1001/jama.2017.13836)] [Medline: [28903154](https://pubmed.ncbi.nlm.nih.gov/28903154/)]
2. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Crit Care Med* 2001 Jul;29(7):1303-1310. [doi: [10.1097/00003246-200107000-00002](https://doi.org/10.1097/00003246-200107000-00002)] [Medline: [11445675](https://pubmed.ncbi.nlm.nih.gov/11445675/)]
3. Novosad SA, Sapiano MR, Grigg C, Lake J, Robyn M, Dumyati G, et al. Vital Signs: Epidemiology of Sepsis: Prevalence of Health Care Factors and Opportunities for Prevention. *MMWR Morb Mortal Wkly Rep* 2016 Aug 26;65(33):864-869 [FREE Full text] [doi: [10.15585/mmwr.mm6533e1](https://doi.org/10.15585/mmwr.mm6533e1)] [Medline: [27559759](https://pubmed.ncbi.nlm.nih.gov/27559759/)]
4. Fleischmann C, Scherag A, Adhikari NKJ, Hartog CS, Tsaganos T, Schlattmann P, International Forum of Acute Care Trialists. Assessment of Global Incidence and Mortality of Hospital-treated Sepsis. Current Estimates and Limitations. *Am J Respir Crit Care Med* 2016 Feb 01;193(3):259-272. [doi: [10.1164/rccm.201504-0781OC](https://doi.org/10.1164/rccm.201504-0781OC)] [Medline: [26414292](https://pubmed.ncbi.nlm.nih.gov/26414292/)]
5. Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, et al. Early Goal-Directed Therapy in the Treatment of Severe Sepsis and Septic Shock. *N Engl J Med* 2001 Nov 08;345(19):1368-1377. [doi: [10.1056/nejmoa010307](https://doi.org/10.1056/nejmoa010307)]
6. Nguyen HB, Corbett SW, Steele R, Banta J, Clark RT, Hayes SR, et al. Implementation of a bundle of quality indicators for the early management of severe sepsis and septic shock is associated with decreased mortality\*. *Critical Care Medicine* 2007;35(4):1105-1112. [doi: [10.1097/01.ccm.0000259463.33848.3d](https://doi.org/10.1097/01.ccm.0000259463.33848.3d)]
7. Sebat F, Musthafa AA, Johnson D, Kramer AA, Shoffner D, Eliason M, et al. Effect of a rapid response system for patients in shock on time to treatment and mortality during 5 years\*. *Critical Care Medicine* 2007;35(11):2568-2575. [doi: [10.1097/01.ccm.0000287593.54658.89](https://doi.org/10.1097/01.ccm.0000287593.54658.89)]
8. Coba V, Whitmill M, Mooney R, Horst HM, Brandt M, Digiovine B, (The Henry Ford Hospital Sepsis Collaborative Group). Resuscitation bundle compliance in severe sepsis and septic shock: improves survival, is better late than never. *J Intensive Care Med* 2011 Jan 10;26(5):304-313. [doi: [10.1177/0885066610392499](https://doi.org/10.1177/0885066610392499)] [Medline: [21220270](https://pubmed.ncbi.nlm.nih.gov/21220270/)]
9. Dellinger RP, Levy MM, Rhodes A, Annane D, Gerlach H, Opal SM, Surviving Sepsis Campaign Guidelines Committee including the Pediatric Subgroup. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. *Crit Care Med* 2013 Feb;41(2):580-637. [doi: [10.1097/CCM.0b013e31827e83af](https://doi.org/10.1097/CCM.0b013e31827e83af)] [Medline: [23353941](https://pubmed.ncbi.nlm.nih.gov/23353941/)]
10. Bone R. Toward an epidemiology and natural history of SIRS (systemic inflammatory response syndrome). *JAMA* 1992;268(24):3452-3455. [Medline: [1460735](https://pubmed.ncbi.nlm.nih.gov/1460735/)]
11. Heim C, Newport DJ, Heit S, Graham YP, Wilcox M, Bonsall R, et al. Pituitary-adrenal and autonomic responses to stress in women after sexual and physical abuse in childhood. *JAMA* 2000 Aug 02;284(5):592-597. [doi: [10.1001/jama.284.5.592](https://doi.org/10.1001/jama.284.5.592)] [Medline: [10918705](https://pubmed.ncbi.nlm.nih.gov/10918705/)]
12. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. *QJM* 2001 Oct;94(10):521-526. [doi: [10.1093/qjmed/94.10.521](https://doi.org/10.1093/qjmed/94.10.521)] [Medline: [11588210](https://pubmed.ncbi.nlm.nih.gov/11588210/)]
13. Finkelsztejn EJ, Jones DS, Ma KC, Pabón MA, Delgado T, Nakahira K, et al. Comparison of qSOFA and SIRS for predicting adverse outcomes of patients with suspicion of sepsis outside the intensive care unit. *Crit Care* 2017 Mar 26;21(1):73 [FREE Full text] [doi: [10.1186/s13054-017-1658-5](https://doi.org/10.1186/s13054-017-1658-5)] [Medline: [28342442](https://pubmed.ncbi.nlm.nih.gov/28342442/)]
14. Rodriguez RM, Greenwood JC, Nuckton TJ, Darger B, Shofer FS, Troeger D, et al. Comparison of qSOFA with current emergency department tools for screening of patients with sepsis for critical illness. *Emerg Med J* 2018 Jun 02;35(6):350-356. [doi: [10.1136/emered-2017-207383](https://doi.org/10.1136/emered-2017-207383)] [Medline: [29720475](https://pubmed.ncbi.nlm.nih.gov/29720475/)]

15. van der Woude SW, van Doormaal FF, Hutten BA, J Nellen F, Holleman F. Classifying sepsis patients in the emergency department using SIRS, qSOFA or MEWS. *Neth J Med* 2018 May;76(4):158-166 [FREE Full text] [Medline: 29845938]
16. Khwannimit B, Bhurayanontachai R, Vattanavanit V. Comparison of the accuracy of three early warning scores with SOFA score for predicting mortality in adult sepsis and septic shock patients admitted to intensive care unit. *Heart Lung* 2019 May;48(3):240-244. [doi: 10.1016/j.hrtlng.2019.02.005] [Medline: 30902348]
17. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med* 2015 Aug 05;7(299):299ra122. [doi: 10.1126/scitranslmed.aab3719] [Medline: 26246167]
18. Calvert J, Desautels T, Chettipally U, Barton C, Hoffman J, Jay M, et al. High-performance detection and early prediction of septic shock for alcohol-use disorder patients. *Ann Med Surg (Lond)* 2016 Jun;8:50-55 [FREE Full text] [doi: 10.1016/j.amsu.2016.04.023] [Medline: 27489621]
19. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med Inform* 2016 Sep 30;4(3):e28 [FREE Full text] [doi: 10.2196/medinform.5909] [Medline: 27694098]
20. 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 6;12(4):e0174708 [FREE Full text] [doi: 10.1371/journal.pone.0174708] [Medline: 28384212]
21. Delahanty RJ, Alvarez J, Flynn LM, Sherwin RL, Jones SS. Development and Evaluation of a Machine Learning Model for the Early Identification of Patients at Risk for Sepsis. *Ann Emerg Med* 2019 Apr;73(4):334-344. [doi: 10.1016/j.annemergmed.2018.11.036] [Medline: 30661855]
22. Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, et al. A computational approach to early sepsis detection. *Comput Biol Med* 2016 Jul 01;74:69-73. [doi: 10.1016/j.combiomed.2016.05.003] [Medline: 27208704]
23. Rothman M, Levy M, Dellinger RP, Jones SL, Fogerty RL, Voelker KG, et al. Sepsis as 2 problems: Identifying sepsis at admission and predicting onset in the hospital using an electronic medical record-based acuity score. *J Crit Care* 2017 Apr;38:237-244 [FREE Full text] [doi: 10.1016/j.jcrc.2016.11.037] [Medline: 27992851]
24. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med* 2018 Nov 22;24(11):1716-1720. [doi: 10.1038/s41591-018-0213-5] [Medline: 30349085]
25. Islam MM, Nasrin T, Walther BA, Wu C, Yang H, Li Y. Prediction of sepsis patients using machine learning approach: A meta-analysis. *Comput Methods Programs Biomed* 2019 Mar;170:1-9. [doi: 10.1016/j.cmpb.2018.12.027] [Medline: 30712598]
26. Calvert J, Saber N, Hoffman J, Das R. Machine-Learning-Based Laboratory Developed Test for the Diagnosis of Sepsis in High-Risk Patients. *Diagnostics (Basel)* 2019 Feb 13;9(1):20 [FREE Full text] [doi: 10.3390/diagnostics9010020] [Medline: 30781800]
27. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med* 2020 Mar 21;46(3):383-400 [FREE Full text] [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]
28. Wang HE, Donnelly JP, Griffin R, Levitan EB, Shapiro NI, Howard G, et al. Derivation of Novel Risk Prediction Scores for Community-Acquired Sepsis and Severe Sepsis\*. *Critical Care Medicine* 2016;44(7):1285-1294. [doi: 10.1097/ccm.0000000000001666]
29. Choy K, Agcaoili C, Halimi K. Impact of community-based education on sepsis. *Crit Care* 2009;13(Suppl 4):P42. [doi: 10.1186/cc8098]
30. Kempker JA, Wang HE, Martin GS. Sepsis is a preventable public health problem. *Crit Care* 2018 May 06;22(1):116 [FREE Full text] [doi: 10.1186/s13054-018-2048-3] [Medline: 29729670]
31. Taplitz RA, Kennedy EB, Bow EJ, Crews J, Gleason C, Hawley DK, et al. Outpatient Management of Fever and Neutropenia in Adults Treated for Malignancy: American Society of Clinical Oncology and Infectious Diseases Society of America Clinical Practice Guideline Update. *JCO* 2018 May 10;36(14):1443-1453. [doi: 10.1200/jco.2017.77.6211]
32. Avery RK, Michaels MG, AST Infectious Diseases Community of Practice. Strategies for safe living following solid organ transplantation-Guidelines from the American Society of Transplantation Infectious Diseases Community of Practice. *Clin Transplant* 2019 Sep 06;33(9):e13519. [doi: 10.1111/ctr.13519] [Medline: 30844096]
33. Office-based Physician Electronic Health Record Adoption. Health IT Dashboard. URL: <https://dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php> [accessed 2020-12-05]
34. Wang H, Donnelly J, Yende S, Levitan E, Shapiro N, Dai Y, et al. Validation of the REGARDS Severe Sepsis Risk Score. *J Clin Med* 2018 Dec 11;7(12):536 [FREE Full text] [doi: 10.3390/jcm7120536] [Medline: 30544923]
35. SNOMED International. URL: <https://www.snomed.org/> [accessed 2020-12-03]
36. Crosby T. How to Detect and Handle Outliers. *Technometrics* 1994 Aug;36(3):315-316. [doi: 10.1080/00401706.1994.10485810]
37. 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(85):2825-2830 [FREE Full text]

38. 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 17;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/)]
39. Breiman L, Last M, Rice J. Random forests: finding quasars. In: *Statistical Challenges in Astronomy*. New York, NY: Springer; 2006:243-254.
40. Lee SI, Lee H, Abbeel P, Ng AY. Efficient L1 Regularized Logistic Regression. *AAAI*. URL: <https://www.aaai.org/Papers/AAAI/2006/AAAI06-064.pdf> [accessed 2021-06-20]
41. Fonti V, Belitser E. Feature Selection using LASSO. *VU Amsterdam*. 2017. URL: [https://beta.vu.nl/nl/Images/werkstuk-fonti\\_tcm235-836234.pdf](https://beta.vu.nl/nl/Images/werkstuk-fonti_tcm235-836234.pdf) [accessed 2021-06-20]
42. Navickas R, Petric V, Feigl AB, Seychell M. Multimorbidity: What do we know? What should we do? *J Comorb* 2016 Feb 17;6(1):4-11 [FREE Full text] [doi: [10.15256/joc.2016.6.72](https://doi.org/10.15256/joc.2016.6.72)] [Medline: [29090166](https://pubmed.ncbi.nlm.nih.gov/29090166/)]
43. Rhee C, Wang R, Zhang Z, Fram D, Kadri SS, Klompas M. Epidemiology of Hospital-Onset Versus Community-Onset Sepsis in U.S. Hospitals and Association With Mortality. *Critical Care Medicine* 2019;47(9):1169-1176. [doi: [10.1097/ccm.0000000000003817](https://doi.org/10.1097/ccm.0000000000003817)]
44. Martin GS, Mannino DM, Moss M. The effect of age on the development and outcome of adult sepsis. *Crit Care Med* 2006 Jan;34(1):15-21. [doi: [10.1097/01.ccm.0000194535.82812.ba](https://doi.org/10.1097/01.ccm.0000194535.82812.ba)] [Medline: [16374151](https://pubmed.ncbi.nlm.nih.gov/16374151/)]
45. Barnato AE, Alexander SL, Linde-Zwirble WT, Angus DC. Racial Variation in the Incidence, Care, and Outcomes of Severe Sepsis. *Am J Respir Crit Care Med* 2008 Feb;177(3):279-284. [doi: [10.1164/rccm.200703-480oc](https://doi.org/10.1164/rccm.200703-480oc)]
46. Mayr FB, Yende S, Linde-Zwirble WT, Peck-Palmer OM, Barnato AE, Weissfeld LA, et al. Infection rate and acute organ dysfunction risk as explanations for racial differences in severe sepsis. *JAMA* 2010 Jun 23;303(24):2495-2503 [FREE Full text] [doi: [10.1001/jama.2010.851](https://doi.org/10.1001/jama.2010.851)] [Medline: [20571016](https://pubmed.ncbi.nlm.nih.gov/20571016/)]
47. Chaudhary NS, Donnelly JP, Wang HE. Racial Differences in Sepsis Mortality at U.S. Academic Medical Center–Affiliated Hospitals\*. *Critical Care Medicine* 2018;46(6):878-883. [doi: [10.1097/ccm.0000000000003020](https://doi.org/10.1097/ccm.0000000000003020)]
48. Donnelly J, Lakkur S, Judd S, Levitan EB, Griffin R, Howard G, et al. Association of Neighborhood Socioeconomic Status With Risk of Infection and Sepsis. *Clin Infect Dis* 2018 Jun 01;66(12):1940-1947 [FREE Full text] [doi: [10.1093/cid/cix1109](https://doi.org/10.1093/cid/cix1109)] [Medline: [29444225](https://pubmed.ncbi.nlm.nih.gov/29444225/)]
49. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*. Washington, DC: National Academies Press (US); 2015.
50. Zhang D, Shen X, Qi X. Resting heart rate and all-cause and cardiovascular mortality in the general population: a meta-analysis. *CMAJ* 2016 Feb 16;188(3):E53-E63 [FREE Full text] [doi: [10.1503/cmaj.150535](https://doi.org/10.1503/cmaj.150535)] [Medline: [26598376](https://pubmed.ncbi.nlm.nih.gov/26598376/)]
51. Wang HE, Griffin R, Judd S, Shapiro NI, Safford MM. Obesity and risk of sepsis: a population-based cohort study. *Obesity (Silver Spring)* 2013 Dec 05;21(12):E762-E769 [FREE Full text] [doi: [10.1002/oby.20468](https://doi.org/10.1002/oby.20468)] [Medline: [23526732](https://pubmed.ncbi.nlm.nih.gov/23526732/)]
52. Walker AMN, Drozd M, Hall M, Patel PA, Paton M, Lowry J, et al. Prevalence and Predictors of Sepsis Death in Patients With Chronic Heart Failure and Reduced Left Ventricular Ejection Fraction. *J Am Heart Assoc* 2018 Oct 16;7(20):e009684. [doi: [10.1161/jaha.118.009684](https://doi.org/10.1161/jaha.118.009684)]
53. Wang HE, Gamboa C, Warnock DG, Muntner P. Chronic kidney disease and risk of death from infection. *Am J Nephrol* 2011 Aug 22;34(4):330-336 [FREE Full text] [doi: [10.1159/000330673](https://doi.org/10.1159/000330673)] [Medline: [21860228](https://pubmed.ncbi.nlm.nih.gov/21860228/)]
54. Inghammar M, Engström G, Ljungberg B, Löfdahl CG, Roth A, Egesten A. Increased incidence of invasive bacterial disease in chronic obstructive pulmonary disease compared to the general population--a population based cohort study. *BMC Infect Dis* 2014 Mar 25;14(1):163 [FREE Full text] [doi: [10.1186/1471-2334-14-163](https://doi.org/10.1186/1471-2334-14-163)] [Medline: [24661335](https://pubmed.ncbi.nlm.nih.gov/24661335/)]
55. Tiwari S, Pratyush DD, Gahlot A, Singh SK. Sepsis in diabetes: A bad duo. *Diabetes Metab Syndr* 2011 Oct;5(4):222-227. [doi: [10.1016/j.dsx.2012.02.026](https://doi.org/10.1016/j.dsx.2012.02.026)] [Medline: [25572769](https://pubmed.ncbi.nlm.nih.gov/25572769/)]
56. Frydrych LM, Bian G, O'Lone DE, Ward PA, Delano MJ. Obesity and type 2 diabetes mellitus drive immune dysfunction, infection development, and sepsis mortality. *J Leukoc Biol* 2018 Aug 01;104(3):525-534. [doi: [10.1002/jlb.5vmr0118-021rr](https://doi.org/10.1002/jlb.5vmr0118-021rr)]
57. Bayley K, Belnap T, Savitz L, Masica A, Shah N, Fleming N. Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied. *Med Care* 2013 Aug;51(8 Suppl 3):S80-S86. [doi: [10.1097/MLR.0b013e31829b1d48](https://doi.org/10.1097/MLR.0b013e31829b1d48)] [Medline: [23774512](https://pubmed.ncbi.nlm.nih.gov/23774512/)]
58. Rhee C, Klompas M. Sepsis trends: increasing incidence and decreasing mortality, or changing denominator? *J Thorac Dis* 2020 Feb;12(Suppl 1):S89-S100 [FREE Full text] [doi: [10.21037/jtd.2019.12.51](https://doi.org/10.21037/jtd.2019.12.51)] [Medline: [32148931](https://pubmed.ncbi.nlm.nih.gov/32148931/)]
59. Rhee C, Zhang Z, Kadri SS, Murphy DJ, Martin GS, Overton E, et al. Sepsis Surveillance Using Adult Sepsis Events Simplified eSOFA Criteria Versus Sepsis-3 Sequential Organ Failure Assessment Criteria\*. *Critical Care Medicine* 2019;47(3):307-314. [doi: [10.1097/ccm.0000000000003521](https://doi.org/10.1097/ccm.0000000000003521)]

## Abbreviations

- AUROC:** area under the receiver operating characteristic curve
- COPD:** chronic obstructive pulmonary disease
- EHR:** electronic health record

**GB:** gradient boosting

**L1-LR:** L1-regularized logistic regression

**LR:** logistic regression

**MAD:** median absolute deviation

**PSJH:** Providence St. Joseph Health

**SCTID:** Systematized Nomenclature of Medicine-Clinical Terms identifier

**SHAP:** Shapley Additive exPlanations

**SNOMED CT:** Systematized Nomenclature of Medicine-Clinical Terms

**SVM:** support vector machine

*Edited by G Eysenbach; submitted 04.05.21; peer-reviewed by J Walsh; comments to author 26.05.21; accepted 02.06.21; published 08.07.21.*

*Please cite as:*

*Lee JY, Molani S, Fang C, Jade K, O'Mahony DS, Kornilov SA, Mico LT, Hadlock JJ*

*Ambulatory Risk Models for the Long-Term Prevention of Sepsis: Retrospective Study*

*JMIR Med Inform 2021;9(7):e29986*

*URL: <https://medinform.jmir.org/2021/7/e29986>*

*doi: [10.2196/29986](https://doi.org/10.2196/29986)*

*PMID: [34086596](https://pubmed.ncbi.nlm.nih.gov/34086596/)*

©Jewel Y Lee, Sevda Molani, Chen Fang, Kathleen Jade, D Shane O'Mahony, Sergey A Kornilov, Lindsay T Mico, Jennifer J Hadlock. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 08.07.2021. 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

# Digital Medical Device Companion (MyIUS) for New Users of Intrauterine Systems: App Development Study

Toeresin Karakoyun<sup>1</sup>; Hans-Peter Podhaisky<sup>2</sup>, PhD; Ann-Kathrin Frenz<sup>3</sup>; Gabriele Schuhmann-Giampieri<sup>4</sup>, MBA, PhD; Thais Ushikusa<sup>5</sup>, MD; Daniel Schröder<sup>6</sup>; Michal Zvolanek<sup>7</sup>, MD; Agnaldo Lopes Da Silva Filho<sup>8</sup>, MD, PhD

<sup>1</sup>eHealth and Medical Software Solutions, Bayer AG, Wuppertal, Germany

<sup>2</sup>Regulatory Medical Device Excellence, Bayer AG, Berlin, Germany

<sup>3</sup>Medical Affairs Statistics, Bayer AG, Wuppertal, Germany

<sup>4</sup>Medical Affairs, Real World Evidence and Outcomes Data Generation, Bayer AG, Berlin, Germany

<sup>5</sup>Medical Affairs, Bayer SA, Sao Paulo, Brazil

<sup>6</sup>BAYOOMED Medical Software Development, BAYOONET AG, Darmstadt, Germany

<sup>7</sup>Medical Affairs, Bayer AG, Berlin, Germany

<sup>8</sup>Department of Gynecology and Obstetrics, Federal University of Minas Gerais, Belo Horizonte, Brazil

**Corresponding Author:**

Toeresin Karakoyun

eHealth and Medical Software Solutions

Bayer AG

eHealth & Medical Software Solutions

Building 0459

Wuppertal, 42096

Germany

Phone: 49 152 23914568

Email: [toeresin.karakoyun@bayer.com](mailto:toeresin.karakoyun@bayer.com)

## Abstract

**Background:** Women choosing a levonorgestrel-releasing intrauterine system may experience changes in their menstrual bleeding pattern during the first months following placement.

**Objective:** Although health care professionals (HCPs) can provide counseling, no method of providing individualized information on the expected bleeding pattern or continued support is currently available for women experiencing postplacement bleeding changes. We aim to develop a mobile phone-based medical app (MyIUS) to meet this need and provide a digital companion to women after the placement of the intrauterine system.

**Methods:** The MyIUS app is classified as a medical device and uses an artificial intelligence-based bleeding pattern prediction algorithm to estimate a woman's future bleeding pattern in terms of intensity and regularity. We developed the app with the help of a multidisciplinary team by using a robust and high-quality design process in the context of a constantly evolving regulatory landscape. The development framework consisted of a phased approach including ideation, feasibility and concept finalization, product development, and product deployment or localization stages.

**Results:** The MyIUS app was considered useful by HCPs and easy to use by women who were consulted during the development process. Following the launch of the sustainable app in selected pilot countries, performance metrics will be gathered to facilitate further technical and feature updates and enhancements. A real-world performance study will also be conducted to allow us to upgrade the app in accordance with the new European Commission Medical Device legislation and to validate the bleeding pattern prediction algorithm in a real-world setting.

**Conclusions:** By providing a meaningful estimation of bleeding patterns and allowing an individualized approach to counseling and discussions about contraceptive method choice, the MyIUS app offers a useful tool that may benefit both women and HCPs. Further work is needed to validate the performance of the prediction algorithm and MyIUS app in a real-world setting.

(JMIR Med Inform 2021;9(7):e24633) doi:[10.2196/24633](https://doi.org/10.2196/24633)

**KEYWORDS**

medical device; levonorgestrel-releasing intrauterine system; mobile medical app; mobile phone

## Introduction

### Background

The importance of digital health and the role of software in clinical care are well recognized [1]. In women's health care, interactive digital tools are becoming increasingly accepted in terms of supporting health care choices and facilitating discussions between women and health care professionals (HCPs) [2]. Furthermore, the popularity of menstrual cycle tracking apps continues to rise, and at present, there are more than 100 *period tracking* apps available, with downloads surpassing 200 million globally, since 2016 [3,4]. Women use these apps for a variety of reasons, including to become more aware of their bodies, to understand how the body reacts at different stages of the menstrual cycle, to be prepared (for the start of menstruation), and to facilitate conversations with their HCP [5].

Globally, over 922 million women of reproductive age rely on some form of contraception [6]. A woman's choice of contraceptive method can depend on a variety of factors, including ease and convenience of use, perceived efficacy, associated costs, and expectations of bleeding pattern [7,8]. Many contraceptive methods, such as oral contraceptive pills, implants, levonorgestrel-releasing intrauterine systems (LNG-IUSs), and copper intrauterine devices are associated with alterations in menstrual bleeding patterns [9-11], and some women cite the potential for unfavorable bleeding patterns (such as frequent or prolonged bleeding) or a fear of menstrual irregularity as key reasons for not choosing particular methods [12-15]. Although LNG-IUSs are associated with reductions in menstrual bleeding over time, it is important to note that in the first 3 months following LNG-IUS placement, bleeding and spotting may increase as a result of the local effect of levonorgestrel on the endometrium, which some women may find unfavorable [16-19]. For women deciding to use LNG-IUSs, the desire for less bleeding or amenorrhea (absence of bleeding) is commonly cited as a key reason for choosing this method [19]; therefore, any bleeding or spotting following placement of the LNG-IUS may be perceived negatively and result in dissatisfaction with the method or concern that there is an underlying issue. This could lead to repeat consultations with their HCP to address concerns or, in some cases, the discontinuation of the method [20-24].

Providing thorough contraceptive counseling can help reduce fear and uncertainty and may encourage women to try a method [25]. Counseling should be tailored to the needs of the individual woman, address concerns and preferences, and provide reassurance regarding potential side effects or complications, as well as allow the woman to set realistic expectations regarding her potential bleeding pattern [26-28]. Each woman's experience of menstrual bleeding is highly individual; in addition, information offered during counseling may not be fully processed, and expectations may not align with the impact that bleeding and spotting changes can have on day-to-day

activities. Indeed, although some studies have shown a beneficial effect of anticipatory counseling on method satisfaction and reductions in discontinuation due to bleeding disturbances [29], it has been suggested that counseling interventions with multiple points of contact may further improve adherence and acceptability of contraceptive methods [30].

Available data from clinical trials provide a valuable source of information on bleeding patterns experienced by women using different LNG-IUSs that can be used to aid method selection in clinical practice and can also be used to provide guidance to women on the bleeding pattern they may expect when initiating an LNG-IUS [18-20,31-34]. However, the experience of bleeding following placement is unique to each woman and may be influenced by a variety of factors. Although information on the most commonly expected bleeding patterns from clinical trials helps to convey the general likelihood of having a certain pattern, there is currently no means available to further tailor this information to the individual woman, and help inform her of her personal expected duration or intensity of bleeding.

A tool that can predict a woman's menstrual bleeding pattern and provide a clinically meaningful output, such as expected duration and intensity of bleeding after device placement, could therefore be a valuable addition to support counseling. Furthermore, after the initial counseling visit and placement of an LNG-IUS, a woman may find that her bleeding pattern is unfavorable and interferes with her quality of life, leading her to seek additional support and reassurance from her HCP to help her manage the situation. Providing an interactive digital app could therefore allow further information to be provided to women in the postplacement period, empowering them to better understand their LNG-IUS and any associated alterations in menstrual bleeding patterns. This may improve confidence and satisfaction with the contraceptive method and encourage continuation, as well as facilitate easier communication between women and their HCPs.

### Rationale

Data gathered through the analysis of daily bleeding diaries used during phase II and phase III clinical trials of LNG-IUS 12 (Kyleena, Bayer AG) indicated that in the initial 90 days following placement of LNG-IUS 12, the most commonly reported menstrual bleeding patterns were prolonged bleeding and irregular bleeding [31,32]. This period of prolonged or irregular bleeding following placement may be perceived as a deterrent to some women when deciding on the contraceptive method or could lead to dissatisfaction and potentially early discontinuation for women using the IUS. Over time, the number of women reporting unfavorable bleeding patterns markedly decreases, with more favorable patterns such as infrequent bleeding and amenorrhea being the most commonly reported patterns at the end of year 1, with menstrual bleeding becoming progressively lighter over the 5-year duration of use [19,32].

It was perceived that providing additional support and information regarding expected future bleeding patterns to women during the initial postplacement interval could provide

reassurance and encourage method persistence. Therefore, an algorithm was developed that allowed the meaningful estimation of future bleeding patterns based on evidence collected on an individual basis through a digital daily bleeding diary [34]. The algorithm represents the first tool of its kind to provide an individualized prediction of future bleeding patterns and could be useful to both women and HCPs.

The artificial intelligence (AI)-based algorithm uses 90-day bleeding diary information to predict future bleeding patterns, including intensity and regularity. After the woman has entered 90 consecutive days of menstrual bleeding information into a daily bleeding diary, a random forest approach is applied to assign her expected bleeding intensity pattern into 1 of 3 categories: predominantly amenorrhea (<5% spotting days or <1% bleeding days within days 91-270 [equivalent to ≤8 spotting days or ≤1 bleeding day]); predominantly spotting (women not belonging to the predominantly amenorrhea cluster and with <5% bleeding days [≤8 bleeding days] within days 91-270); predominantly bleeding (all other women, ie, ≥5% bleeding days [≥8 bleeding days] within days 91-270). A logistic regression model is then used to estimate the probability of the woman having a regular cycle. The probability of correct classification using the model is high (>70%), and the generated bleeding categories are considered informative [34].

## Objective

To facilitate the use of the algorithm in routine clinical practice, it is important to generate an interface that allows the input of a woman's bleeding diary information and combines it with a clear and easy-to-understand report of the algorithm output. Therefore, our aim is to develop a mobile medical app (MyIUS), which integrates the collection of daily menstrual bleeding information, the AI bleeding pattern prediction algorithm, and a report of the predicted bleeding category. The app is intended to support IUS users during the initial postplacement period and provide a prediction of future bleeding profiles, facilitate effective counseling, and encourage users to persist with IUS device use. This paper describes the development of MyIUS as a sustainable app in the context of an evolving regulatory landscape.

## Methods

### Development Approach Context

Mobile medical apps are defined as software that can be executed on a mobile platform, which can be used either as an accessory to a regulated medical device or to transform a mobile platform into a regulated medical device [35]. International guidance for developers of mobile medical apps is available from the World Health Organization and European Commission, and guidance on a national level is available from bodies such as the US Food and Drug Administration and UK National Institute for Health and Care Excellence [36-38]. However, there is currently no comprehensive framework for the development process; guidance varies between countries, authorities, and institutions and addresses differing aspects of app design, from data protection principles and technical considerations through to defining medical purposes [39].

The spread of different sources of guidance across agencies means that it can be challenging to ascertain the most appropriate development approach to ensure compliance with all relevant standards and regulations [39]. Furthermore, apps are often global products; therefore, developers frequently need to navigate complex compliance requirements involving different agencies, regulations, and guidelines. It is also important to consider the language and cultural aspects of the app. In addition to language translation, some countries have different measurement units and different numeric and date and time formats. These aspects should be addressed in any development plan to ensure that the app will be accessible and suitable for use in different countries and is sustainable over time.

### Regulatory Context

The MyIUS app is classified as a medical device, which is defined as an instrument, apparatus, appliance or software for a specific medical purpose or purposes that does not achieve its principal intended action by pharmacological, immunological, or metabolic means [37,40]. The MyIUS software is intended to monitor menstrual bleeding and spotting. This means that under the forthcoming European Commission Medical Device legislation (due to be launched in May 2021), the MyIUS software will be designated, according to Rule 11 of Medical Device Regulation (MDR) EU 2017/745, as a moderate risk or Class IIa medical device [41]. This moderate risk classification corresponds to the software safety Class A of IEC 62304.

In terms of medical device software, the new MDR, EU 2017/745, represents a significant upgrade of the default classification. Although the default classification for software according to previous legislation (Medical Device Directive [MDD] 93/42/EEC) is a Class I medical device that allows a self-declaration process to demonstrate conformity with the medical device legislation (with no notified body approval necessary for the technical file), the new default classification of medical device software as Class IIa in the forthcoming legislation requires an active approval process of a notified body based on a review of a technical file in order to obtain a declaration of conformity (European Conformity mark). To meet the regulatory requirements for this medical device category, it is essential to demonstrate conformity with general safety and performance requirements, to provide clinical evidence, and to apply a documented design-control process based on a certified quality management system.

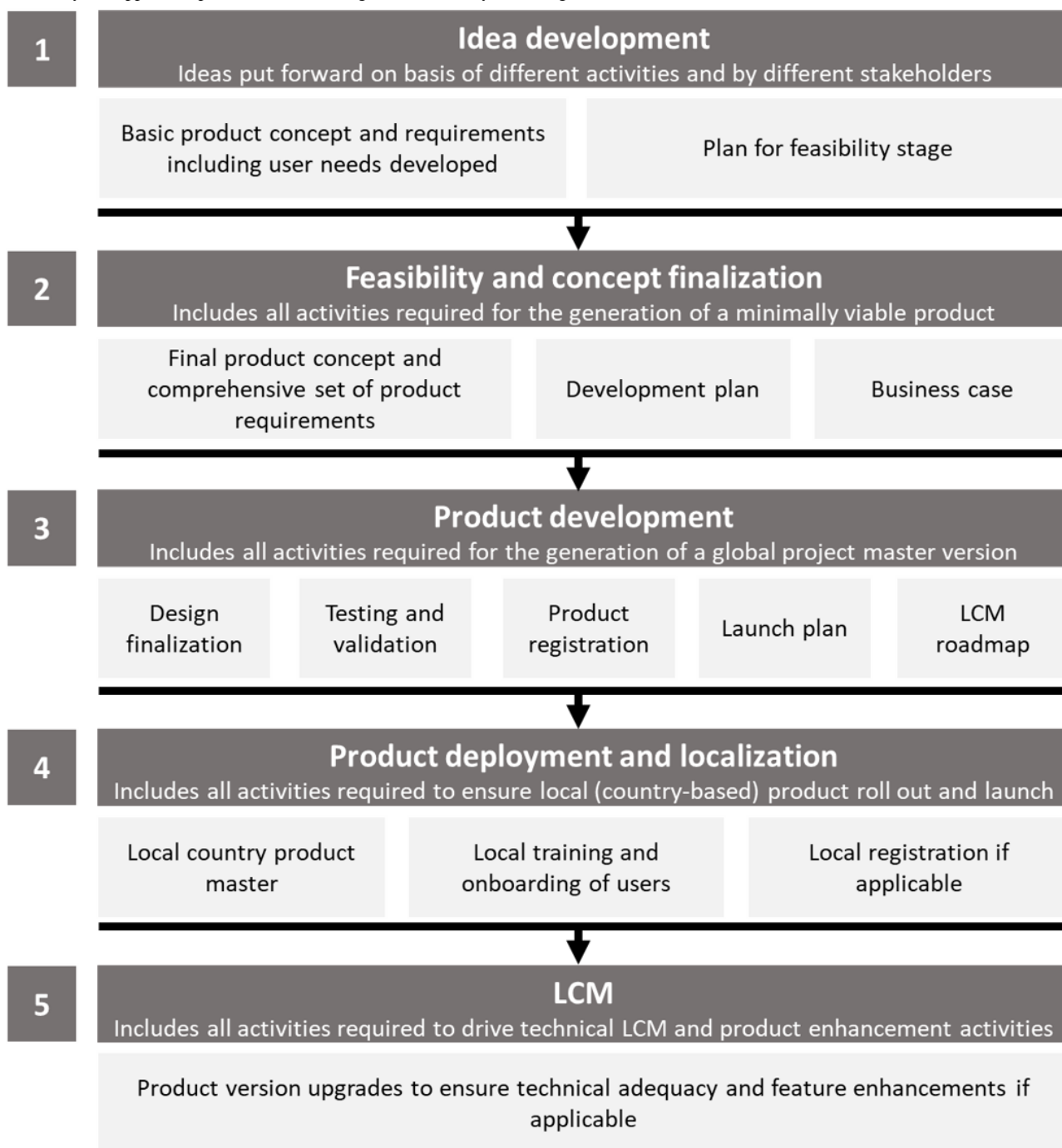
To address these stricter requirements for software in the new European legislation, following the launch of the MyIUS app, users will be invited to participate in a real-world performance study sponsored by Bayer AG. This will allow the collection of complementary evidence in a real-world setting, beyond the data provided by Bayer's previous clinical trial program, in order to support the transition from a Class I medical device as described by MDD 93/42/EEC to a Class IIa medical device, as described by MDR EU 2017/745 [37,41,42].

The rapidly growing field of medical software apps has led to a constantly evolving regulatory landscape, requiring developers to explore innovative pathways in app design and development. Our medical software product development framework consisted

of a phased approach with four key stages, followed by life cycle management comprising a postmarket surveillance plan. The development stages included ideation, feasibility and

concept finalization, product development, and product deployment and localization (Figure 1).

Figure 1. MyIUS app development framework stages. LCM: life cycle management.



**Idea Development**

Individual interviews were conducted with a panel of 40 obstetricians and gynecologists in Germany and Brazil to identify the areas of need and inform the development of the

MyIUS app. A concept document describing the idea of the MyIUS app was shared with individual panel members. Web-assisted telephone interviews were conducted to gain feedback on the perceived utility and benefits of the app (Textbox 1).



**Textbox 1.** Perceived benefits and utility of the MyIUS app for health care professionals and intrauterine system users identified during interviews with obstetricians and gynecologists.

**For health care professionals (HCPs)**

- Could support individualized management and discussion of relevant options with women
- May help HCPs to give more accurate advice to patients about what to expect over the next 3 months
- Patients could be less likely to return to HCPs to discuss bleeding issues, as they know what to expect
- Report tool could provide a realistic view of the woman's bleeding pattern
- May facilitate more accurate and efficient discussions by reducing reliance on woman's recollection and subjective description of bleeding

**For intrauterine system (IUS) users**

- IUS users have a personal support tool that they can interact with throughout their postplacement journey
- May help women feel supported and monitored by their HCP throughout the postplacement period
- Digital tool may be better accepted and used by women than a printed leaflet or other such material
- Could improve awareness of bleeding patterns and help to normalize changes after IUS placement
- May increase motivation to continue with IUSs

Insights from these interviews revealed that in routine clinical practice, there is a need for a tool that can support counseling around expected menstrual bleeding changes beyond the initial counseling visit and provide additional information to women during the first 3 to 6 months after placement of the IUS, when they may experience alterations in their normal bleeding pattern. The HCPs interviewed also stated that the tool should be simple to use for easy onboarding and should be self-explanatory, to avoid the need for users to request additional information or explanation from the HCP. For users, there is a perceived need to provide an additional method of support in the initial postplacement period and improve knowledge and understanding of bleeding patterns during this time. The MyIUS app was developed to meet these needs.

The MyIUS app is a collaborative project between Bayer AG (sponsors and developers of the AI algorithm), BAYOOCARE GmbH (legal manufacturers of the app), and BAYOONET AG (ISO 13485–certified software developer of the app) with Concentrix Global Services GmbH providing first- and second-level support. A multidisciplinary team was established to ensure an efficient, high-quality, and compliant design process, which included expertise from the Medical Software Product Lead; eHealth Systems Engineer; Law Patents and Compliance Department; Clinical Development, Medical Affairs, Regulatory Affairs and Clinical Medical Devices divisions; Pharmacovigilance and Pharmaco-device vigilance experts; Medical Software Quality Expert, app development partner, and commercial teams, as well as key insights from the developers of the AI bleeding pattern prediction algorithm. Internal teams from Bayer AG collaborated with external partners from the BAYOONET division team at BAYOONET AG. Development teams worked in an agile way to provide efficient and robust design and development capabilities.

The key requirements of the MyIUS app identified at this stage were to collect baseline parameters and daily bleeding

information for at least 90 days after placement of the IUS, accompany the LNG-IUS user from placement through at least 90 days postplacement offering useful information, and to provide a prediction of bleeding profile for the next 6 months with respect to intensity and regularity based on the collected data, which could facilitate communication between the IUS user and their HCP.

### Feasibility and Concept Finalization

The MyIUS app is intended for users of Bayer's LNG-IUS (Mirena, Kyleena, and Jaydess or also known as Skyla) following placement. Users may download the app onto their smartphones to monitor and predict the effect of LNG-IUS on menstrual bleeding based on their input of daily bleeding information. The app is not intended to make any diagnostic decision or act as a substitute for evidence-based counseling but instead will act as a *digital companion* or user support tool for women and as a source of bleeding information for HCPs that may facilitate a more individualized counseling approach.

High-level minimum viable stakeholder requirements were defined for the first version; these are presented in [Textbox 2](#). The key stakeholders identified included the end user (main stakeholder), regulatory personnel and regulatory bodies, and quality (both medical and digital) management personnel. A detailed concept and development plan was generated to include all functional and nonfunctional aspects of design, such as verification, validation, usability engineering, and development of the app. These included specifications for stakeholder requirements and software requirements, architecture documentation, module or unit test specification, and software integration test specification. Risk assessment and postmarket surveillance will be conducted after the launch of the app to gain valuable insights into stakeholder requirements and potential issues that may need to be addressed.

**Textbox 2.** High-level minimum viable user requirements identified for the MyIUS app.

<p><b>App setup and access</b></p> <ul style="list-style-type: none"> <li>• Unlocking of app with access code</li> <li>• User to enter baseline data</li> <li>• Access to frequently asked questions and informative video</li> <li>• This step should be intuitive and take &lt;15 minutes to complete</li> </ul> <p><b>Interaction with the app after placement of the intrauterine system (IUS)</b></p> <ul style="list-style-type: none"> <li>• Bleeding diary recording bleeding, spotting, or no bleeding</li> <li>• Reminder function</li> <li>• Allow backfilling</li> <li>• Motivational function and gamification (collection of knowledge gems)</li> <li>• Should include counseling snippets (knowledge gems) for the first 3 months</li> </ul> <p><b>Post-90-day interaction with the app</b></p> <ul style="list-style-type: none"> <li>• IUS user submits data after 90 days</li> <li>• Result: user assigned to one of three bleeding categories, with description of most likely bleeding pattern</li> <li>• Reminder to schedule follow-up appointment with health care professional</li> <li>• Support to motivate user to continue entering bleeding diary data</li> <li>• PDF copy of bleeding calendar display</li> </ul> <p><b>Other</b></p> <ul style="list-style-type: none"> <li>• Export and import data function (in case of phone switch)</li> <li>• Frequently asked questions</li> <li>• Imprint</li> <li>• User support material</li> <li>• User manual or instructions for use</li> <li>• Adverse event reporting via link to appropriate external site</li> </ul>
---

## Product Development

To facilitate integration into the MyIUS app, the original AI bleeding pattern prediction algorithm described by Frenz et al [34] was redeveloped in R 3.6.0 (The R Foundation). Educational content was generated in collaboration with medical affairs team members to ensure the scientific accuracy and validity of the material.

To optimize app design, a usability engineering process was followed in accordance with IEC 62366 [43], which included context analysis, use specification, primary operating functions, and hazardous scenarios (characteristics related to safety and hazard identification). Ongoing iterative testing was performed throughout the development process to optimize the final product. Enhancements identified by usability testing were implemented during software development.

A prototype was developed based on the minimum viable user requirements specified in [Textbox 2](#). Visuals were generated for the setup pages and profile options, menu bar, settings page,

support page, legal notice, bleeding report for HCPs, main *home* screen, reminders for when recording of days or recording of bleeding intensity (none, spotting, or bleeding) is missed, calendar, frequently asked questions, prediction day screen (including the report of expected bleeding pattern generated by the algorithm), user feedback page, and educational content pop-ups (so called *knowledge gems*). Within the scope of a formative usability study, the prototype of the app was presented to a panel of 8 women. The panel was considered broadly representative of the envisioned end user of the MyIUS app; women were between the ages of 23 and 48 years and included a mixture of IUS, contraceptive pill, copper intrauterine device, and condom users as well as parous and nulliparous women. All women reported that they would use the app in their daily lives, with most (5/8, 63%) agreeing that the app was “very good” and “user friendly.” Users were asked what they would like to see in the final version of the app and what changes they would make to the existing prototype ([Textbox 3](#)). Suggested improvements and desires for the final version were considered for the next stages of development.

**Textbox 3.** Suggestions for improvement and desires for final version of the MyIUS app collected from the end-user panel.

<b>Improvements</b>
<ul style="list-style-type: none"> <li>• More colorful design</li> <li>• Use consistent wording and terminology throughout</li> <li>• Space for additional notes or information to be entered (eg, exercise)</li> <li>• Provide more information on the different topics including bleeding and contraception</li> <li>• Option to mark days without entry as nonbleeding days</li> </ul>
<b>Desires for final version</b>
<ul style="list-style-type: none"> <li>• Include additional information about the intrauterine system itself and practical information (eg, how to check threads)</li> <li>• Ensure app remains easy to access and use, with the same ability to rapidly enter bleeding data</li> <li>• Reminders to add daily bleeding diary information</li> <li>• Help section in the app that includes instructions on how to use</li> </ul>

Following the completion of the development, verification testing and comprehensive validation testing, including user acceptance testing (summative study), were carried out. These included testing the primary operating functions and hazardous scenarios. A panel of 15 representative end users was engaged virtually (because of restrictions in place as a result of the COVID-19 pandemic) and consisted of women from the United States, Germany, Brazil, and Chile, aged between 20 and 50 years. Women were asked to perform specific tasks and navigate through the app, and their performance was assessed according to predefined criteria ([Multimedia Appendix 1](#)). Feedback on

app functionality and overall design was also requested from the users ([Table 1](#)).

Representative examples of final wireframes for the MyIUS app, which demonstrate key aspects of app design and highlight the changes that were implemented based on feedback from users, are presented in [Figure 2](#).

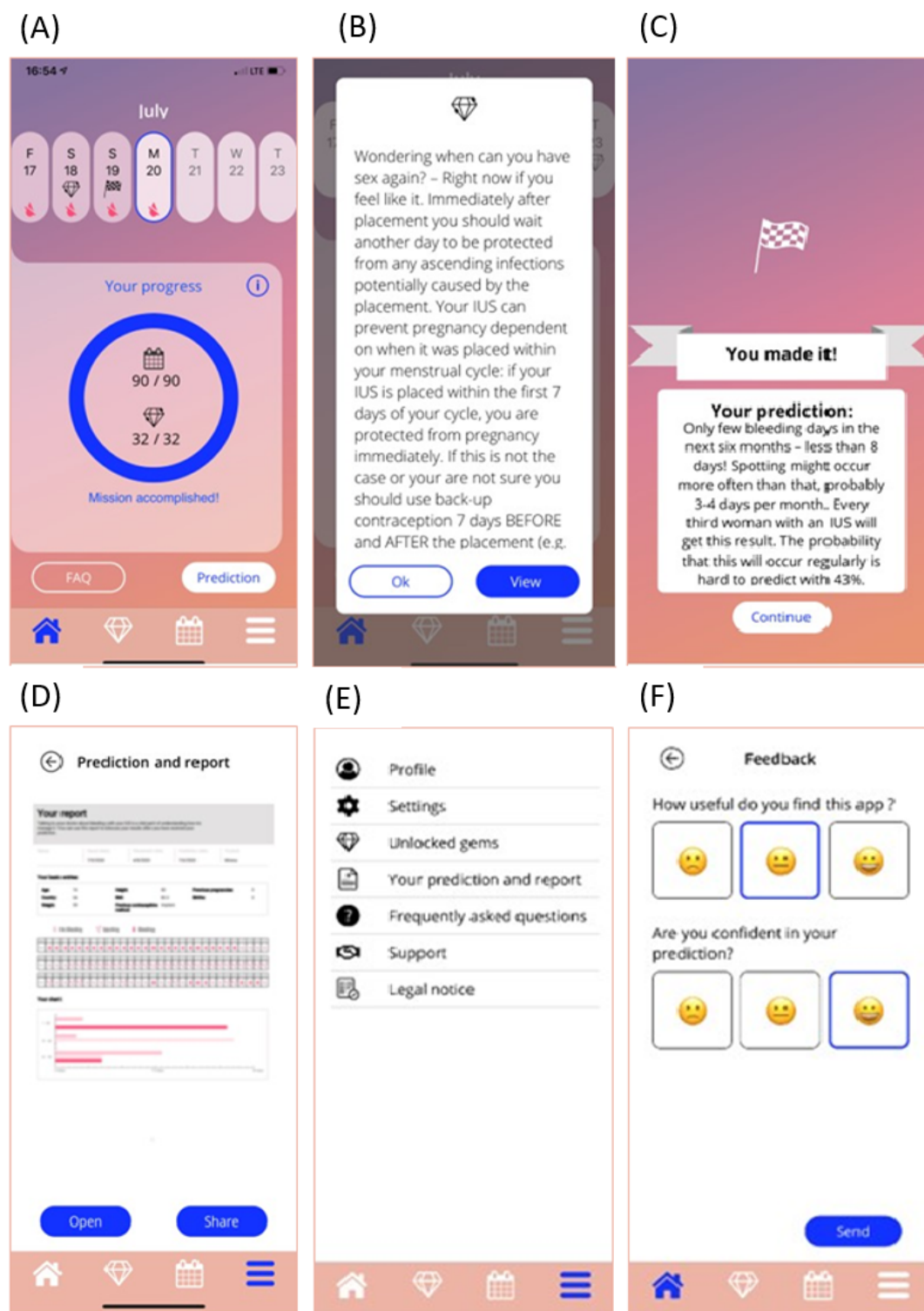
Product registration, regulatory submission, and self-certification (according to current MDD 93/42/EEC legislation) were handled by BAYOONET AG.

**Table 1.** Feedback to specific questions regarding MyIUS app in later development stages.

Question	Positive response, n (%) <sup>a</sup>
Would use the app in daily life	13 (93)
Number of setup questions appropriate	11 (79)
Gem concept is clear	11 (79)
Would recommend the app to a friend	10 (71)
Understand the home screen	10 (71)

<sup>a</sup>One woman from the panel did not respond to questions; therefore, positive responses were from a total of 14.

**Figure 2.** Representative examples of MyIUS app wireframes showing the (A) home screen, (B) pop-up of knowledge gem, (C) bleeding prediction screen, (D) bleeding prediction report for health care professionals, (E) menu screen, and (F) feedback screen.



**Product Deployment and Localization**

To ensure that the app can be easily localized and upgraded, all text used within the graphical user interface is managed separately from the interface, meaning that the addition of a new language will not require additional programming; instead, a translation of all text will be provided. Local language translation will be performed by a professional translation

agency. Transcripts will then be reviewed by legal, medical, and regulatory teams in each country and, in addition, accuracy, colloquial language, and comprehension will be further checked by local language speakers, including an expert HCP. The approved local language transcript will then be implemented by a technical team at BAYOOCARE GmbH.

The app code can incorporate predefined formats (as per device settings) without the need for user interaction. For specific



settings where the user might be free to choose within the app, specific conversion logic and rounding are implemented. Local country teams will also ensure compliance of the local country's product master with the specific data privacy laws of that country.

Local country product masters, local registrations, local training, and local user onboarding have been implemented in specific pilot countries, including Germany (first pilot country), Mexico, France, Spain, Portugal, and the United Kingdom. Launch of the final sustainable app in additional countries will follow.

## Results

### Final Product

The app is available free of charge in the App Store (Apple Inc) and Google Play Store (Google LLC). Once downloaded, the user will need to activate the app using the code supplied by their HCP alongside the LNG-IUS prescription. The global product master has been confirmed to comply with the General Data Protection Regulation to ensure the protection of individual privacy. After a user downloads the app, the data are stored locally on the user's smartphone. When 90 days of bleeding diary information has been entered, the data are then encrypted and sent as a package to a BAYOONET-managed Amazon Web Services cloud server that runs the AI bleeding pattern prediction algorithm and returns an output to the woman's device. All data are anonymized and not linked to mobile number, name, or any other personal identifier. No data are stored on the BAYOONET AG server after the algorithm is run.

### Launch and Life Cycle Management

Launch of the MyIUS app in Germany, the first pilot country, took place in early July 2020. The collection of usage metrics, such as the number of downloads, is ongoing and will be used to inform launches in additional pilot countries (such as Mexico, Spain, and Portugal). Following the global launch of the sustainable app, product enhancement activities, including upgrades to ensure technical adequacy and feature enhancements, if applicable, will be performed on a regular basis.

### Real-World Data Collection

To generate complementary evidence to support the upgrade of the MyIUS app to a Class IIa medical device under the new European Commission regulation [41], a real-world performance study is planned. The study will be sponsored and conducted by Bayer AG.

Users of the app will be given the choice to participate in a real-world performance study, and participants will need to provide informed e-consent via the app before enrollment. Participants in the study will use the app to enter 90 days of bleeding diary data and will receive their individual bleeding pattern prediction, and they will then be asked to continue to submit bleeding diary entries for a further 180 days. Data will be gathered in compliance with General Data Protection Regulation; data will be anonymized, and no personal identifiers or mobile numbers will be collected. Data will be saved locally

on the user's smartphone or device, and no data will be stored on the BAYOONET AG server after the algorithm is run.

In addition to generating evidence to allow the upgrade of the MyIUS app, data collected in the performance study will also allow the validation of the AI bleeding pattern prediction algorithm in a real-world setting. The prediction algorithm was developed and trained using data from controlled clinical trials in Kyleena and cross-validated using data from clinical studies of Mirena and Jaydess. The MyIUS app therefore offers an opportunity to gather insights into the performance of the algorithm in the general population and allow additional confirmation of its utility as well as identify further improvements.

## Discussion

### Principal Findings

The MyIUS app was created using a robust and compliant development process that sought insights from HCPs and end users to ensure an app that was of high quality and fit-for-purpose. The end result is an app that will provide a companion to LNG-IUS users in the postplacement period, allowing them to monitor their bleeding pattern and receive an accurate prediction of their expected future duration and intensity of menstrual bleeding.

LNG-IUSs are long-acting contraceptive methods that can be used for 3-5 years. Although some women experience an initial increase in bleeding and spotting during the first 90 days of use, at 12 months, LNG-IUSs are generally associated with a decrease in bleeding days compared with baseline [16-18], although it should be noted that bleeding patterns are considered to be dose dependent and therefore can vary between different LNG-IUSs. More women using the higher dose LNG-IUS 20 (Mirena) experience amenorrhea for example, whereas the users of the lower dose LNG-IUS 8 (Jaydess) report a higher number of bleeding and spotting days than those using LNG-IUS 20 or LNG-IUS 12 [19]. Accordingly, counseling should cover both short-term and long-term bleeding patterns, as well as the benefits, side effects, and risks to help women set realistic expectations [19]. Offering comprehensive counseling on these aspects can improve user satisfaction and continuation with LNG-IUSs [27,29]. However, it is important to note that providing counseling does not completely mitigate the risk of a woman discontinuing with a method as a result of dissatisfaction [44]. Although the discontinuation rate with LNG-IUSs is low overall [33,45,46], discontinuation can have important consequences. Women who discontinue LNG-IUS use may switch to user-dependent or short-acting methods such as condoms, injectables, or oral contraceptive pills or may not switch methods promptly (within 3 months), leaving them at higher risk of unintended pregnancy [44,47,48].

In this regard, MyIUS provides women using IUSs with a means to monitor their bleeding pattern, with the added benefit of generating a meaningful output in the form of a prediction of future bleeding, which could encourage persistence with the method. Through gamification and the provision of knowledge insights, the MyIUS app may help to improve the experience

of women in the first months following LNG-IUS placement. The MyIUS app could also provide reassurance to women who experience altered bleeding patterns following LNG-IUS placement and help them feel more in control of the situation. The accurate prediction (probability of correct classification >70%) of future bleeding patterns should also allow women to form realistic expectations about the amount and duration of bleeding they are likely to experience, which can allow them to have better discussions with their HCP regarding their contraceptive method moving forward [34].

Women's choice of contraception is often influenced by whether HCPs mention or recommend a specific method [49,50]. There are various reasons why HCPs may or may not recommend a specific method, and HCPs' knowledge of bleeding patterns has been found to be strongly associated with the provision of LNG-IUSs, with those who are unfamiliar with the potential bleeding pattern alterations being less likely to include LNG-IUSs in their counseling [51]. By providing an additional tool to support HCPs with counseling and educate women regarding bleeding patterns, MyIUS could help encourage HCPs to include LNG-IUSs in discussions with women, providing women with a greater choice of contraceptive options that may suit their needs. Reports generated from the MyIUS app may also facilitate discussions of contraception and bleeding patterns between HCPs and women and allow information to be tailored to the individual. Furthermore, tracking apps can also reduce the number of clinical appointments needed and can decrease workloads for HCPs [3].

### Limitations

The MyIUS app was tested by a panel of 23 end users from four different countries (Germany, Brazil, the United States, and Chile) as part of the development process; however, as attitudes, beliefs, digital literacy levels, and personal preferences vary between person to person and country to country, the opinions of the panel may not be reflective of all potential end users. Further insights gathered through the real-world validation study, app store ratings, and user feedback from within the app will be beneficial to confirm the utility of the app in a wider population of users. This will also help inform further updates and improvements to the app in the future.

Initial testing and validation of the AI bleeding pattern prediction algorithm were performed on bleeding diary data from clinical trials of LNG-IUS 12, LNG-IUS 20, and LNG-IUS 8. The planned real-world data study will therefore be essential to collect evidence on the performance of the algorithm in a much wider cohort of women under normal use conditions. Furthermore, the real-world evidence generated will be needed to upgrade the app to a Class IIa medical device under new EU legislation.

Finally, thus far, the AI algorithm and MyIUS app have only been tested and validated in women using LNG-IUSs for contraception. In clinical practice, HCPs may use LNG-IUSs, such as LNG-IUS 20, for other indications such as treating heavy menstrual bleeding; therefore, further investigation is

needed to determine the performance of the algorithm and the utility of the app for women using LNG-IUSs for indications in addition to contraception.

### Comparison With Prior Work

The use of personal, digital health informatics is becoming increasingly widespread, and self-tracking of menstruation using apps offers a convenient and easily accessible way for women to manage their own health, allowing them to compare past and average cycles and help identify regularity or irregularity [3,5,52]. The feeling of reward gained by entering data into an app and receiving immediate feedback is suggested to increase motivation and encourage persistence with medical interventions [53]. Features such as gamification and gaining useful knowledge also add to a sense of satisfaction. These aspects may contribute to the popularity of menstrual cycle tracking apps such as Flo (Flo Health Inc), used by over 140 million women worldwide and Clue (Biowink GmbH), used by over 15 million (metrics taken from respective app pages in the App Store and Google Play Store, February 2021).

MyIUS has been developed using a compliant, robust design process involving input from a multidisciplinary team working in an agile way to support the development of a validated mobile medical device. The MyIUS app is the first digital tool designed to support women who choose LNG-IUSs as their method of contraception by using daily menstrual bleeding diary information to provide an accurate prediction of future menstrual bleeding patterns on an individual basis. The app will provide educational insights to users regarding LNG-IUSs and menstrual bleeding patterns and is intended to act as a digital companion for women after placement of an LNG-IUS.

### Conclusions

The MyIUS app has been designed to be simple to use and provides a companion to women during the initial period after IUS placement as well as generating meaningful estimates of bleeding patterns. By tracking menstrual bleeding patterns and estimating expected future patterns, the app may help to facilitate discussions about contraception and bleeding patterns between HCPs and women. The app could also provide reassurance to women who experience altered bleeding patterns following LNG-IUS placement and help them feel more in control of the situation. Furthermore, HCPs may use the information provided by menstrual bleeding diary entries and the AI bleeding pattern prediction algorithm to personalize and enhance counseling about possible bleeding with LNG-IUSs, helping women to set realistic expectations, potentially reducing discontinuation and increasing method satisfaction.

As the desire for customized health care that fits individuals' unique needs and preferences increases, the MyIUS app offers a first step in making this a reality in the contraceptive counseling and decision-making process for HCPs and women. Further information from user feedback and the planned real-world validation study of the bleeding pattern prediction algorithm are required to inform future refinements and confirm the value of MyIUS for women and HCPs.

## Acknowledgments

Development of the AI-based algorithm and MyIUS app was conducted and sponsored by Bayer AG, Berlin, Germany. Editorial support for this paper was provided by Highfield Communication, Oxford, United Kingdom, with funding from Bayer AG.

## Conflicts of Interest

TK, HPP, AKF, GSG, and MZ are employees of Bayer AG. TU is an employee of Bayer SA, Brazil. DS is an employee of BAYOONET AG, Germany. ALSF is a member of the Heavy Menstrual Bleeding: Evidence-Based Learning for Best Practice group, a panel of independent physicians with an expert interest in heavy menstrual bleeding, the formation of which was facilitated by Bayer AG. He also acted as a consultant for Bayer AG and received consultancy honoraria.

## Multimedia Appendix 1

Acceptance criteria and results for usability engineering analysis.

[\[DOCX File, 13 KB - medinform\\_v9i7e24633\\_app1.docx\]](#)

## References

1. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. *NPJ Digit Med* 2020;3:14 [FREE Full text] [doi: [10.1038/s41746-019-0212-z](https://doi.org/10.1038/s41746-019-0212-z)] [Medline: [32047860](https://pubmed.ncbi.nlm.nih.gov/32047860/)]
2. Dehlendorf C, Fitzpatrick J, Steinauer J, Swiader L, Grumbach K, Hall C, et al. Development and field testing of a decision support tool to facilitate shared decision making in contraceptive counseling. *Patient Educ Couns* 2017 Jul;100(7):1374-1381. [doi: [10.1016/j.pec.2017.02.009](https://doi.org/10.1016/j.pec.2017.02.009)] [Medline: [28237522](https://pubmed.ncbi.nlm.nih.gov/28237522/)]
3. Bull JR, Rowland SP, Scherwitzl EB, Scherwitzl R, Danielsson KG, Harper J. Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles. *NPJ Digit Med* 2019;2:83 [FREE Full text] [doi: [10.1038/s41746-019-0152-7](https://doi.org/10.1038/s41746-019-0152-7)] [Medline: [31482137](https://pubmed.ncbi.nlm.nih.gov/31482137/)]
4. Earle S, Marston HR, Hadley R, Banks D. Use of menstruation and fertility app trackers: a scoping review of the evidence. *BMJ Sex Reprod Health* 2020 Apr 06;47(2):90-101. [doi: [10.1136/bmjsex-2019-200488](https://doi.org/10.1136/bmjsex-2019-200488)] [Medline: [32253280](https://pubmed.ncbi.nlm.nih.gov/32253280/)]
5. Epstein DA, Lee NB, Kang JH, Agapie E, Schroeder J, Pina LR, et al. Examining menstrual tracking to inform the design of personal informatics tools. *Proc SIGCHI Conf Hum Factor Comput Syst* 2017 May 02;2017:6876-6888 [FREE Full text] [doi: [10.1145/3025453.3025635](https://doi.org/10.1145/3025453.3025635)] [Medline: [28516176](https://pubmed.ncbi.nlm.nih.gov/28516176/)]
6. Contraceptive use by method. United Nations Department of Economic and Social Affairs. 2019. URL: <https://tinyurl.com/2pabv8y2> [accessed 2021-02-15]
7. Wyatt KD, Anderson RT, Creedon D, Montori VM, Bachman J, Erwin P, et al. Women's values in contraceptive choice: a systematic review of relevant attributes included in decision aids. *BMC Womens Health* 2014 Feb 13;14(1):28 [FREE Full text] [doi: [10.1186/1472-6874-14-28](https://doi.org/10.1186/1472-6874-14-28)] [Medline: [24524562](https://pubmed.ncbi.nlm.nih.gov/24524562/)]
8. Beckert V, Aqua K, Bechtel C, Cornago S, Kallner HK, Schulze A, et al. Insertion experience of women and health care professionals in the Kyleena Satisfaction Study. *Eur J Contracept Reprod Health Care* 2020 Jun;25(3):182-189. [doi: [10.1080/13625187.2020.1736547](https://doi.org/10.1080/13625187.2020.1736547)] [Medline: [32223466](https://pubmed.ncbi.nlm.nih.gov/32223466/)]
9. Polis CB, Hussain R, Berry A. There might be blood: a scoping review on women's responses to contraceptive-induced menstrual bleeding changes. *Reprod Health* 2018 Jun 26;15(1):114 [FREE Full text] [doi: [10.1186/s12978-018-0561-0](https://doi.org/10.1186/s12978-018-0561-0)] [Medline: [29940996](https://pubmed.ncbi.nlm.nih.gov/29940996/)]
10. Tolley E, Loza S, Kafafi L, Cummings S. The impact of menstrual side effects on contraceptive discontinuation: findings from a longitudinal study in Cairo, Egypt. *Int Fam Plan Perspect* 2005 Mar;31(1):15-23 [FREE Full text] [doi: [10.1363/3101505](https://doi.org/10.1363/3101505)] [Medline: [15888405](https://pubmed.ncbi.nlm.nih.gov/15888405/)]
11. Mansour D, Korver T, Marintcheva-Petrova M, Fraser IS. The effects of Implanon on menstrual bleeding patterns. *Eur J Contracept Reprod Health Care* 2008 Jun;13 Suppl 1:13-28. [doi: [10.1080/13625180801959931](https://doi.org/10.1080/13625180801959931)] [Medline: [18330814](https://pubmed.ncbi.nlm.nih.gov/18330814/)]
12. Buhling KJ, Hauck B, Dermout S, Ardaens K, Marions L. Understanding the barriers and myths limiting the use of intrauterine contraception in nulliparous women: results of a survey of European/Canadian healthcare providers. *Eur J Obstet Gynecol Reprod Biol* 2014 Dec;183:146-154 [FREE Full text] [doi: [10.1016/j.ejogrb.2014.10.020](https://doi.org/10.1016/j.ejogrb.2014.10.020)] [Medline: [25461369](https://pubmed.ncbi.nlm.nih.gov/25461369/)]
13. Hauck B, Costescu D. Barriers and misperceptions limiting widespread use of intrauterine contraception among Canadian women. *J Obstet Gynaecol Can* 2015 Jul;37(7):606-616. [doi: [10.1016/S1701-2163\(15\)30198-5](https://doi.org/10.1016/S1701-2163(15)30198-5)] [Medline: [26366817](https://pubmed.ncbi.nlm.nih.gov/26366817/)]
14. Goldstuck ND. Reducing barriers to the use of the intrauterine contraceptive device as a long acting reversible contraceptive. *Afr J Reprod Health* 2014 Dec;18(4):15-25. [Medline: [25854089](https://pubmed.ncbi.nlm.nih.gov/25854089/)]
15. Clark LR, Barnes-Harper KT, Ginsburg KR, Holmes WC, Schwarz DF. Menstrual irregularity from hormonal contraception: a cause of reproductive health concerns in minority adolescent young women. *Contraception* 2006 Sep;74(3):214-219. [doi: [10.1016/j.contraception.2006.03.026](https://doi.org/10.1016/j.contraception.2006.03.026)] [Medline: [16904414](https://pubmed.ncbi.nlm.nih.gov/16904414/)]
16. Goldthwaite LM, Creinin MD. Comparing bleeding patterns for the levonorgestrel 52 mg, 19.5 mg, and 13.5 mg intrauterine systems. *Contraception* 2019 Aug;100(2):128-131. [doi: [10.1016/j.contraception.2019.03.044](https://doi.org/10.1016/j.contraception.2019.03.044)] [Medline: [31051118](https://pubmed.ncbi.nlm.nih.gov/31051118/)]



17. Bednarek PH, Jensen JT. Safety, efficacy and patient acceptability of the contraceptive and non-contraceptive uses of the LNG-IUS. *Int J Womens Health* 2010 Aug 09;1:45-58 [FREE Full text] [doi: [10.2147/ijwh.s4350](https://doi.org/10.2147/ijwh.s4350)] [Medline: [21072274](https://pubmed.ncbi.nlm.nih.gov/21072274/)]
18. Andersson K, Odland V, Rybo G. Levonorgestrel-releasing and copper-releasing (Nova T) IUDs during five years of use: a randomized comparative trial. *Contraception* 1994 Jan;49(1):56-72. [doi: [10.1016/0010-7824\(94\)90109-0](https://doi.org/10.1016/0010-7824(94)90109-0)] [Medline: [8137626](https://pubmed.ncbi.nlm.nih.gov/8137626/)]
19. Beckert V, Ahlers C, Frenz A, Gerlinger C, Bannemerschult R, Lukkari-Lax E. Bleeding patterns with the 19.5 mg LNG-IUS, with special focus on the first year of use: implications for counselling. *Eur J Contracept Reprod Health Care* 2019 Aug;24(4):251-259 [FREE Full text] [doi: [10.1080/13625187.2019.1630817](https://doi.org/10.1080/13625187.2019.1630817)] [Medline: [31223042](https://pubmed.ncbi.nlm.nih.gov/31223042/)]
20. Weisberg E, Bateson D, McGeechan K, Mohapatra L. A three-year comparative study of continuation rates, bleeding patterns and satisfaction in Australian women using a subdermal contraceptive implant or progestogen releasing-intrauterine system. *Eur J Contracept Reprod Health Care* 2014 Feb;19(1):5-14. [doi: [10.3109/13625187.2013.853034](https://doi.org/10.3109/13625187.2013.853034)] [Medline: [24229367](https://pubmed.ncbi.nlm.nih.gov/24229367/)]
21. Bahamondes L, Brache V, Meirik O, Ali M, Habib N, Landoulsi S, WHO Study Group on Contraceptive Implants for Women. A 3-year multicentre randomized controlled trial of etonogestrel- and levonorgestrel-releasing contraceptive implants, with non-randomized matched copper-intrauterine device controls. *Hum Reprod* 2015 Nov;30(11):2527-2538. [doi: [10.1093/humrep/dev221](https://doi.org/10.1093/humrep/dev221)] [Medline: [26409014](https://pubmed.ncbi.nlm.nih.gov/26409014/)]
22. Bateson D, Harvey C, Trinh L, Stewart M, Black KI. User characteristics, experiences and continuation rates of copper intrauterine device use in a cohort of Australian women. *Aust N Z J Obstet Gynaecol* 2016 Dec;56(6):655-661. [doi: [10.1111/ajo.12534](https://doi.org/10.1111/ajo.12534)] [Medline: [27704541](https://pubmed.ncbi.nlm.nih.gov/27704541/)]
23. Dickerson LM, Diaz VA, Jordon J, Davis E, Chirina S, Goddard JA, et al. Satisfaction, early removal, and side effects associated with long-acting reversible contraception. *Fam Med* 2013;45(10):701-707 [FREE Full text] [Medline: [24347187](https://pubmed.ncbi.nlm.nih.gov/24347187/)]
24. Diedrich JT, Desai S, Zhao Q, Secura G, Madden T, Peipert JF. Association of short-term bleeding and cramping patterns with long-acting reversible contraceptive method satisfaction. *Am J Obstet Gynecol* 2015 Jan;212(1):50-58 [FREE Full text] [doi: [10.1016/j.ajog.2014.07.025](https://doi.org/10.1016/j.ajog.2014.07.025)] [Medline: [25046805](https://pubmed.ncbi.nlm.nih.gov/25046805/)]
25. Diamond-Smith N, Campbell M, Madan S. Misinformation and fear of side-effects of family planning. *Cult Health Sex* 2012;14(4):421-433. [doi: [10.1080/13691058.2012.664659](https://doi.org/10.1080/13691058.2012.664659)] [Medline: [22390371](https://pubmed.ncbi.nlm.nih.gov/22390371/)]
26. Dehlendorf C, Levy K, Kelley A, Grumbach K, Steinauer J. Women's preferences for contraceptive counseling and decision making. *Contraception* 2013 Aug;88(2):250-256 [FREE Full text] [doi: [10.1016/j.contraception.2012.10.012](https://doi.org/10.1016/j.contraception.2012.10.012)] [Medline: [23177265](https://pubmed.ncbi.nlm.nih.gov/23177265/)]
27. Dehlendorf C, Krajewski C, Borrero S. Contraceptive counseling: best practices to ensure quality communication and enable effective contraceptive use. *Clin Obstet Gynecol* 2014 Dec;57(4):659-673 [FREE Full text] [doi: [10.1097/GRF.0000000000000059](https://doi.org/10.1097/GRF.0000000000000059)] [Medline: [25264697](https://pubmed.ncbi.nlm.nih.gov/25264697/)]
28. Villavicencio J, Allen RH. Unscheduled bleeding and contraceptive choice: increasing satisfaction and continuation rates. *Open Access J Contracept* 2016;7:43-52 [FREE Full text] [doi: [10.2147/OAJC.S85565](https://doi.org/10.2147/OAJC.S85565)] [Medline: [29386936](https://pubmed.ncbi.nlm.nih.gov/29386936/)]
29. Backman T, Huhtala S, Luoto R, Tuominen J, Rauramo I, Koskenvuo M. Advance information improves user satisfaction with the levonorgestrel intrauterine system. *Obstet Gynecol* 2002 Apr;99(4):608-613. [doi: [10.1016/s0029-7844\(01\)01764-1](https://doi.org/10.1016/s0029-7844(01)01764-1)] [Medline: [12039121](https://pubmed.ncbi.nlm.nih.gov/12039121/)]
30. Halpern V, Lopez LM, Grimes DA, Stockton LL, Gallo MF. Strategies to improve adherence and acceptability of hormonal methods of contraception. *Cochrane Database Syst Rev* 2013 Oct 26(10):CD004317. [doi: [10.1002/14651858.CD004317.pub4](https://doi.org/10.1002/14651858.CD004317.pub4)] [Medline: [24163097](https://pubmed.ncbi.nlm.nih.gov/24163097/)]
31. Gemzell-Danielsson K, Schellschmidt I, Apter D. A randomized, phase II study describing the efficacy, bleeding profile, and safety of two low-dose levonorgestrel-releasing intrauterine contraceptive systems and Mirena. *Fertil Steril* 2012 Mar;97(3):616-622. [doi: [10.1016/j.fertnstert.2011.12.003](https://doi.org/10.1016/j.fertnstert.2011.12.003)] [Medline: [22222193](https://pubmed.ncbi.nlm.nih.gov/22222193/)]
32. Nelson AL. LNG-IUS 12: a 19.5 levonorgestrel-releasing intrauterine system for prevention of pregnancy for up to five years. *Expert Opin Drug Deliv* 2017 Sep;14(9):1131-1140. [doi: [10.1080/17425247.2017.1353972](https://doi.org/10.1080/17425247.2017.1353972)] [Medline: [28696796](https://pubmed.ncbi.nlm.nih.gov/28696796/)]
33. Nelson A, Apter D, Hauck B, Schmelter T, Rybowski S, Rosen K, et al. Two low-dose levonorgestrel intrauterine contraceptive systems: a randomized controlled trial. *Obstet Gynecol* 2013 Dec;122(6):1205-1213. [doi: [10.1097/AOG.000000000000019](https://doi.org/10.1097/AOG.000000000000019)] [Medline: [24240244](https://pubmed.ncbi.nlm.nih.gov/24240244/)]
34. Frenz A, Ahlers C, Beckert V, Gerlinger C, Friede T. Predicting menstrual bleeding patterns with levonorgestrel-releasing intrauterine systems. *Eur J Contracept Reprod Health Care* 2021 Feb;26(1):48-57 [FREE Full text] [doi: [10.1080/13625187.2020.1843015](https://doi.org/10.1080/13625187.2020.1843015)] [Medline: [33269954](https://pubmed.ncbi.nlm.nih.gov/33269954/)]
35. Policy for device software functions and mobile medical applications. US Food and Drug Administration. 2019. URL: <https://www.fda.gov/media/80958/download> [accessed 2021-02-15]
36. 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 2021-02-15]
37. Guidelines relating to the application of the council directive 93/42/EEC on medical devices. European Commission. 2010. URL: <https://ec.europa.eu/docsroom/documents/10337/attachments/1/translations/en/renditions/pdf> [accessed 2021-04-30]



38. Evidence standards framework for digital health technologies: user guide. National Institute for Health and Care Excellence. 2019. URL: <https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies> [accessed 2021-02-15]
39. Ferretti A, Ronchi E, Vayena E. From principles to practice: benchmarking government guidance on health apps. *Lancet Digit Health* 2019 Jun;1(2):55-57 [FREE Full text] [doi: [10.1016/S2589-7500\(19\)30027-5](https://doi.org/10.1016/S2589-7500(19)30027-5)] [Medline: [33323230](https://pubmed.ncbi.nlm.nih.gov/33323230/)]
40. Regulation (EU) 2017/745 of the European Parliament and of the Council. European Commission. 2017. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745> [accessed 2021-02-15]
41. Regulation EU 2017/745. European Commission. 2020. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02017R0745-20200424> [accessed 2021-02-15]
42. Guidance on clinical evaluation (MDR) and performance evaluation (IVDR) of medical device software. European Commission. 2020. URL: <https://ec.europa.eu/docsroom/documents/40323> [accessed 2021-02-15]
43. IEC 62366-1 medical devices - part 1: applicability of usability engineering to medical devices. International Electrotechnical Commission. 2015. URL: <https://www.iso.org/obp/ui/#iso:std:iec:62366:-1:ed-1:v1:en> [accessed 2021-02-15]
44. Ali MM, Cleland J, Shah IH. Causes and consequences of contraceptive discontinuation: evidence from 60 demographic and health surveys. 2012. URL: [https://apps.who.int/iris/bitstream/handle/10665/75429/9789241504058\\_eng.pdf](https://apps.who.int/iris/bitstream/handle/10665/75429/9789241504058_eng.pdf) [accessed 2021-02-15]
45. Backman T, Huhtala S, Blom T, Luoto R, Rauramo I, Koskenvuo M. Length of use and symptoms associated with premature removal of the levonorgestrel intrauterine system: a nation-wide study of 17,360 users. *Br J Obstet Gynaecol* 2000 Mar;107(3):335-339. [doi: [10.1111/j.1471-0528.2000.tb13228.x](https://doi.org/10.1111/j.1471-0528.2000.tb13228.x)] [Medline: [10740329](https://pubmed.ncbi.nlm.nih.gov/10740329/)]
46. Gemzell-Danielsson K, Apter D, Hauck B, Schmelter T, Rybowski S, Rosen K, et al. The effect of age, parity and body mass index on the efficacy, safety, placement and user satisfaction associated with two low-dose levonorgestrel intrauterine contraceptive systems: subgroup analyses of data from a phase III trial. *PLoS One* 2015;10(9):e0135309 [FREE Full text] [doi: [10.1371/journal.pone.0135309](https://doi.org/10.1371/journal.pone.0135309)] [Medline: [26378938](https://pubmed.ncbi.nlm.nih.gov/26378938/)]
47. Simmons RG, Sanders JN, Geist C, Gawron L, Myers K, Turok DK. Predictors of contraceptive switching and discontinuation within the first 6 months of use among Highly Effective Reversible Contraceptive Initiative Salt Lake study participants. *Am J Obstet Gynecol* 2019 Apr;220(4):1-12. [doi: [10.1016/j.ajog.2018.12.022](https://doi.org/10.1016/j.ajog.2018.12.022)]
48. Barden-O'Fallon J, Speizer IS, Calhoun LM, Corroon M. Women's contraceptive discontinuation and switching behavior in urban Senegal, 2010-2015. *BMC Womens Health* 2018 Feb 05;18(1):35 [FREE Full text] [doi: [10.1186/s12905-018-0529-9](https://doi.org/10.1186/s12905-018-0529-9)] [Medline: [29402320](https://pubmed.ncbi.nlm.nih.gov/29402320/)]
49. Bitzer J, Cupanik V, Fait T, Gemzell-Danielsson K, Grob P, Oddens BJ, et al. Factors influencing women's selection of combined hormonal contraceptive methods after counselling in 11 countries: results from a subanalysis of the CHOICE study. *Eur J Contracept Reprod Health Care* 2013 Oct;18(5):372-380. [doi: [10.3109/13625187.2013.819077](https://doi.org/10.3109/13625187.2013.819077)] [Medline: [23941311](https://pubmed.ncbi.nlm.nih.gov/23941311/)]
50. Caetano C, Blikendaal S, Engler Y, Lombardo M. From awareness to usage of long-acting reversible contraceptives: results of a large European survey. *Int J Gynaecol Obstet* 2020 Dec;151(3):366-376 [FREE Full text] [doi: [10.1002/ijgo.13363](https://doi.org/10.1002/ijgo.13363)] [Medline: [32852798](https://pubmed.ncbi.nlm.nih.gov/32852798/)]
51. Harper CC, Blum M, de Bocanegra HT, Darney PD, Speidel JJ, Policar M, et al. Challenges in translating evidence to practice: the provision of intrauterine contraception. *Obstet Gynecol* 2008 Jun;111(6):1359-1369. [doi: [10.1097/AOG.0b013e318173fd83](https://doi.org/10.1097/AOG.0b013e318173fd83)] [Medline: [18515520](https://pubmed.ncbi.nlm.nih.gov/18515520/)]
52. Levy J, Romo-Avilés N. "A good little tool to get to know yourself a bit better": a qualitative study on users' experiences of app-supported menstrual tracking in Europe. *BMC Public Health* 2019 Sep 03;19(1):1213 [FREE Full text] [doi: [10.1186/s12889-019-7549-8](https://doi.org/10.1186/s12889-019-7549-8)] [Medline: [31481043](https://pubmed.ncbi.nlm.nih.gov/31481043/)]
53. Starič KD, Trajković V, Belani H, Vitagliano A, Bukovec P. Smart phone applications for self-monitoring of the menstrual cycle: a review and content analysis. *Clin Exp Obstet Gynecol* 2019;46(5):731-735. [doi: [10.12891/ceog4830.2019](https://doi.org/10.12891/ceog4830.2019)]

---

## Abbreviations

- AI:** artificial intelligence  
**HCP:** health care professional  
**LNG-IUS:** levonorgestrel-releasing intrauterine system  
**MDD:** Medical Device Directive  
**MDR:** Medical Device Regulation
-

*Edited by C Lovis; submitted 28.09.20; peer-reviewed by J Galvez-Olortegui, Z Reis; comments to author 05.12.20; revised version received 04.03.21; accepted 16.04.21; published 13.07.21.*

*Please cite as:*

*Karakoyun T, Podhaisky HP, Frenz AK, Schuhmann-Giampieri G, Ushikusa T, Schröder D, Zvolanek M, Lopes Da Silva Filho A*  
*Digital Medical Device Companion (MyIUS) for New Users of Intrauterine Systems: App Development Study*

*JMIR Med Inform 2021;9(7):e24633*

*URL: <https://medinform.jmir.org/2021/7/e24633>*

*doi: [10.2196/24633](https://doi.org/10.2196/24633)*

*PMID: [34255688](https://pubmed.ncbi.nlm.nih.gov/34255688/)*

©Toeresin Karakoyun, Hans-Peter Podhaisky, Ann-Kathrin Frenz, Gabriele Schuhmann-Giampieri, Thais Ushikusa, Daniel Schröder, Michal Zvolanek, Agnaldo Lopes Da Silva Filho. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 13.07.2021. 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

# Predicting Antituberculosis Drug–Induced Liver Injury Using an Interpretable Machine Learning Method: Model Development and Validation Study

Tao Zhong<sup>1\*</sup>, BSc; Zian Zhuang<sup>2,3,4\*</sup>, BSc; Xiaoli Dong<sup>5,6</sup>, PhD; Ka Hing Wong<sup>5,6</sup>, PhD; Wing Tak Wong<sup>5,6</sup>, PhD; Jian Wang<sup>1</sup>, BSc; Daihai He<sup>2,4</sup>, PhD; Shengyuan Liu<sup>1</sup>, PhD

<sup>1</sup>Department of Tuberculosis Control, Shenzhen Nanshan Center for Chronic Disease Control, Shenzhen, China

<sup>2</sup>Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

<sup>3</sup>Department of Biostatistics, University of California, Los Angeles, CA, United States

<sup>4</sup>Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China

<sup>5</sup>Research Institute for Future Food, The Hong Kong Polytechnic University, Hong Kong, China

<sup>6</sup>Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hong Kong, China

\*these authors contributed equally

**Corresponding Author:**

Shengyuan Liu, PhD

Department of Tuberculosis Control

Shenzhen Nanshan Center for Chronic Disease Control

Hua Ming Road No 7

Nanshan District

Shenzhen, 518000

China

Phone: 86 13543301395

Email: [jfk@sznsmbj.com](mailto:jfk@sznsmbj.com)

**Related Article:**

This is a corrected version. See correction statement: <https://medinform.jmir.org/2021/8/e32415>

## Abstract

**Background:** Tuberculosis (TB) is a pandemic, being one of the top 10 causes of death and the main cause of death from a single source of infection. Drug-induced liver injury (DILI) is the most common and serious side effect during the treatment of TB.

**Objective:** We aim to predict the status of liver injury in patients with TB at the clinical treatment stage.

**Methods:** We designed an interpretable prediction model based on the XGBoost algorithm and identified the most robust and meaningful predictors of the risk of TB-DILI on the basis of clinical data extracted from the Hospital Information System of Shenzhen Nanshan Center for Chronic Disease Control from 2014 to 2019.

**Results:** In total, 757 patients were included, and 287 (38%) had developed TB-DILI. Based on values of relative importance and area under the receiver operating characteristic curve, machine learning tools selected patients' most recent alanine transaminase levels, average rate of change of patients' last 2 measures of alanine transaminase levels, cumulative dose of pyrazinamide, and cumulative dose of ethambutol as the best predictors for assessing the risk of TB-DILI. In the validation data set, the model had a precision of 90%, recall of 74%, classification accuracy of 76%, and balanced error rate of 77% in predicting cases of TB-DILI. The area under the receiver operating characteristic curve score upon 10-fold cross-validation was 0.912 (95% CI 0.890-0.935). In addition, the model provided warnings of high risk for patients in advance of DILI onset for a median of 15 (IQR 7.3-27.5) days.

**Conclusions:** Our model shows high accuracy and interpretability in predicting cases of TB-DILI, which can provide useful information to clinicians to adjust the medication regimen and avoid more serious liver injury in patients.

**KEYWORDS**

accuracy; drug; drug-induced liver injury; high accuracy; injury; interpretability; interpretation; liver; machine learning; model; prediction; treatment; tuberculosis; XGBoost algorithm

## Introduction

Tuberculosis (TB) is an infectious disease caused by the bacillus *Mycobacterium tuberculosis*. It is one of the top 10 causes of death worldwide and the leading cause of death from a single infectious disease [1]. In 2019, approximately 10 million people were diagnosed with TB and 1.4 million people died worldwide [1]. To prevent the spread of pulmonary TB, timely and effective anti-TB treatment is very important [2]. First-line anti-TB drugs include pyrazinamide (PZA), ethambutol (EMB), isoniazid (INH), and rifampin (RIF) [3-6]. When treating patients with TB, drug-induced liver injury (DILI) is the most frequent and serious side effect [7-10]. Among various populations, the incidence of TB-DILI ranges from 2.3% to 27.7% during anti-TB therapy [11-14]. Researchers have suggested that anti-TB drugs are hepatotoxic [11,15-18].

TB-DILI may result from direct toxic injury to hepatocytes by anti-TB drugs or their metabolites or immune-mediated liver injury and induction of hepatocyte apoptosis caused by anti-TB drugs that trigger multiple inflammatory immune pathways [11,19]. TB-DILI is characterized by a transient mild elevation of transaminases or acute hepatitis [20]. Fulminant hepatic failure is likely to develop in severe cases, whereas chronic hepatitis occurs in a minority of patients.

Currently, clinical liver tests usually include biochemical parameters of blood, such as transaminases including alanine transaminase (ALT), alkaline phosphatase, bilirubin, lactate dehydrogenase, and albumin, along with liver imaging and histopathologic evaluation. It is difficult to distinguish DILI from non-DILI on the basis of these indicators, since test results are largely consistent in DILI and non-DILI detection. In addition, clinical markers commonly used at present, accounting for neither differences in type and mechanisms of action of hepatotoxic drugs nor individual patients' characteristics, only facilitate evaluation based on toxicity outcomes [21]. Therefore, identification of predictors at clinical stages and risk predictors of TB-DILI among patients has become an urgent and necessary task.

Previous studies have shown that TB-DILI is associated with some demographic characteristics and underlying chronic disease [12,22-26]. Patterson et al [27] suggested that an increase in pretreatment ALT and the gradient of ALT changes increase the risk of late TB-DILI. Thus, in addition to the cumulative anti-TB drug dose, ALT levels and demographic variable such as age, gender, education level, income, and BMI were included in our model as predictors. Various models are used to identify drugs associated with the risk of DILI at the preclinical stage [28]. Machine learning models have demonstrated strong predictive power and retained a simple form for communication with researchers [29-39]. XGBoost is a boosting ensemble machine learning algorithm that integrates a few classification

and regression trees models to form a strong classifier [40,41]. It performs well in dealing with nonlinear and complex relationships among variables [42]. We designed an interpretable prediction model by using the XGBoost algorithm and identified the most robust and meaningful predictors of the risk of TB-DILI. Then, using these discriminative predictors, the machine learning model built an interpretable decision tree to provide early warning signals before TB-DILI occurs, so as to help clinicians adjust the medication plan in time and potentially reduce the possibility of TB-DILI. In this study, we retrospectively assessed 757 patients with TB who were registered for treatment in Nanshan District (Shenzhen, China) from 2014 to 2019.

## Methods

### Data

We extracted data on 757 pulmonary TB cases registered in the Hospital Information System of Shenzhen Nanshan Center for Chronic Disease Control from 2014 to 2019, including those that are smear-positive and undergoing initial treatment. Some patients did not have continuous treatment or were initially discharged from hospital and subsequently rehospitalized, resulting in the recorded treatment duration exceeding the normal range and unclear cumulative dosage of anti-TB drugs. Such abnormal cases are not able to contribute to predictions among patients receiving regular treatment. Thus, we selected 300 days as a time-window empirically on the basis of the typical course of TB treatment [1]. We excluded cases of TB-DILI that were recorded 300 days after the start of the anti-TB treatments. In total, data from 743 patients were finally included in the model. We defined patients as positive DILI cases in accordance with the American Thoracic Society criteria [11]: in the presence of hepatitis symptoms, the increase in ALT levels was 3-fold the normal upper limit, and in the absence of hepatitis symptoms, this increase was 5-fold the normal upper limit.

Patients' demographic and clinical data included gender, age, weight, education level, income, height, hepatitis B status, diabetes status, cumulative anti-TB drug dose, and ALT levels. For patients who did not develop TB-DILI, we collected their total amount of prescribed anti-TB medication as of the latest hepatic examination. For patients who developed TB-DILI, we recorded their cumulative dose of anti-TB medication as of the time when TB-DILI was detected. In addition, we measured the patient's most recent ALT levels before the last hepatic examination, and the average rate of change of the last 2 ALT levels tested before the final liver function test. We calculated the cumulative dose of each drug separately (PZA, RFP, EMB, and INH) for combination drug therapy.

Upon initiation of therapy, the patients were segregated to form the training and validation data sets. The data of patients



admitted before April 2019 (607 patients and 186 smear-positive cases) and after April 2019 (136 patients and 95 smear-positive cases) were included in the training and validation data sets, respectively.

### Descriptive Statistics

Descriptive statistics were calculated for positive and TB-DILI cases. Demographic and laboratory data of the 2 groups were compared using 2-sample *t* tests for normally distributed continuous variables, the Kruskal–Wallis rank sum test for nonnormally distributed continuous variables, and chi-square tests for categorical variables. Missing values were omitted when tested for differences. [Multimedia Appendix 1](#) shows the proportion of missing values for each variable.

### Prediction Model

We used the XGBoost algorithm for the prediction model [41]. XGBoost is a high-performance machine learning algorithm based on the tree boosting system [43-47]. It uses a sparsity-aware learning algorithm to process sparse data and weighted quantile sketch to approximate tree learning [41]. Since the decision tree is a simple classifier composed of hierarchically organized dichotomous determinations, its structure also demonstrates good interpretability [48-50]. In addition, the model can deal with missing values well. When the model searches for the best candidate split criteria for tree growth, they will also assign a default direction for the missing values on those nodes [41]. The interpretable criteria and high tolerance for missing data in the decision tree make the model robust and meaningful when dealing with clinical data. To obtain a model that can be conveniently applied in a clinical setting, we attempted to reduce the complexity of the model as much as possible. Hence, we choose the single-tree XGBoost algorithm as the prediction model.

To build the model, we first included all demographic and clinical data as predictors. The dependent variable is DILI status, which is a binary outcome. We trained the single-tree XGBoost algorithm with the training set. By considering each feature's contribution for each tree in the model, we determined their relative importance to the tree model [51]. We repeated stratified 10-fold cross-validation 100 times to model on the training data set to obtain the mean value of each feature's relative importance. Then, we arranged the top 10 predictors in accordance with their relative importance. The predictors were added into the model individually in descending order of relative importance to form 10 candidate models. We repeated stratified 10-fold cross-validation 100 times to the candidate models on the basis of the training data set to obtain the mean area under the receiver operating characteristic curve (AUC) and selected the model with the maximum AUC as the final model. Then, we trained the selected model with the whole training data set to obtain the interpretable decision tree. The detailed process of the stratified *k*-fold cross-validation and the parameters set in model is provided in [Multimedia Appendix 1](#).

### Evaluation of Model Performance

We trained the model with the whole training data set and applied the model on the validation data set. We then evaluated

the prediction results on the basis of the confusion matrix, which is a specific table to visualize the performance of a classification model [52]. In accordance with the confusion matrix, we calculated the value of the following evaluation indicators: precision, recall, F1 value, classification accuracy, and balanced error rate. Detailed descriptions of the formulae for the indicators are provided in [Multimedia Appendix 1](#). To determine whether the model can send early an warning signal in time, we also calculated the duration from the timepoint when model sent the warning signal to the actual date of TB-DILI diagnosis among incorrectly classified cases. Meanwhile, we compared the performance (AUC) of the single-tree XGBoost algorithm with that of the multitree XGBoost algorithm, logistic regression, single-tree random forest algorithm, and multitree random forest algorithm through 10-fold cross-validation using the whole data set. We determined 95% CI values for AUC values with the DeLong method [53]. We applied selected variables to train the single-tree XGBoost model since variable selection is part of the whole algorithm. The complete data set was applied to train the other models. In addition, we applied multiple imputation by chained equations [54] to address missing data for logistic regression.

### Sensitivity Analysis

We selected 250 days and 350 days as alternative time windows to filter data. Then, we trained the model and compared the selected predictors. Performance (AUC) of the original model and that of 2 alternative models were also compared on the basis of the whole data set through 10-fold cross-validation. All analyses were performed with R (version 4.0.4, The R Foundation). The codes used in this study can be found in the GitHub repository [55].

## Results

In total, 743 patients were included in the analysis, of whom 281 (37.8%) and 462 (62.2%) were classified as TB-DILI-positive and -negative, respectively. [Table 1](#) shows the descriptive statistics. The median age of patients was 30 (IQR 25-45) years, and 484 (65.1%) patients were male. Most patients (*n*=272, 43.5%) had a bachelor's degree or higher education level. Median weight of the patients was 56 (IQR 50-63) kg and their median height was 168 (IQR 160-173) cm. In total, 24 (3.2%) patients had hepatitis B, and 69 (9.3%) patients had diabetes. The proportion of male patients who had DILI (*n*=281, 74.0%) was significantly higher than that of patients who did not have DILI (*n*=276, 59.7%). The most recently determined ALT level and average rate of change of the last 2 ALT measures of patients with DILI (27.0 U/L, IQR 17.0-34.0 U/L and 0.27 U/[Lday], IQR 0.0-0.6 U/[Lday], respectively) were significantly higher than those of patients who did not have DILI (11.0 U/L, IQR 8.3-16.0 U/L and 0.0 U/[Lday], IQR -0.1 to 0.1 U/[Lday], respectively). [Figure 1](#) shows the number of TB-DILI cases on each day after the initiation of anti-TB treatment. The median time from treatment to the onset of DILI is 27 (IQR 15-48) days.

**Table 1.** Demographic and clinical characteristics of patients (N=743).

Characteristics	Overall	Negative cases (n=462)	Positive cases (n=281)	P value
Males, n (%)	484 (65.1)	276 (59.7)	208 (74.0)	<.001
Age <sup>a</sup> (years), median (IQR)	30 (25-45)	30 (25-45)	31 (25-44)	.62
Weight <sup>a</sup> (kg), median (IQR)	56.0 (50.0-63.0)	55.00 (49.0-63.0)	57.0 (51.5-63.0)	.06
<b>Education level, n (%)</b>				.55
Lower than middle school	201 (32.2)	123 (33.1)	78 (30.8)	
Middle school	152 (24.3)	93 (25.0)	59 (23.3)	
Bachelor's degree or higher	272 (43.5)	156 (41.9)	116 (45.8)	
Income <sup>a</sup> (RMB <sup>b</sup> ), median (IQR)	500,000 (300,000-800,000)	500,000 (300,000-800,000)	600,000 (500,000-1,000,000)	.002
Height <sup>a</sup> (cm), median (IQR)	168.0 (160.0-173.0)	167.0 (160.0-172.0)	168.0 (162.0-173.0)	.03
Hepatitis B, n (%)	24 (3.2)	16 (3.5)	8 (2.8)	.81
Diabetes, n (%)	69 (9.3)	48 (10.4)	21 (7.5)	.23
BMI <sup>a</sup> , median (IQR)	20.0 (18.5-22.1)	19.9 (18.4-22.0)	20.2 (18.7-22.2)	.39
Pyrazinamide dose <sup>a</sup> (g), median (IQR)	16.8 (3.0-60.0)	24.0 (3.1-87.9)	5.4 (3.0-25.6)	<.001
Rifampicin dose <sup>a</sup> (g), median (IQR)	13.5 (1.3-67.5)	40.5 (5.5-94.5)	3.2 (1.2-12.6)	<.001
Ethambutol dose <sup>a</sup> (g), median (IQR)	18.7 (2.2-91.1)	50.3 (4.4-139.2)	5.3 (2.2-18.8)	<.001
Isoniazid dose <sup>a</sup> (g), median (IQR)	8.1 (0.8-37.0)	22.8 (3.6-58.3)	2.0 (0.6-6.5)	<.001
Recent alanine transaminase measurement <sup>a,c</sup> (U/L), median (IQR)	13.0 (10.0-23.0)	11.0 (8.3-16.0)	27.0 (17.0-34.0)	<.001
Rate of change in alanine transaminase levels <sup>a,d</sup> (U/[Lday]), median (IQR)	0.0 (-0.1 to 0.1)	0.00 (-0.1 to 0.1)	0.27 (0.0 to 0.6)	<.001

<sup>a</sup>Nonnormally distributed variables.

<sup>b</sup>1 RMB=US \$0.15.

<sup>c</sup>Patients' most recently determined alanine transaminase level before the latest hepatic examination.

<sup>d</sup>Average rate of change of the patients' last 2 alanine transaminase measures before the final liver function test (increment divided by the duration).

**Figure 1.** Days from tuberculosis treatment to the onset of drug-induced liver injury among the patients in our study. DILI: drug-induced liver injury, TB: tuberculosis.

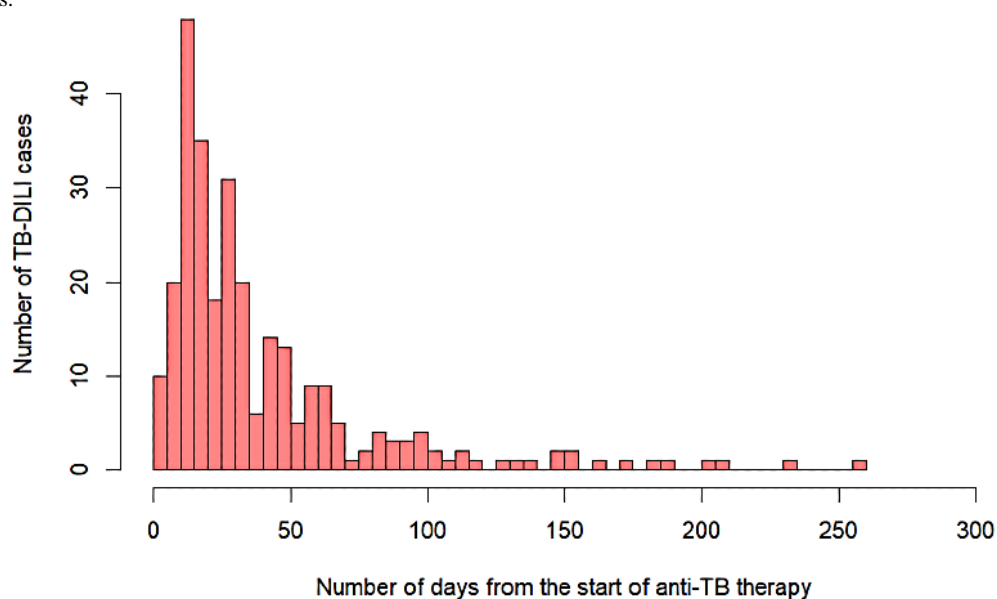
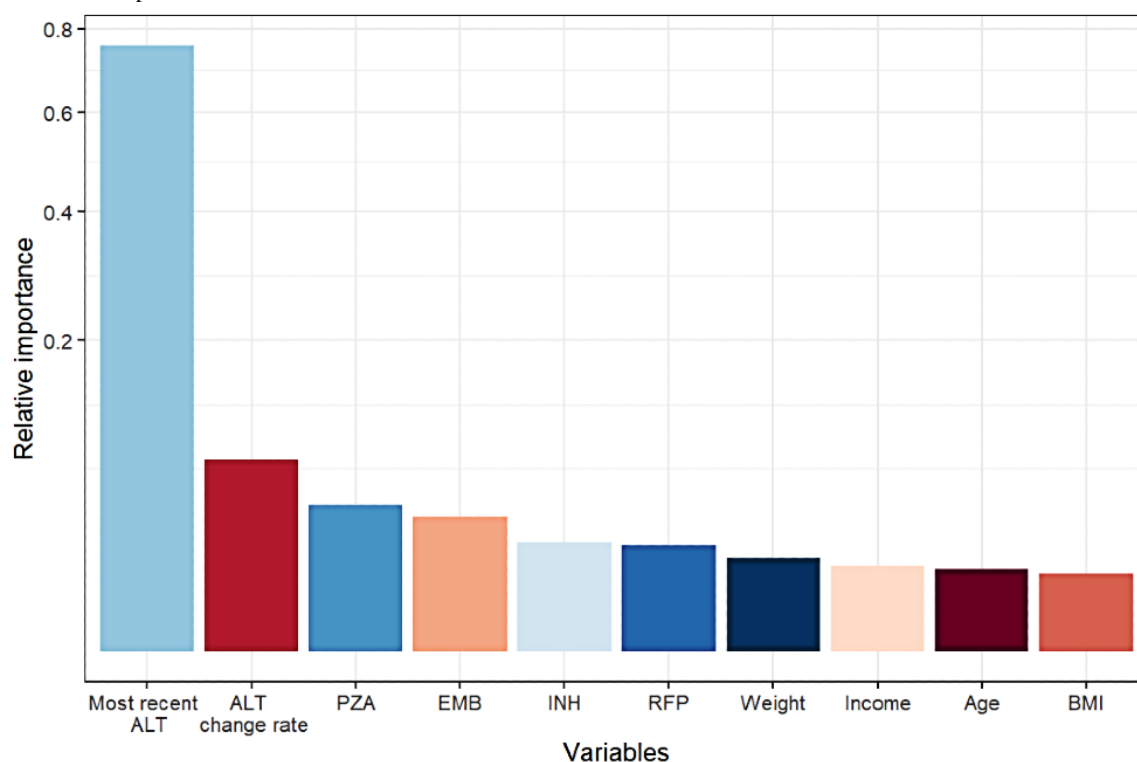


Figure 2 shows the top 10 important variables selected by the single-tree XGBoost model. The most recent ALT levels were found to be the most important factor in the prediction process. We added 10 variables in the model individually to form 10 candidate models. After 10-fold cross-validation 100 times with the training and testing data sets, the model with 4 variables had the maximum AUC value (Table 2). Thus, we selected the model with 4 variables (the most recent ALT measure, average

rate of change of the last 2 ALT measures, cumulative dose of PZA, and cumulative dose of EMB) as the final model. Then, we trained the selected model with the whole training data set. Figure 3 shows the content of a single decision tree of the model. The decision process starts from the most recent ALT test value, and then dichotomous determinations are made at each node in the decision tree; this process ends with outputting predictions (high or low risk of DILI).

**Figure 2.** Top 10 important variables selected by the single-tree XGBoost model. ALT: alanine transaminase, EMB: ethambutol, INH: isoniazid, PZA: pyrazinamide, RFP: rifampicin.



**Table 2.** Summary of AUC<sup>a</sup> values for candidate model.

Candidate model	Variables, n	AUC, mean (SD)
1	1	0.908 (0.043)
2	2	0.912 (0.040)
3	3	0.913 (0.041)
4 (selected model)	4	0.918 (0.040)
5	5	0.917 (0.040)
6	6	0.915 (0.040)
7	7	0.913 (0.040)
8	8	0.913 (0.041)
9	9	0.912 (0.041)
10	10	0.911 (0.041)

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

**Figure 3.** Detailed overview of the single decision tree of the model. The decision process starts from the left (most recent ALT measure) and ends at the right ("Yes": high risk of drug-induced liver injury or "No": low risk of drug-induced liver injury). Dichotomous determinations are made at every node in the decision tree. Cumulative doses of PZA and EMB are referenced. Black paths are the default direction for missing values. ALT: alanine transaminase, EMB: ethambutol, PZA: pyrazinamide, RFP: rifampicin.

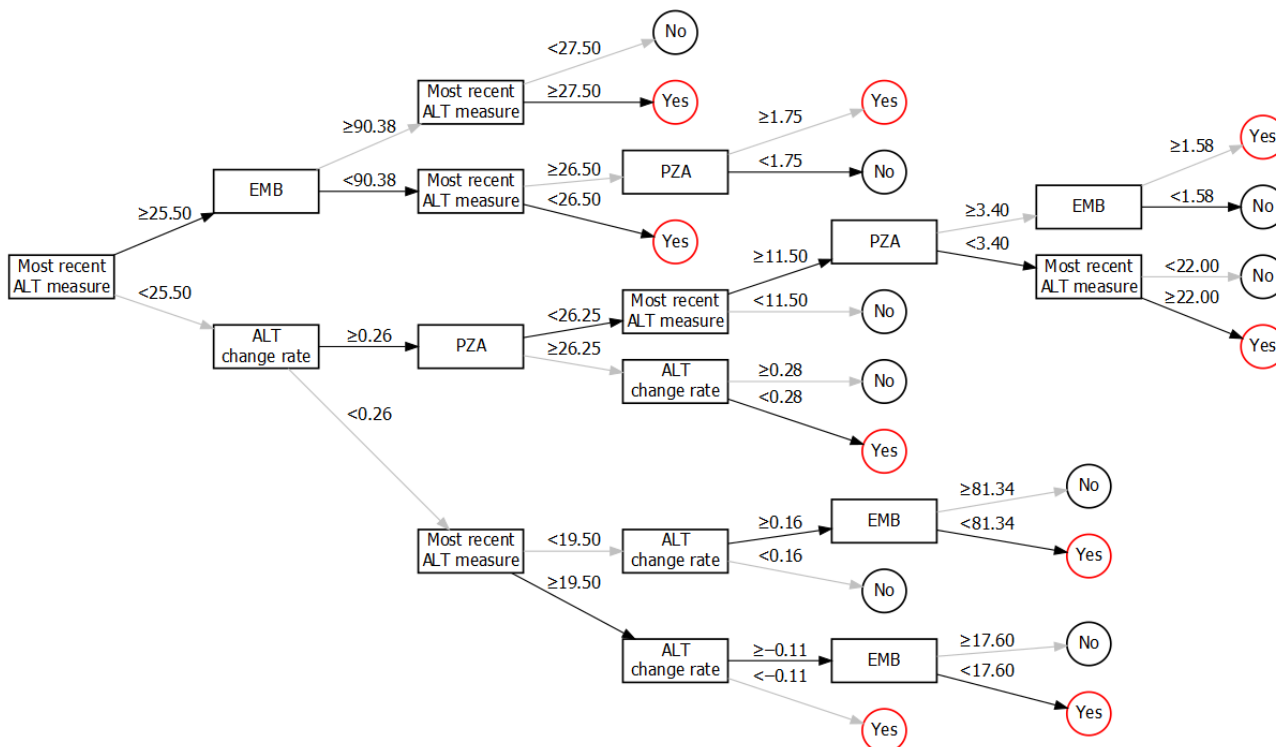


Table 3 summarizes the performance of the model on basis of the validation data set (136 cases). A total of 70 cases of DILI were correctly predicted, and 33 negative cases were successfully predicted. The model had a precision of 90%, recall of 74%, classification accuracy of 76%, balanced error rate of 77%, and F1 value of 81%. For correctly predicted cases, the median number of days between DILI onset and the provision of warnings of high risk by the model for the patients was 15 (IQR 7.3-27.5) days (Figure 4). Multimedia Appendix 1 shows a comparison of the performance of the single-tree XGBoost model and the multitree XGBoost model, logistic regression model, and multi- or single-tree random forest model on the whole data set, based on the receiver operating characteristic curve and the AUC. The multitree XGBoost model performed the best (AUC=0.940, 95% CI 0.924-0.956). The single-tree XGBoost model had an AUC of 0.912 (95% CI 0.890-0.935),

which was very similar to that of the multitree model and higher than that of the rest of the models.

Table 4 shows the AUC values for candidate models under different time windows upon sensitivity analysis. Both final models under different time windows included the 4 most important predictors, same as those of our original model. The most recent ALT measure, average rate of change of patients' last 2 ALT measures, and cumulative dose of PZA were identified as the best predictors in all 3 models. Nevertheless, our original model also selected the cumulative dose of EMB as an important predictor and, while being trained on the basis of 250-day and 350-day time windows, selected cumulative doses of RFP and INH, respectively. The performance of the 3 models is summarized in Multimedia Appendix 1, which shows that all 3 models have similar patterns of the receiver operating characteristic curve and provided largely consistent AUC values.

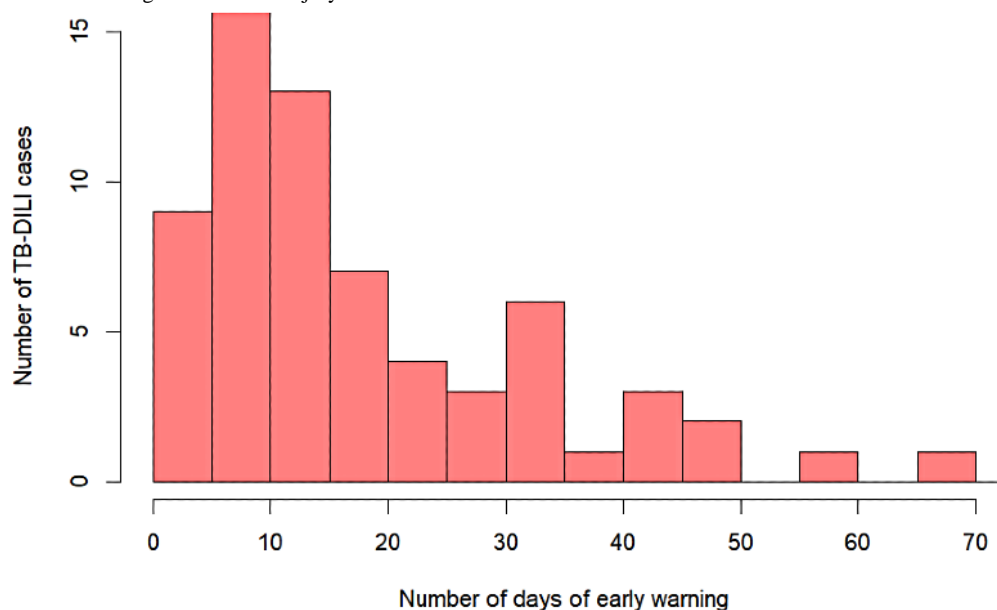
**Table 3.** Model performance<sup>a</sup> with the validation data set.

Prediction or reference model	Yes	No
Yes, n	70	8
No, n	25	33

<sup>a</sup>Precision=90%, recall=74%, F1 value=81%, classification accuracy=76%, and balanced error rate=77%.



**Figure 4.** Number of days between the onset of drug-induced liver injury and the model providing warnings of high risk for patients with TB-DILI. TB-DILI: tuberculosis with drug-induced liver injury.



**Table 4.** Summary of AUC<sup>a</sup> values for candidate models upon sensitivity analysis.

Candidate model	Variables, n	AUC of the model with a 250-day time window, mean (SD)	AUC of the model with a 350-day time window, mean (SD)
1	1	0.910 (0.040)	0.913 (0.042)
2	2	0.916 (0.039)	0.911 (0.040)
3	3	0.920 (0.039)	0.915 (0.041)
4 (selected model)	4	0.922 (0.039)	0.915 (0.041)
5	5	0.921 (0.039)	0.915 (0.041)
6	6	0.918 (0.038)	0.915 (0.042)
7	7	0.918 (0.039)	0.915 (0.041)
8	8	0.917 (0.039)	0.913 (0.041)
9	9	0.916 (0.039)	0.913 (0.041)
10	10	0.916 (0.039)	0.912 (0.041)

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

## Discussion

### Principal Findings

Anti-TB drugs are one of the most common and effective means of treating TB in the clinical setting and can effectively control disease progression among patients with TB. Nevertheless, studies have suggested that patients are likely to develop DILI during the treatment process owing to the hepatotoxicity of anti-TB drugs [11,15-18] and long duration of TB treatment [56]. Clinicians often have difficulties in predicting the efficacy of anti-TB treatment as well as liver injury status in patients with TB. To identify reliable and accurate predictors and better predict DILI during TB treatment, we built the single-tree XGBoost machine learning model and selected variables with significant effects. Our model can provide suggestions to clinicians to adjust their medication regimens in a timely manner to avoid causing more severe liver injury. To our knowledge,

this is the first time that XGBoost model has been applied to predict DILI at the clinical treatment stage.

Interestingly, the proportion of TB-DILI cases is significantly higher among men compared than among women (Table 1). This result is consistent with that of Chang et al [23], which suggested that males were 2.1-fold more likely to develop hepatotoxicity than females after being adjusted for age. We found that patients with DILI had significantly higher values for the most recent ALT measure and higher mean rates of change between the 2 most recent ALT measures than those without DILI (Table 1). Singanayagam et al [57] also demonstrated the association between pretreatment ALT and 2-week on-treatment ALT levels in patients with DILI.

Based on the results of variable selection, the significant predictors for predicting DILI are the most recent ALT measure, average change rate of the last 2 ALT measures, and the cumulative doses of PZA and EMB. According to the decision

tree (Figure 3), the decision process of our model was to initially focus on the most recent ALT measure of a patient and then comprehensively evaluate the rate of change in ALT levels and the cumulative intake dose of both PZA and EMB to make predictions. A previous study [58] reported that the initial concentration of PZA and its metabolites are associated with hepatotoxicity [58]. Cao et al [59] suggested that combination therapy with PZA, INH, and RIF is likely to increase the risk of hepatotoxicity compared to monotherapy with INH and RIF. In addition, the addition of a low dose of PZA to a regimen of INH, RIF, and EMB did not significantly increase the incidence of DILI in the first 2 months of anti-TB therapy [60]. In the branches of our single-tree model (Figure 3), thresholds to determine whether the cumulative dose of anti-TB drugs contributed to the development of DILI under different situations were also provided.

Currently, various machine learning algorithms have been assessed for early detection of DILI and have shown to have a high prediction accuracy [61,62]. Xu et al [63] proposed a deep learning model, which achieved a classification accuracy of 86.9% in external validation for DILI prediction after training with a set of 475 samples. Dominic et al [64] combined mechanistic detection of hepatic safety with a Bayesian machine learning algorithm to build the model, which has a balanced accuracy of 86%, sensitivity of 87%, and specificity of 85%, thus improving the prediction of DILI risk. In addition, the XGBoost model was applied to increase the specificity of mass TB screening [65]. Our model also demonstrated the high prediction accuracy and interpretability of the XGBoost model at the clinical treatment stage.

Compared with alternative models, the multitree XGBoost model performed the best, as revealed from the AUC value upon cross-validation (Multimedia Appendix 1). The single-tree XGBoost model displayed similar performance to that of the multitree XGBoost model. Since the single-tree model is easier to interpret, takes up fewer computing resources, and provides predictions in a shorter period of time, the single-tree XGBoost model is more suitable in the clinical setting. In addition, the single-tree XGBoost model performed better than multitree random forest model and single-tree random forest model. The

logistic regression model is also interpretable. Nevertheless, since linear models cannot directly process missing values, the missing clinical data could affect the performance of logistic regression. For multiple imputation, additional assumptions and prior information are required, which is likely to complicate the process and reduce the robustness of the model. In addition, sensitivity analysis has shown that our model has a consistently high prediction accuracy when trained with different time windows. Therefore, the single-tree XGBoost model is the most appropriate among all candidate models.

### Limitations

Our model also has some limitations of note. First, although the model identified the most meaningful predictors for the risk of TB-DILI, pathological conclusions should be made cautiously since the model was entirely driven by the data input. The model needs to be adjusted accordingly when the data are updated. Second, there is a lack of validation for our model on other data sets. Future studies could further explore these issues by applying the model in a combined larger data set. Inputting more data is likely to contribute to the identification of more effective predictors and generate higher prediction accuracy. In addition, it is also necessary to validate the model's performance on an imbalanced data set to determine whether a further reweighting or resampling is needed to improve prediction accuracy.

### Conclusions

We developed a single-tree XGBoost model, which demonstrated the patients' most recent ALT measure, average rate of change of patients' 2 latest ALT measures, and cumulative doses of PZA and EMB as the best predictors for assessing the DILI risk. In the validation data set, the model displayed high accuracy (precision=90%, recall=74%, classification accuracy=76%, and balanced error rate=77%) and interpretability in predicting the TB-DILI cases. In addition, the median number of days between the model providing warnings of high risk among patients and DILI onset is 15 (IQR 7.3-27.5) days, which suggests that it is possible for clinicians to adjust the medication regimen by referring to the model's prediction and avoid causing more serious liver injury.

### Acknowledgments

Funding was obtained from Shenzhen Science and Technology Innovation Commission: Research on Early Warning Model of Drug-induced Liver Injury in Tuberculosis Patients Based on Machine Learning (award# JCYJ20190809153201668) and the Sanming Project of Medicine in Shenzhen (award# SZSM201603029).

### Conflicts of Interest

None declared.

Multimedia Appendix 1  
Supplementary material.

[DOCX File, 5878 KB - [medinform\\_v9i7e29226\\_app1.docx](#) ]

### References

1. OECD and World Health Organization. Tuberculosis. In: Health at a Glance: Asia/Pacific 2020. Measuring Progress Towards Universal Health Coverage. Paris: OECD Publishing; 2020.
2. Albert RK, Iseman M, Sbarbaro JA, Stage A, Pierson DJ. Monitoring patients with tuberculosis for failure during and after treatment. *Am Rev Respir Dis* 1976 Dec;114(6):1051-1060. [doi: [10.1164/arrd.1976.114.6.1051](https://doi.org/10.1164/arrd.1976.114.6.1051)] [Medline: [827221](https://pubmed.ncbi.nlm.nih.gov/827221/)]
3. Isa S, Ebonyi A, Shehu N, Idoko P, Anejo-Okopi J, Simji G, et al. Antituberculosis drugs and hepatotoxicity among hospitalized patients in Jos, Nigeria. *Int J Mycobacteriol* 2016 Mar;5(1):21-26 [FREE Full text] [doi: [10.1016/j.ijmyco.2015.10.001](https://doi.org/10.1016/j.ijmyco.2015.10.001)] [Medline: [26927986](https://pubmed.ncbi.nlm.nih.gov/26927986/)]
4. Shakya R, Rao BS, Shrestha B. Incidence of hepatotoxicity due to antitubercular medicines and assessment of risk factors. *Ann Pharmacother* 2004 Jun;38(6):1074-1079. [doi: [10.1345/aph.1D525](https://doi.org/10.1345/aph.1D525)] [Medline: [15122004](https://pubmed.ncbi.nlm.nih.gov/15122004/)]
5. An H, Wu X, Wang Z, Xu J, Zheng S, Wang K. The clinical characteristics of anti-tuberculosis drug induced liver injury in 2457 hospitalized patients with tuberculosis in China. *Afr J Pharm Pharmacol* 2013;7(13):710-714. [doi: [10.5897/AJPP2013.2963](https://doi.org/10.5897/AJPP2013.2963)]
6. Dheda K, Barry CE, Maartens G. Tuberculosis. *Lancet* 2016 Mar;387(10024):1211-1226. [doi: [10.1016/S0140-6736\(15\)00151-8](https://doi.org/10.1016/S0140-6736(15)00151-8)]
7. Jeong I, Park JS, Cho YJ, Yoon HI, Song J, Lee CT, et al. Drug-induced hepatotoxicity of anti-tuberculosis drugs and their serum levels. *J Korean Med Sci* 2015 Feb;30(2):167-172 [FREE Full text] [doi: [10.3346/jkms.2015.30.2.167](https://doi.org/10.3346/jkms.2015.30.2.167)] [Medline: [25653488](https://pubmed.ncbi.nlm.nih.gov/25653488/)]
8. Anand A, Seth A, Paul M, Puri P. Risk Factors of Hepatotoxicity During Anti-tuberculosis Treatment. *Med J Armed Forces India* 2006 Jan;62(1):45-49 [FREE Full text] [doi: [10.1016/S0377-1237\(06\)80155-3](https://doi.org/10.1016/S0377-1237(06)80155-3)] [Medline: [27407844](https://pubmed.ncbi.nlm.nih.gov/27407844/)]
9. Tostmann A, Boeree M, Aarnoutse R, de Lange WCM, van der Ven AJAM, Dekhuijzen R. Antituberculosis drug-induced hepatotoxicity: concise up-to-date review. *J Gastroenterol Hepatol* 2008 Feb;23(2):192-202. [doi: [10.1111/j.1440-1746.2007.05207.x](https://doi.org/10.1111/j.1440-1746.2007.05207.x)] [Medline: [17995946](https://pubmed.ncbi.nlm.nih.gov/17995946/)]
10. Wondwossen A, Waqtoola C, Gemeda A. Incidence of antituberculosis-drug-induced hepatotoxicity and associated risk factors among tuberculosis patients in Dawro Zone, South Ethiopia: A cohort study. *Int J Mycobacteriol* 2016 Mar;5(1):14-20 [FREE Full text] [doi: [10.1016/j.ijmyco.2015.10.002](https://doi.org/10.1016/j.ijmyco.2015.10.002)] [Medline: [26927985](https://pubmed.ncbi.nlm.nih.gov/26927985/)]
11. Saukkonen JJ, Cohn DL, Jasmer RM, Schenker S, Jereb JA, Nolan CM, ATS (American Thoracic Society) Hepatotoxicity of Antituberculosis Therapy Subcommittee. An official ATS statement: hepatotoxicity of antituberculosis therapy. *Am J Respir Crit Care Med* 2006 Oct 15;174(8):935-952. [doi: [10.1164/rccm.200510-1666ST](https://doi.org/10.1164/rccm.200510-1666ST)] [Medline: [17021358](https://pubmed.ncbi.nlm.nih.gov/17021358/)]
12. Abbara A, Chitty S, Roe JK, Ghani R, Collin SM, Ritchie A, et al. Drug-induced liver injury from antituberculous treatment: a retrospective study from a large TB centre in the UK. *BMC Infect Dis* 2017 Mar 24;17(1):231 [FREE Full text] [doi: [10.1186/s12879-017-2330-z](https://doi.org/10.1186/s12879-017-2330-z)] [Medline: [28340562](https://pubmed.ncbi.nlm.nih.gov/28340562/)]
13. Singanayagam A, Sridhar S, Dhariwal J, Abdel-Aziz D, Munro K, Connell DW, et al. A comparison between two strategies for monitoring hepatic function during antituberculous therapy. *Am J Respir Crit Care Med* 2012 Mar 15;185(6):653-659. [doi: [10.1164/rccm.201105-0850OC](https://doi.org/10.1164/rccm.201105-0850OC)] [Medline: [22198973](https://pubmed.ncbi.nlm.nih.gov/22198973/)]
14. Yee D, Valiquette C, Pelletier M, Parisien I, Rocher I, Menzies D. Incidence of serious side effects from first-line antituberculosis drugs among patients treated for active tuberculosis. *Am J Respir Crit Care Med* 2003 Jun 01;167(11):1472-1477. [doi: [10.1164/rccm.200206-626OC](https://doi.org/10.1164/rccm.200206-626OC)] [Medline: [12569078](https://pubmed.ncbi.nlm.nih.gov/12569078/)]
15. Steele MA, Burk RF, DesPrez RM. Toxic hepatitis with isoniazid and rifampin. A meta-analysis. *Chest* 1991 Feb;99(2):465-471. [doi: [10.1378/chest.99.2.465](https://doi.org/10.1378/chest.99.2.465)] [Medline: [1824929](https://pubmed.ncbi.nlm.nih.gov/1824929/)]
16. Yew W, Leung C. Antituberculosis drugs and hepatotoxicity. *Respirology* 2006 Nov;11(6):699-707. [doi: [10.1111/j.1440-1843.2006.00941.x](https://doi.org/10.1111/j.1440-1843.2006.00941.x)] [Medline: [17052297](https://pubmed.ncbi.nlm.nih.gov/17052297/)]
17. Younossian AB, Rochat T, Ketterer J, Wacker J, Janssens J. High hepatotoxicity of pyrazinamide and ethambutol for treatment of latent tuberculosis. *Eur Respir J* 2005 Sep;26(3):462-464 [FREE Full text] [doi: [10.1183/09031936.05.00006205](https://doi.org/10.1183/09031936.05.00006205)] [Medline: [16135729](https://pubmed.ncbi.nlm.nih.gov/16135729/)]
18. Shu C, Lee C, Lee M, Wang J, Yu C, Lee L. Hepatotoxicity due to first-line anti-tuberculosis drugs: a five-year experience in a Taiwan medical centre. *Int J Tuberc Lung Dis* 2013 Jul;17(7):934-939. [doi: [10.5588/ijtld.12.0782](https://doi.org/10.5588/ijtld.12.0782)] [Medline: [23743313](https://pubmed.ncbi.nlm.nih.gov/23743313/)]
19. Honglan Z. Prevention of anti-tuberculosis drug-induced liver injury and therapeutic drugs selection. *J Pract Med* 2020;36(24):3307-3311. [doi: [10.3969/j.issn.1006-5725.2020.24.001](https://doi.org/10.3969/j.issn.1006-5725.2020.24.001)]
20. Chinese Medical Association: Tuberculosis Branch. Guidelines for the diagnosis and treatment of anti-tuberculosis drug-induced liver injury (2019 edition). *Chin J Tuberculosis Respir* 2019;042(005):343-356. [doi: [10.3760/cma.j.issn.1001-0939.2019.05.007](https://doi.org/10.3760/cma.j.issn.1001-0939.2019.05.007)]
21. Peifang S, Qiusha P, Ling Y. Drug-induced liver injury biomarkers. *World Chin Med* 2020;15(23):37-44. [doi: [10.3969/j.issn.1673-7202.2020.23.004](https://doi.org/10.3969/j.issn.1673-7202.2020.23.004)]
22. Hosford JD, von Fricken ME, Lauzardo M, Chang M, Dai Y, Lyon JA, et al. Hepatotoxicity from antituberculous therapy in the elderly: a systematic review. *Tuberculosis (Edinb)* 2015 Mar;95(2):112-122 [FREE Full text] [doi: [10.1016/j.tube.2014.10.006](https://doi.org/10.1016/j.tube.2014.10.006)] [Medline: [25595441](https://pubmed.ncbi.nlm.nih.gov/25595441/)]
23. Chang KC, Leung CC, Yew WW, Lau TY, Tam CM. Hepatotoxicity of pyrazinamide: cohort and case-control analyses. *Am J Respir Crit Care Med* 2008 Jun 15;177(12):1391-1396. [doi: [10.1164/rccm.200802-355OC](https://doi.org/10.1164/rccm.200802-355OC)] [Medline: [18388355](https://pubmed.ncbi.nlm.nih.gov/18388355/)]

24. Lammert C, Imler T, Teal E, Chalasani N. Patients With Chronic Liver Disease Suggestive of Nonalcoholic Fatty Liver Disease May Be at Higher Risk for Drug-Induced Liver Injury. *Clin Gastroenterol Hepatol* 2019 Dec;17(13):2814-2815. [doi: [10.1016/j.cgh.2018.12.013](https://doi.org/10.1016/j.cgh.2018.12.013)] [Medline: [30580093](https://pubmed.ncbi.nlm.nih.gov/30580093/)]
25. Chitturi S, Farrell G. Drug-Induced Liver Disease. In: Schiff ER, Maddrey WC, Sorrell MF, editors. *Schiff's Diseases of the Liver* (11th edition). New Delhi: Wiley-Blackwell; 2011:703-783.
26. Tweed CD, Wills GH, Crook AM, Dawson R, Diacon AH, Louw CE, et al. Liver toxicity associated with tuberculosis chemotherapy in the REMoxTB study. *BMC Med* 2018 Mar 28;16(1):46 [FREE Full text] [doi: [10.1186/s12916-018-1033-7](https://doi.org/10.1186/s12916-018-1033-7)] [Medline: [29592805](https://pubmed.ncbi.nlm.nih.gov/29592805/)]
27. Patterson B, Abbara A, Collin S, Henderson M, Shehata M, Gorgui-Naguib H, et al. Predicting drug-induced liver injury from anti-tuberculous medications by early monitoring of liver tests. *J Infect* 2021 Feb;82(2):240-244. [doi: [10.1016/j.jinf.2020.09.038](https://doi.org/10.1016/j.jinf.2020.09.038)] [Medline: [33271167](https://pubmed.ncbi.nlm.nih.gov/33271167/)]
28. Chen M, Bisgin H, Tong L, Hong H, Fang H, Borlak J, et al. Toward predictive models for drug-induced liver injury in humans: are we there yet? *Biomark Med* 2014;8(2):201-213 [FREE Full text] [doi: [10.2217/bmm.13.146](https://doi.org/10.2217/bmm.13.146)] [Medline: [24521015](https://pubmed.ncbi.nlm.nih.gov/24521015/)]
29. Quiroz JC, Feng Y, Cheng Z, Rezazadegan D, Chen P, Lin Q, et al. Development and Validation of a Machine Learning Approach for Automated Severity Assessment of COVID-19 Based on Clinical and Imaging Data: Retrospective Study. *JMIR Med Inform* 2021 Feb 11;9(2):e24572 [FREE Full text] [doi: [10.2196/24572](https://doi.org/10.2196/24572)] [Medline: [33534723](https://pubmed.ncbi.nlm.nih.gov/33534723/)]
30. Hou C, Zhong X, He P, Xu B, Diao S, Yi F, et al. Predicting Breast Cancer in Chinese Women Using Machine Learning Techniques: Algorithm Development. *JMIR Med Inform* 2020 Jun 08;8(6):e17364 [FREE Full text] [doi: [10.2196/17364](https://doi.org/10.2196/17364)] [Medline: [32510459](https://pubmed.ncbi.nlm.nih.gov/32510459/)]
31. Sandhu S, Lin AL, Brajer N, Sperling J, Ratliff W, Bedoya AD, et al. Integrating a Machine Learning System Into Clinical Workflows: Qualitative Study. *J Med Internet Res* 2020 Nov 19;22(11):e22421 [FREE Full text] [doi: [10.2196/22421](https://doi.org/10.2196/22421)] [Medline: [33211015](https://pubmed.ncbi.nlm.nih.gov/33211015/)]
32. Hong S, Lee S, Lee J, Cha WC, Kim K. Prediction of Cardiac Arrest in the Emergency Department Based on Machine Learning and Sequential Characteristics: Model Development and Retrospective Clinical Validation Study. *JMIR Med Inform* 2020 Aug 04;8(8):e15932 [FREE Full text] [doi: [10.2196/15932](https://doi.org/10.2196/15932)] [Medline: [32749227](https://pubmed.ncbi.nlm.nih.gov/32749227/)]
33. Fujihara K, Matsubayashi Y, Harada Yamada M, Yamamoto M, Iizuka T, Miyamura K, et al. Machine Learning Approach to Decision Making for Insulin Initiation in Japanese Patients With Type 2 Diabetes (JDDM 58): Model Development and Validation Study. *JMIR Med Inform* 2021 Jan 27;9(1):e22148 [FREE Full text] [doi: [10.2196/22148](https://doi.org/10.2196/22148)] [Medline: [33502325](https://pubmed.ncbi.nlm.nih.gov/33502325/)]
34. Minerali E, Foil DH, Zorn KM, Lane TR, Ekins S. Comparing Machine Learning Algorithms for Predicting Drug-Induced Liver Injury (DILI). *Mol Pharm* 2020 Jul 06;17(7):2628-2637 [FREE Full text] [doi: [10.1021/acs.molpharmaceut.0c00326](https://doi.org/10.1021/acs.molpharmaceut.0c00326)] [Medline: [32422053](https://pubmed.ncbi.nlm.nih.gov/32422053/)]
35. Sakatis MZ, Reese MJ, Harrell AW, Taylor MA, Baines IA, Chen L, et al. Preclinical strategy to reduce clinical hepatotoxicity using in vitro bioactivation data for >200 compounds. *Chem Res Toxicol* 2012 Oct 15;25(10):2067-2082. [doi: [10.1021/tx300075j](https://doi.org/10.1021/tx300075j)] [Medline: [22931300](https://pubmed.ncbi.nlm.nih.gov/22931300/)]
36. Nakayama S, Atsumi R, Takakusa H, Kobayashi Y, Kurihara A, Nagai Y, et al. A zone classification system for risk assessment of idiosyncratic drug toxicity using daily dose and covalent binding. *Drug Metab Dispos* 2009 Sep;37(9):1970-1977. [doi: [10.1124/dmd.109.027797](https://doi.org/10.1124/dmd.109.027797)] [Medline: [19487250](https://pubmed.ncbi.nlm.nih.gov/19487250/)]
37. Chen M, Borlak J, Tong W. High lipophilicity and high daily dose of oral medications are associated with significant risk for drug-induced liver injury. *Hepatology* 2013 Jul;58(1):388-396. [doi: [10.1002/hep.26208](https://doi.org/10.1002/hep.26208)] [Medline: [23258593](https://pubmed.ncbi.nlm.nih.gov/23258593/)]
38. Williams DP, Lasic SE, Foster AJ, Semenova E, Morgan P. Predicting Drug-Induced Liver Injury with Bayesian Machine Learning. *Chem Res Toxicol* 2020 Jan 21;33(1):239-248. [doi: [10.1021/acs.chemrestox.9b00264](https://doi.org/10.1021/acs.chemrestox.9b00264)] [Medline: [31535850](https://pubmed.ncbi.nlm.nih.gov/31535850/)]
39. Semenova E, Williams D, Afzal A, Lasic S. A Bayesian neural network for toxicity prediction. *Comput Toxicol* 2020 Nov;16:100133 [FREE Full text] [doi: [10.1016/j.comtox.2020.100133](https://doi.org/10.1016/j.comtox.2020.100133)]
40. De'ath G, Fabricius K. Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology* 2000 Nov;81(11):3178-3192 [FREE Full text] [doi: [10.1890/0012-9658\(2000\)081\[3178:cartap\]2.0.co;2](https://doi.org/10.1890/0012-9658(2000)081[3178:cartap]2.0.co;2)]
41. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; New York, NY. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
42. Ogunleye A, Wang Q. 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/)]
43. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Statist* 2001 Oct 1;29(5):1189-1232. [doi: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451)]
44. Zhang X, Li T, Wang J, Li J, Chen L, Liu C. Identification of Cancer-Related Long Non-Coding RNAs Using XGBoost With High Accuracy. *Front Genet* 2019;10:735 [FREE Full text] [doi: [10.3389/fgene.2019.00735](https://doi.org/10.3389/fgene.2019.00735)] [Medline: [31456817](https://pubmed.ncbi.nlm.nih.gov/31456817/)]
45. Le NQK, Do DT, Chiu F, Yapp EKY, Yeh H, Chen C. XGBoost Improves Classification of MGMT Promoter Methylation Status in IDH1 Wildtype Glioblastoma. *J Pers Med* 2020 Sep 15;10(3):128 [FREE Full text] [doi: [10.3390/jpm10030128](https://doi.org/10.3390/jpm10030128)] [Medline: [32942564](https://pubmed.ncbi.nlm.nih.gov/32942564/)]



46. Parsa A, Movahedi A, Taghipour H, Derrible S, Mohammadian A. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accid Anal Prev* 2020 Mar;136:105405. [doi: [10.1016/j.aap.2019.105405](https://doi.org/10.1016/j.aap.2019.105405)] [Medline: [31864931](https://pubmed.ncbi.nlm.nih.gov/31864931/)]
47. Mitchell R, Frank E. Accelerating the XGBoost algorithm using GPU computing. *PeerJ Comput Sci* 2017;3:e127. [doi: [10.7717/peerj-cs.127](https://doi.org/10.7717/peerj-cs.127)]
48. Li Y, Yang L, Yang B, Wang N, Wu T. Application of interpretable machine learning models for the intelligent decision. *Neurocomputing* 2019 Mar;333:273-283 [FREE Full text] [doi: [10.1016/j.neucom.2018.12.012](https://doi.org/10.1016/j.neucom.2018.12.012)]
49. Yan L, Zhang H, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell* 2020 May 14;2(5):283-288. [doi: [10.1038/s42256-020-0180-7](https://doi.org/10.1038/s42256-020-0180-7)]
50. Bi Y, Xiang D, Ge Z, Li F, Jia C, Song J. An Interpretable Prediction Model for Identifying N-Methylguanosine Sites Based on XGBoost and SHAP. *Mol Ther Nucleic Acids* 2020 Dec 04;22:362-372 [FREE Full text] [doi: [10.1016/j.omtn.2020.08.022](https://doi.org/10.1016/j.omtn.2020.08.022)] [Medline: [33230441](https://pubmed.ncbi.nlm.nih.gov/33230441/)]
51. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H. xgboost: eXtreme Gradient Boosting. R package version 0.4-2. URL: <https://mran.microsoft.com/snapshot/2020-07-15/web/packages/xgboost/vignettes/xgboost.pdf> [accessed 2021-07-12]
52. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv. Preprint posted online October 11, 2020.
53. 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. [Medline: [3203132](https://pubmed.ncbi.nlm.nih.gov/3203132/)]
54. Azur M, Stuart E, Frangakis C, Leaf P. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011 Mar;20(1):40-49 [FREE Full text] [doi: [10.1002/mpr.329](https://doi.org/10.1002/mpr.329)] [Medline: [21499542](https://pubmed.ncbi.nlm.nih.gov/21499542/)]
55. Zhuang Z. TB-Dili-forecast-model. GitHub. URL: <https://github.com/Larryzza/TB-Dili-forecast-model> [accessed 2021-07-12]
56. Huai C, Wei Y, Li M, Zhang X, Wu H, Qiu X, et al. Genome-Wide Analysis of DNA Methylation and Antituberculosis Drug-Induced Liver Injury in the Han Chinese Population. *Clin Pharmacol Ther* 2019 Dec;106(6):1389-1397. [doi: [10.1002/cpt.1563](https://doi.org/10.1002/cpt.1563)] [Medline: [31247120](https://pubmed.ncbi.nlm.nih.gov/31247120/)]
57. Singanayagam A, Sridhar S, Dhariwal J, Abdel-Aziz D, Munro K, Connell DW, et al. A comparison between two strategies for monitoring hepatic function during antituberculous therapy. *Am J Respir Crit Care Med* 2012 Mar 15;185(6):653-659. [doi: [10.1164/rccm.201105-0850OC](https://doi.org/10.1164/rccm.201105-0850OC)] [Medline: [22198973](https://pubmed.ncbi.nlm.nih.gov/22198973/)]
58. Shih T, Pai C, Yang P, Chang W, Wang N, Hu OY. A novel mechanism underlies the hepatotoxicity of pyrazinamide. *Antimicrob Agents Chemother* 2013 Apr;57(4):1685-1690 [FREE Full text] [doi: [10.1128/AAC.01866-12](https://doi.org/10.1128/AAC.01866-12)] [Medline: [23357778](https://pubmed.ncbi.nlm.nih.gov/23357778/)]
59. Cao J, Mi Y, Shi C, Bian Y, Huang C, Ye Z, et al. First-line anti-tuberculosis drugs induce hepatotoxicity: A novel mechanism based on a urinary metabolomics platform. *Biochem Biophys Res Commun* 2018 Mar 04;497(2):485-491. [doi: [10.1016/j.bbrc.2018.02.030](https://doi.org/10.1016/j.bbrc.2018.02.030)] [Medline: [29454961](https://pubmed.ncbi.nlm.nih.gov/29454961/)]
60. Horita N, Miyazawa N, Yoshiyama T, Kojima R, Ishigatsubo Y, Kaneko T. Currently Used Low-Dose Pyrazinamide Does Not Increase Liver-Injury in the First Two Months of Tuberculosis Treatment. *Intern Med* 2015;54(18):2315-2320 [FREE Full text] [doi: [10.2169/internalmedicine.54.5533](https://doi.org/10.2169/internalmedicine.54.5533)] [Medline: [26370854](https://pubmed.ncbi.nlm.nih.gov/26370854/)]
61. Minerali E, Foil DH, Zorn KM, Lane TR, Ekins S. Comparing Machine Learning Algorithms for Predicting Drug-Induced Liver Injury (DILI). *Mol Pharm* 2020 Jul 06;17(7):2628-2637 [FREE Full text] [doi: [10.1021/acs.molpharmaceut.0c00326](https://doi.org/10.1021/acs.molpharmaceut.0c00326)] [Medline: [32422053](https://pubmed.ncbi.nlm.nih.gov/32422053/)]
62. Chierici M, Francescato M, Bussola N, Jurman G, Furlanello C. Predictability of drug-induced liver injury by machine learning. *Biol Direct* 2020 Feb 13;15(1):3 [FREE Full text] [doi: [10.1186/s13062-020-0259-4](https://doi.org/10.1186/s13062-020-0259-4)] [Medline: [32054490](https://pubmed.ncbi.nlm.nih.gov/32054490/)]
63. Xu Y, Dai Z, Chen F, Gao S, Pei J, Lai L. Deep Learning for Drug-Induced Liver Injury. *J Chem Inf Model* 2015 Oct 26;55(10):2085-2093. [doi: [10.1021/acs.jcim.5b00238](https://doi.org/10.1021/acs.jcim.5b00238)] [Medline: [26437739](https://pubmed.ncbi.nlm.nih.gov/26437739/)]
64. Williams DP, Lazic SE, Foster AJ, Semenova E, Morgan P. Predicting Drug-Induced Liver Injury with Bayesian Machine Learning. *Chem Res Toxicol* 2020 Jan 21;33(1):239-248. [doi: [10.1021/acs.chemrestox.9b00264](https://doi.org/10.1021/acs.chemrestox.9b00264)] [Medline: [31535850](https://pubmed.ncbi.nlm.nih.gov/31535850/)]
65. Septiandri AA, Aditiawarman A, Tjong R, Burhan E, Shankar A. Improving Mass TB Screening Using Cost-Sensitive Machine Learning Classification. 2017 Presented at: American Thoracic Society 2017 International Conference; May 19-24, 2017; Washington, DC.

## Abbreviations

- ALT:** alanine transaminase
- AUC:** area under the receiver operating characteristic curve
- DILI:** drug-induced liver injury
- EMB:** ethambutol
- INH:** isoniazid
- PZA:** pyrazinamide
- RIF:** rifampicin
- TB:** tuberculosis

*Edited by G Eysenbach; submitted 30.03.21; peer-reviewed by T Chen, K Turner; comments to author 15.04.21; revised version received 12.05.21; accepted 16.05.21; published 20.07.21.*

*Please cite as:*

Zhong T, Zhuang Z, Dong X, Wong KH, Wong WT, Wang J, He D, Liu S

*Predicting Antituberculosis Drug-Induced Liver Injury Using an Interpretable Machine Learning Method: Model Development and Validation Study*

*JMIR Med Inform 2021;9(7):e29226*

URL: <https://medinform.jmir.org/2021/7/e29226>

doi: [10.2196/29226](https://doi.org/10.2196/29226)

PMID: [34283036](https://pubmed.ncbi.nlm.nih.gov/34283036/)

©Tao Zhong, Zian Zhuang, Xiaoli Dong, Ka Hing Wong, Wing Tak Wong, Jian Wang, Daihai He, Shengyuan Liu. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 20.07.2021. 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

# Predicting Writing Styles of Web-Based Materials for Children's Health Education Using the Selection of Semantic Features: Machine Learning Approach

Wenxiu Xie<sup>1</sup>, MSc; Meng Ji<sup>2</sup>, DPhil; Yanmeng Liu<sup>2</sup>, MA; Tianyong Hao<sup>3</sup>, DPhil; Chi-Yin Chow<sup>1</sup>, DPhil

<sup>1</sup>Department of Computer Science, City University of Hong Kong, Hong Kong, China (Hong Kong)

<sup>2</sup>School of Languages and Cultures, The University of Sydney, Sydney, Australia

<sup>3</sup>School of Computer Science, South China Normal University, Guangzhou, China

**Corresponding Author:**

Meng Ji, DPhil

School of Languages and Cultures

The University of Sydney

Room 635 A18

Brennan MacCallum Building

Sydney, 2006

Australia

Phone: 61 449858887

Email: [christine.ji@sydney.edu.au](mailto:christine.ji@sydney.edu.au)

## Abstract

**Background:** Medical writing styles can have an impact on the understandability of health educational resources. Amid current web-based health information research, there is a dearth of research-based evidence that demonstrates what constitutes the best practice of the development of web-based health resources on children's health promotion and education.

**Objective:** Using authoritative and highly influential web-based children's health educational resources from the Nemours Foundation, the largest not-for-profit organization promoting children's health and well-being, we aimed to develop machine learning algorithms to discriminate and predict the writing styles of health educational resources on children versus adult health promotion using a variety of health educational resources aimed at the general public.

**Methods:** The selection of natural language features as predictor variables of algorithms went through initial automatic feature selection using ridge classifier, support vector machine, extreme gradient boost tree, and recursive feature elimination followed by revision by education experts. We compared algorithms using the automatically selected (n=19) and linguistically enhanced (n=20) feature sets, using the initial feature set (n=115) as the baseline.

**Results:** Using five-fold cross-validation, compared with the baseline (115 features), the Gaussian Naive Bayes model (20 features) achieved statistically higher mean sensitivity ( $P=.02$ ; 95% CI -0.016 to 0.1929), mean specificity ( $P=.02$ ; 95% CI -0.016 to 0.199), mean area under the receiver operating characteristic curve ( $P=.02$ ; 95% CI -0.007 to 0.140), and mean macro F1 ( $P=.006$ ; 95% CI 0.016-0.167). The statistically improved performance of the final model (20 features) is in contrast to the statistically insignificant changes between the original feature set (n=115) and the automatically selected features (n=19): mean sensitivity ( $P=.13$ ; 95% CI -0.1699 to 0.0681), mean specificity ( $P=.10$ ; 95% CI -0.1389 to 0.4017), mean area under the receiver operating characteristic curve ( $P=.008$ ; 95% CI 0.0059-0.1126), and mean macro F1 ( $P=.98$ ; 95% CI -0.0555 to 0.0548). This demonstrates the importance and effectiveness of combining automatic feature selection and expert-based linguistic revision to develop the most effective machine learning algorithms from high-dimensional data sets.

**Conclusions:** We developed new evaluation tools for the discrimination and prediction of writing styles of web-based health resources for children's health education and promotion among parents and caregivers of children. User-adaptive automatic assessment of web-based health content holds great promise for distant and remote health education among young readers. Our study leveraged the precision and adaptability of machine learning algorithms and insights from health linguistics to help advance this significant yet understudied area of research.

(*JMIR Med Inform* 2021;9(7):e30115) doi:[10.2196/30115](https://doi.org/10.2196/30115)

**KEYWORDS**

online health education; health educational resource development; machine learning; health linguistics

## Introduction

### Background

Web-based health education and promotion has become increasingly popular among all age groups [1]. Although existing research on web-based health educational materials has focused on adults or general readers, there is an increasing body of research on the assessment and evaluation of web-based educational resources on children's health [2,3]. Clinical and academic research shows that effective writing styles can have an impact on the understanding and reception of medical and health educational resources for different reader groups [4-6]. There is a pressing need to investigate the writing style of web-based health resources on children's health promotion and education for the main readers of such materials as parents and child caregivers to ensure information relevance and acceptability. The Agency for Healthcare Research and Quality is the lead federal agency charged with improving the safety and quality of America's health care system, including pediatric health care products and services [7]. The Agency for Healthcare Research and Quality has developed the Patient Education Materials Assessment Tool (PEMAT) to ensure the development and delivery of quality health care products and services. Key assessment criteria of PEMAT include health information understandability, relevance, and actionability [8,9].

Much of the current research has focused on exploring these assessment dimensions separately using long-standing readability tools [10-13] or machine learning algorithms of natural language features [14-16] using features such as general medical vocabularies, consumer medical vocabulary, natural language features such as a part of speech features [17-19], and other metadata [20]. Furthermore, many of these data-intensive and data-driven studies did not consider insights from research fields directly relevant to health educational resource development and evaluation. The lack of model interpretability has largely limited the applicability of such computational research in practical health education. How to effectively link linguistic research, health education, and machine learning modeling needs to be addressed.

The core question of our study is to develop machine learning models to discriminate and predict what constitutes a suitable writing style of web-based health resources on children's health promotion and education. Research-based evidence is needed to inform and improve the current practice of web-based health educational resource development on health issues related to the promotion of children's health and well-being for readers such as parents, caregivers of children, and teenagers. Our study aims to assess the writing styles of web-based health resources on children's health through an integrated, holistic approach, that is, the development of machine learning models to evaluate whether the content and the writing style of a piece of web-based health educational material is more related to children's health promotion and education, or more for the general public. The underlying hypothesis of our study is that the content and writing

style of high-quality web-based health educational resources vary with the intended readership, which is based on the principles of clinically developed guidelines for health educational resource assessment such as PEMAT [21-23] and health educational research findings in support of user-oriented health communication styles [24-31].

### Data Sets and Feature Extraction

#### *Corpus Data Collection and Classification*

The Nemours Foundation is the world's largest nonprofit organization dedicated to improving the health and well-being of children, and the website of the Foundation has high-quality health education resources developed by medical experts and experienced health educators purposefully for different readerships including parents, children (aged  $\leq 13$  years) and teenagers (aged 13-20 years) [32]. Given the inherent difficulties of conducting large-scale surveys of web-based health educational materials among young children, we used high-quality, authoritative, and children-oriented health materials on the KidsHealth website [33] as the training data to develop machine learning algorithms to predict the relevance and suitability of health education resources for young children with English as the native language. The entire data set contains around 200 children-oriented health texts and 800 adult health texts that we collected on websites developed by nonprofit health organizations and intended for the public, such as the World Health Organization (Multimedia Appendix 1 presents some of the websites used).

#### *Text Screening Criterion*

For the selection of health information for the general public, the main screening criteria were that the websites must have been certified by the Health on the Net Foundation, an international accreditation authority of web-based health information, and they must have been developed by health authorities to provide accurate health educational information. These included governmental health organizations, accredited nonprofit health organizations engaged in health promotion and education, or national or regional associations of specific disease prevention and control. We carefully screened a total of 200 children's health readings from the website of Nemours KidsHealth [33] as one of the most authoritative children health education websites, accredited by the Health on the Net Foundation [34] for its authority (details of the editorial team and the site team are clearly stated), justifiability (health information is complete and provided in an objective, balanced, and transparent manner), and transparency (the site is easy to use, and its mission is clear). The intended readers were clearly the parents and caregivers of children, as shown in the user-specific website structure. It should be noted that there was a clear imbalance between the two sets of health texts, which reflects the reality of web-based health educational resources, as children-oriented health materials are much less than adult-oriented health resources.



### Corpus Annotation of Semantic Features

We annotated the health texts using the semantic tagging system developed by the University of Lancaster, United Kingdom [35]. The annotated health texts contained 115 semantic features under 21 lexical categories—A: general or abstract terms; B: the body and individual; C: arts and crafts; E: emotions; F: food and farming; G: government and politics; H: architecture, housing, and the home; I: money, commerce, and industry; K: entertainment, sports, or games; L: live and living things; M: movement, location, travel, and transport; N: numbers and measurement; O: substances, materials, objects, and equipment; P: education; Q: language or communication; S: social actions, states, and process; T: time, W: world and environment; X: psychological actions, states, and processes; Y: science or technology; Z: names and grammars. Although the University of Lancaster Semantic Annotation System (USAS) was developed for general English studies, it has wide applications in specialist language studies, including health education and information. It is one of the most commonly used English semantic annotation systems.

Our study chose USAS purposefully, as we aimed to select linguistic and semantic features that may be used for developing machine learning algorithms to predict the semantic relevance and suitability of web-based health information among children. The semantic features described earlier are more suitable for analyzing and modeling the content relevance of health information. Many current studies use grammatical or syntactic features to develop machine learning algorithms for health information evaluation. However, grammatical, syntactic, morphological, or other types of structural or functional linguistic features cannot be used to study the contents of health information. The relevance of health information content for specific populations is largely underexplored in current health informatics using natural language processing and machine learning. Our study took advantage of the extensive English semantic coverage of USAS and developed algorithms using a small number of semantic features (20 from the original 115 semantic features) that measured diverse dimensions of the relevance and suitability of web-based health contents for English-speaking young children: approaches to medical knowledge acquisition; assessment of health situations; describing efforts; complexity of actions; attention, stress, or emphasis on key points; and finally, communicative interactivity. All these dimensions of health information relevance and suitability for young readers are supported and represented by semantic features incorporated in the comprehensive annotation system of USAS.

### Statistical Analysis

Table 1 shows the Mann-Whitney *U* test of linguistic features as statistically significant features in web-based health education texts on the education of children's versus adults' health. The results show that children-oriented and adult-oriented health resources had statistically significant differences in the originally annotated semantic features ( $n=115$ ). In addition to the two-tailed *P* values, the effect sizes (Cohen *d*) of the independent sample two-tailed *t* test were produced to measure the statistical differences between the two sets of health texts. As the mean

differences were taken between health texts for children and adult health promotion, a positive Cohen *d* effect size indicated that a certain semantic feature is a characteristic feature of children-oriented health resources. A negative Cohen *d* effect size suggested that a semantic feature is more significant in health educational resources intended for the public.

A number of semantic features were identified as characteristic of adult-oriented health resources: semantic features that had large negative Cohen *d* effect sizes (above 0.5) included B2 health and disease ( $P<.001$ ; Cohen  $d=-0.802$ ); B3 medicine and medical treatment ( $P<.001$ ; Cohen  $d=-0.800$ ); Z2 geographical names ( $P<.001$ ; Cohen  $d=-0.674$ ); Z3 other proper names ( $P<.001$ ; Cohen  $d=-0.594$ ); M7 places ( $P<.001$ ; Cohen  $d=-0.587$ ); Y1 science and technology generally ( $P<.001$ ; Cohen  $d=-0.522$ ); Z99 out-of-dictionary rare expressions ( $P<.001$ ; Cohen  $d=-0.776$ ); A15 safety or danger ( $P<.001$ ; Cohen  $d=-0.543$ ); and S1 social actions, states, and processes ( $P<.001$ ; Cohen  $d=-0.547$ ). Semantic features with medium effect sizes (Cohen  $d=-0.5$  to  $0.3$ ) were related to social processes, money, religion, and numeracy: G1 government, politics, and election ( $P<.001$ ; Cohen  $d=-0.496$ ); W3 geographical terms ( $P<.001$ ; Cohen  $d=-0.414$ ); L1 life and living things ( $P<.001$ ; Cohen  $d=-0.391$ ); I1 money generally ( $P<.001$ ; Cohen  $d=-0.370$ ); L2 living creature ( $P<.001$ ; Cohen  $d=-0.362$ ); S5 social groups and affiliation ( $P<.001$ ; Cohen  $d=-0.356$ ); S9 religion ( $P=.001$ ; Cohen  $d=-0.324$ ); and N1 numbers ( $P=.006$ ; Cohen  $d=-0.315$ ).

Textual features that were statistically significant in children-oriented health texts reflected the different cognitive processing of health information and health communication styles between children and adults. Semantic features that had a large Cohen *d* effect size ( $0.5-0.9$ ) for children-oriented health texts included words indicating simple actions and steps: M1 moving, coming, and going ( $P<.001$ ; Cohen  $d=0.547$ ); M2 putting, taking pulling, and pushing ( $P<.001$ ; Cohen  $d=0.517$ ); E2 emotional expressions of like or dislike feelings ( $P<.001$ ; Cohen  $d=0.556$ ); X3 sensory words describing sight, taste, feel, and touch feelings ( $P<.001$ ; Cohen  $d=0.684$ ); S4 kinships ( $P<.001$ ; Cohen  $d=0.713$ ); X8 expressions describing efforts, attempts, and resolution ( $P<.001$ ; Cohen  $d=0.803$ ); and words of textual coherence or logical structure—Z8 pronouns ( $P<.001$ ; Cohen  $d=0.907$ ); Z6 negative expression ( $P<.001$ ; Cohen  $d=0.764$ ); and Z7 conditional expressions ( $P<.001$ ; Cohen  $d=0.575$ ).

There were two semantic categories related to emphasis, stress, and attention: A14 focusing subjuncts that draw attention to or focus on ( $P=.04$ ; Cohen  $d=0.519$ ) and A13 words as maximizers, boosters, approximators, and compromisers ( $P<.001$ ; Cohen  $d=0.645$ ). Semantic features that were identified as characteristic features of children-oriented health reading of a medium Cohen *d* effect size ( $0.3-0.5$ ) included F1 food-related expressions ( $P<.001$ ; Cohen  $d=0.493$ ); O1 substances and materials generally ( $P<.001$ ; Cohen  $d=0.49$ ); B1 terms relating to the human body and bodily processes ( $P=.002$ ; Cohen  $d=0.362$ ); O4 physical attributes ( $P<.001$ ; Cohen  $d=0.348$ ); and E4 expressions of happiness or sadness ( $P<.001$ ; Cohen  $d=0.493$ ).



The large number of semantic features of statistical significance ( $P < .05$ ) and medium-to-large effect sizes (Cohen  $d$  0.3-0.9) needed to be further reduced to a smaller set of textual features to ensure the stability, efficiency, and convenience of any empirical assessment tool to be developed. The following sections will elaborate on machine learning–assisted automatic feature selection, followed by a review and revision of the empirical analytical instrument from the perspective of user-adaptive health resource design and health linguistics. The final machine learning model aims to provide high-precision automated predictions of the suitability of web-based health educational resources for young readers.

Machine learning algorithms are known for their lack of interpretability compared with statistical models. Through the successive permutation of the predictor features in the final algorithm (Gaussian Naive Bayes [GNB]), we calculated the impact of individual features on the performance of the algorithm, that is, its sensitivity and specificity. Two sets of semantic features were identified as significant contributors to the prediction of children- versus adult-oriented health educational resources. Each set of features that emerged in the process of algorithm development represented a balanced combination of semantic classes, which were statistically significant features in children- or adult-oriented materials.

**Table 1.** Semantic feature of health educational texts.

Semantic features	Children-oriented, mean (SD)	Adult-oriented, mean (SD)	Statistical difference		Effect size (Cohen $d$ )
			Mann-Whitney $U$ test	$P$ value <sup>a</sup>	
A5: evaluation: good or bad	5.65 (7.267)	4.1 (4.994)	67510.0	.17	0.340
A15: safety or danger	0.230 (1.020)	1.560 (3.950)	56287.0	<.001	–0.543
B2: health and disease	7.910 (13.792)	22.45 (30.619)	41001.0	<.001	–0.802
B3: medicine and medical treatment	4.360 (8.392)	12.46 (17.280)	46443.5	<.001	–0.800
F1: food	10.30 (25.407)	3.490 (13.801)	51368.0	<.001	0.491
M1: moving, coming, going	5.27 (7.399)	2.92 (5.259)	52775.0	<.001	0.547
S1: social actions, states, and processes	1.850 (2.738)	3.820 (6.090)	54876.5	<.001	–0.547
S2: people	12.42 (15.635)	10.22 (16.519)	65131.5	.04	0.207
S4: kin	2.860 (4.221)	1.070 (3.247)	52886.5	<.001	0.713
S5: groups and affiliation	1.500 (3.672)	2.520 (4.771)	58355.0	<.001	–0.356
S8: helping or hindering	5.140 (6.315)	6.920 (9.634)	62823.5	.007	–0.318
S9: religion and the supernatural	0.140 (0.738)	0.440 (1.587)	64677.0	.001	–0.324
T1: time	11.3 (12.639)	12.94 (15.022)	67348.0	.16	–0.181
X3: sensory	4.920 (7.606)	2.020 (4.618)	50469.0	<.001	0.684
X9: ability	1.88 (3.619)	1.83 (3.612)	69246.0	.37	0.019
Z2: geographical names	0.550 (1.496)	3.120 (6.184)	45505.5	<.001	–0.674
Z6: negative	5.840 (5.958)	3.080 (4.861)	51392.0	<.001	0.764
Z8: pronoun	59.79 (53.287)	31.56 (38.830)	46155.0	<.001	0.907
Z99: unmatched expressions	13.74 (17.037)	37.58 (49.684)	39069.0	<.001	–0.776

<sup>a</sup>Asymptotic significance (two-tailed).

## Methods

We applied machine learning algorithms to learn the important features for detecting the writing styles of web-based health educational resources on children's health promotion and education. Recursive feature elimination (RFE), ridge classifier (RC), extreme gradient boosting (XGBoost) [36], and support vector machine (SVM) [37] were used to assist in automatic feature selection. RFE is commonly used with SVM (denoted as RFE\_SVM) to build a model and remove unimportant features [38]. In addition to linear models such as SVM, tree-based models are also an effective method to learn feature importance, and XGBoost was used as the learning estimator

of RFE (denoted as RFE\_XGB). For algorithms RC, SVM, and RFE, we used the implementation in scikit-learn [39]. For XGBoost, we used the Python package xgboost [40].

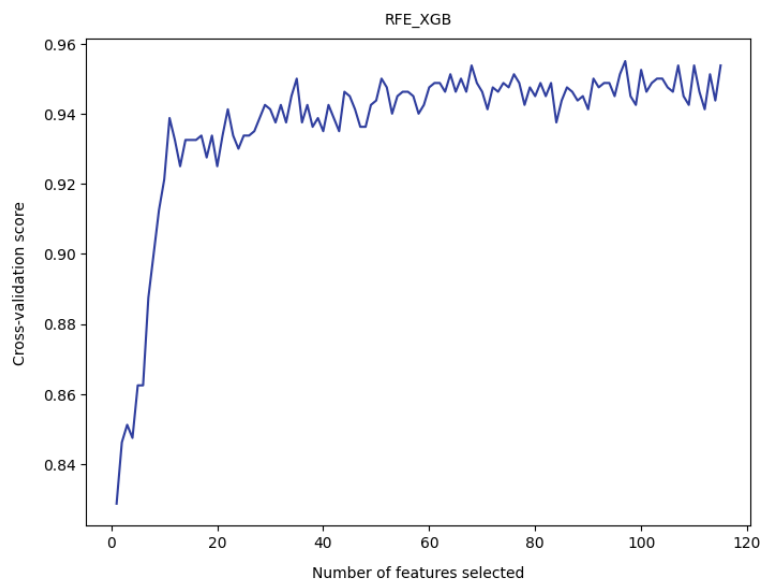
For the RC and RFE algorithms, scikit-learn has built-in cross-validation variants *RidgeClassifierCV* and *RFECV*, which perform leave-one-out five-fold cross-validation to search for the best hyper-parameters and select the best cross-validated features, respectively. For SVM, which only needs to tune the regularization parameter  $C$ , we applied the commonly used *GridSearchCV* for hyperparameter tuning. The *GridSearchCV* algorithm performs an exhaustive search over specified parameter values to determine the best and cross-validated parameter values of the model. For XGBoost, which has nine

hyper-parameters including some continuous ones, we applied *RandomizedSearchCV*, which performs a randomized search over parameters and samples a fixed number of parameter settings from the specified distribution. We set the number of parameter settings  $n\_iter$  of *Randomized SearchCV* as 300. The hyperparameter  $n\_iter$  defines the number of parameter settings that are sampled. With a large value of  $n\_iter$ , the algorithm was able to find better hyper-parameters from a large parameter setting with high quality. The fine-tuned results of the better hyper-parameters are shown in [Multimedia Appendix 2](#). For the hyper-parameters that were not listed, we used the default values in the model.

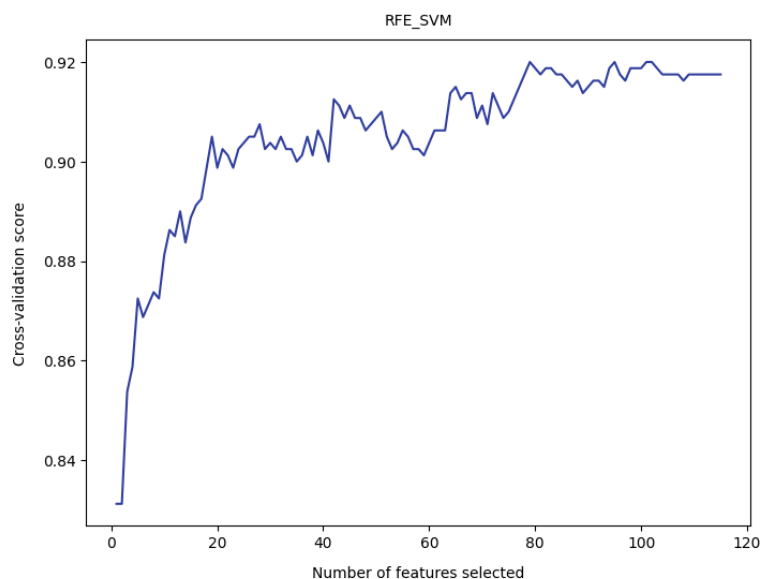
We applied RFE\_SVM and RFE\_XGB to evaluate the cross-validation score when increasing the number of selected features. The automatic tuning results of the number of features selected by cross-validation are shown in [Multimedia Appendix](#)

2. As shown in the results ([Figures 1 and 2](#)), both the SVM and XGBoost model gained a nearly stable cross-validation score greater than 0.9 when the number of selected features was equal to or greater than 40. This result indicated that when only 40 features were used, the model was still able to achieve good performance, and adding more features did not help much. As a result, we applied 40 as a threshold to select the top 40 important features learned by RC and XGBoost. The details of the selected top 40 features of RC and XGBoost are shown in [Multimedia Appendix 2](#) and [Figures 3 and 4](#). RFE\_SVM learned 95 features, eliminating 20 unimportant features from all 115 features. For the RFE\_XGB, 97 features were selected, and 18 unimportant features were eliminated. Finally, the intersected 19 features from the RC, XGBoost, RFE\_SVM, and RFE\_XGB were selected as automatically learned features from the machine learning algorithms.

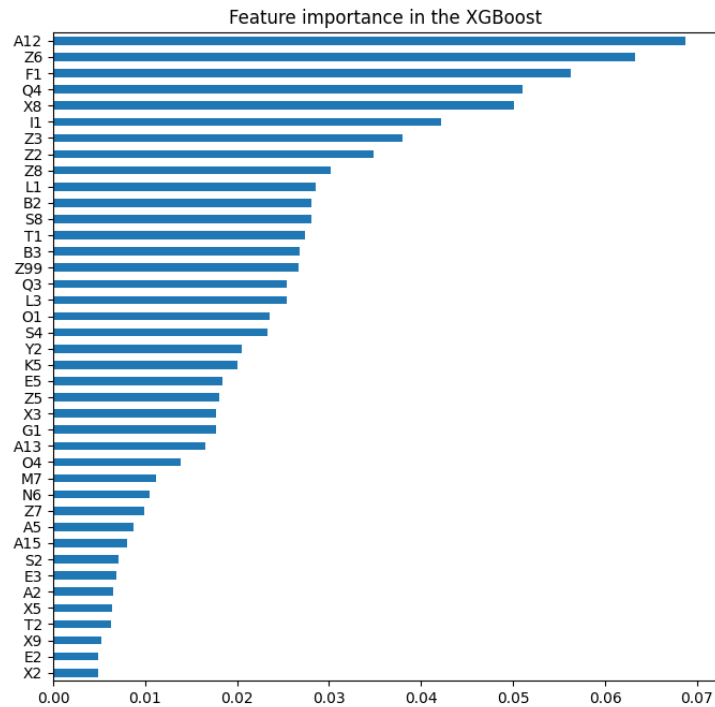
**Figure 1.** Automatic tuning of the number of features selected with cross-validation of RFE\_XGB.



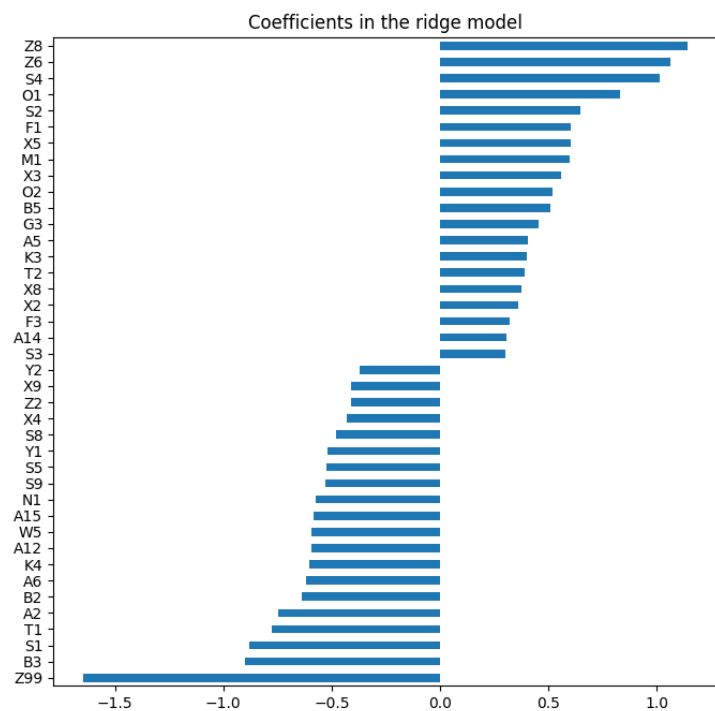
**Figure 2.** Automatic tuning of the number of features selected with cross-validation of RFE\_SVM.



**Figure 3.** Automatic feature importance ranking using extreme gradient boost tree. XGBoost: extreme gradient boosting.



**Figure 4.** Automatic feature importance ranking using the ridge classifier.



## Results

### Feature Selection Results

Table 2 shows the performance of the three machine learning classifiers on the testing data, which were largely similar in terms of overall model accuracy, macro average F1, and F1 for adult- and children-oriented health readings. The top semantic features in the initial automatic feature selection were as follows (for a detailed description of these codes, see the USAS):

- RC: Z99, B3, S1, T1, A2, B2, A6, K4, A12, W5, A15, N1, S9, S5, Y1, S8, X4, Z2, X9, Y2, S3, A14, F3, X2, X8, T2, K3, A5, G3, B5, O2, X3, M1, X5, F1, S2, O1, S4, Z6, Z8
- XGB Tree: X2, E2, X9, T2, X5, A2, E3, S2, A15, A5, Z7, N6, M7, O4, A13, G1, X3, Z5, E5, K5, Y2, S4, O1, L3, Q3, Z99, B3, T1, S8, B2, L1, Z8, Z2, Z3, I1, X8, Q4, F1, Z6, A12
- RFE using SVM as the feature scoring algorithm: A2, A3, A4, A5, A6, A7, A9, A10, A11, A12, A13, A14, A15, B1, B2, B3, B5, C1, E4, E5, E6, F1, F3, F4, G1, G2, G3, H2,

- H3, H4, H5, I1, I2, I3, I4, K1, K2, K3, K4, K5, L1, L2, L3, M1, M3, M4, M5, M6, M8, N1, N2, N3, N4, N5, N6, O2, O3, O4, P1, Q1, Q2, Q3, Q4, S1, S2, S3, S4, S5, S6, S7, S8, S9, T1, T2, T4, W1, W2, W4, W5, X2, X3, X4, X5, X6, X7, X9, Y1, Y2, Z1, Z2, Z3, Z4, Z6, Z8, Z99
- RFE using XGB as the feature scoring algorithm: A1, A2, A3, A4, A5, A6, A7, A8, A9, A10, A11, A12, A13, A14, A15, B1, B2, B3, B4, B5, C1, E1, E2, E3, E4, E6, F1, F2,
  - G1, H3, H4, I1, I2, I3, I4, K3, K4, K5, K6, L2, L3, M1, M2, M3, M4, M5, M6, M7, N1, N3, N4, N5, N6, O1, O2, O3, O4, Q1, Q2, Q3, Q4, S1, S2, S3, S4, S5, S6, S7, S8, S9, T1, T2, T3, T4, W1, W3, W4, X2, X3, X5, X6, X7, X8, X9, Y1, Y2, Z0, Z1, Z2, Z3, Z4, Z5, Z6, Z7, Z8, Z9, Z99
  - The common 19 features of the four feature selection algorithms are as follows: Z8, S2, S8, F1, A5, S4, X3, M1, T1, S5, S9, Z99, A15, S1, X9, Z6, B2, Z2, B3.

**Table 2.** Classifiers used for automatic feature selection.

Classifier and text class	Accuracy	Macro average F1 <sup>a</sup>	Precision	Recall	F1
<b>Ridge classifier</b>	0.925	0.89			
Adult-oriented readings			0.99	0.91	0.95
Children-oriented readings			0.74	0.97	0.84
<b>SVM<sup>b</sup></b>	0.93	0.89			
Adult-oriented readings			0.95	0.96	0.96
Children-oriented readings			0.84	0.8	0.82
<b>XGB<sup>c</sup></b>	0.94	0.90			
Adult-oriented readings			0.95	0.98	0.96
Children-oriented readings			0.91	0.78	0.84

<sup>a</sup>F1 = 2 × [(precision × recall) / (precision + recall)].

<sup>b</sup>SVM: support vector machine.

<sup>c</sup>XGB: extreme gradient boosting.

Table 3 shows the comparison of the performance of algorithms using the original 115 features as predictor variables and the automatically selected 19 semantic features as predictor variables. With GNB classifier, we reduced the predictor variables from 115 to 19, the mean sensitivity (of the five folds of data) decreased from 0.685 to 0.634 (0.074%), the mean specificity increased from 0.771 to 0.903 (17.04%), and the mean area under the receiver operating characteristic curve (AUC) increased from 0.822 to 0.982 (7.21%). Similar patterns were observed with K-nearest neighbor (KNN). Mean sensitivity decreased from 0.973 to 0.943 when the predictor variables reduced in number. By contrast, the model mean specificity increased by 33.7% from 0.526 to 0.703 and the mean AUC increased by 3.79% from 0.901 to 0.935. This suggested that for some algorithms such as GNB and KNN, feature selection can increase the model efficiency, at least partially. However, with XGB, both mean sensitivity and mean specificity decreased by around 0.5%, resulting in a decrease of mean AUC of 0.95%. The decrease in the mean sensitivity, mean specificity, and mean AUC of XGB and the decrease in mean specificity of GNB and KNN using automatically selected features indicated that further linguistic revision was needed. Linguistic review of the automatically selected features will ascertain whether the automatically selected features were linguistically meaningful and explainable.

Features that were deemed linguistically irrelevant or unexplainable will be replaced by semantic features that are highly relevant and significant for health language studies. Incorporating insights from language studies into automatic feature selection will help in the development of adaptive and

interpretable machine learning algorithms. Increasing the interpretability and practical usability of algorithms can be achieved at the stage of the linguistic review of automatically selected feature sets.

We eliminated S9, T1, S2, and Z2 and added X8, A12, A11, A13, and A14. These were the semantic features that were highly relevant to health linguistics. X8 are terms depicting the level of effort and resolution. This is a statistically significant feature of children's educational resources ( $P < .001$ ; Cohen  $d = 0.803$ ). Typical words of X8 were tried, fights, hard, fighting, try, and struggling, which were prevalent in health educational resources for children to describe bodily reactions to diseases and viruses. In contrast, adult-oriented health education resources were abundant in words and expressions of A12, which were abstract terms denoting the varying levels of difficulties: challenge, adversity, and complexity. The independent  $t$  test showed that A12 was a characteristic semantic feature of general health materials ( $P < .001$ ; Cohen  $d = -0.234$ ). A11 included abstract terms denoting importance or significance and abstract terms denoting noticeability or markedness. Typical words of A11 were main, significant, important, serious, principal, emergency, distinctive, urgent, crucial, and emergencies that were abundant in adult health educational resources ( $P < .001$ ; Cohen  $d = -0.0348$ ). A13 included words such as maximizers, boosters, approximators, and compromisers ( $P < .001$ ; Cohen  $d = 0.645$ ). Typical words of A13 were very, almost, more, as, about, up, to, approximately, fully, even, and enormously, which were prevalent in children's health education resources. Finally, A14 focused on subjuncts that drew attention to or to focus upon ( $P = .04$ ; Cohen  $d = 0.519$ ). Typical words of

A14 were especially, just, and only, which were highly frequent in children's health educational readings.

**Table 3.** Performance of classifiers using 115 (originally tagged) and 19 (automatically selected) features.

Classifier and feature sets	Sensitivity, mean (SD)	Specificity, mean (SD)	AUC <sup>a</sup> , mean (SD)
<b>GNB<sup>b</sup></b>			
All 115 features	0.685 (0.125)	0.771 (0.116)	0.822 (0.062)
Automatically selected 19 features	0.634 (0.074)	0.903 (0.063)	0.882 (0.054)
<b>KNN<sup>c</sup></b>			
All 115 features	0.973 (0.013)	0.526 (0.096)	0.901 (0.032)
Automatically selected 19 features	0.943 (0.028)	0.703 (0.048)	0.935 (0.023)
<b>XGB<sup>d</sup></b>			
All 115 features	0.982 (0.01)	0.766 (0.059)	0.978 (0.012)
Automatically selected 19 features	0.970 (0.019)	0.737 (0.051)	0.970 (0.016)

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

<sup>b</sup>GNB: Gaussian Naive Bayes.

<sup>c</sup>KNN: K-nearest neighbor algorithm.

<sup>d</sup>XGB: extreme gradient boosting.

Table 4 shows the linguistic profiling framework we developed for the revised set of semantic features. It includes the 15 automatically selected features and the manually added five features based on their relevance for health linguistic and language studies, as well as their function as statistically significant, large characteristic features of children- versus adult-oriented health educational readings. The linguistic framework for comparing health texts intended for these two distinct readerships contained three key dimensions that were cognitive abilities, social context of health issues, and user-adaptive health communication style. Under each dimension, there were several contrastive semantic features which help to distinguish health readings for different readers. Within the dimension of cognitive abilities, four semantic features reflect the different scope of health knowledge of children versus adults. For example, F1 food-related words and expressions (creams, peanuts, spread, appetite, foods, salt, sugar, meal, pasta, and rice), and X3 sensory expressions describing taste, color, sight, feel, and sound of things (hearing, see, notice, scented, hear, watch, sound, smell, colorful, etc) were prevalent in children's health readings as their main approach to health knowledge acquisition. In contrast, more abstract, complex, rare, difficult words were characteristic features of adult health readings—B2: medicine (medical, condition, disorder, stroke, tumor, injury, illness, health, miscarriage, infertility, etc); B3: medical treatment (neurological, diagnosed, computed tomography, cure, scan, medicinal, analgesic, healing, diagnosis, drugs, etc), and Z99: complex, out-of-dictionary words (cyclones, aldosterone, noncancerous, vestibulocochlear, neurofibromatosis, tinnitus, muskrat, ondatra, zibethicus, herbivore, alkanes, esters, aldehydes, etc).

Children and adults also use different approaches to assess health events and situations: A5 words that evaluate events in terms of good or bad and false or true were more prevalent in children's readings with typical words such as wrong, right, better, good, true, positive, improved, greater, ok, and best. In

contrast, A15 words that assess health situations in terms of safety, risk, and harm were more prevalent in adult health readings with typical expressions that we found in the corpus: at-risk, safe, dangerous, exposures, hazard, safety, insurance, warning, alert, and alarming. X9 terms describing success and failure, gains and losses, and benefits and risks were also prevalent in adult health materials. This finding aligns well with the latest research on health communication using the Prospect Theory [41], which highlights the human propensity to maximize benefits and minimize risks, including in health care and medical settings. Typical words of X9 included effective, successful, lose, achieve, gains, go wrong, overcome, solve, cope, and competent. The complexity of actions is another important feature of health education reading [42,43]. In children's health readings, simple actions and verbs describing the direction of movements were prevalent—typical words in M1 were moving, coming, and going, get, follow, step, and steps. In contrast, the mean frequency of S8 words describing levels of help, obstacles, and hindrance was statistically higher in adult health readings such as stop, prevent, cooperate, benefits, resistance, protect, protecting, support, supporting, and help.

We also identified predictor features that are relevant to the social context of health issues [44]. This dimension includes two sets of semantic features of interpersonal relations and the socioeconomic contexts of health issues. For example, S4 words of kinships (family, parents, siblings, relatives, children, household, families, etc) were more common in children's health readings, whereas S5 words of people's social groups and affiliation were prevalent in adult health educational readings such as network, loneliness, community, member, partnership, and alliance. Another important semantic feature is S1 terms related to participation, involvement, entitlement, and eligibility or describing personality traits such as strength, weakness, vulnerability, and disadvantaged. Typical words of S1 were vulnerable, self-esteem, meeting, helplessness, social, and contacts, which were highly frequent in adult health readings.



We could not find an equivalent semantic feature class in children’s health readings to match S1 as a characteristic of adult health readings.

The health communicative style is another key dimension of semantic features [30]. We found that an effective communicative style is particularly relevant for children-oriented health educational readings [45]. For example, to match the machine learning–selected feature of A11 terms describing importance and priority, we added two functionally equivalent semantic features that were prevalent in children’s health readings to help increase the emphasis and stress on the key health messages of the texts: A13 and A14. Both were mostly

adverbs describing the degree, levels, extent, severity of objects, and events. For example, typical words in A13 were very, almost, more, as, about, up, to, approximately, fully, even, enormously; and typical words of A14 were especially, just, and only. These words stand in contrast with A11 words that characterize the prioritization and importance attribution among adults: main, significant, important, serious, principal, emergency, distinctive, urgent, crucial, and emergencies. Finally, terms that help increase the logical coherence of health readings were highly frequent in children’s health readings but not in adult readings. These include Z8, the use of pronouns (it, this, who, that, you, what, we, they, their, which, your, our, and anything), and Z6, the use of negative expressions.

**Table 4.** Revised linguistic evaluation framework with final 20 features.

Dimensions of linguistic analyses	Texts on children’s health	Texts on adults’ health
<b>Cognitive abilities</b>		
Scope of health knowledge	<ul style="list-style-type: none"> <li>F1 (food)</li> <li>X3 (sensory: taste, sound, and touch)</li> </ul>	<ul style="list-style-type: none"> <li>B2 (medicine); B3 (medical treatment)</li> <li>Z99 (complex and out-of-dictionary words)</li> </ul>
Assessment of situations	<ul style="list-style-type: none"> <li>A5 (good or bad and true or false)</li> </ul>	<ul style="list-style-type: none"> <li>A15 (safety or danger)</li> <li>X9 (success or failure, gains or loss, and benefits or risks)</li> </ul>
Describing efforts	<ul style="list-style-type: none"> <li>X8 (level of efforts or resolution)</li> </ul>	<ul style="list-style-type: none"> <li>A12 (level of difficulty)</li> </ul>
Complexity of actions	<ul style="list-style-type: none"> <li>M1 (actions of movement)</li> </ul>	<ul style="list-style-type: none"> <li>S8 (level of help or hindrance)</li> </ul>
<b>The social context of health issues</b>		
Interpersonal relations	<ul style="list-style-type: none"> <li>S4 (kin)</li> </ul>	<ul style="list-style-type: none"> <li>S5 (social groups and affiliation)</li> </ul>
Socioeconomic context	<ul style="list-style-type: none"> <li>N/A<sup>a</sup></li> </ul>	<ul style="list-style-type: none"> <li>S1 (terms related to participation, involvement, entitlement, eligibility; or describing personality traits such as strength, weakness, vulnerability, and disadvantaged)</li> </ul>
<b>Communicative style</b>		
Attention emphasis and stress	<ul style="list-style-type: none"> <li>A13 (degree)</li> <li>A14 (particularizers)</li> </ul>	<ul style="list-style-type: none"> <li>A11 (importance)</li> </ul>
Logical coherence	<ul style="list-style-type: none"> <li>Z8 (pronouns)</li> <li>Z6 (negative)</li> </ul>	<ul style="list-style-type: none"> <li>N/A</li> </ul>

<sup>a</sup>N/A: not applicable.

Table 5 shows features in the linguistic evaluation framework for a binary logistic regression analysis (enter) with children-oriented health resources as the reference class. The statistical result aligns with the linguistic analysis well: 10 semantic features had negative unstandardized coefficients and less than 1 odds ratio, suggesting that with the increase of values in these features, the odds of the health text being a children-oriented health reading were higher than those of the health text being an adult health reading. For example, the odds ratio of Z6 negative expressions ( $P<.001$ ) was 0.778 (95% CI 0.69-0.876), which means that with the increase of one Z6 word, the odds of the health text being an adult health reading reduced by a mean of 22.2%. The odds ratio of S4 (words describing kinships;  $P<.001$ ) was 0.823 (95% CI 0.746-0.907), meaning

with the increase of one word of S4 (such as parents, siblings, grandparents, etc), the odds of the health text being a children’s reading was 17.7% higher than those of the health text being an adult-oriented health reading. X8 ( $P=.07$ ), A14 ( $P=.66$ ), M1 ( $P=.17$ ), and A13 ( $P=.39$ ) were statistically insignificant predictor variables. Similarly, 10 semantic features were identified as characteristic features of adult health readings: A11, B2, B3, Z99, X9, S8, S5, S1, A12, A15. A11 and X9 were statistically insignificant predictor variables. The odds ratio of A15 was 1.945 (95% CI 1.335-2.833), which means that with the increase of one word of A15 (words evaluating safety, danger, or risks of health events), the odds of the health text being an adult reading was 94.5% higher than those of the text being a children-oriented health reading.

**Table 5.** Predictor variables of binary logistic regression (children=0; adult=1).

Relevance of semantic features to outcomes	Values			
	$\beta$ (SE)	Wald test	<i>P</i> value	OR <sup>a</sup> (95% CI)
<b>Semantic features related to higher ORs of health texts on children's health</b>				
Z6	-0.252 (0.061)	16.966	<.001	0.778 (0.690-0.876)
X8	-0.228 (0.127)	3.233	.07	0.796 (0.621-1.021)
S4	-0.195 (0.050)	15.351	<.001	0.823 (0.746-0.907)
X3	-0.134 (0.033)	16.715	<.001	0.875 (0.820-0.933)
A5	-0.104 (0.038)	7.418	.006	0.902 (0.837-0.971)
A14	-0.063 (0.144)	0.192	.66	0.939 (0.707-1.246)
M1	-0.054 (0.039)	1.927	.17	0.948 (0.878-1.022)
F1	-0.038 (0.011)	11.374	.001	0.963 (0.942-0.984)
A13	-0.036 (0.042)	0.744	.39	0.965 (0.889-1.047)
Z8	-0.021 (0.008)	7.589	.006	0.979 (0.964-0.994)
<b>Semantic features related to higher ORs of health texts on adults' health</b>				
A11	0.030 (0.086)	0.124	.73	1.031 (0.871-1.219)
B2	0.032 (0.013)	6.397	.01	1.032 (1.007-1.058)
B3	0.066 (0.019)	12.425	<.001	1.068 (1.030-1.108)
Z99	0.067 (0.011)	35.849	<.001	1.069 (1.046-1.093)
X9	0.118 (0.064)	3.400	.07	1.126 (0.993-1.277)
S8	0.162 (0.040)	16.137	<.001	1.176 (1.087-1.273)
S5	0.248 (0.057)	19.056	<.001	1.281 (1.146-1.432)
S1	0.279 (0.085)	10.703	.001	1.322 (1.118-1.562)
A12	0.297 (0.102)	8.573	.003	1.346 (1.103-1.642)
A15	0.665 (0.192)	12.003	.001	1.945 (1.335-2.833)

<sup>a</sup>OR: odds ratio.

### Performance Comparison of Classifiers Using Three Sets of Features

Tables 6-10 show the results of the comparison of GNB algorithms developed using the originally tagged multidimensional feature set (n=115), automatically selected feature set (n=19), and linguistically enhanced feature set (n=20). Table 7 shows that both the automatically selected and the linguistically enhanced feature set achieved statistically improved AUC over the original high-dimensional feature set: automatically selected ( $P=.008$ ) and linguistically enhanced ( $P=.02$ ), significant at the adjusted  $P=.17$  using Bonferroni correction. The difference in AUC between the two streamlined feature sets was not statistically significant ( $P=.56$ ). In terms of model sensitivity, the automatically selected feature set did

not achieve statistically significant improvement over the OR feature set ( $P=.13$ ) but the linguistically enhanced feature set did ( $P=.01$ ). The sensitivity of the linguistically enhanced feature set was also statistically improved over the automatically selected feature set ( $P<.001$ ). In terms of model specificity, the automatically selected feature set did not improve over the OR feature set ( $P=.10$ ), but the linguistically enhanced feature set did ( $P=.01$ ). The specificity between the automatically selected and linguistically enhanced feature sets did not differ significantly ( $P=.53$ ). Finally, in terms of macro F1, which provides a balanced assessment of the model performance, the automatically selected feature set did not improve over the baseline OR feature set ( $P=.98$ ). The linguistically enhanced feature set improved significantly over the OR feature set ( $P=.006$ ) and automatically selected feature set ( $P=.001$ ).

**Table 6.** Performance of machine learning models using different sets of features as predictors.

Algorithm	AUC <sup>a</sup> , mean (SD)	Sensitivity, mean (SD)	Specificity, mean (SD)	Macro F1 <sup>b</sup> , mean (SD)
115 features	0.8224 (0.0617)	0.6848 (0.1252)	0.7714 (0.1161)	0.6336 (0.080)
19 features (automatic selection)	0.8817 (0.0539)	0.6339 (0.0743)	0.9029 (0.0626)	0.6333 (0.067)
20 features (linguistic review)	0.8888 (0.0315)	0.7733 (0.076)	0.8629 (0.0843)	0.7248 (0.0451)

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

<sup>b</sup>F1 =  $2 \times [(\text{precision} \times \text{recall}) / (\text{precision} + \text{recall})]$ .

**Table 7.** Pairwise corrected resampled *t* test of area under the receiver operating characteristic curve differences using three sets of features as predictors.

Pair	Description	Mean difference (%)	95% CI of mean difference	<i>P</i> value (two-tailed)
1	19 features versus 115 features	7.2096	0.0059 to 0.1126	.008 <sup>a</sup>
2	20 features versus 115 features	8.0729	-0.0071 to 0.1399	.02 <sup>a</sup>
3	20 features versus 19 features	0.8052	-0.0421 to 0.0563	.56

<sup>a</sup>*P* value significant at .0167 (Bonferroni correction).

**Table 8.** Pairwise corrected resampled *t* test of sensitivity differences using three sets of features as predictors.

Pair	Description	Mean difference (%)	95% CI of the mean difference	<i>P</i> value (two-tailed)
1	19 features versus 115 features	-7.4336	-0.1699 to 0.0681	.13
2	20 features versus 115 features	12.9204	-0.016 to 0.1929	.011 <sup>a</sup>
3	20 features versus 19 features	21.9885	0.1048 to 0.174	<.001 <sup>a</sup>

<sup>a</sup>*P* value significant at .0167 (Bonferroni correction).

**Table 9.** Pairwise corrected resampled *t* test of specificity differences using three sets of features as predictors.

Pair	Description	Mean difference (%)	95% CI of the mean difference	<i>P</i> value (two-tailed)
1	19 features versus 115 features	17.037	-0.1389 to 0.4017	.10
2	20 features versus 115 features	11.8519	-0.0163 to 0.1991	.01 <sup>a</sup>
3	20 features versus 19 features	-4.4304	-0.2923 to 0.2123	.53

<sup>a</sup>*P* value significant at .0167 (Bonferroni correction).

**Table 10.** Pairwise corrected resampled *t* test of macro F1 differences using three sets of features as predictors.

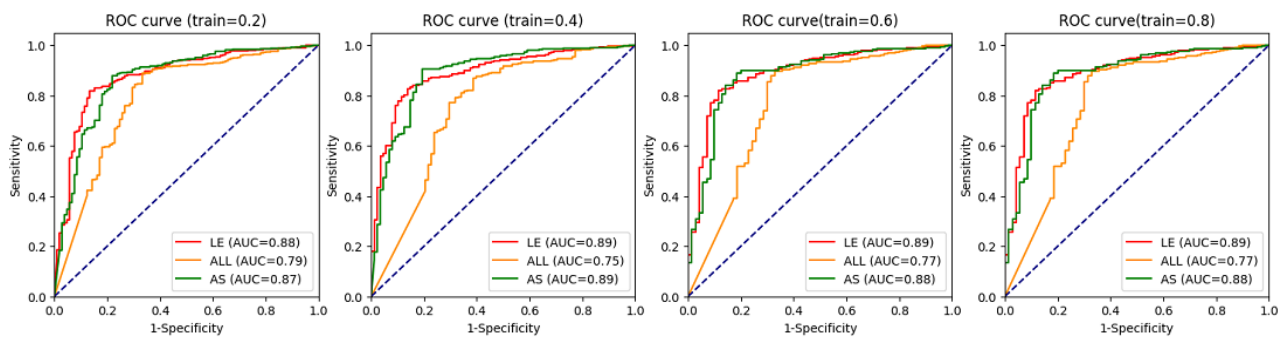
Pair	Description	Mean difference (standardized; %)	95% CI of the mean difference	<i>P</i> value (two-tailed)
1	19 features versus 115 features	-0.0539	-0.0555 to 0.0548	.98
2	20 features versus 115 features	14.3813	0.0158 to 0.1665	.006 <sup>a</sup>
3	20 features versus 19 features	14.4430	0.0422 to 0.1407	.001

<sup>a</sup>*P* value significant at .0167 (Bonferroni correction).

We also tested the scalability and effectiveness of the 20 linguistically enhanced features (Figure 5). We compared the performance with 115 initial all features (ALL) and 19 automatically selected features. The data were randomly divided into a training set and test set with different split rates of 0.2, 0.4, 0.6, and 0.8. The performance was evaluated using receiver operating characteristic curve and AUC metrics. As shown in Figure 5, the model using linguistically enhanced features always yielded the best performance with a stable AUC score

of 0.89 with the different training data set size. Moreover, when using only 20% data for training (train=0.2), the model using linguistically enhanced features still achieved a much higher performance than the baseline (using ALL features), demonstrating its effectiveness and potential for scalability. Thus, incorporating both linguistic features and machine learning features can better help in the interpretation and auto-learning of health educational materials.

**Figure 5.** Scalability and effectiveness of the 20 linguistically enhanced features. AS: automatically selected; AUC: area under the receiver operating characteristic curve; LE: linguistically enhanced; ROC: receiver operating characteristic curve.



## Discussion

### Principal Findings

Our study illustrated machine learning–assisted selection of textual features to develop new algorithms to predict the content and writing style of credible web-based resources for children’s health education and promotion among the parents and caregivers of young children. We used high-quality health educational resources developed by influential children’s health promotion and educational organizations as training data. We illustrated that feature selection to reduce high-dimensional feature sets is an effective method for improving the efficiency of machine learning algorithms, as shown by the improved performance of the AUC of the model using automatically selected features ( $n=19$ ) as predictor variables over the originally tagged feature set ( $n=115$ ;  $P=.008$ ). However, specificity, sensitivity, and macro F1 did not improve when using the automatically selected feature set. We then refined automatic feature selection by incorporating linguistic insights from health linguistics and user-oriented health communication. The linguistically enhanced features led to a statistically significant improvement in sensitivity; macro F1 over the automatically selected feature set: sensitivity ( $P<.001$ ) and macro F1 ( $P=.001$ ); and statistically significant improvement of AUC, sensitivity, specificity, and macro F1 over the original high-dimensional feature set: AUC ( $P=.02$ ), sensitivity ( $P=.01$ ), specificity ( $P=.01$ ), and macro F1 ( $P=.006$ ).

Machine learning algorithms were known for their lack of interpretability. Through the successive permutation of the linguistically enhanced predictor variables in the developed GNB algorithm, we explored the individual impact of each feature on the model’s sensitivity and specificity. Two sets of semantic features emerged as large contributors to the model’s ability to predict the suitability of health educational resources for adults and children, respectively. We found the final algorithm interpretable using the linguistic profiling framework developed for those automatically selected features. For the prediction of adult-oriented health education readings, that is, features highly relevant for the sensitivity of the model, 11 semantic features were identified as large contributors as indicated by the decrease of sensitivity in their absence: X3 (–9.4%; words of sensory: taste, sound, touch, sight, smell, etc),

S4 (–8.93%; kinships), Z99 (–8.78%; complex words), A14 (–7.99%; focusing subjuncts that draw attention to or to focus upon), Z8 (–6.9%) (pronouns), A11 (–6.11%; terms describing importance and priority), S1 (–5.96%; terms of participation, involvement, entitlement, and eligibility or describing personality traits such as strength, weakness, vulnerability, and disadvantaged), A5 (–5.94%; words of evaluating good or bad or true or false), B3 (–5.33%; medical treatment), S8 (–4.86%; words describing levels of help, obstacles, and hindrance), X9 (–0.31%; success or failure; gains or loss; or benefits or risks).

For the prediction of health education readings on children’s health, that is, features highly relevant for the specificity of the model, 10 semantic features were identified as large contributors, as shown by the decrease in model specificity with the successive permutation of these features (Figure 6): X8 (–24.5%; words describing efforts and resolution), F1 (–23.18%; food-related words), S5 (–14.57%; social groups and affiliation), A15 (–9.93%; words evaluating safety and danger), M1 (–9.27%; movement words), B2 (–9.27%; medicine), Z6 (–8.61%; negative), A13 (–5.96%; degree), A12 (–2.65%; difficulty), and X9 (–0.66%; success or failure; gains or loss; and benefits or risks).

It is worth noting that features identified as key contributors to model sensitivity were not necessarily features that were statistically significant in adult-oriented health readings (Table 1). For example, X3, S4, A14, Z8, and A5 were statistically significant in children’s health resources, which however had large impacts on the model sensitivity (Figure 7). Similarly, S5, A15, B2, A12, and X9 were statistically significant features of adult health materials but they also had an impact on model specificity, which is the ability of the machine learning algorithm to predict health texts as children-oriented health materials. This led to our interpretation that the newly developed algorithm represents a balanced mix of linguistically relevant, meaningful semantic features that were statistically significant in either children or adult health materials. Thus, the approach to outcome prediction of machine learning differs significantly from that of statistical analysis. However, our study demonstrated that both statistical and linguistic insights can improve the performance of machine learning–assisted feature selection and subsequent prediction.

Figure 6. Impact of selected features on mean sensitivity.

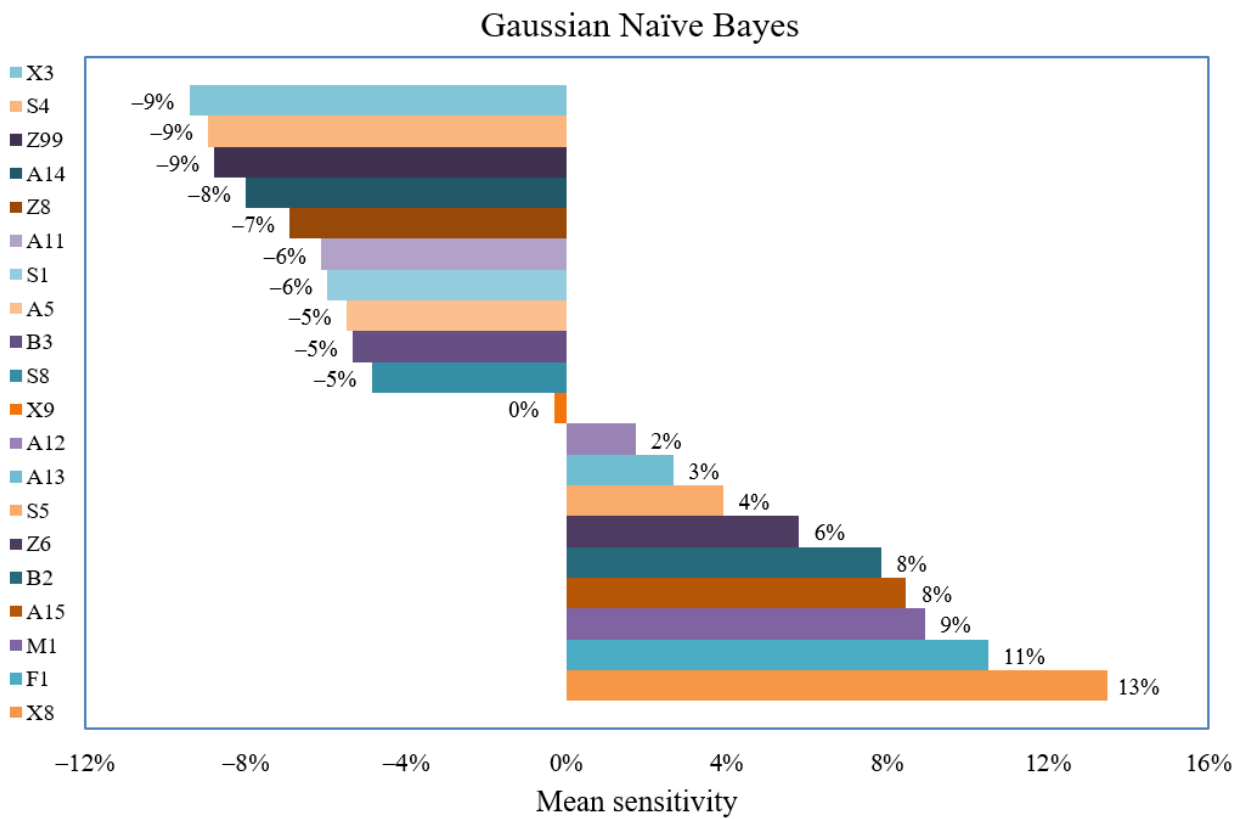
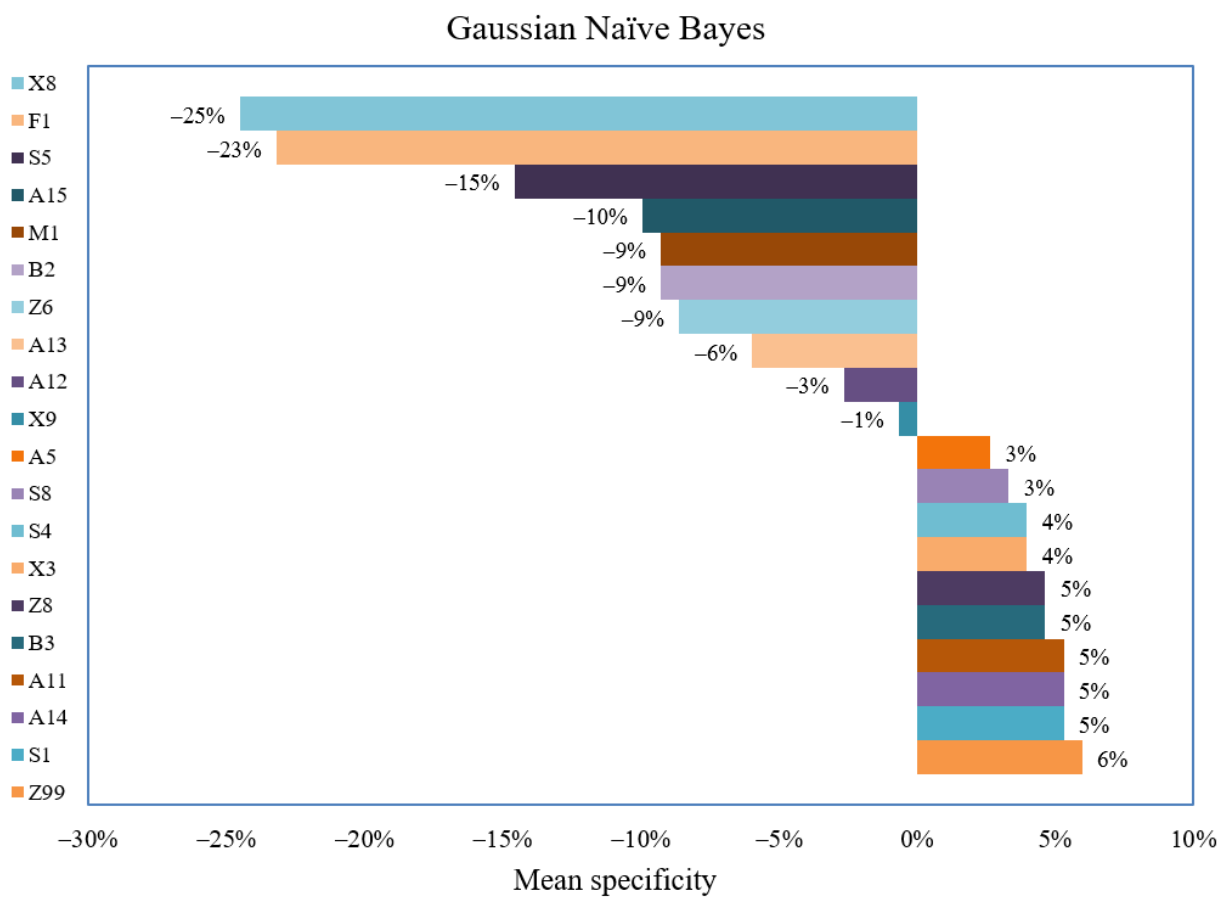


Figure 7. Impact of selected features on mean specificity.





## Limitations and Future Research

The size of the training data set was relatively small, with a couple hundred texts of children-oriented health readings. However, this reflects the reality, as children's health educational resources are much less than adult health readings. As a result, the model specificity was consistently lower than the model sensitivity. In addition, in the linguistic evaluation framework (Table 4), the structure was not well balanced. Items were not complete for all evaluation subcategories, such as health communication styles. Further studies are required to fill the research gaps that emerged in this study.

## Conclusions

Our study has shown that children-oriented and adult-oriented health educational readings in English have distinct semantic

features that can be effectively exploited to develop machine learning algorithms with proven discriminatory accuracy. Specifically, we identified three large sets of semantic features related to the varying cognitive approaches to health information acquisition, the social contexts of health issues, and user-adaptive health communication styles. Machine learning is known to lack interpretability. Our study developed algorithms that are interpretable from the perspective of linguistics and user-oriented health information assessment. Thus, our study shows that a more integrated approach to computerized health information assessment combining insights from fields such as linguistics and health education can help harness the power of machine learning to advance applied social and health research.

## Conflicts of Interest

None declared.

### Multimedia Appendix 1

Health on the Net Foundation–certified websites used.

[DOCX File , 14 KB - [medinform\\_v9i7e30115\\_app1.docx](#) ]

### Multimedia Appendix 2

Core parameters of the hyper-parameter tuning of ridge classifier, extreme gradient boosting, support vector machine, and recursive feature elimination.

[DOCX File , 16 KB - [medinform\\_v9i7e30115\\_app2.docx](#) ]

## References

1. Hu Y, Sundar SS. Effects of online health sources on credibility and behavioral intentions. *Commun Res* 2009 Nov 25;37(1):105-132. [doi: [10.1177/0093650209351512](#)]
2. Ghaddar SF, Valerio MA, Garcia CM, Hansen L. Adolescent health literacy: the importance of credible sources for online health information. *J Sch Health* 2012 Jan;82(1):28-36. [doi: [10.1111/j.1746-1561.2011.00664.x](#)] [Medline: [22142172](#)]
3. Kubb C, Foran HM. Online health information seeking by parents for their children: systematic review and agenda for further research. *J Med Internet Res* 2020 Aug 25;22(8):e19985 [FREE Full text] [doi: [10.2196/19985](#)] [Medline: [32840484](#)]
4. Michielutte R, Bahnson J, Dignan M, Schroeder E. The use of illustrations and narrative text style to improve readability of a health education brochure. *J Cancer Educ* 1992;7(3):251-260. [doi: [10.1080/08858199209528176](#)] [Medline: [1419592](#)]
5. Pakhchanian H, Yuan M, Raiker R, Waris S, Geist C. *Semin Ophthalmol* 2021 May 11:1-6. [doi: [10.1080/08820538.2021.1919721](#)] [Medline: [33975496](#)]
6. Hart S. Patient education accessibility. *Medical Writ* 2015 Dec 23;24(4):190-194. [doi: [10.1179/2047480615Z.000000000321](#)]
7. Scanlon MC, Harris JM, Levy F, Sedman A. Evaluation of the agency for healthcare research and quality pediatric quality indicators. *Pediatrics* 2008 Jun 01;121(6):1723-1731. [doi: [10.1542/peds.2007-3247](#)] [Medline: [18474532](#)]
8. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ Couns* 2014 Sep;96(3):395-403 [FREE Full text] [doi: [10.1016/j.pec.2014.05.027](#)] [Medline: [24973195](#)]
9. Shoemaker SJ, Wolf MS, Brach C. The Patient Education Materials Assessment Tool (PEMAT) and user's guide. Agency for Healthcare Research and Quality. URL: <https://www.ahrq.gov/health-literacy/patient-education/pemat.html> [accessed 2021-06-30]
10. Visser K, Slattery M, Stewart V. Help or hinder? An assessment of the accessibility, usability, reliability and readability of disability funding website information for Australian mental health consumers. *Health Soc Care Community* 2020 Oct 13:13192. [doi: [10.1111/hsc.13192](#)] [Medline: [33051906](#)]
11. Bai XY, Zhang YW, Li J, Li Y, Qian JM. Online information on Crohn's disease in Chinese: an evaluation of its quality and readability. *J Dig Dis* 2019 Nov;20(11):596-601. [doi: [10.1111/1751-2980.12822](#)] [Medline: [31583816](#)]
12. Scott BB, Johnson AR, Doval AF, Tran BN, Lee BT. Readability and understandability analysis of online materials related to abdominal aortic aneurysm repair. *Vasc Endovascular Surg* 2020 Mar;54(2):111-117. [doi: [10.1177/1538574419879855](#)] [Medline: [31607232](#)]

13. Mac OA, Thayre A, Tan S, Dodd RH. Web-based health information following the renewal of the cervical screening program in Australia: evaluation of readability, understandability, and credibility. *J Med Internet Res* 2020 Jun 26;22(6):e16701 [FREE Full text] [doi: [10.2196/16701](https://doi.org/10.2196/16701)] [Medline: [32442134](https://pubmed.ncbi.nlm.nih.gov/32442134/)]
14. François T, Miltsakaki E. Do NLP and machine learning improve traditional readability formulas? In: Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations. 2012 Presented at: First Workshop on Predicting and Improving Text Readability for Target Reader Populations; June 2012; Montréal, Canada URL: <https://www.aclweb.org/anthology/W12-2207/>
15. Petersen SE, Ostendorf M. A machine learning approach to reading level assessment. *Comput Speech Lang* 2009 Jan;23(1):89-106. [doi: [10.1016/j.csl.2008.04.003](https://doi.org/10.1016/j.csl.2008.04.003)]
16. Zowalla R, Wiesner M, Pfeifer D. Automatically assessing the expert degree of online health content using SVMs. *Stud Health Technol Inform* 2014;202:48-51. [Medline: [25000012](https://pubmed.ncbi.nlm.nih.gov/25000012/)]
17. Palotti J, Zuccon G, Hanbury A. Consumer health search on the web: study of web page understandability and its integration in ranking algorithms. *J Med Internet Res* 2019 Jan 30;21(1):e10986 [FREE Full text] [doi: [10.2196/10986](https://doi.org/10.2196/10986)] [Medline: [30698536](https://pubmed.ncbi.nlm.nih.gov/30698536/)]
18. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PLoS One* 2010 Nov;5(11):e14118 [FREE Full text] [doi: [10.1371/journal.pone.0014118](https://doi.org/10.1371/journal.pone.0014118)] [Medline: [21124761](https://pubmed.ncbi.nlm.nih.gov/21124761/)]
19. Beaunoyer E, Arsenault M, Lomanowska AM, Guitton MJ. Understanding online health information: evaluation, tools, and strategies. *Patient Educ Couns* 2017 Feb;100(2):183-189. [doi: [10.1016/j.pec.2016.08.028](https://doi.org/10.1016/j.pec.2016.08.028)] [Medline: [27595436](https://pubmed.ncbi.nlm.nih.gov/27595436/)]
20. Albalawi Y, Nikolov NS, Buckley J. Trustworthy health-related tweets on social media in Saudi Arabia: tweet metadata analysis. *J Med Internet Res* 2019 Oct 08;21(10):e14731 [FREE Full text] [doi: [10.2196/14731](https://doi.org/10.2196/14731)] [Medline: [31596242](https://pubmed.ncbi.nlm.nih.gov/31596242/)]
21. Arslan D, Tutar MS, Kozanhan B. Evaluating the readability, understandability, and quality of online materials about chest pain in children. *Eur J Pediatr* 2020 Dec;179(12):1881-1891 [FREE Full text] [doi: [10.1007/s00431-020-03772-8](https://doi.org/10.1007/s00431-020-03772-8)] [Medline: [32894353](https://pubmed.ncbi.nlm.nih.gov/32894353/)]
22. Arnold M, Reichard A, Gutman K, Westermann L, Sanchez V. Cross-cultural adaptation of hearing loss self-management patient education materials: development of the Caja de Instrumentos de Pérdida Auditiva. *Am J Audiol* 2020 Dec 09;29(4):691-700 [FREE Full text] [doi: [10.1044/2020\\_aja-19-00120](https://doi.org/10.1044/2020_aja-19-00120)]
23. Brega AG, Freedman MA, LeBlanc WG, Barnard J, Mabachi NM, Cifuentes M, et al. Using the health literacy universal precautions toolkit to improve the quality of patient materials. *J Health Commun* 2015;20 Suppl 2:69-76 [FREE Full text] [doi: [10.1080/10810730.2015.1081997](https://doi.org/10.1080/10810730.2015.1081997)] [Medline: [26513033](https://pubmed.ncbi.nlm.nih.gov/26513033/)]
24. Verhoeven F, Steehouder MF, Hendrix RM, Van Gemert-Pijnen JE. From expert-driven to user-oriented communication of infection control guidelines. *Int J Hum Comput* 2010 Jun;68(6):328-343 [FREE Full text] [doi: [10.1016/j.ijhcs.2009.07.003](https://doi.org/10.1016/j.ijhcs.2009.07.003)]
25. Berggren UJ, Gunnarsson E. User - oriented mental health reform in Sweden: featuring 'professional friendship'. *Disabil Soc* 2010 Jul 23;25(5):565-577 [FREE Full text] [doi: [10.1080/09687599.2010.489303](https://doi.org/10.1080/09687599.2010.489303)]
26. Gonzales AL, Hancock JT, Pennebaker JW. Language style matching as a predictor of social dynamics in small groups. *Commun Res* 2009 Nov 04;37(1):3-19. [doi: [10.1177/0093650209351468](https://doi.org/10.1177/0093650209351468)]
27. Rains SA. Language style matching as a predictor of perceived social support in computer-mediated interaction among individuals coping with illness. *Commun Res* 2015 Jan 13;43(5):694-712. [doi: [10.1177/0093650214565920](https://doi.org/10.1177/0093650214565920)]
28. Lord SP, Sheng E, Imel ZE, Baer J, Atkins DC. More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client. *Behav Ther* 2015 May;46(3):296-303 [FREE Full text] [doi: [10.1016/j.beth.2014.11.002](https://doi.org/10.1016/j.beth.2014.11.002)] [Medline: [25892166](https://pubmed.ncbi.nlm.nih.gov/25892166/)]
29. Ireland ME, Slatcher RB, Eastwick PW, Scissors LE, Finkel EJ, Pennebaker JW. Language style matching predicts relationship initiation and stability. *Psychol Sci* 2011 Jan 13;22(1):39-44. [doi: [10.1177/0956797610392928](https://doi.org/10.1177/0956797610392928)] [Medline: [21149854](https://pubmed.ncbi.nlm.nih.gov/21149854/)]
30. Lundebjerg NE, Trucil DE, Hammond EC, Applegate WB. When it comes to older adults, language matters: Journal of the American Geriatrics Society adopts modified American Medical Association style. *J Am Geriatr Soc* 2017 Jul;65(7):1386-1388 [FREE Full text] [doi: [10.1111/jgs.14941](https://doi.org/10.1111/jgs.14941)] [Medline: [28568284](https://pubmed.ncbi.nlm.nih.gov/28568284/)]
31. AMA Manual of Style Committee. *AMA Manual of Style: A Style Guide for Authors and Editors*, 10th Ed. New York: Oxford University Press; 2007.
32. Nemours. URL: <https://www.nemours.org/welcome.html> [accessed 2021-06-30]
33. KidsHealth. URL: <https://kidshealth.org/en/kids/> [accessed 2021-04-30]
34. Health On the Net. URL: <https://www.hon.ch/en/> [accessed 2021-06-30]
35. Rayson P, Archer D, Piao S, McEnery T. The UCREL semantic analysis system. In: Proceedings of the Beyond Named Entity Recognition Semantic Labeling for NLP Tasks Workshop in LREC'04. 2004 Presented at: Workshop: Beyond Named Entity Recognition Semantic Labeling for NLP Tasks in LREC'04; January 2004; Lisbon, Portugal.
36. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H. Xgboost: extreme gradient boosting. R Package Version 0.4-2 2015:1-4 [FREE Full text]
37. Jakkula V. Tutorial on support vector machine (SVM). School of EECS, Washington State University. 2006. URL: [https://www.semanticscholar.org/paper/Tutorial-on-Support-Vector-Machine-\(SVM\)-Jakkula/7cc83e98367721bfb908a8f703ef5379042c4bd9](https://www.semanticscholar.org/paper/Tutorial-on-Support-Vector-Machine-(SVM)-Jakkula/7cc83e98367721bfb908a8f703ef5379042c4bd9) [accessed 2021-06-30]

38. Brownlee J. Recursive Feature Elimination (RFE) for feature selection in Python. Machine Learning Mastery. URL: <https://machinelearningmastery.com/rfe-feature-selection-in-python/> [accessed 2021-06-30]
39. 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;12:2825-2830 [FREE Full text]
40. xgboost 1.4.2. XGBoost Python Package. 2021. URL: <https://pypi.org/project/xgboost/> [accessed 2021-07-07]
41. Edwards K. Prospect theory: a literature review. Int Rev Financial Anal 1996 Jan;5(1):19-38. [doi: [10.1016/S1057-5219\(96\)90004-6](https://doi.org/10.1016/S1057-5219(96)90004-6)]
42. Lipari M, Berlie H, Saleh Y, Hang P, Moser L. Understandability, actionability, and readability of online patient education materials about diabetes mellitus. Am J Health Syst Pharm 2019 Jan 25;76(3):182-186. [doi: [10.1093/ajhp/zxy021](https://doi.org/10.1093/ajhp/zxy021)] [Medline: [31408087](https://pubmed.ncbi.nlm.nih.gov/31408087/)]
43. Ramos CL, Williams J, Bababekov Y, Chang D, Carter B, Jones P. Assessing the understandability and actionability of online neurosurgical patient education materials. World Neurosurg 2019 Oct;130:588-597. [doi: [10.1016/j.wneu.2019.06.166](https://doi.org/10.1016/j.wneu.2019.06.166)] [Medline: [31260846](https://pubmed.ncbi.nlm.nih.gov/31260846/)]
44. Sorensen G, Nagler E, Pawar P, Gupta P, Pednekar M, Wagner G. Lost in translation: The challenge of adapting integrated approaches for worker health and safety for low- and middle-income countries. PLoS One 2017;12(8):e0182607 [FREE Full text] [doi: [10.1371/journal.pone.0182607](https://doi.org/10.1371/journal.pone.0182607)] [Medline: [28837688](https://pubmed.ncbi.nlm.nih.gov/28837688/)]
45. Logan R, Siegel E, editors. K-12 health education, health communication, and health literacy: strategies to improve lifelong health. In: Health Literacy in Clinical Practice and Public Health. Amsterdam, Netherlands: IOS Press; 2020:1-616.

## Abbreviations

**AUC:** area under the receiver operating characteristic curve  
**GNB:** Gaussian Naive Bayes  
**KNN:** K-nearest neighbor  
**PEMAT:** Patient Education Materials Assessment Tool  
**RC:** ridge classifier  
**RFE:** recursive feature elimination  
**SVM:** support vector machine  
**USAS:** University of Lancaster Semantic Annotation System  
**XGBoost:** extreme gradient boosting

*Edited by G Eysenbach; submitted 01.05.21; peer-reviewed by J Lei, M Rodrigues; comments to author 21.05.21; revised version received 22.05.21; accepted 15.06.21; published 22.07.21.*

*Please cite as:*

*Xie W, Ji M, Liu Y, Hao T, Chow CY*

*Predicting Writing Styles of Web-Based Materials for Children's Health Education Using the Selection of Semantic Features: Machine Learning Approach*

*JMIR Med Inform 2021;9(7):e30115*

*URL: <https://medinform.jmir.org/2021/7/e30115>*

*doi: [10.2196/30115](https://doi.org/10.2196/30115)*

*PMID: [34292167](https://pubmed.ncbi.nlm.nih.gov/34292167/)*

©Wenxiu Xie, Meng Ji, Yanmeng Liu, Tianyong Hao, Chi-Yin Chow. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org/>), 22.07.2021. 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

# Patient Representation From Structured Electronic Medical Records Based on Embedding Technique: Development and Validation Study

Yanqun Huang<sup>1,2</sup>, BSc; Ni Wang<sup>1,2</sup>, BSc; Zhiqiang Zhang<sup>1,2</sup>, BSc; Honglei Liu<sup>1,2</sup>, PhD; Xiaolu Fei<sup>3</sup>, PhD; Lan Wei<sup>3</sup>, PhD; Hui Chen<sup>1,2</sup>, PhD

<sup>1</sup>School of Biomedical Engineering, Capital Medical University, Beijing, China

<sup>2</sup>Beijing Key Laboratory of Fundamental Research on Biomechanics in Clinical Application, Capital Medical University, Beijing, China

<sup>3</sup>Information Center, Xuanwu Hospital, Capital Medical University, Beijing, China

**Corresponding Author:**

Hui Chen, PhD

School of Biomedical Engineering

Capital Medical University

No 10, Xitoutiao, Youanmenwai, Fengtai District

Beijing, 100069

China

Phone: 86 1083911545

Email: [chenhui@ccmu.edu.cn](mailto:chenhui@ccmu.edu.cn)

## Abstract

**Background:** The secondary use of structured electronic medical record (sEMR) data has become a challenge due to the diversity, sparsity, and high dimensionality of the data representation. Constructing an effective representation for sEMR data is becoming more and more crucial for subsequent data applications.

**Objective:** We aimed to apply the embedding technique used in the natural language processing domain for the sEMR data representation and to explore the feasibility and superiority of the embedding-based feature and patient representations in clinical application.

**Methods:** The entire training corpus consisted of records of 104,752 hospitalized patients with 13,757 medical concepts of disease diagnoses, physical examinations and procedures, laboratory tests, medications, etc. Each medical concept was embedded into a 200-dimensional real number vector using the Skip-gram algorithm with some adaptive changes from shuffling the medical concepts in a record 20 times. The average of vectors for all medical concepts in a patient record represented the patient. For embedding-based feature representation evaluation, we used the cosine similarities among the medical concept vectors to capture the latent clinical associations among the medical concepts. We further conducted a clustering analysis on stroke patients to evaluate and compare the embedding-based patient representations. The Hopkins statistic, Silhouette index (SI), and Davies-Bouldin index were used for the unsupervised evaluation, and the precision, recall, and F1 score were used for the supervised evaluation.

**Results:** The dimension of patient representation was reduced from 13,757 to 200 using the embedding-based representation. The average cosine similarity of the selected disease (subarachnoid hemorrhage) and its 15 clinically relevant medical concepts was 0.973. Stroke patients were clustered into two clusters with the highest SI (0.852). Clustering analyses conducted on patients with the embedding representations showed higher applicability (Hopkins statistic 0.931), higher aggregation (SI 0.862), and lower dispersion (Davies-Bouldin index 0.551) than those conducted on patients with reference representation methods. The clustering solutions for patients with the embedding-based representation achieved the highest F1 scores of 0.944 and 0.717 for two clusters.

**Conclusions:** The feature-level embedding-based representations can reflect the potential clinical associations among medical concepts effectively. The patient-level embedding-based representation is easy to use as continuous input to standard machine learning algorithms and can bring performance improvements. It is expected that the embedding-based representation will be helpful in a wide range of secondary uses of sEMR data.

(*JMIR Med Inform* 2021;9(7):e19905) doi:[10.2196/19905](https://doi.org/10.2196/19905)



**KEYWORDS**

electronic medical records; Skip-gram; feature representation; patient representation; stroke

**Introduction**

The past decade has witnessed an explosion in the amount of digital information stored in electronic medical records (EMRs), which contain massive quantities of information on the clinical history of patients. The wide secondary use of this information for various clinical applications has become a prevalent trend [1], helping to make diagnostic decisions [2-4], predict patient outcomes [5-8], and provide treatment recommendations [9-11].

As we all know, the method of data representation is becoming more and more crucial for the performance of data applications [12,13]. Recently, many researchers have made preliminary attempts to convert different types of medical data to vectors by representation learning. They have then applied EMR data with these representations to clinical tasks [6,14,15], making more effective use of medical data and improving performance in the predictive analyses. Cui et al [6] compared the performances of three distributed representation methods (ie, Skip-gram, Continuous Bag-of-Words, and latent semantic analysis) for the prediction of hospital cost and length of stay (LOS). Ning et al [15] trained vector representations for medical concepts from biomedical journal articles through Skip-gram and proposed a fully automated feature extraction method for disease phenotyping based on the medical concept vector representation. Moreover, some researchers learned patient representation through deep learning [3,5,12]. Zhe Wang et al [5] designed a feature rearrangement representation based on the convolutional neural network for heart failure mortality prediction. Lei Wang et al [3] used autoencoder, an unsupervised deep learning algorithm, to generate lower-dimensional representations from EMR data in various predictive tasks such as readmission prediction and pneumonia prediction. A similar study [12] used the recurrent neural network-based denoising autoencoder to encode patient records into low-dimensional and dense vectors for heart failure prediction.

However, there are still challenges in the representation of structured EMR (sEMR) data containing high-dimensional and diverse features. Such features as demographic characteristics, disease diagnoses, physical examinations and procedures, and laboratory tests may have discrete or continuous values, making it difficult to reveal the latent relations among them. Moreover, it is difficult to make full use of every available feature (laboratory tests, for example) due to the unavoidable missing values. It is worth exploring how to deal with the patient records with features that are unequal in length.

Therefore, in this study, we leveraged a distributed embedding technique originated in natural language processing (NLP), the Skip-gram algorithm, with several adaptive changes to obtain effective representations from the sEMR data. The feature representation was evaluated by the dimension reduction visualization method and feature correlation analysis method. We further conducted clustering analyses on patients expressed with the proposed representations to evaluate the representation scheme. We aimed to explore the feasibility and superiority of

the embedding-based representations in data mining tasks for sEMR data.

**Methods****Study Data and Data Preprocessing**

The sEMR data of 144,375 hospital admissions for 104,752 patients were collected from Xuanwu Hospital, Capital Medical University, Beijing, China, between January 2014 and December 2016. Patients' features were grouped into seven major categories: demographic characteristics, hospital admission and discharge, utilization of medical resources, disease diagnoses (identified by International Classification of Disease, Tenth Revision [ICD-10] code), examination and procedures undergone (identified by International Classification of Diseases, Ninth Revision, Clinical Modification [ICD-9-CM] code), laboratory tests, and medications (Table S1 in [Multimedia Appendix 1](#)). They were maintained for each hospital stay. If a patient had multiple hospitalizations or multiple laboratory tests, only the first hospitalization or laboratory test was included. Patients' personal information was completely removed from the data set before we could access the data remotely, ensuring the data were used in an anonymous and safe manner. The study and data use were approved by the Human Research Ethics Committees of the hospital.

Data analysis concentrated on a certain disease would be more targeted and specific because a certain group of patients may have similar characteristics. Stroke is a severe disease with high prevalence, high mortality, and high disability [16,17]. It is meaningful and crucial to mine the knowledge hidden in the data for stroke diagnosis and treatment. Thus, we focused on stroke patients for representation evaluation. In the data set, there were 8232 records involving adult patients with a primary diagnosis of stroke (ICD-10 codes I60 to I64, I66, and I67.8 [18]). Among them, 1397 patients had a primary diagnosis of hemorrhagic stroke (HS; ICD-10 codes I60 to I62) and 6835 of ischemic stroke (IS; ICD-10 codes I63, I64, I66, and I67.8).

Because the Skip-gram algorithm required discrete inputs, values of continuous features were binned into several discrete values. Age was grouped into <18, 18-34, 35-44, 45-59, and ≥60 years. Each laboratory test item was categorized into 2 classes (normal and abnormal) or 3 classes (high, medium, and low) according to the clinical laboratory test references. Other continuous features were grouped into 4 percentile bins (quartiles), each containing one-fourth of all samples (Table S1 in [Multimedia Appendix 1](#)). Therefore, a feature had several discrete values called medical concepts. For example, the feature "sex" had two concepts, male and female. If a patient record was considered as a sentence, the medical concepts in the record were then considered as words in the sentence. All the records composed the training corpus. Features involved in the representation included demographic characteristics, hospital admission, utilization of medical resources, disease diagnoses, physical examinations and procedures, laboratory tests, and medications. Features related to the patient's outcomes,



including LOS, hospital cost, and the discharging route, were used to evaluate the patient representations; thus, they were excluded from both training corpora. The full corpus consisted of 13,757 unique medical concepts derived from 104,752 patient records, while a subcorpus consisted of 3769 unique medical concepts derived from records of 8232 stroke patients.

Medical concepts were initially encoded in one-hot vectors, where the dimension of the one-hot vector equaled the number of distinct concepts in the data set. In the one-hot code scheme, a vocabulary of all the distinct medical concepts in the corpus was generated first; then, each medical concept was represented as a 0-1 vector, where the index of the target concept in the vocabulary was set to 1, and all the others were set to 0.

### Embedding-Based Representation

We used the Skip-gram algorithm [19] to learn the representation of medical features. The Skip-gram algorithm can map words into a low-dimensional real number space where the relevant words were located closely. Assuming that similar words may share similar contexts, the Skip-gram algorithm predicted surrounding words of the current (target) word. The same context prediction was repeated as the target word moving to the next. The goal of the Skip-gram algorithm was to maximize the following average log probability:

$$\frac{1}{T} \sum_{t=1}^T \sum_{j=-c}^c \log p(w_{t+j} | w_t)$$

where  $T$  was the length of the sentence that contained the target word,  $c$  (set to 5 in this study) was the size of the training context (called window size),  $w_t$  and  $w_{t+j}$  denoted the target word and the  $j$ th neighboring words before or after the target word in the training context window,  $v$  represented the  $d$ -dimensional ( $d$  was set to 200 in this study) real number vector of the word, and  $W$  (13,757 and 3769 for full corpus and stroke corpus in this study) was the total number of words in the corpus.

Unlike a natural language sentence with a relatively fixed word order, a medical concept's location in a record was appointed manually. It was difficult to assume that the more relevant the concepts were, the closer they were located in a record. Therefore, medical concepts relevant to the target concept might not appear in the training context window in Equation (1). To reduce the impact of the concept sequence on the Skip-gram algorithm, we used the shuffling mechanism [14] to rearrange the order of medical concepts within each record in the corpus randomly. The shuffled corpus was then used for training embedding vectors. The shuffling-training process was performed 20 times, resulting in 20 embedding vectors corresponding to one medical concept. The average of these vectors was considered the final embedding representation of the concept. Because a patient could only take one medical concept for a certain feature, the feature was therefore represented as an embedding vector. After training with the Skip-gram algorithm, a patient who had  $k$  medical concepts for  $k$  features would have  $k$  real number vectors. The average of these vectors was considered the embedding-based representation for the patient.

## Evaluation of the Representation Schemes

### Evaluation of the Feature Representation

The feature representation was first evaluated visually by mapping the  $d$ -dimensional real number vector space into a two-dimensional space using the t-distributed stochastic neighbor embedding (t-SNE) algorithm [14,20]. We used the software Python 3.7 and the `sklearn.manifold.TSNE` tool for the visualization. The t-SNE algorithm's main parameters were as follows: dimension of the embedded space=2, perplexity=30, learning rate=200, number of iterations=1000, gradient calculation method=Barnes-Hut, and angle=0.5. We compared the reduction visualization of medical concepts' vectors training with different corpora. For the purpose of clarity, 441 diagnosis concepts that occurred in at least ten records in the stroke corpus were mapped into the two-dimensional space. They were divided into 14 categories according to the Clinical Classifications Software code [21] for further analysis.

The embedding-based feature representation was then evaluated on how it could capture the latent association among features. We identified the 10 closest medical concepts from each of the diagnosis, laboratory test, physical examination and procedure, medication, and other feature categories in the low-dimensional embedding space for two index diagnosis concepts: subarachnoid hemorrhage (SAH) and occlusion and stenosis of middle cerebral artery (OSMCA). The similarities between the index diagnosis concepts and others were measured by cosine similarity, which was suitable for numerical vectors.

### Evaluation of the Patient Representation

The distributed embedding technique had the advantage of revealing the potential relevance among samples [19], and the unsupervised clustering analysis was a machine learning task that depended more on the sample relevance. Therefore, clustering analysis was used for determining whether the proposed patient representation had a certain advantage in revealing the potential relevance among patients, thus making the clustering solution more aggregative. For the purpose of comparison, 6 embedding-based patient representation schemes and 2 reference schemes were employed. Four embedding-based representations were generated using the initial full corpus, the initial stroke corpus, the shuffled full corpora, and the shuffled stroke corpora as the training corpus. Additionally, to explore the impacts of the numbers of features included in the training context on the representations, we also designed two representation learning schemes that used the initial full and stroke corpora with the maximum window sizes of 255 and 224, respectively. The maximum window size was the length of the record that had the most medical concepts in the corpus. Two commonly used data representation methods were used as the reference methods; one was the multi-hot representation, which was the bitwise summations of one-hot codes for all features, and the other was the mixture of multi-hot codes for discrete features and original values for continuous features. In the mixture representation, we selected 59 laboratory tests in at least 90% of stroke patients. Missing values in the laboratory tests were interpolated using the median of the corresponding laboratory tests. Figure S1 in [Multimedia Appendix 1](#) depicts

the representation schemes used in this study with simple examples.

We conducted k-means clustering analyses on the stroke patients, using cosine distance for the embedding-based representations, Jaccard distance [22] for the multi-hot representation, and Jaccard distance (for discrete features) plus cosine distance (for continuous features) for the mixture representation. We evaluated clustering solutions by Hopkins statistics [23], Silhouette index (SI) [24], and Davies-Bouldin index (DBI) [24]. Hopkins statistics describe the uniformity of data for clustering, while the SI validates the consistency within clusters, and the DBI measures the average similarity between each cluster and the one that most resembles it. The values of Hopkins statistics and DBI range from 0 to 1, while the value of SI ranges from -1 to 1. Higher Hopkins statistics and SI and lower DBI suggest better clustering results. SI was also used to compare k-means clustering solutions for different values of k to determine the optimal number of clusters in this study. The features related to the patient's outcomes were compared to identify the differences between the clusters. Clustering

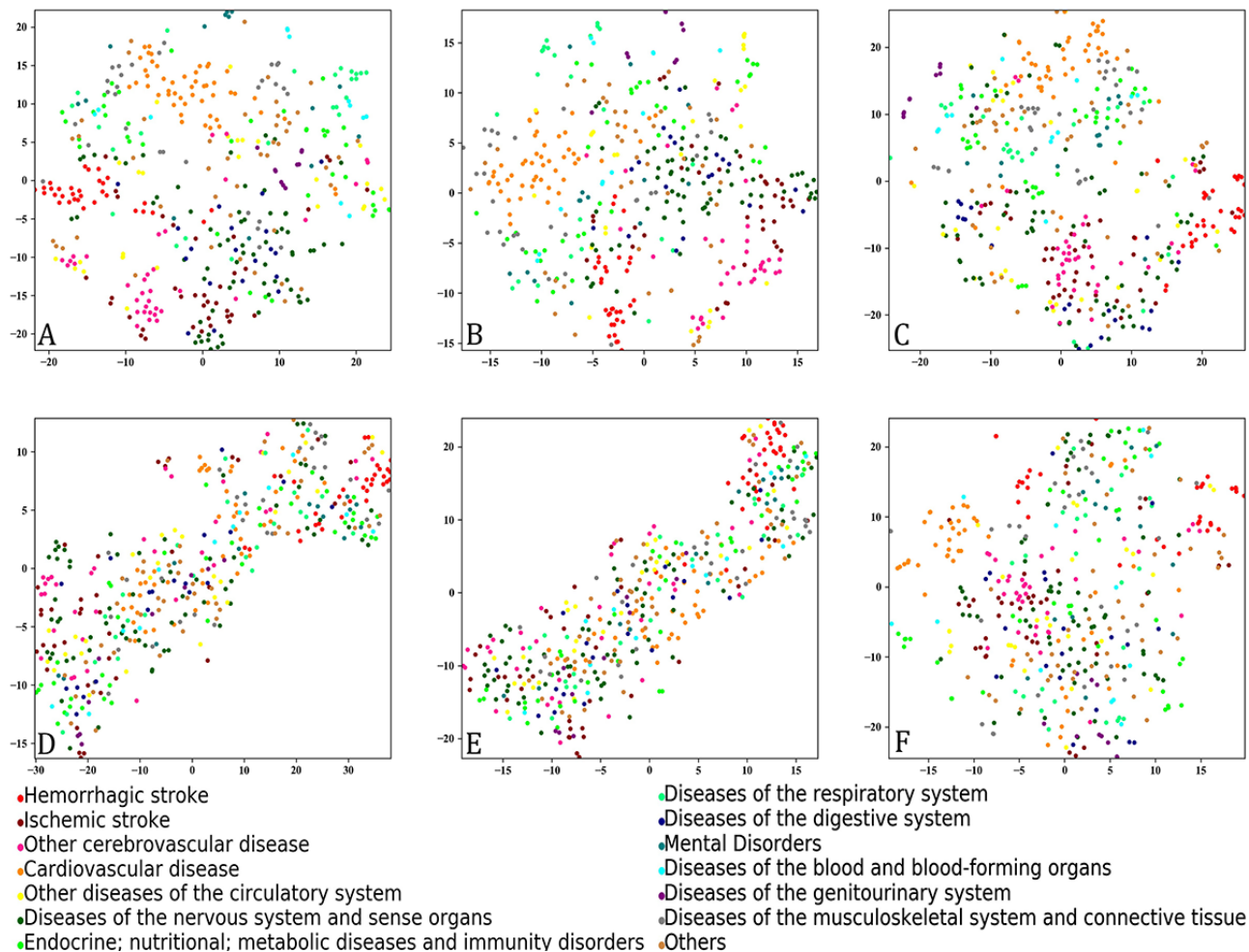
solutions were also assessed concerning the clinical characteristics, including demographic characteristics, utilization of medical resources, disease diagnoses, laboratory tests, procedures, and patient outcomes. Differences in these clinical features were compared among clusters by statistical tests, aiming to confirm whether the knowledge discovered by the clustering analyses was consistent with the clinical facts or new to the medical domain.

## Results

### Feature Representation Visualization

Figure 1 shows the embedding vectors for disease concepts trained with different corpora in the two-dimension space. Vectors for disease concepts trained with the stroke corpus (Figures 1D and E) were more concentrated than those with the full corpus (Figures 1A and B). Further, disease vectors trained with the shuffled full corpus (Figure 1A) showed stronger disease aggregation compared with those trained with the initial full corpus (Figure 1B) and those trained with the full corpus using the maximum window size (Figure 1C).

**Figure 1.** Embedding vectors of diagnosis concepts in the t-distributed stochastic neighbor embedding space. The embedding vectors were trained by Skip-gram algorithm with a window size of 5 from (A) the shuffled full corpora, (B) the initial full corpus, (D) the shuffled stroke corpora, and (E) the initial stroke corpus, with a window size of 255 from (C) the initial full corpus, and with a window size of 224 from (F) the initial stroke corpus.



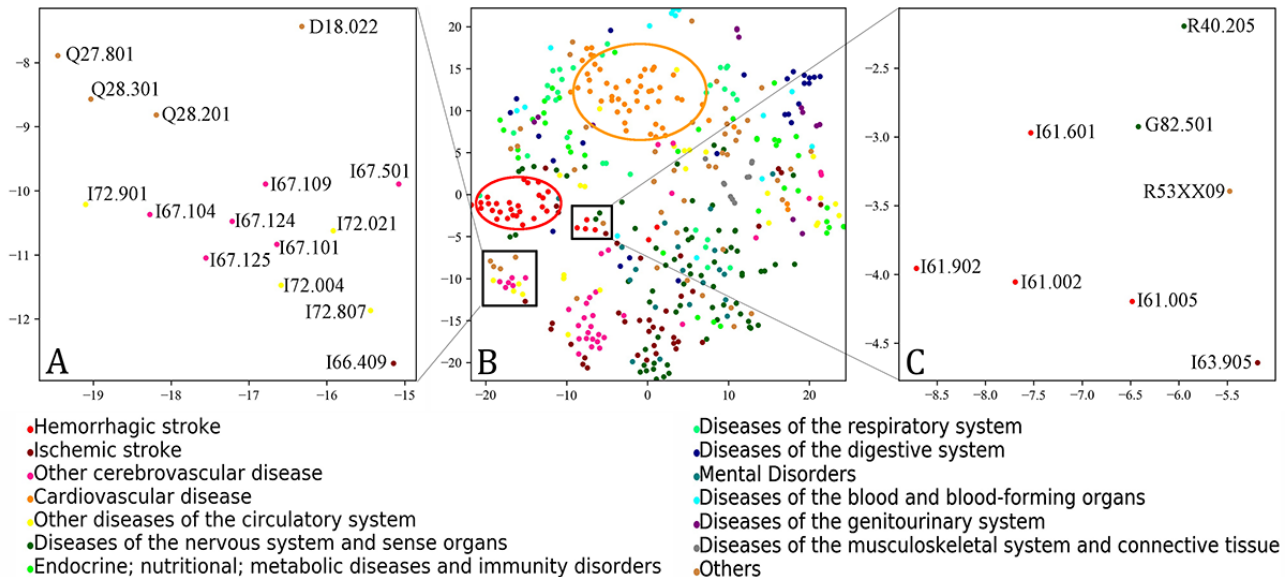
As shown in Figure 2B, most of the disease concepts related to hemorrhagic stroke (red dots) and cardiovascular disease (orange dots) were concentrated inside the red and orange circles,

respectively, suggesting that disease concepts of the same categories were more likely to come together in the embedding space. Further, the nearby medical concepts in the embedding

space were usually clinically correlated. For example, in the embedding space, the disease concepts coded by different ICD-10 codes but in the same rough disease category were able to gather together (eg, circulatory system disease with ICD-10 codes beginning with Q2 and I in Figure 2A). Additionally, as shown in Figure 2C, diseases of abnormal perception symptoms

and signs such as coma (ICD-10 code, R40.205), tetraplegia (G82.501), and malaise and fatigue (R53xx09) were adjacent to cerebrovascular diseases of intracerebral hemorrhage (I61.002, I61.005, I61.601, and I61.902) and cerebral infarction (I63.905). This was consistent with the clinical facts [25].

**Figure 2.** Visualization of the selected 441 diagnosis concepts in the embedding space. (A) and (C) are the locally enlarged areas in black rectangle boxes of (B), in which the embedding vectors were trained from the full corpus with the concept shuffling and were projected to a place by the t-distributed stochastic neighbor embedding technique.



### Features Correlation Analysis

Table 1 lists the 3 closest medical concepts (from different feature categories) to two cerebrovascular diseases: SAH and OSMCA. Among medical concepts of different categories, even if heterogeneous, clinically relevant concepts could be identified by the cosine similarity among concept vectors. For example, the closest laboratory tests to SAH were red and turbid cerebrospinal fluid, consistent with clinical fact. Moreover, the embedding vectors could reveal more detailed information about the medical concepts in the same rough category. For two diagnosis concepts of typical cerebrovascular diseases, SAH and OSMCA, the closest procedures were aneurysm clipping

and percutaneous drug-eluting stent implantation, which were usually used for treating SAH and OSMCA in clinical, respectively. Besides, the closest concepts to the same index concept were not precisely the same when their representations were training with the full and stroke corpus, but both were clinically relevant to the index concept. We also noticed that the cosine similarities between the index concept and their closest concepts in the stroke corpus were larger than in the full corpus. For example, the average of the similarities of the disease SAH and its 15 closest medical concepts were 0.910 and 0.973 in the full and the stroke corpus, respectively. Table S2 in Multimedia Appendix 1 shows the 10 medical concepts closest to SAH and OSMCA for each feature category.

**Table 1.** The 15 closest medical concepts whose embedding representations were trained with the full corpus and the stroke corpus of the disease concepts subarachnoid hemorrhage and the occlusion and stenosis of middle cerebral artery.

Category	Occlusion and stenosis of middle cerebral artery		Subarachnoid hemorrhage			
	Closest concept <sup>a</sup>	Similarity	Closest concept <sup>a</sup>	Similarity	Closest concept <sup>b</sup>	Similarity
Disease diagnoses	Occlusion and stenosis of anterior cerebral artery	0.964	Subarachnoid hemorrhage from anterior communicating artery	0.932	Subarachnoid hemorrhage from posterior communicating artery	0.976
	Occlusion and stenosis of multiple and bilateral cerebral arteries	0.962	Subarachnoid hemorrhage from posterior communicating artery	0.929	Subarachnoid hemorrhage from anterior communicating artery	0.975
	Occlusion and stenosis of posterior cerebral artery	0.958	Bronchitis, not specified as acute or chronic	0.925	Aneurysm	0.971
Laboratory tests	Platelet aggregation test with turbidimetry: high	0.915	Cerebrospinal fluid color: red	0.933	Cerebrospinal fluid transparency: turbid	0.975
	Plasma protein C: high	0.914	Cerebrospinal fluid transparency: turbid	0.904	Cerebrospinal fluid color: blood color	0.959
	Platelet aggregation test with turbidimetry: low	0.910	Cerebrospinal fluid color: orange	0.863	White blood cell count in cerebrospinal fluid: high	0.958
Procedures	Percutaneous drug-eluting stent implantation	0.861	Embolization of intracranial aneurysm	0.985	Embolization of intracranial aneurysm	0.986
	Percutaneous drug-eluting stent implantation of subclavian artery	0.848	Aneurysm clipping	0.974	Aneurysm clipping	0.974
	Transcranial angioplasty	0.822	Embolization of intracranial vessels	0.960	Skull titanium plate placement	0.965
Medications	Probucol tablet	0.938	Hypertonic sodium chloride hydroxyethyl starch 40 injection	0.938	Tramadol	0.987
	Songling Xuemaikang capsule <sup>c</sup>	0.924	Nimodipine	0.895	Fasudil	0.983
	Yufeng Ningxin Drop Pills <sup>c</sup>	0.920	Fructose sodium diphosphate injection	0.894	Dezocine injection	0.982
Others	Allergic to metformin	0.858	Neurosurgery ICU <sup>d</sup>	0.924	ICU of Neurosurgery department	0.976
	Allergic to vinpocetine	0.852	Ventilator utilization	0.796	Discharge department: Neurosurgery department	0.964
	Allergic to iopromide	0.852	Discharge department: Neurosurgery department	0.796	Admission department: Neurosurgery department	0.962

<sup>a</sup>Embedding vectors of concepts were trained with the full corpora.

<sup>b</sup>Embedding vectors of concepts were trained with the stroke corpora.

<sup>c</sup>Traditional Chinese medication.

<sup>d</sup>ICU: intensive care unit.

## Patient Clustering Analysis

In the k-means clustering analyses, the optimal k was determined to be 2, where the corresponding SI value was the highest when k changed from 2 to 15 in each of the representation schemes (Figure S2 in [Multimedia Appendix 1](#)). The greatest values of

Hopkins statistics (0.931) and SI (0.862) and the lowest value of DBI (0.551) were seen in the clustering solution in which patients were represented by the embedding vectors ([Table 2](#)), suggesting that patients with the embedding vectors could be clustered with higher uniformity and aggregation and lower dispersion.



**Table 2.** Clustering performance on interval evaluation indexes based on various patient representations.

Representation schemes	Parameters for training			Cluster evaluation indexes		
	Corpus used	Corpus with shuffling	Window size	Hopkins statistic	Silhouette index	Davies-Bouldin index
Embedding-based representation	Full	Yes	5	0.922	0.783	1.067
	Stroke	Yes	5	0.913	0.862 <sup>a</sup>	0.551 <sup>b</sup>
	Full	No	5	0.903	0.685	1.711
	Stroke	No	5	0.925	0.672	1.382
	Full	No	255	0.922	0.783	1.065
	Stroke	No	224	0.931 <sup>c</sup>	0.790	0.772
Multi-hot representation <sup>d</sup>	N/A <sup>e</sup>	N/A	N/A	0.813	0.233	3.236
Mixture representation <sup>f</sup>	N/A	N/A	N/A	0.918	0.141	4.157

<sup>a</sup>Highest value of the Silhouette index.

<sup>b</sup>Lowest value of the Davies-Bouldin index.

<sup>c</sup>Highest value of the Hopkins statistic.

<sup>d</sup>Multi-hot representation: representation method of the combinations of one-hot codes.

<sup>e</sup>N/A: not applicable.

<sup>f</sup>Mixture representation: representation method of the combination of multi-hot codes for discrete features and real numbers for continuous values of age and laboratory tests.

Among the 8 clustering solutions, cluster 1 contained an average of 6869 (range 6214-7704) patients, of whom 92.2% (range 85.5%-95.7%) had a primary diagnosis of IS. Cluster 2 contained an average of 1363 (range 528-2018) patients, of whom 63.1% (range 51.2%-74.5%) had a primary diagnosis of HS. Therefore, we used IS as the label of patients in cluster 1 and HS as the label of patients in cluster 2. Among the embedding-based representations, the representation trained with the shuffled full corpus reached the greatest F1 scores of 0.944 and 0.717 for clusters 1 and 2, respectively (Table 3). In this clustering solution, 95.0% (6495/6835) of the IS patients and 69.4% (970/1397) of the HS patients were correctly grouped into clusters 1 and 2, respectively. Among the patients (340/6835, 5.0%) with a primary diagnosis of IS who were grouped into cluster 2, 9.4% (32/340) of them had HS as the secondary diagnosis. Meanwhile, among the patients (427/1397, 30.6%) with a primary diagnosis of HS who were grouped into cluster 1, 48.9% (209/427) of them had IS as the secondary

diagnosis. In this situation, the clustering performance might be underestimated.

Between clusters 1 and 2 of stroke patients represented by the embedding vectors learned from shuffled full corpus, there were significant differences in mortality rate (45/6922, 0.65% vs 91/1310, 6.95%,  $P<.001$ ), cost per hospital stay (17.7 vs 113.0 thousand yuan renminbi,  $P<.001$ ), and LOS (9.8 vs 12.6 days,  $P<.001$ ). Patients in cluster 2 occupied more medical resources than those in cluster 1 concerning the ventilator (544/1310, 41.5% vs 105/6922, 1.5%,  $P<.001$ ) and intensive care unit (1025/1310, 78.2% vs 353/6922, 5.1%,  $P<.001$ ). This might partially be linked to the fact that patients in cluster 2 usually also had such acute diseases as pneumonia (189/1310, 14.4% vs 318/6922, 4.6%,  $P<.001$ ), while patients in cluster 1 had chronic diseases like paralysis (3735/6922, .54.0% vs 119/1310, 9.1%,  $P<.001$ ). Table S3 in Multimedia Appendix 1 depicts more comparisons.



**Table 3.** Clustering performance on interval evaluation indexes based on various patient representations.

Representation	Parameters for training			True label	Cluster 1 patients, n	Cluster 2 patients, n	Evaluation indexes		
	Corpus used	Shuffle	Window size				Precision	Recall	F1 score
Embedding-based	Full	Yes	5	IS <sup>a</sup>	6495	340	0.938	0.950	0.944 <sup>b</sup>
				HS <sup>c</sup>	427	970	0.740	0.694	0.717 <sup>d</sup>
	Stroke	Yes	5	IS	6530	305	0.928	0.955	0.942
				HS	506	891	0.745	0.638	0.687
	Full	No	5	IS	6587	248	0.855	0.964	0.906
				HS	1117	280	0.530	0.200	0.291
	Stroke	No	5	IS	6472	363	0.903	0.947	0.924
				HS	699	698	0.658	0.500	0.568
	Full	No	255	IS	6305	530	0.927	0.922	0.925
				HS	493	904	0.630	0.647	0.639
	Stroke	No	224	IS	6378	457	0.932	0.933	0.932
				HS	467	930	0.671	0.666	0.668
Multi-hot <sup>e</sup>	N/A <sup>f</sup>	N/A	N/A	IS	5874	961	0.938	0.859	0.897
				HS	388	1009	0.512	0.722	0.599
Mixture <sup>g</sup>	N/A	N/A	N/A	IS	5945	890	0.957	0.870	0.911
				HS	269	1128	0.559	0.807	0.661

<sup>a</sup>IS: ischemic stroke.

<sup>b</sup>Highest F1 score for cluster 1.

<sup>c</sup>HS: hemorrhagic stroke.

<sup>d</sup>Highest F1 score for cluster 2.

<sup>e</sup>Multi-hot: representation method of the combinations of one-hot codes.

<sup>f</sup>N/A: not applicable.

<sup>g</sup>Mixture: representation method of the combination of multi-hot codes for discrete features and real numbers for continuous values of age and laboratory tests.

## Discussion

### Principal Findings

Representation for structured medical data is critical for data mining tasks in the medical domain [3,5,6,14]. The one-hot code scheme is a simple and widely used representation. However, it may be unsuitable for the complex and diverse EMR data due to its high dimensionality and sparsity. Analyses of massive one-hot coded data may require greater computational power because of not only their high-dimensional and sparse nature but also the unclear potential relevance of the data [26]. Therefore, many studies have focused on effective and efficient data representation. In this study, we adopted an embedding-based method derived from NLP techniques to represent the structured patient data. The proposed representations brought a deep and intuitive insight into associations among medical concepts and a great performance improvement in a similarity-based data mining task.

The distributed embedding representations have the merits of low dimensionality and the capability for revealing the latent relationship among the represented objects [19]. Thus, the embedding-based or deep learning-based representation has

been widely used in various applications, especially in the clinical NLP domains, to represent unstructured medical texts, including biomedical publications [27], clinical notes [28], and radiology reports [29-31]. With these representations, researchers could perform feature engineering with less expert effort and transform raw texts into low-dimensional dense vectors with clinical meanings and further identify implicit patterns in patients. Inspired by the representation learning from the unstructured medical data, researchers adopted these representation methods for structured medical data, including medical codes such as diagnosis codes, procedure codes, and drug codes [5,32], laboratory tests [12], and time-related data, which was informative for patients [1,12,33,34].

In this study, we borrowed the idea from this originally text-oriented technique and applied it to sEMR data with diverse patient features. We embedded each medical concept into a low-dimensional real number vector using the Skip-gram algorithm. Both the visualized and quantitative analyses showed that the embedding-based feature representation provided a relatively clear understanding of the associations and connections among the medical concepts, which were consistent with medical knowledge and clinical practice. On the other hand, clustering solutions on patients represented with

embedding vectors showed a better clustering nature than those expressed with multi-hot vectors. The embedding-based representation showed advantages in dimension reduction and in the convenience of numerical computation and association mining in this study.

An informative representation was usually derived from different modalities and medical data sources, such as cross-sectional and longitudinal data, and quantitative indexes and narrative notes. In this study, demographic characteristics, diagnoses, physical examinations and procedures, laboratory tests, medications, and hospital admission and discharge were all brought into the feature representation learning. A particular and unavoidable characteristic of laboratory tests was that patients might take different laboratory test items according to the need for diagnosis and treatment. This must result in lots of missing values for laboratory tests. The joint use of the discretization of continuous values and the Skip-gram algorithm solved the problem, making all the available features to be fully used. Clustering analyses showed that patients represented by embedding vectors were more likely to cluster together than those represented in the original form, where about three-quarters of laboratory tests were dropped due to missing values. It may partially attribute to the inclusion of all the features and discretization of the continuous features.

When using the Skip-gram algorithm for representation learning for sEMR data, several adaptive changes had been made. First, we applied the shuffling mechanism when building the training corpus to reduce the impact of the concept order on the coverage of the training context. Glicksberg et al [14] randomly shuffled the medical concepts within a time interval. We further extended the idea of shuffling concepts. The medical concepts were rearranged randomly within a patient record 20 times. The resulting 20 embedded vectors for each medical concept trained with different shuffled corpora were then averaged as the final embedding vector. Results from several evaluation tasks showed that the shuffling-based representation at both the feature and patient level had a more satisfactory performance compared with their not shuffling-based counterparts.

In the Skip-gram algorithm, the range of training context was also crucial to the algorithm performance. For the same reason as for using a shuffling method, we set the window size to the maximum to have the training context covered the most neighboring concepts. However, there was no outstanding performance improvement in the clustering task. The finding was consistent with other studies [33,35] that found performances got worse as the window size increased. It indicated that wide training context might introduce redundant information or even noise to the training. Besides, the corpus used in the Skip-gram algorithm was also linked to the

performance improvement on the clustering task in this study. Stroke patients whose representations were trained with the corpus including all the patient records were clustered into two groups with higher aggregation and lower dispersion than those whose representations were trained with the corpus including only stroke patients' records. The finding was similar to that of a study by Yanshan Wang et al [27] that the embedding-based representation from the public domain corpora showed more satisfactory results in biomedical information retrieval than from the biomedical domain corpora.

### Limitations

Our study had some limitations. First, we did not use time-oriented patient records with critical importance for evaluating the patient course and prognosis. The history of medical events may affect future medical events; these medical sequence data are crucial for clinical diagnosis and treatment. Rich time-oriented data, including time-series features in an inpatient record and the temporality between multiple inpatient records, were used for learning patient representations by some algorithms targeted at sequence data, such as recurrent neural network [12], time-aware attention method [33], Deepr [36], and Patient2Vec [37]. Those time-related representations, which captured patients' sequential information from a longitudinal perspective, could be used for supervised prediction tasks [12,36,37], and the unsupervised task-like disease clustering analyses at the feature level [33]. In contrast, we just took the cross-sectional data with diverse feature types of patients, focusing on a certain hospitalization's static characteristics, into the effective representations by Skip-gram algorithm at both feature and patient levels. Second, a patient representation was just a simple average of the embedding vectors for features with equal weights. This may be not completely consistent with the fact that clinical features may have different importance to the diagnosis of a specific disease. Last, we only evaluated the effectiveness of the embedding-based patient representation with clustering analysis. The proposed patient representation needs more validation in various clinically meaningful tasks.

### Conclusions

In this study, we applied an embedding technique in learning the patient representation from sEMR data with different types of clinical features. With the original Skip-gram algorithm's adaptive changes, the embedding-based representations could somehow reflect the potential associations among features and patients. The performance improvement in a clinically meaningful clustering task suggested the proposed patient representation's effectiveness and efficiency. It is expected that the embedding-based representation will be helpful in a wide range of secondary uses of EMR data.

---

### Acknowledgments

This work was supported by the National Natural Science Foundation of China (grants 81971707 and 81701792).

---

### Conflicts of Interest

None declared.

Multimedia Appendix 1  
Supplementary materials.

[DOCX File , 351 KB - [medinform\\_v9i7e19905\\_app1.docx](#) ]

## References

1. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Health Inform* 2018 Dec;22(5):1589-1604. [doi: [10.1109/JBHI.2017.2767063](#)] [Medline: [29989977](#)]
2. Huang Z, Dong W, Duan H, Liu J. A Regularized Deep Learning Approach for Clinical Risk Prediction of Acute Coronary Syndrome Using Electronic Health Records. *IEEE Trans Biomed Eng* 2018 May;65(5):956-968. [doi: [10.1109/TBME.2017.2731158](#)] [Medline: [28742027](#)]
3. Wang L, Tong L, Davis D, Arnold T, Esposito T. The application of unsupervised deep learning in predictive models using electronic health records. *BMC Med Res Methodol* 2020 Feb 26;20(1):37 [FREE Full text] [doi: [10.1186/s12874-020-00923-1](#)] [Medline: [32101147](#)]
4. He J, Hu Y, Zhang X, Wu L, Waitman LR, Liu M. Multi-perspective predictive modeling for acute kidney injury in general hospital populations using electronic medical records. *JAMIA Open* 2019 Apr;2(1):115-122 [FREE Full text] [doi: [10.1093/jamiaopen/ooy043](#)] [Medline: [30976758](#)]
5. Wang Z, Zhu Y, Li D, Yin Y, Zhang J. Feature rearrangement based deep learning system for predicting heart failure mortality. *Comput Methods Programs Biomed* 2020 Feb 06;191:105383. [doi: [10.1016/j.cmpb.2020.105383](#)] [Medline: [32062185](#)]
6. Cui L, Xie X, Shen Z. Prediction task guided representation learning of medical codes in EHR. *J Biomed Inform* 2018 Aug;84:1-10 [FREE Full text] [doi: [10.1016/j.jbi.2018.06.013](#)] [Medline: [29928997](#)]
7. Xiao C, Ma T, Dieng AB, Blei DM, Wang F. Readmission prediction via deep contextual embedding of clinical concepts. *PLoS One* 2018;13(4):e0195024 [FREE Full text] [doi: [10.1371/journal.pone.0195024](#)] [Medline: [29630604](#)]
8. Barbieri S, Kemp J, Perez-Concha O, Kotwal S, Gallagher M, Ritchie A, et al. Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and Describing Patients-at-Risk. *Sci Rep* 2020 Jan 24;10(1):1111 [FREE Full text] [doi: [10.1038/s41598-020-58053-z](#)] [Medline: [31980704](#)]
9. Wang Y, Tian Y, Tian L, Qian Y, Li J. An electronic medical record system with treatment recommendations based on patient similarity. *J Med Syst* 2015 May;39(5):55. [doi: [10.1007/s10916-015-0237-z](#)] [Medline: [25762458](#)]
10. Kruser JM, Benjamin BT, Gordon EJ, Michelson KN, Wunderink RG, Holl JL, et al. Patient and Family Engagement During Treatment Decisions in an ICU: A Discourse Analysis of the Electronic Health Record. *Crit Care Med* 2019 Jun;47(6):784-791. [doi: [10.1097/CCM.0000000000003711](#)] [Medline: [30896465](#)]
11. Santoro SL, Bartman T, Cua CL, Lemle S, Skotko BG. Use of Electronic Health Record Integration for Down Syndrome Guidelines. *Pediatrics* 2018 Sep;142(3):2017-4119 [FREE Full text] [doi: [10.1542/peds.2017-4119](#)] [Medline: [30154119](#)]
12. Ruan T, Lei L, Zhou Y, Zhai J, Zhang L, He P, et al. Representation learning for clinical time series prediction tasks in electronic health records. *BMC Med Inform Decis Mak* 2019 Dec 17;19(Suppl 8):259 [FREE Full text] [doi: [10.1186/s12911-019-0985-7](#)] [Medline: [31842854](#)]
13. Oh W, Steinbach MS, Castro MR, Peterson KA, Kumar V, Caraballo PJ, et al. Evaluating the Impact of Data Representation on EHR-Based Analytic Tasks. *Stud Health Technol Inform* 2019 Aug 21;264:288-292 [FREE Full text] [doi: [10.3233/SHTI190229](#)] [Medline: [31437931](#)]
14. Glicksberg B, Miotto R, Johnson K, Shameer K, Li L, Chen R. Automated disease cohort selection using word embeddings from Electronic Health Records. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2018;23:145-156. [doi: [10.1142/9789813235533\\_0014](#)] [Medline: [29218877](#)]
15. Ning W, Chan S, Beam A, Yu M, Geva A, Liao K, et al. Feature extraction for phenotyping from semantic and knowledge resources. *J Biomed Inform* 2019 Mar;91:103122 [FREE Full text] [doi: [10.1016/j.jbi.2019.103122](#)] [Medline: [30738949](#)]
16. GBD 2016 Neurology Collaborators. Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol* 2019 May;18(5):459-480 [FREE Full text] [doi: [10.1016/S1474-4422\(18\)30499-X](#)] [Medline: [30879893](#)]
17. Pana TA, Wood AD, Mamas MA, Clark AB, Bettencourt-Silva JH, McLernon DJ, Norfolk and Norwich Stroke and TIA Register Steering Committee Collaborators. Myocardial infarction after acute ischaemic stroke: Incidence, mortality and risk factors. *Acta Neurol Scand* 2019 Sep;140(3):219-228. [doi: [10.1111/ane.13135](#)] [Medline: [31140583](#)]
18. Chen H, Shi L, Wang N, Han Y, Lin Y, Dai M, et al. Analysis on geographic variations in hospital deaths and endovascular therapy in ischaemic stroke patients: an observational cross-sectional study in China. *BMJ Open* 2019 Jun 24;9(6):e029079 [FREE Full text] [doi: [10.1136/bmjopen-2019-029079](#)] [Medline: [31239305](#)]
19. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. *J. 2013 Dec Presented at: Neural Information Processing Systems, 26; 2013; Lake Tahoe, Nevada.*
20. Laurens VDM, Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research* 2008 Jan;9(86):2579-2605.
21. Agency for Healthcare Research and Quality. HCUP clinical classification software (CCS) for ICD-9CM. URL: <https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp> [accessed 2020-12-12]

22. Hier DB, Kopel J, Brint SU, Wunsch DC, Olbricht GR, Azizi S, et al. Evaluation of standard and semantically-augmented distance metrics for neurology patients. *BMC Med Inform Decis Mak* 2020 Aug 26;20(1):203 [FREE Full text] [doi: [10.1186/s12911-020-01217-8](https://doi.org/10.1186/s12911-020-01217-8)] [Medline: [32843023](https://pubmed.ncbi.nlm.nih.gov/32843023/)]
23. Qiu B, Cao X. Clustering boundary detection for high dimensional space based on space inversion and Hopkins statistics. *Knowledge-Based Systems* 2016 Apr;98:216-225. [doi: [10.1016/j.knosys.2016.01.035](https://doi.org/10.1016/j.knosys.2016.01.035)]
24. Vendramin L, Campello RJGB, Hruschka ER. Relative clustering validity criteria: A comparative overview. *Statistical Anal Data Mining* 2010 Jun 30;3(4):209-235. [doi: [10.1002/sam.10080](https://doi.org/10.1002/sam.10080)]
25. Koga M, Iguchi Y, Ohara T, Tahara Y, Fukuda T, Noguchi T, et al. Acute ischemic stroke as a complication of Stanford type A acute aortic dissection: a review and proposed clinical recommendations for urgent diagnosis. *Gen Thorac Cardiovasc Surg* 2018 Aug;66(8):439-445. [doi: [10.1007/s11748-018-0956-4](https://doi.org/10.1007/s11748-018-0956-4)] [Medline: [29948797](https://pubmed.ncbi.nlm.nih.gov/29948797/)]
26. Singh N, Garg N, Pant J. A Comprehensive Study of Challenges and Approaches for Clustering High Dimensional Data. *IJCA* 2014 Apr 18;92(4):7-10. [doi: [10.5120/15995-4844](https://doi.org/10.5120/15995-4844)]
27. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A Comparison of Word Embeddings for the Biomedical Natural Language Processing. *J Biomed Inform* 2018 Sep 11:12-20. [doi: [10.1016/j.jbi.2018.09.008](https://doi.org/10.1016/j.jbi.2018.09.008)] [Medline: [30217670](https://pubmed.ncbi.nlm.nih.gov/30217670/)]
28. Weng W, Waghlikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Med Inform Decis Mak* 2017 Dec 01;17(1):155-155 [FREE Full text] [doi: [10.1186/s12911-017-0556-8](https://doi.org/10.1186/s12911-017-0556-8)] [Medline: [29191207](https://pubmed.ncbi.nlm.nih.gov/29191207/)]
29. Liu H, Zhang Z, Xu Y, Wang N, Huang Y, Yang Z, et al. Use of BERT (Bidirectional Encoder Representations from Transformers)-Based Deep Learning Method for Extracting Evidences in Chinese Radiology Reports: Development of a Computer-Aided Liver Cancer Diagnosis Framework. *J Med Internet Res* 2021 Jan 12;23(1):e19689 [FREE Full text] [doi: [10.2196/19689](https://doi.org/10.2196/19689)] [Medline: [33433395](https://pubmed.ncbi.nlm.nih.gov/33433395/)]
30. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language Processing in Radiology: A Systematic Review. *Radiology* 2016 May;279(2):329-343. [doi: [10.1148/radiol.16142770](https://doi.org/10.1148/radiol.16142770)] [Medline: [27089187](https://pubmed.ncbi.nlm.nih.gov/27089187/)]
31. Liu H, Xu Y, Zhang Z, Wang N, Huang Y, Hu Y, et al. A Natural Language Processing Pipeline of Chinese Free-Text Radiology Reports for Liver Cancer Diagnosis. *IEEE Access* 2020 Aug;8:159110-159119. [doi: [10.1109/access.2020.3020138](https://doi.org/10.1109/access.2020.3020138)]
32. Choi Y, Chiu CY, Sontag D. Learning Low-Dimensional Representations of Medical Concepts. *AMIA Jt Summits Transl Sci Proc* 2016;2016:41-50 [FREE Full text] [Medline: [27570647](https://pubmed.ncbi.nlm.nih.gov/27570647/)]
33. Cai X, Gao J, Ngiam KY, Ooi BC, Zhang Y, Yuan X. Medical Concept Embedding with Time-Aware Attention. 2018 Jul Presented at: The 27th International Joint Conference on Artificial Intelligence; 2018; Stockholm, Sweden.
34. Pham T, Tran T, Phung D, Venkatesh S. Predicting healthcare trajectories from medical records: A deep learning approach. *J Biomed Inform* 2017 May;69:218-229 [FREE Full text] [doi: [10.1016/j.jbi.2017.04.001](https://doi.org/10.1016/j.jbi.2017.04.001)] [Medline: [28410981](https://pubmed.ncbi.nlm.nih.gov/28410981/)]
35. Ling W, Tsvetkov Y, Amir S, Fernandez R, Lin CC. Not all contexts are created equal: better word representations with variable attention. 2015 Sep Presented at: 2015 Conference on Empirical Methods in Natural Language Processing; 2015; Lisbon, Portugal. [doi: [10.18653/v1/d15-1161](https://doi.org/10.18653/v1/d15-1161)]
36. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S.  $\text{DeepPr}$ : A Convolutional Net for Medical Records. *IEEE J Biomed Health Inform* 2017 Dec;21(1):22-30. [doi: [10.1109/JBHI.2016.2633963](https://doi.org/10.1109/JBHI.2016.2633963)] [Medline: [27913366](https://pubmed.ncbi.nlm.nih.gov/27913366/)]
37. Zhang J, Kowsari K, Harrison JH, Lobo JM, Barnes LE. Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record. *IEEE Access* 2018;6:65333-65346. [doi: [10.1109/access.2018.2875677](https://doi.org/10.1109/access.2018.2875677)]

## Abbreviations

**DBI:** Davies-Bouldin index

**EMR:** electronic medical record

**HS:** hemorrhagic stroke

**ICD-10:** International Classification of Diseases, Tenth Revision

**ICD-9-CM:** International Classification of Diseases, Ninth Revision, Clinical Modification

**IS:** ischemic stroke

**LOS:** length of stay

**NLP:** natural language processing

**OSMCA:** occlusion and stenosis of middle cerebral artery

**SAH:** subarachnoid hemorrhage

**sEMR:** structured electronic medical record

**SI:** Silhouette index

**t-SNE:** t-distributed stochastic neighbor embedding

*Edited by C Lovis; submitted 06.05.20; peer-reviewed by J Lei, S Barbieri; comments to author 26.10.20; revised version received 18.12.20; accepted 05.06.21; published 23.07.21.*

*Please cite as:*

*Huang Y, Wang N, Zhang Z, Liu H, Fei X, Wei L, Chen H*

*Patient Representation From Structured Electronic Medical Records Based on Embedding Technique: Development and Validation Study*

*JMIR Med Inform 2021;9(7):e19905*

*URL: <https://medinform.jmir.org/2021/7/e19905>*

*doi: [10.2196/19905](https://doi.org/10.2196/19905)*

*PMID: [34297000](https://pubmed.ncbi.nlm.nih.gov/34297000/)*

©Yanqun Huang, Ni Wang, Zhiqiang Zhang, Honglei Liu, Xiaolu Fei, Lan Wei, Hui Chen. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 23.07.2021. 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

# A Machine Learning–Based Algorithm for the Prediction of Intensive Care Unit Delirium (PRIDE): Retrospective Study

Sujeong Hur<sup>1,2\*</sup>, MS; Ryoung-Eun Ko<sup>3\*</sup>, MD; Junsang Yoo<sup>4</sup>, PhD; Juhyung Ha<sup>5</sup>; Won Chul Cha<sup>2,6,7</sup>, MD; Chi Ryang Chung<sup>3</sup>, MD, PhD

<sup>1</sup>Department of Patient Experience Management Part, Samsung Medical Center, Seoul, Republic of Korea

<sup>2</sup>Department of Digital Health, SAIHST, Sungkyunkwan University, Seoul, Republic of Korea

<sup>3</sup>Department of Critical Care Medicine and Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

<sup>4</sup>Department of Nursing, College of Nursing, Sahmyook University, Seoul, Republic of Korea

<sup>5</sup>Department of Computer Science, Indiana University, Bloomington, IN, United States

<sup>6</sup>Department of Emergency Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

<sup>7</sup>Digital Innovation Center, Samsung Medical Center, Seoul, Republic of Korea

\*these authors contributed equally

**Corresponding Author:**

Chi Ryang Chung, MD, PhD

Department of Critical Care Medicine and Medicine

Samsung Medical Center

Sungkyunkwan University School of Medicine

81 Irwon-ro, Gangnam-gu

Seoul, 06351

Republic of Korea

Phone: 82 2 3410 3430

Fax: 82 2 2148 7088

Email: [chiryang.chung@gmail.com](mailto:chiryang.chung@gmail.com)

## Abstract

**Background:** Delirium frequently occurs among patients admitted to the intensive care unit (ICU). There is limited evidence to support interventions to treat or resolve delirium in patients who have already developed delirium. Therefore, the early recognition and prevention of delirium are important in the management of critically ill patients.

**Objective:** This study aims to develop and validate a delirium prediction model within 24 hours of admission to the ICU using electronic health record data. The algorithm was named the Prediction of ICU Delirium (PRIDE).

**Methods:** This is a retrospective cohort study performed at a tertiary referral hospital with 120 ICU beds. We only included patients who were 18 years or older at the time of admission and who stayed in the medical or surgical ICU. Patients were excluded if they lacked a Confusion Assessment Method for the ICU record from the day of ICU admission or if they had a positive Confusion Assessment Method for the ICU record at the time of ICU admission. The algorithm to predict delirium was developed using patient data from the first 2 years of the study period and validated using patient data from the last 6 months. Random forest (RF), Extreme Gradient Boosting (XGBoost), deep neural network (DNN), and logistic regression (LR) were used. The algorithms were externally validated using MIMIC-III data, and the algorithm with the largest area under the receiver operating characteristics (AUROC) curve in the external data set was named the PRIDE algorithm.

**Results:** A total of 37,543 cases were collected. After patient exclusion, 12,409 remained as our study population, of which 3816 (30.8%) patients experienced delirium incidents during the study period. Based on the exclusion criteria, out of the 96,016 ICU admission cases in the MIMIC-III data set, 2061 cases were included, and 272 (13.2%) delirium incidents occurred. The average AUROCs and 95% CIs for internal validation were 0.916 (95% CI 0.916-0.916) for RF, 0.919 (95% CI 0.919-0.919) for XGBoost, 0.881 (95% CI 0.878-0.884) for DNN, and 0.875 (95% CI 0.875-0.875) for LR. Regarding the external validation, the best AUROC were 0.721 (95% CI 0.72-0.721) for RF, 0.697 (95% CI 0.695-0.699) for XGBoost, 0.655 (95% CI 0.654-0.657) for DNN, and 0.631 (95% CI 0.631-0.631) for LR. The Brier score of the RF model is 0.168, indicating that it is well-calibrated.

**Conclusions:** A machine learning approach based on electronic health record data can be used to predict delirium within 24 hours of ICU admission. RF, XGBoost, DNN, and LR models were used, and they effectively predicted delirium. However, with the potential to advise ICU physicians and prevent ICU delirium, prospective studies are required to verify the algorithm's performance.

(*JMIR Med Inform* 2021;9(7):e23401) doi:[10.2196/23401](https://doi.org/10.2196/23401)

## KEYWORDS

clinical prediction; delirium; electronic health record; intensive care unit; machine learning

## Introduction

Delirium, defined as acute brain dysfunction characterized by disturbances of awareness, attention, and cognition with a fluctuating course linked with an underlying medical condition, frequently occurs among patients admitted to intensive care units (ICUs) [1]. Up to 80% of critically ill patients affected by delirium are at an increased risk of requiring ventilation for a substantially long duration, high hospital and ICU mortality, and long-term cognitive impairment. The medical care for these patients also results in increased medical costs [2-4].

There is currently limited evidence to support interventions to treat or resolve delirium in patients who have already developed delirium [5]. Therefore, the early recognition and prevention of delirium are indispensable for patients with a high risk of developing delirium. Previous studies have shown that a proportion of the cases of delirium may be avoidable [6]. Accordingly, several prediction models have been developed to predict delirium in patients who may benefit from delirium prevention [7-9]. The models developed thus far focus on predicting delirium during the entire ICU stay using predisposing clinical features obtained within 24 hours of ICU admission or immediately upon ICU admission. Considering that ICU patients experience dynamic changes in medical conditions within the initial 24 hours after ICU admission, these models are limited because they focus on predicting only the long-term occurrence of delirium during the entire ICU stay.

Furthermore, these prediction models only include variables that have already been identified as risk factors for delirium in other studies [7,9,10].

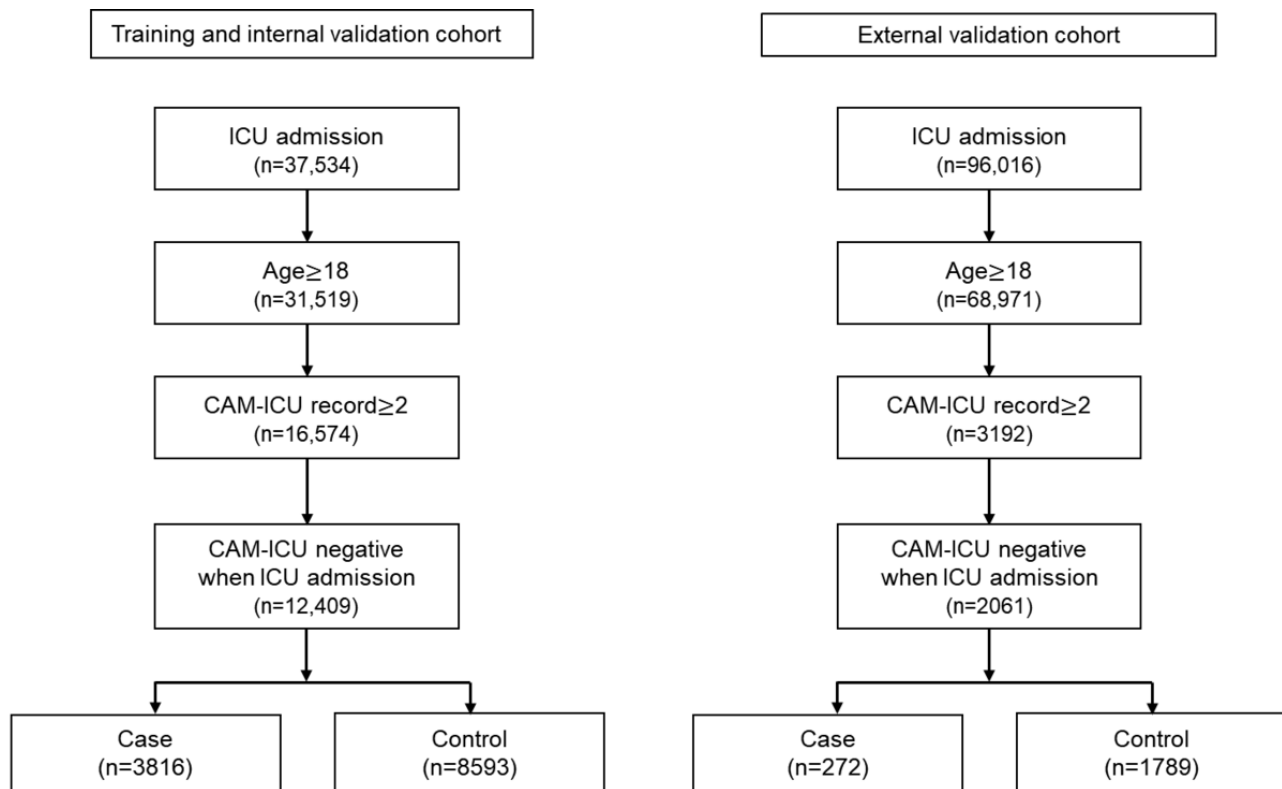
Therefore, we developed a machine learning-based model for the early prediction of delirium among medical and surgical ICU patients using electronic health record (EHR) data. This prediction model uses data obtained within 4 hours of ICU admission to predict delirium within 24 hours after ICU admission.

## Methods

### Study Setting and Population

We conducted a retrospective study of all critically ill patients admitted to the ICUs of the Samsung Medical Center (a 1989-bed university-affiliated, tertiary referral hospital in Seoul, South Korea) from July 1, 2016, to August 31, 2019. We only included patients who were 18 years or older at the time of admission and who stayed in the medical or surgical ICU. Patients were excluded if they lacked a Confusion Assessment Method for the ICU (CAM-ICU) record from the day of ICU admission or if they had a positive CAM-ICU record at the time of ICU admission. The flow diagram in [Figure 1](#) shows the patient selection process. The study protocol was removed from all identifiers and approved by the SMC (Samsung Medical Center) Institutional Review Board (IRB No. 2020-02-026), as all identifiers were removed. The IRB approval form is presented in [Multimedia Appendix 1](#).

**Figure 1.** Flow diagram of the participant selection process. CAM-ICU: confusion assessment method for the intensive care unit; ICU: intensive care unit.



### Source of Data

This study used data from the Clinical Data Warehouse Darwin-C database of the SMC and the Medical Information Mart for Intensive Care III (MIMIC-III) database (v1.4). The SMC data set was used for the derivation and validation cohort, and the MIMIC-III data set was used for the external validation cohort. The MIMIC-III database is a clinical database consisting of data from more than 38,000 ICU patients (medical, surgical, trauma-surgical, coronary, and cardiac-surgery data) admitted to Beth Israel Deaconess Medical center (Boston, MA) from June 2001 to October 2012 [11]. The MIMIC-III database can be accessed upon obtaining approval from its administrators.

### Outcome

To screen for delirium, all ICU patients were assessed with the CAM-ICU [12]. The primary outcome of the study was the prediction of the occurrence of delirium within 24 hours of ICU admission. Delirium was defined as a negative CAM-ICU result obtained within the first 4 hours, and a positive CAM-ICU result obtained between 4 and 24 hours of ICU admission. In our institute, CAM-ICU results were obtained 3 times a day, and a senior nurse rechecked the recorded CAM-ICU scores.

### Predictor Variables

We used clinical characteristics, ICU admission category (medical or surgical), primary cause of admission (respiratory, cardiovascular, gastrointestinal, neurology, perioperative, nephrology, metabolic, or trauma), primary diagnosis, vital signs, prescription medications, and laboratory test results as the predictor variables. All variables were extracted from the EHR data set.

### Feature Selection and Data Processing

We first extracted all relevant variables for the prediction model from other studies. Next, 2 clinical experts (CRC and REK) reviewed the relevant variables and selected the crucial ones based on previous clinical studies and clinical relevance. We then further restricted the variables depending on whether they could be automatically extracted from EHRs and had low missing rates. Finally, for the external validation in the MIMIC-III data set, we selected variables found in both SMC and MIMIC-III. The MIMIC-III data set shows the final variables used as input for model development. The list of variables used is shown in [Textbox 1](#), and the missing rate in the variable list is presented in [Multimedia Appendix 2](#).

**Textbox 1.** Variables used for model development.

<p><b>General information</b></p> <ul style="list-style-type: none"> <li>• Age, sex, and invasive mechanical ventilation</li> </ul> <p><b>Admission category</b></p> <ul style="list-style-type: none"> <li>• Medical intensive care unit (ICU) or surgical ICU</li> </ul> <p><b>Reason for ICU admission</b></p> <ul style="list-style-type: none"> <li>• Respiratory, cardiovascular, gastrointestinal, neurology, perioperative, nephrology, metabolic, and trauma</li> </ul> <p><b>Vital signs</b></p> <ul style="list-style-type: none"> <li>• Systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, peripheral capillary oxygen saturation, and Glasgow Coma Scale (eye, verbal, and motor)</li> </ul> <p><b>Comorbidity</b></p> <ul style="list-style-type: none"> <li>• Charlson Comorbidity Index</li> </ul> <p><b>Laboratory tests</b></p> <ul style="list-style-type: none"> <li>• Complete blood count: white blood count, hemoglobin, hematocrit, platelet count, and erythrocyte sedimentation rate</li> <li>• Coagulation: prothrombin time (INR) and activated partial thromboplastin time</li> <li>• Chemistry: Total protein, albumin, total bilirubin, aspartate aminotransferase, alanine aminotransferase, glucose fasting, blood urea nitrogen, creatinine, phosphorus, sodium, potassium, magnesium, calcium (ionized), C-reactive protein quantitative, and lactic acid</li> <li>• Arterial Blood Gas Analysis: pH, PaCO<sub>2</sub>, PaO<sub>2</sub>, HCO<sub>3</sub>, and O<sub>2</sub> Saturation</li> </ul> <p><b>Medications</b></p> <ul style="list-style-type: none"> <li>• Antibiotics, anticholinergic and antipsychotics, benzodiazepines, miscellaneous antidepressants, anxiolytics, sedatives and hypnotics, vasopressors, opiate agonists, opiate antagonists, cholinergic agents, and steroids</li> </ul>
---

With regard to the general data processing, we first processed invalid values by eliminating them. Invalid values include extreme outliers in numerical values (for example, numerical values for vitals are eliminated using certain rules (ie, heart rate values should be between 0 and 300). Second, we processed numerical values by normalizing and scaling them. We performed standard normalization and min-max scaling such that the final numerical values were between 0 and 1. Finally, we processed the missing values. Missing values in the numerical data were filled with mean values, whereas missing values in categorical data were left blank such that the dummy variables were all equal to 0.

For certain variables with temporal information, such as vital values and laboratory test results, we determined statistical values such as the mean, standard deviation, min, max, and the closest values to the ICU admission to ensure multiple rows of numerical values can be summarized into one. Subsequently, to reduce the number of features necessary to train the model, we picked only one of the statistical values according to the feature importance of the random forest (RF). For example, there were initially multiple values for diastolic blood pressure (DBP) with respect to time. We calculated statistical values

such as the mean, standard deviation, min, max, and the latest DBP values. Finally, we only selected the mean DBP because it was the most important among the statistical values of the DBP according to the feature importance of the RF.

### Model Development and Validation

We split the data set into a development data set and a data set. For the development data set, we used the data obtained between July 1, 2016, and December 31, 2018. For the validation set, we used the data obtained between January 1, 2019, and August 31, 2019. Of the 37,543 admitted cases, 12,409 cases were selected in this study. These were divided into the development set (n=9589, 77.3%) and the internal validation set (n=2820, 22.7%). Among the 9589 cases in the development data set, there were 3060 (31.9%) cases of delirium, and among the 2820 cases in the validation data set, there were 756 (26.8%) cases of delirium. We did not apply specific methodology (eg, undersampling) to resolve the outcome imbalance problem because it was not extreme.

We employed RF, extreme gradient boosting (XGBoost), deep neural network (DNN), and logistic regression (LR) as the candidate prediction models.

## Parameter Tuning

We also used an automated machine learning called the Tree-based Pipeline Optimization Tool for model selection and parameter searching [13].

For DNN, we used 512, 256, and 128 neurons for hidden layers, ReLU function for activation function in hidden layers, sigmoid function for activation function in the output layer, and binary cross-entropy function as the loss function. For XGBoost, we used a tree booster with 100 estimators, the learning rate as 0.1, and the subsample ratio as 0.75.

## External Validation

After development and internal validation, we performed the external validation of our delirium prediction model using the MIMIC-III database. The validation set was extracted from the MIMIC-III database, which included patients with at least two CAM-ICU records obtained within at least 24 hours.

The model with the highest area under the receiver operating characteristics (AUROC) curve in the external validation was named the PRIDE (Prediction of ICU Delirium) algorithm.

## Statistical Analysis

Continuous variables are presented in terms of means and SD, and categorical variables are presented in terms of their frequencies and percentages. The performances of the different models were compared using the AUROC, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) at the threshold. In the internal validation, model performance was evaluated through the average and 95% CI of the AUROCs. Additionally, we used a calibration curve and the Brier score to test the reliability of our

model. To determine the clinically relevant threshold, we used a decision curve.

We employed the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) statement to report the results of our prediction model. Data processing, statistical analysis, and the development and validation of the machine learning algorithms were performed using R version 3.6.2 [14] and Python version 3.6.8 [15].

The source code has been made available on Github [16].

## Results

### Study Population

During the study period, a total of 37,543 cases were collected. Patients who were 18 years or older at the time of ICU admission were included. Cases with less than two CAM-ICU records after admission to the ICU and those with a positive CAM-ICU upon ICU admission were excluded. After patient exclusion, 12,409 remained as our study population. The case group consisted of 3816 (30.8%) patients who experienced delirium incidents during the study period. With regard to the MIMIC-III (external validation) data set, patients younger than 18 years of age, those with less than two CAM-ICU records recorded within 24 hours, and those with positive CAM-ICU records upon ICU admission were excluded. Based on the exclusion criteria, out of the 96,016 ICU admission cases, 2061 cases were included, and 272 (13.2%) delirium incidents occurred.

Baseline characteristics of the training and test sets of the SMC and MIMIC-III data sets are shown in [Table 1](#).



**Table 1.** Baseline characteristics of data sets.

Characteristics	Development		Internal validation		External validation (MIMIC-III <sup>a</sup> )	
	Case (n=3060)	Control (n=6529)	Case (n=756)	Control (n=2064)	Case (n=272)	Control (n=1789)
Age (years), mean (SD)	65.4 (14.4)	61.1 (13.3)	65.3 (14.7)	59.7 (14.0)	82.5 (59.7)	73.7 (52.3)
Sex (male), n (%)	1994 (65.2)	4227 (64.7)	466 (61.6)	1190 (57.7)	137 (50.4)	925 (51.7)
<b>Admission category, n (%)</b>						
Medical	1431 (46.8)	1639 (25.1)	293 (38.8)	344 (16.7)	151 (55.5)	1,141(63.8)
Surgical	1629 (53.2)	4890 (74.9)	1,720 (83.3)	463 (61.2)	121 (44.5)	648(36.2)
<b>Reason for ICU<sup>b</sup> admission, n (%)</b>						
Respiratory	692 (22.6)	374 (5.7)	166 (22.0)	71 (3.4)	34 (12.5)	247 (13.8)
Cardiovascular	480 (15.7)	1166(17.9)	105 (13.9)	283 (13.7)	73 (26.8)	561 (31.4)
Gastrointestinal	145 (4.7)	139 (2.1)	28 (3.7)	25 (1.2)	55 (20.2)	379 (21.2)
Neurology	102 (3.3)	115 (1.8)	46 (6.1)	174 (8.4)	57 (21.0)	270 (15.1)
Peri-operation	1168 (38.2)	4278 (65.5)	320 (42.3)	1448 (70.2)	14 (5.1)	81 (4.5)
Nephrology	83 (2.7)	78 (1.2)	17 (2.2)	17 (0.8)	4 (1.5)	54 (3.0)
Metabolic	15 (0.5)	6 (0.1)	1 (0.1)	0 (0.0)	5 (1.8)	85 (4.8)
Hematology	15 (0.5)	22 (0.3)	4 (0.5)	6 (0.3)	5 (1.8)	33 (1.8)
Trauma	10 (0.3)	10 (0.2)	2 (0.3)	0 (0.0)	25 (9.2)	79 (4.4)
Others	350 (11.4)	341 (5.2)	67 (8.9)	40 (1.9)	—	—
Initial SOFA <sup>c</sup> , mean (SD)	7.1 (3.6)	3.2 (2.7)	7.0 (3.7)	2.8 (2.5)	5.4 (3.3)	3.1 (2.4)
Vasopressor <sup>d</sup> , n (%)	1,145 (37.4)	793 (12.1)	292 (38.6)	175 (8.5)	38 (14.0)	71 (4.0)
Invasive mechanical ventilator, n (%)	1,900 (62.1)	1,165 (17.8)	456 (60.3)	344 (16.7)	19 (7.0)	73 (4.1)
CCI <sup>e</sup> , mean (SD)	0.9 (2.2)	0.3 (1.3)	1.1 (2.5)	0.4 (1.4)	3.2 (1.8)	2.8 (1.8)
<b>Comorbidity, n (%)</b>						
Heart disease	149 (4.9)	123 (1.9)	17 (2.2)	20 (1.0)	33 (12.1)	232 (13.0)
Stroke	98 (3.2)	33 (0.5)	29 (3.8)	24 (1.2)	16 (5.9)	57 (3.2)
Malignancy	434 (14.2)	319 (4.9)	80 (10.6)	60 (2.9)	5 (1.8)	12 (0.7)
Renal failure	71 (2.3)	99 (1.5)	16 (2.1)	26 (1.3)	17 (12.1)	106 (12.7)
Liver disease	146 (4.8)	92 (1.4)	23 (3.0)	15 (0.7)	38 (14.0)	168 (9.4)
Dementia	55 (1.8)	9 (0.1)	17 (2.2)	6 (0.3)	—	—
<b>Vital signs, mean (SD)</b>						
Systolic BP <sup>f</sup>	125.8 (24.7)	127.4 (21.7)	128.6 (26.4)	130.8 (22.2)	118.5 (25.0)	119.5 (22.7)
Diastolic BP	71.6 (15.5)	74.3 (14.1)	72.3 (17.9)	73.3 (14.7)	63.3 (15.9)	63.9 (15.4)
Heart rate	85.4 (20.7)	80.7 (16.6)	84.5 (20.5)	80.6 (16.1)	89.9 (20.3)	85.8 (19.4)
Respiratory rate	19.3 (3.9)	18.3 (2.5)	18.8 (3.7)	17.7 (2.5)	20.5 (6.5)	19.2 (5.4)
SpO <sub>2</sub> <sup>g</sup>	96.1 (5.5)	96.9 (4.3)	96.5 (4.0)	97.4 (2.1)	95.9 (3.6)	96.2 (3.1)
Body temperature (°C)	36.7 (0.8)	36.6 (0.5)	36.7 (0.8)	36.6 (0.4)	36.8 (0.7)	36.8 (0.6)
<b>ABGA<sup>h</sup>, mean (SD)</b>						
pH	7.4 (0.1)	7.4 (0.1)	7.4 (0.1)	7.4 (0.1)	7.4 (0.1)	7.4 (0.1)
PaCO <sub>2</sub>	35.5 (14.1)	36.2 (7.5)	35.3 (11.3)	36.4 (6.2)	42.0 (12.4)	41.1 (12.1)
PaO <sub>2</sub>	119.8 (83.4)	184.4 (110.8)	122.0 (82.5)	190.4 (104.8)	131.9 (106.0)	125.0 (67.9)
HCO <sub>3</sub>	22.0 (5.5)	23.1 (3.6)	22.0 (5.3)	23.5 (2.9)	24.1 (5.4)	24.2 (4.7)

<sup>a</sup>MIMIC-III: Medical Information Mart for Intensive Care III.

<sup>b</sup>ICU: intensive care unit.

<sup>c</sup>SOFA: sequential organ failure assessment.

<sup>d</sup>Vasopressor: epinephrine, norepinephrine, dobutamine, dopamine, vasopressin.

<sup>e</sup>CCI: Charlson Comorbidity Index.

<sup>f</sup>BP: blood pressure.

<sup>g</sup>SpO<sub>2</sub>: peripheral capillary oxygen saturation.

<sup>h</sup>ABGA: arterial blood gas analysis.

## Internal Validation

The average AUROCs and 95% CIs for internal validation were 0.919 (95% CI 0.919-0.919) for XGBoost, 0.916 (95% CI 0.916-0.916) for RF, 0.881 (95% CI 0.878-0.884) for DNN, and 0.875 (95% CI 0.875-0.875) for LR. For each model, we selected the highest value of specificity among sensitivities over 0.9 as the cut-off point for the threshold. The best model for the internal validation was XGBoost, with an AUROC of 0.919 (95% CI 0.919-0.919). Its sensitivity, specificity, PPV, and NPV were 0.904 (95% CI 0.904-0.905), 0.731 (95% CI 0.729-0.732), 0.565 (95% CI 0.563-0.566), and 0.952 (95% CI 0.952-0.952), respectively.

## External Validation

For the external validation, the average AUROCs and 95% CI were 0.721 (95% CI 0.72-0.721) for RF, 0.697 (95% CI

0.695-0.699) for XGBoost, 0.655 (95% CI 0.654-0.657) for DNN, and 0.631 (95% CI 0.631-0.631) for LR. For the external validation on the MIMIC-III database, the model with the best AUROC was the RF model, with an AUROC of 0.721 and a sensitivity, specificity, PPV, and NPV of 0.91 (95% CI 0.909-0.912), 0.27 (95% CI 0.266-0.273), 0.159 (95% CI 0.159-0.16), and 0.952 (95% CI 0.951-0.953), respectively. A comparison of the performances of all of the models is shown in [Table 2](#), and the ROC curves are shown in [Figure 2](#).

For the external validation with MIMIC-III, we only selected variables that could be found both in SMC and MIMIC-III. As a result, only 59 variables were selected. The variables were categorized into general information, flowsheet, laboratory test results, and prescription of medication. The most important variable was the use of invasive mechanical ventilation in the general information. The importance of each final variable used in model development is shown in [Figure 3](#).

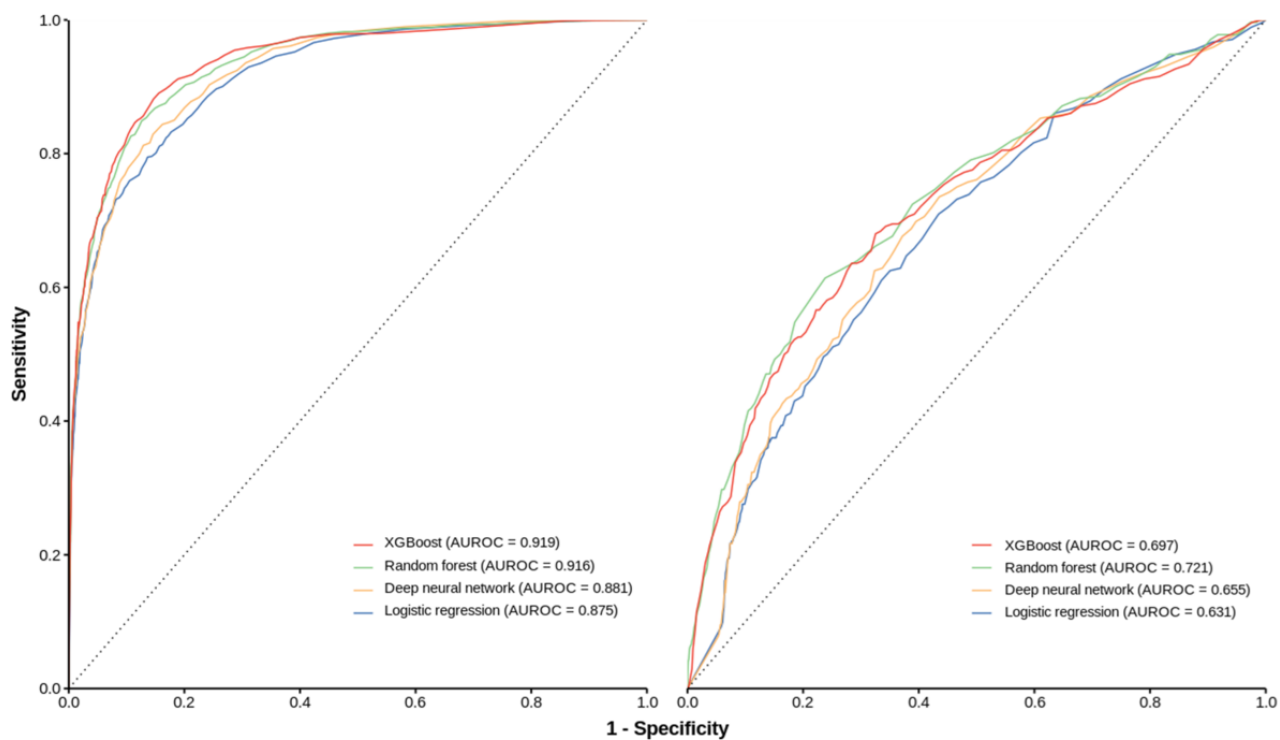
**Table 2.** Predictive performance of each model.

Model and data set	AUROC <sup>a</sup> , mean (95% CI)	Sensitivity, mean (95% CI)	Specificity, mean (95% CI)	Positive predictive value, mean (95% CI)	Negative predictive value, mean (95% CI)
<b>Random forest</b>					
Internal data set	0.916 (0.916-0.916)	0.904 (0.904-0.905)	0.746 (0.744-0.747)	0.579 (0.578-0.580)	0.953 (0.952-0.953)
External data set	0.721 (0.720-0.721)	0.910 (0.909-0.912)	0.270 (0.266-0.273)	0.159 (0.159-0.160)	0.952 (0.951-0.953)
<b>XGBoost<sup>b</sup></b>					
Internal data set	0.919 (0.919-0.919)	0.904 (0.904-0.905)	0.731 (0.729-0.732)	0.565 (0.563-0.566)	0.952 (0.952-0.952)
External data set	0.697 (0.695-0.699)	0.908 (0.906-0.909)	0.250 (0.245-0.255)	0.156 (0.155-0.156)	0.946 (0.945-0.947)
<b>Deep neural network</b>					
Internal data set	0.881 (0.878-0.884)	0.906 (0.905-0.907)	0.622 (0.608-0.635)	0.485 (0.477-0.492)	0.944 (0.943-0.945)
External data set	0.655 (0.654-0.657)	0.907 (0.905-0.908)	0.197 (0.192-0.201)	0.147 (0.146-0.147)	0.932 (0.931-0.933)
<b>Logistic regression</b>					
Internal data set	0.875 (0.875-0.875)	0.901 (0.901-0.901)	0.605 (0.605-0.605)	0.469 (0.469-0.469)	0.940 (0.940-0.940)
External data set	0.631 (0.631-0.631)	0.904 (0.904-0.904)	0.155 (0.155-0.155)	0.140 (0.140-0.140)	0.914 (0.914-0.914)

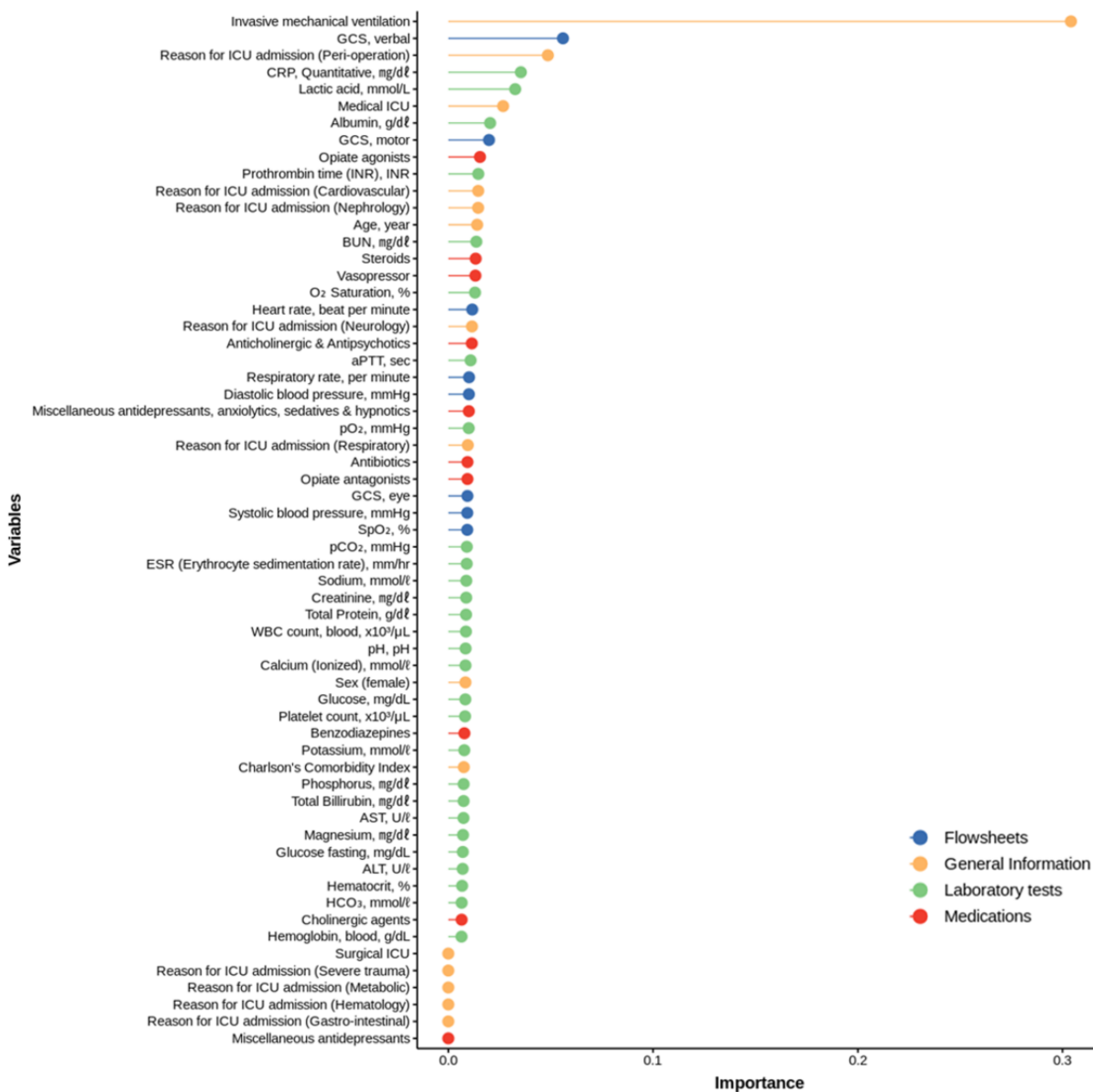
<sup>a</sup>AUROC: area under the receiver operating characteristic curve.

<sup>b</sup>XGBoost: extreme gradient boosting.

**Figure 2.** Receiver operating characteristic curves for all the prediction of intensive care unit delirium models. AUROC: area under the receiver operating characteristic curve; XGBoost: extreme gradient boosting.



**Figure 3.** Variable importance of the prediction of intensive care unit delirium model. ALT: alanine aminotransferase; aPTT: activated partial thromboplastin time; AST: aspartate aminotransferase; BUN: blood urea nitrogen; CRP: C-reactive protein; GCS: Glasgow Coma Scale; ICU: intensive care unit; INR: international normalized ratio.



### Model Assessment

For further model evaluation, calibration and decision curve analyses were performed. The Brier score for the XGBoost model with regard to predicting delirium was 0.094 for the internal validation data set, indicating that our model is reliable. The best model for external validation is RF, with the Brier score of 0.168. A Brier score of 0 indicates a perfect calibration, and the closest the value is to 0, the better model calibration. The calibration plot is shown in Multimedia Appendix 3. The decision curve analysis showed that the net benefit was useful for determining the threshold. For the PRIDE algorithm, the threshold for delirium prediction was selected as 0.13, and at this cut-off point, the net benefit was 0.234. The PRIDE model has a wide range of threshold probabilities and offers reasonable

clinical applicability. The decision curve analysis is presented in Multimedia Appendix 4.

### Discussion

#### Principal Results

We have demonstrated that the proposed delirium prediction model, which employs a machine learning algorithm with EHR data, can predict the development of delirium in medical and surgical ICU patients. In addition to our internal validation, we externally validated our findings using the MIMIC-III patient database. With the PRIDE model, we showed that delirium prediction models could be automated exclusively using risk factors derived from EHR data. The three main results of our study are as follows: (1) the model predicted delirium within the first 24 hours of ICU admission by only using data collected

within the first 4 hours after ICU admission, (2) all variables were extracted from EMR data obtained from both medical and surgical intensive care patients, and (3) the model showed acceptable performance with regard to the external validation data set.

Among the various departments in a hospital, the incidence of delirium is the highest in the ICU, and it is well-documented that delirium occurs in 25% of critically ill adults in ICUs within the first 24 hours after admission [17-19]. This data shows that the early prediction of delirium upon initial ICU admission is crucial. Furthermore, the early prediction of the development of delirium can help clinicians make clinical decisions at an optimal time and provide preventive and personalized care with nondrug interventions for high-risk patients. Examples of such care are cognitive stimulation, orientation improvement, and early mobilization [20].

### Comparison With Prior Work

Owing to the prevalence of delirium in patients admitted to ICUs, the routine use of preventive measures for delirium is recommended. However, previous studies have shown that clinicians' predictions of the development of delirium are less accurate than those of ICU delirium prediction models [7]. Thus, delirium prediction models developed using machine learning can support clinicians in the early recognition of delirium, thereby immensely benefiting patients at high risk of delirium [21]. Furthermore, although several risk prediction models have been proposed, they are based on the manual evaluation of individual risk factors, and thus, may be challenging to implement [7,22,23]. Hence, in practice, automated models are preferable and more feasible. For these reasons, the implementation of automated tools for predicting the risk of delirium development using data extracted from EHR would improve clinical practices with regard to ICU management. Furthermore, the EHR-based prediction model uses a pipeline that automatically extracts variables and calculates models containing enough variables.

Previous studies have used several risk factors for delirium in ICUs, including age, severity score, cause of admission, usage of sedative agents, and laboratory results. In contrast with previous studies, the PRIDE model includes several additional variables such as vital signs (heart rate and blood pressure) and medication information that is excluded from EHRs. These differences allow our model to predict delirium incidents within 4 hours of ICU admission only using EHR data. Further, the PRIDE model did not include a severity score because this can only be obtained after 24 hours of ICU admission; in addition, since this information is separate from EMR data, using a severity score would require further efforts by the clinician. A few reports have also presented EMR-based machine learning models to predict delirium [24,25]. Whereas the prediction models presented in these reports are for all hospital-admitted

patients, in this study, we developed a versatile model specifically for ICU patients at risk of delirium.

The strength of our study is the EMR-driven model that was both internally and externally validated, using SMC and MIMIC-III data, respectively. Although our result showed lower accuracy with external data than internal data, this result can be improved if the missing rate of key features decreases. In the case of CAM-ICU, 96% was missing in MIMIC-III. In addition, a decrease in accuracy with an external database was not uncommon in literature [26]. For example, a study predicting serious bacterial infections among fevers in children reported that the AUC of the external data was 0.26 lower than the internal data [27].

In clinical settings, missing values occur for various reasons. To handle missing data, we used mean values in the numerical data. We left the missing values in the categorical data blank such that the dummy variables were all equal to 0 method. Recently, deep learning-based advanced techniques, such as long short-term memory and recurrent neural network, were also introduced to impute missing data, and by employing these methods, they could improve model performances [28]. When choosing a missing handling method, knowing the missing pattern can improve model performance and work better when applied to clinical applications.

### Limitations

There are potential limitations to our study that should be acknowledged. First, our study was retrospectively performed and validated. Prospective interventional studies are needed to verify the performance of the model and to reconfirm its clinical usefulness. Second, a selection bias might exist because we selected variables available in all cohorts, and this study was conducted in a retrospective manner. Furthermore, we excluded patients without CAM-ICU data (47% of the total number of ICU-admitted patients). In this regard, it should be noted that the purpose of this study was to develop a readily available model; therefore, we only selected the variables that could be used commonly in all cohorts. Finally, although the CAM-ICU tool is regarded as highly sensitive and specific to the detection of ICU delirium, it has critical limitations. As it only has binary labels, we cannot access the degree of delirium exacerbation. Furthermore, it is recorded in a "point-in-time" manner; thus, there may be some patients whose CAM-ICU tests were missed because they were completed outside the study's time frame [29,30].

### Conclusions

We have developed and validated the delirium prediction model, which can predict the occurrence of delirium within 24 hours of ICU admission, using clinical data obtained in the first 4 hours after ICU admission. The PRIDE algorithm has acceptable AUCROC and sensitivity; thus, it has the potential to help advise ICU physicians and prevent ICU delirium.



## Acknowledgments

This research was supported by a Korea Health Technology R&D Project grant through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant HI19C0275). REK, WCC, and CRC contributed to the conception and study design. SH, JY, JH, WCC, and CRC performed the data analysis and interpretation. SH, REK, and CRC drafted the manuscript for intellectual content. SH, REK, WCC, JY, JH, and CRC revised the manuscript. All authors have read and approved the final manuscript.

## Authors' Contributions

SH and REK are co-first authors of this study. CRC (chiryang.chung@gmail.com) and WCC (docchaster@gmail.com) are co-corresponding authors for this article.

## Conflicts of Interest

None declared.

### Multimedia Appendix 1

Institutional review board approval form.

[[PDF File \(Adobe PDF File\), 712 KB - medinform\\_v9i7e23401\\_app1.pdf](#)]

### Multimedia Appendix 2

The missing rate in the variable list.

[[XLSX File \(Microsoft Excel File\), 12 KB - medinform\\_v9i7e23401\\_app2.xlsx](#)]

### Multimedia Appendix 3

Calibration curve of the best model.

[[PNG File , 113 KB - medinform\\_v9i7e23401\\_app3.png](#)]

### Multimedia Appendix 4

Decision curve of the best model. PRIDE: prediction of intensive care unit delirium model.

[[PNG File , 134 KB - medinform\\_v9i7e23401\\_app4.png](#)]

## References

1. Cavallazzi R, Saad M, Marik PE. Delirium in the ICU: An overview. *Ann Intensive Care* 2012;2(1):49. [doi: [10.1186/2110-5820-2-49](https://doi.org/10.1186/2110-5820-2-49)]
2. Salluh JIF, Wang H, Schneider EB, Nagaraja N, Yenokyan G, Damluji A, et al. Outcome of delirium in critically ill patients: systematic review and meta-analysis. *BMJ* 2015 Jun 03;350(may19 3):h2538-h2538 [FREE Full text] [doi: [10.1136/bmj.h2538](https://doi.org/10.1136/bmj.h2538)] [Medline: [26041151](https://pubmed.ncbi.nlm.nih.gov/26041151/)]
3. Pandharipande PP, Girard TD, Jackson JC, Morandi A, Thompson JL, Pun BT, BRAIN-ICU Study Investigators. Long-term cognitive impairment after critical illness. *N Engl J Med* 2013 Oct 03;369(14):1306-1316 [FREE Full text] [doi: [10.1056/NEJMoa1301372](https://doi.org/10.1056/NEJMoa1301372)] [Medline: [24088092](https://pubmed.ncbi.nlm.nih.gov/24088092/)]
4. Milbrandt EB, Deppen S, Harrison PL, Shintani AK, Speroff T, Stiles RA, et al. Costs associated with delirium in mechanically ventilated patients. *Crit Care Med* 2004 Apr;32(4):955-962. [doi: [10.1097/01.ccm.0000119429.16055.92](https://doi.org/10.1097/01.ccm.0000119429.16055.92)] [Medline: [15071384](https://pubmed.ncbi.nlm.nih.gov/15071384/)]
5. Serafim RB, Bozza FA, Soares M, do Brasil PEA, Tura BR, Ely EW, et al. Pharmacologic prevention and treatment of delirium in intensive care patients: A systematic review. *J Crit Care* 2015 Aug;30(4):799-807. [doi: [10.1016/j.jcrc.2015.04.005](https://doi.org/10.1016/j.jcrc.2015.04.005)] [Medline: [25957498](https://pubmed.ncbi.nlm.nih.gov/25957498/)]
6. Brummel NE, Girard TD. Preventing delirium in the intensive care unit. *Crit Care Clin* 2013 Jan;29(1):51-65 [FREE Full text] [doi: [10.1016/j.ccc.2012.10.007](https://doi.org/10.1016/j.ccc.2012.10.007)] [Medline: [23182527](https://pubmed.ncbi.nlm.nih.gov/23182527/)]
7. van den Boogaard M, Pickkers P, Slooter AJC, Kuiper MA, Spronk PE, van der Voort PHJ, et al. Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: observational multicentre study. *BMJ* 2012 Feb 09;344(feb09 3):e420-e420 [FREE Full text] [doi: [10.1136/bmj.e420](https://doi.org/10.1136/bmj.e420)] [Medline: [22323509](https://pubmed.ncbi.nlm.nih.gov/22323509/)]
8. Chen Y, Du H, Wei B, Chang X, Dong C. Development and validation of risk-stratification delirium prediction model for critically ill patients: A prospective, observational, single-center study. *Medicine (Baltimore)* 2017 Jul;96(29):e7543 [FREE Full text] [doi: [10.1097/MD.0000000000007543](https://doi.org/10.1097/MD.0000000000007543)] [Medline: [28723773](https://pubmed.ncbi.nlm.nih.gov/28723773/)]
9. Wassenaar A, van den Boogaard M, van Achterberg T, Slooter AJC, Kuiper MA, Hoogendoorn ME, et al. Multinational development and validation of an early prediction model for delirium in ICU patients. *Intensive Care Med* 2015 Jun 18;41(6):1048-1056 [FREE Full text] [doi: [10.1007/s00134-015-3777-2](https://doi.org/10.1007/s00134-015-3777-2)] [Medline: [25894620](https://pubmed.ncbi.nlm.nih.gov/25894620/)]

10. Wassenaar A, Schoonhoven L, Devlin JW, van Haren FMP, Slooter AJC, Jorens PG, et al. Delirium prediction in the intensive care unit: comparison of two delirium prediction models. *Crit Care* 2018 May 05;22(1):114 [FREE Full text] [doi: [10.1186/s13054-018-2037-6](https://doi.org/10.1186/s13054-018-2037-6)] [Medline: [29728150](https://pubmed.ncbi.nlm.nih.gov/29728150/)]
11. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016 May 24;3:160035 [FREE Full text] [doi: [10.1038/sdata.2016.35](https://doi.org/10.1038/sdata.2016.35)] [Medline: [27219127](https://pubmed.ncbi.nlm.nih.gov/27219127/)]
12. Ely EW, Margolin R, Francis J, May L, Truman B, Dittus R, et al. Evaluation of delirium in critically ill patients: Validation of the confusion assessment method for the intensive care unit (CAM-ICU). *Critical Care Medicine* 2001;29(7):1370-1379. [doi: [10.1097/00003246-200107000-00012](https://doi.org/10.1097/00003246-200107000-00012)]
13. Olson R, Moore J. TPOT: A tree-based pipeline optimization tool for automating machine learning. In: *Automated Machine Learning* Springer; May 18, 2019:A.
14. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/> [accessed 2020-04-03]
15. van Rossum G, Drake F. Python reference manual. Netherlands: Centrum voor Wiskunde en Informatica Amsterdam; Apr 30, 1995.
16. juhha1 / Delirium Prediction. Github. URL: [https://github.com/juhha1/Delirium\\_Prediction](https://github.com/juhha1/Delirium_Prediction) [accessed 2021-07-19]
17. Inouye SK, Westendorp RG, Saczynski JS. Delirium in elderly people. *The Lancet* 2014 Mar;383(9920):911-922. [doi: [10.1016/s0140-6736\(13\)60688-1](https://doi.org/10.1016/s0140-6736(13)60688-1)]
18. Ely EW, Shintani A, Truman B, Speroff T, Gordon SM, Harrell FE, et al. Delirium as a predictor of mortality in mechanically ventilated patients in the intensive care unit. *JAMA* 2004 Apr 14;291(14):1753-1762. [doi: [10.1001/jama.291.14.1753](https://doi.org/10.1001/jama.291.14.1753)] [Medline: [15082703](https://pubmed.ncbi.nlm.nih.gov/15082703/)]
19. Serafim RB, Dutra MF, Saddy F, Tura B, de Castro JEC, Villarinho LC, et al. Delirium in postoperative nonventilated intensive care patients: Risk factors and outcomes. *Ann Intensive Care* 2012 Dec 31;2(1):51 [FREE Full text] [doi: [10.1186/2110-5820-2-51](https://doi.org/10.1186/2110-5820-2-51)] [Medline: [23272945](https://pubmed.ncbi.nlm.nih.gov/23272945/)]
20. Schweickert WD, Pohlman MC, Pohlman AS, Nigos C, Pawlik AJ, Esbrook CL, et al. Early physical and occupational therapy in mechanically ventilated, critically ill patients: A randomised controlled trial. *The Lancet* 2009 May;373(9678):1874-1882. [doi: [10.1016/s0140-6736\(09\)60658-9](https://doi.org/10.1016/s0140-6736(09)60658-9)]
21. Siddiqi N. Predicting delirium: Time to use delirium risk scores in routine practice? *Age Ageing* 2016 Jan 13;45(1):9-10. [doi: [10.1093/ageing/afv183](https://doi.org/10.1093/ageing/afv183)] [Medline: [26764390](https://pubmed.ncbi.nlm.nih.gov/26764390/)]
22. Inouye SK, Viscoli CM, Horwitz RI, Hurst LD, Tinetti ME. A predictive model for delirium in hospitalized elderly medical patients based on admission characteristics. *Ann Intern Med* 1993 Sep 15;119(6):474-481. [doi: [10.7326/0003-4819-119-6-199309150-00005](https://doi.org/10.7326/0003-4819-119-6-199309150-00005)] [Medline: [8357112](https://pubmed.ncbi.nlm.nih.gov/8357112/)]
23. Kishi T, Hirota T, Matsunaga S, Iwata N. Antipsychotic medications for the treatment of delirium: A systematic review and meta-analysis of randomised controlled trials. *J Neurol Neurosurg Psychiatry* 2016 Jul 04;87(7):767-774. [doi: [10.1136/jnnp-2015-311049](https://doi.org/10.1136/jnnp-2015-311049)] [Medline: [26341326](https://pubmed.ncbi.nlm.nih.gov/26341326/)]
24. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and validation of an electronic health record-based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. *JAMA Netw Open* 2018 Aug 03;1(4):e181018 [FREE Full text] [doi: [10.1001/jamanetworkopen.2018.1018](https://doi.org/10.1001/jamanetworkopen.2018.1018)] [Medline: [30646095](https://pubmed.ncbi.nlm.nih.gov/30646095/)]
25. Corradi JP, Thompson S, Mather JF, Waszynski CM, Dicks RS. Prediction of incident delirium using a random forest classifier. *J Med Syst* 2018 Nov 14;42(12):261. [doi: [10.1007/s10916-018-1109-0](https://doi.org/10.1007/s10916-018-1109-0)] [Medline: [30430256](https://pubmed.ncbi.nlm.nih.gov/30430256/)]
26. Shung D, Simonov M, Gentry M, Au B, Laine L. Machine learning to predict outcomes in patients with acute gastrointestinal bleeding: A systematic review. *Dig Dis Sci* 2019 Aug 4;64(8):2078-2087. [doi: [10.1007/s10620-019-05645-z](https://doi.org/10.1007/s10620-019-05645-z)] [Medline: [31055722](https://pubmed.ncbi.nlm.nih.gov/31055722/)]
27. Bleeker S, Moll H, Steyerberg E, Donders A, Derksen-Lubsen G, Grobbee D, et al. External validation is necessary in prediction research. *Journal of Clinical Epidemiology* 2003 Sep;56(9):826-832. [doi: [10.1016/s0895-4356\(03\)00207-5](https://doi.org/10.1016/s0895-4356(03)00207-5)]
28. Han-Gyu K, Gil-Jin J, Ho-Jin C, Minho K, Young-Won K, Jaehun C. Recurrent neural networks with missing information imputation for medical examination data prediction. 2017 Feb 13 Presented at: 2017 IEEE International Conference on Big Data and Smart Computing (BigComp); February 13-16, 2017; Jeju, South Korea URL: <https://ieeexplore.ieee.org/abstract/document/7881685> [doi: [10.1109/bigcomp.2017.7881685](https://doi.org/10.1109/bigcomp.2017.7881685)]
29. van Eijk MM, van den Boogaard M, van Marum RJ, Benner P, Eikelenboom P, Honing ML, et al. Routine use of the confusion assessment method for the intensive care unit. *Am J Respir Crit Care Med* 2011 Aug;184(3):340-344. [doi: [10.1164/rccm.201101-0065oc](https://doi.org/10.1164/rccm.201101-0065oc)]
30. Neto AS, Nassar AP, Cardoso SO, Manetta JA, Pereira VG, Espósito DC, et al. Delirium screening in critically ill patients. *Critical Care Medicine* 2012;40(6):1946-1951. [doi: [10.1097/ccm.0b013e31824e16c9](https://doi.org/10.1097/ccm.0b013e31824e16c9)]

## Abbreviations

**AUROC:** area under the receiver operating characteristic curve

**CAM-ICU:** confusion assessment method for the intensive care unit

**DBP:** diastolic blood pressure  
**DNN:** deep neural network  
**EHR:** electronic health record  
**ICU:** intensive care unit  
**IRB:** institutional review board  
**LR:** logistic regression  
**MIMIC-III:** Medical Information Mart for Intensive Care III  
**NPV:** negative predictive value  
**PRIDE:** Prediction of ICU Delirium  
**PPV:** positive predictive value  
**RF:** random forest  
**SMC:** Samsung Medical Center  
**XGBoost:** extreme gradient boosting

*Edited by C Lovis; submitted 12.08.20; peer-reviewed by M Syed, A Azzam, YR Park; comments to author 26.08.20; revised version received 10.10.20; accepted 07.06.21; published 26.07.21.*

*Please cite as:*

*Hur S, Ko RE, Yoo J, Ha J, Cha WC, Chung CR*

*A Machine Learning–Based Algorithm for the Prediction of Intensive Care Unit Delirium (PRIDE): Retrospective Study*

*JMIR Med Inform 2021;9(7):e23401*

*URL: <https://medinform.jmir.org/2021/7/e23401>*

*doi: [10.2196/23401](https://doi.org/10.2196/23401)*

*PMID: [34309567](https://pubmed.ncbi.nlm.nih.gov/34309567/)*

©Sujeong Hur, Ryoung-Eun Ko, Junsang Yoo, Juhyung Ha, Won Chul Cha, Chi Ryang Chung. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 26.07.2021. 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

# Candidemia Risk Prediction (CanDETEC) Model for Patients With Malignancy: Model Development and Validation in a Single-Center Retrospective Study

Junsang Yoo<sup>1\*</sup>, RN, PhD; Si-Ho Kim<sup>2\*</sup>, MD; Sujeong Hur<sup>3,4</sup>, RN, MS; Juhyung Ha<sup>5</sup>; Kyungmin Huh<sup>6</sup>, MD; Won Chul Cha<sup>4,7,8</sup>, MD

<sup>1</sup>Department of Nursing, College of Nursing, Sahmyook University, Seoul, Republic of Korea

<sup>2</sup>Division of Infectious Disease, Samsung Changwon Hospital, Sungkyunkwan University School of Medicine, Changwon, Republic of Korea

<sup>3</sup>Department of Patient Experience Management, Samsung Medical Center, Seoul, Republic of Korea

<sup>4</sup>Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea

<sup>5</sup>Department of Computer Science, Indiana University Bloomington, Bloomington, IN, United States

<sup>6</sup>Division of Infectious Disease, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

<sup>7</sup>Department of Emergency Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

<sup>8</sup>Digital Innovation Center, Samsung Medical Center, Seoul, Republic of Korea

\*these authors contributed equally

**Corresponding Author:**

Won Chul Cha, MD

Department of Emergency Medicine

Samsung Medical Center

Sungkyunkwan University School of Medicine

81, Irwon-ro, Gangnam-gu

Seoul, 06351

Republic of Korea

Phone: 82 2 3410 2053

Email: [wc.cha@samsung.com](mailto:wc.cha@samsung.com)

**Related Article:**

This is a corrected version. See correction statement: <https://medinform.jmir.org/2022/1/e36385>

## Abstract

**Background:** Appropriate empirical treatment for candidemia is associated with reduced mortality; however, the timely diagnosis of candidemia in patients with sepsis remains poor.

**Objective:** We aimed to use machine learning algorithms to develop and validate a candidemia prediction model for patients with cancer.

**Methods:** We conducted a single-center retrospective study using the cancer registry of a tertiary academic hospital. Adult patients diagnosed with malignancies between January 2010 and December 2018 were included. Our study outcome was the prediction of candidemia events. A stratified undersampling method was used to extract control data for algorithm learning. Multiple models were developed—a combination of 4 variable groups and 5 algorithms (auto-machine learning, deep neural network, gradient boosting, logistic regression, and random forest). The model with the largest area under the receiver operating characteristic curve (AUROC) was selected as the *Candida* detection (CanDETEC) model after comparing its performance indexes with those of the Candida Score Model.

**Results:** From a total of 273,380 blood cultures from 186,404 registered patients with cancer, we extracted 501 records of candidemia events and 2000 records as control data. Performance among the different models varied (AUROC 0.771- 0.889), with all models demonstrating superior performance to that of the Candida Score (AUROC 0.677). The random forest model performed the best (AUROC 0.889, 95% CI 0.888-0.889); therefore, it was selected as the CanDETEC model.

**Conclusions:** The CanDETEC model predicted candidemia in patients with cancer with high discriminative power. This algorithm could be used for the timely diagnosis and appropriate empirical treatment of candidemia.

(*JMIR Med Inform* 2021;9(7):e24651) doi:[10.2196/24651](https://doi.org/10.2196/24651)

## KEYWORDS

candidemia; precision medicine; supervised machine learning; decision support systems, clinical; infection control; decision support; machine learning; development; validation; prediction; risk; model

## Introduction

Candidemia is a representative nosocomial bloodstream infection that contributes to the mortality of immunocompromised patients; it has been shown to occur in 3% of patients in intensive care and 20% of immunosuppressed patients [1]. In addition, owing to a compromised immunity from chemotherapy or malignancy itself, patients with cancer have been reported as the most vulnerable hosts to candidemia [2-4].

Significant mortality has been reported over several decades. In studies from the 1980s, the mortality rate in patients with cancer found to have candidemia exceeded 50% [5-7]. High mortality rates, ranging from 30% to 51%, have also been reported in studies after 2010 [3,4,8]. The mortality rate of candidemia was significantly higher than that of bacteremia [3].

Early empirical treatment is important for patients with candidemia. A retrospective study [8] showed that patients with candidemia whose antifungal treatment was initiated 12 hours after onset of candidemia had a hospital mortality rate twice that of patients whose antifungal treatment was initiated within 12 hours of onset. Despite evidence on the need for early treatment of candidemia, only a small number of patients, receive timely antifungal treatment because of the difficulty of early diagnosis [9].

There has been an unmet clinical need—the coexistence of timeliness, high reliability, and cost-effectiveness in candidemia diagnosis. Blood culture is the reference standard for the candidemia diagnosis [10]. Due to its inherent nature, obtaining

results can take a median of 2 to 3 days. Thus, this delay constitutes one challenge in the problem of timely diagnosis [11]. Multiple statistical models, such as the Candida Score, have been developed for the early prediction of candidemia [12-14]. However, such models have neither been tested on unseen data sets during development nor shown consistent performance in subsequent external validation studies [15-17]. The new T2Candida molecular test, which combines magnetic resonance with molecular diagnostics, is useful for the detection of candidemia in a very short amount of time and with high accuracy [18]; however, the high cost of the T2Candida molecular test is a barrier to its wide application in clinical practice [19,20].

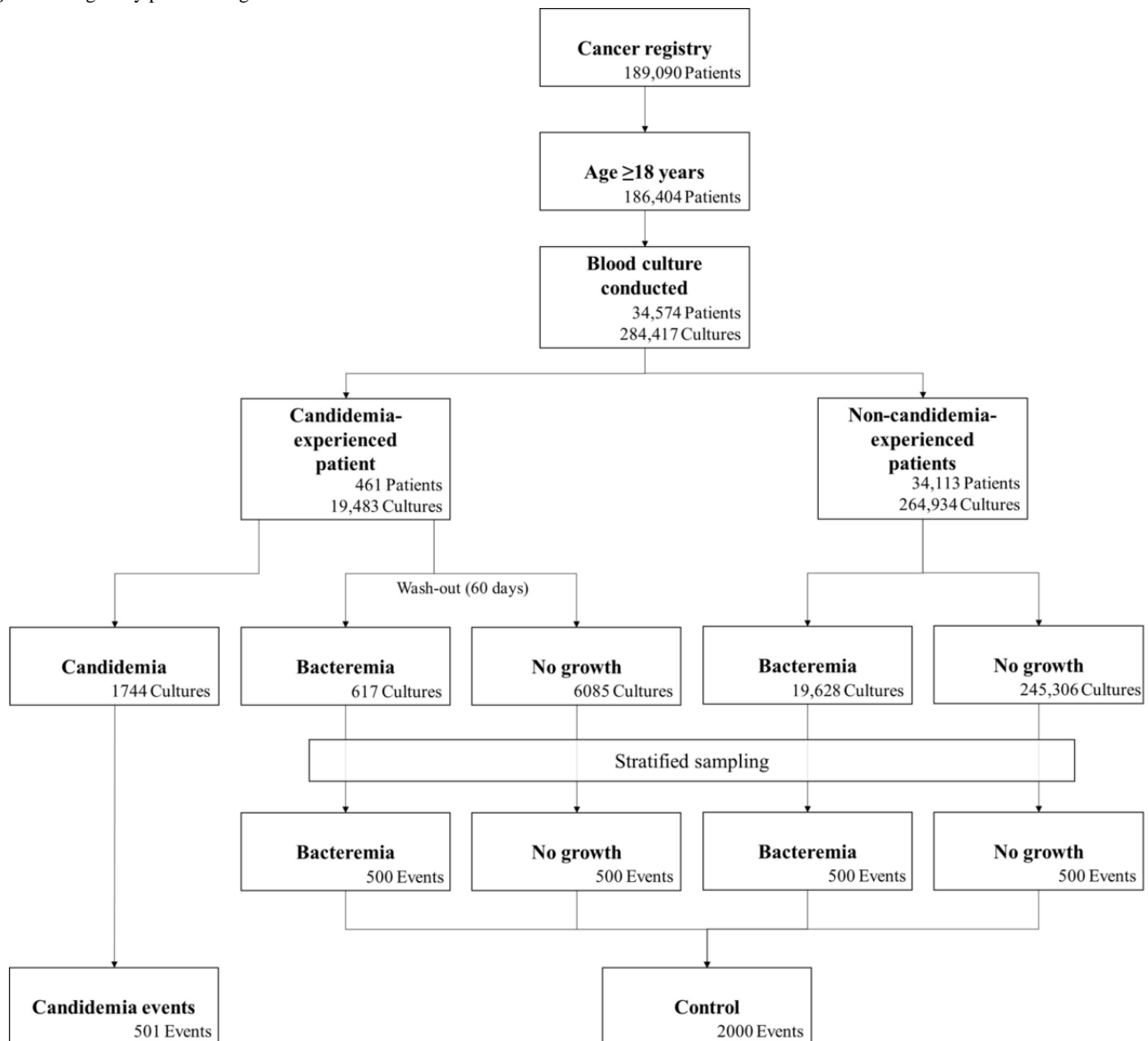
Electronic health records allow efficient extraction and integration of clinical data, and the development of machine learning algorithms for critical care has been vigorously researched [21,22]. We aimed to develop a candidemia prediction model for patients with cancer using machine learning algorithms.

## Methods

### Study Population

This study was conducted in a 1950-bed tertiary academic single hospital in Seoul, Republic of Korea. The study population included adult patients ( $\geq 18$  years old) diagnosed with a malignancy and from whom blood cultures had been obtained between January 2010 and December 2018 after diagnosis. The data set used in this analysis was extracted from the cancer registry and clinical data warehouse of the study site. The selection process is shown in [Figure 1](#).



**Figure 1.** Eligibility process diagram.

## Outcome

We defined *candidemia event* as a positive culture for any *Candida* species in more than one blood sample. If candidemia was detected in a follow-up culture within 7 days, the subsequent event was merged with previous candidemia events. Because the algorithm was designed to predict candidemia events at the time of blood culture extraction, data were processed at the level of each event rather than at the patient level.

A data set with a low candidemia prevalence can cause difficulties in model training [23,24]. To solve this problem, stratified undersampling was conducted at a 1:4 ratio in 4 different subsets (used as a control data): (1) events of bacteremia in patients who experienced candidemia; (2) events of negative blood culture in patients who experienced candidemia; (3) events of bacteremia in patients who had not experienced candidemia; and (4) events of negative blood culture in patients who had not experienced candidemia. In the control subsets with patients who had experienced candidemia,

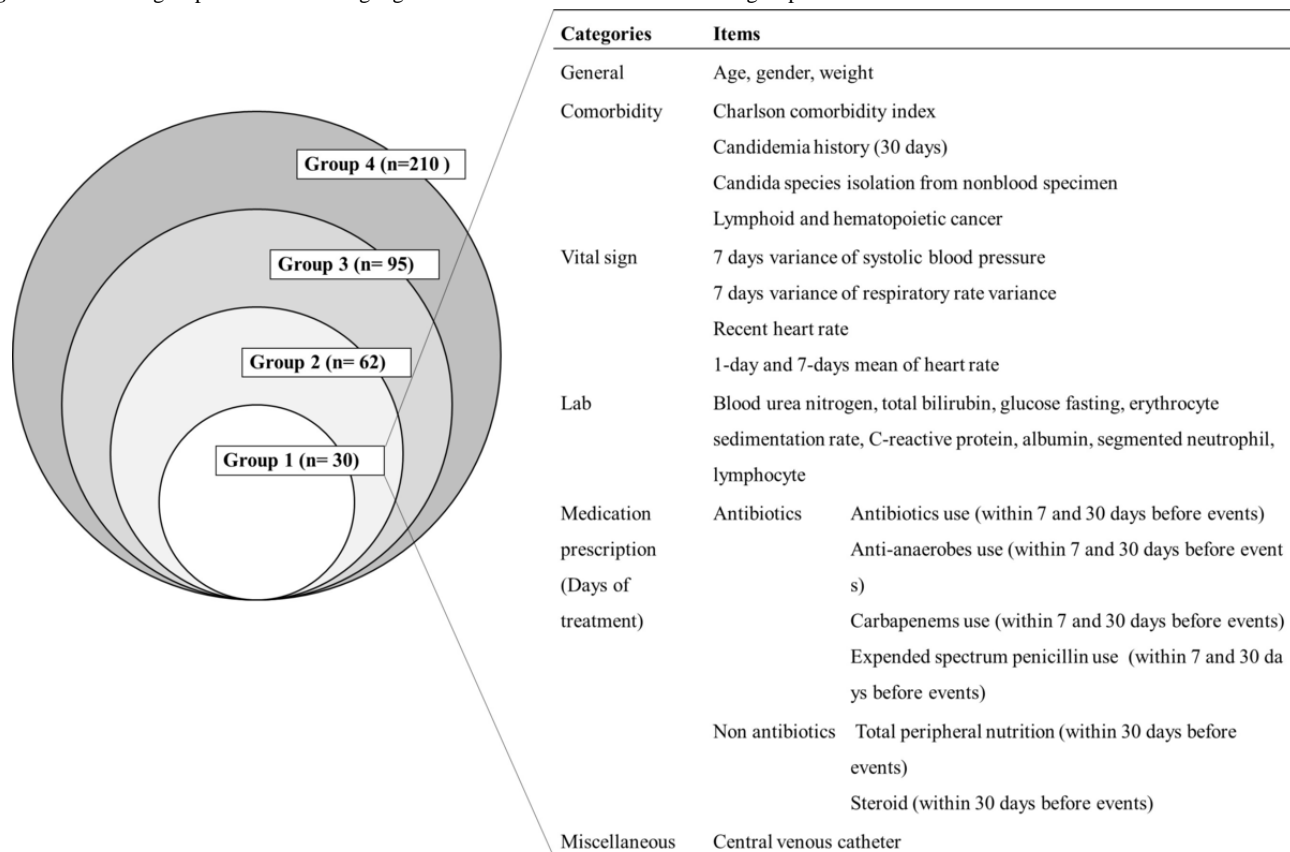
a 60-day wash-out period was imposed from the day of onset of the candidemia event.

## Model Development

### Stage 1: Feature Selection and Preprocessing

Upon review of clinical-domain literature on candidemia risk factors, we identified 210 variables [2,13,14,25-27] that have been widely used in the development of machine learning algorithms in other clinical fields. We extracted data for these variables from the electronic health records of the study site. Each variable was classified into 4 groups based on variable importance, clinical importance, auto-extractability, and missing rate (Figure 2 and Multimedia Appendix 1). We gradually eliminated features: variable group 4 had 210 variables, whereas variable group 1 (higher variable importance, higher clinical importance, better extractability, and less missing values) had only 30 variables. In order to prevent the algorithm from learning from postdiagnostic data, we removed candidemia diagnostic code and antifungal agent prescription information from the input data.

**Figure 2.** Variable groups used for training algorithms and a detailed list of variable group 1.



To impute missing data, we used 2 serial methods. We used the carry-forward method to fill empty bins with the most recent value. This method reflects the workflow of the clinician in recording new data as the patients' condition changes; it is also easy and simple. In the case of missing values that were not imputed by the carry-forward method, average values were used. All numerical variables were normalized.

**Stage 2: Data Partitioning**

The data set was divided into training and test sets (7:3 ratio) using stratified sampling, matched by case-control group, binned age, Charlson comorbidity index, and sex.

**Stage 3: Model Development**

A total of 20 models were developed using a combination of 5 algorithms (logistic regression, deep neural network, random forest, gradient boosting, and automated machine learning) algorithms and 4 variable groups. For each model, 100 different development trials were conducted by changing random seeds to prevent selective performance reporting in the subsequent model evaluation stage. We used 2 methods for model selection and parameter optimization—an automated machine learning tool called the tree-based pipeline optimization tool [28], which helps identify the best prediction model and the best parameters for each variable group by using genetic algorithms and cross-validated performance on the training set (Multimedia Appendix 2), and a simple grid search, which is used to pick the parameters with the best performance.

**Stage 4. Model Evaluation**

Each model was evaluated using a test set (unseen data). Area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value, negative predictive value, and F1 score were used to measure performance. Under the assumption that the algorithm was to be used as a screening tool, sensitivity >0.90 with the highest F1 score, was determined as the threshold for classifying the risk group. One-way analysis of variance was conducted to compare AUROC values. Statistical significance was set at .05. Subsequently, the Bonferroni test was conducted with  $\alpha=.0025$  for each of the 20 outcomes combining the algorithm types and variable groups.

The model with the highest AUROC was selected as the Candida detection (CanDETEC) algorithm. To examine how well the predicted risk correlated with the observed risk, we generated a linear regression model using 10 bins based on the predicted risk of the CanDETEC model. In general, if the coefficients and intercept of the linear regression model were close to 1 and 0, respectively, the model was considered well calibrated.

Candida Score [13], a traditional statistical candidemia prediction model, was used as the standard. We compared the CanDETEC model and the rounded Candida Score:  $1 \times (\text{total parenteral nutrition}) + 1 \times (\text{surgery}) + 1 \times (\text{multifocal candida colonization}) + 2 \times (\text{severe sepsis})$  [13]. Performance indexes were calculated for Candida Score >3. We also employed a net benefit index to compare both models as well as to determine whether the chosen threshold could have beneficial clinical implications.

## Statistical Programming

We used R (version 3.6.2; The R Project) and Python (version 3.6.8) for data preprocessing, statistical analysis, visualization, development, and validation of machine learning algorithms [29,30]. The sample preprocessed data set (20 records) and code for developing our models are available [31].

## Results

### Study Population

A total of 186,404 adult patients were in the cancer registry. A total of 273,380 blood cultures were obtained from 34,574 patients after cancer diagnosis, with 1744 (0.6%) *Candida* species isolated from blood cultures. A total of 501 candidemia events were identified in 461 patients. The most predominant species of *Candida* were *C. albicans* (164/501, 32.7%), *C. tropicalis* (162/501, 32.3%), and *C. glabrata* (96/501, 19.2%) (Multimedia Appendix 3). We found 40 repeat candidemia

events, and the median interval between events was 21.4 days. (IQR 14.5-57.4 days). Of the 271,636 blood cultures with a negative candidemia result, 2000 were extracted as control data.

Hematologic malignancy was the most common malignancy (Table 1). Several clinical factors were significantly associated with case events. Patients with candidemia received longer treatment with steroids, parenteral nutrition, and antibiotics within 30 days before candidemia. Furthermore, antianaerobic therapy (mean 10.1 days) was twice as long ( $P<.001$ ) in those with candidemia than that in controls (5.5 days). In addition, patients with candidemia had higher C-reactive protein, bilirubin, fasting glucose, blood urea nitrogen, and lactic acid levels. The 30-day all-cause mortality rate was significantly higher ( $P<.001$ ) among those with candidemia (264/501, 52.7%) than among controls (303/2000, 15.2%). There was no difference ( $P=.18$ ) in the use of antifungal agents in candidemia (70/501, 14.0%) and control (233/2000, 11.6%) records (Table 2).

**Table 1.** General characteristics of the study data set.

Characteristic	Candidemia events (n=501)	Control (n=2000)	P value
<b>Sex, n (%)</b>			.32
Male	297 (59.3)	1237 (61.9)	
Female			
Age (years), mean (SD)	59.5 (14.4)	55.9 (14.9)	<.001
Hospital stay before culture (days), mean (SD)	23.0 (25.0)	22.2 (65.4)	.66
<b>Comorbidity, n (%)</b>			
Hepatic disease	60 (12.0)	298 (14.9)	.11
Cardiovascular disease	68 (13.6)	203 (10.2)	.03
Endocrine disease	54 (10.8)	198 (9.9)	.62
Digestive disease	36 (7.2)	117 (5.8)	.31
Respiratory disease	20 (4.0)	79 (4.0)	>.999
Other disease	11 (2.2)	38 (1.9)	.80
Charlson Comorbidity Index, mean (SD)	4.7 (2.3)	4.2 (2.2)	<.001
<b>Cancer origin site, n (%)</b>			
Lymphoid or hematopoietic	231 (46.1)	1116 (55.8)	<.001
Digestive	153 (30.5)	551 (27.6)	.20
Respiratory	43 (8.6)	113 (5.7)	.02
Female genital	23 (4.6)	49 (2.5)	.02
Other	38 (7.6)	122 (6.1)	.27
Multiple primary	44 (8.8)	235 (11.8)	.07
Metastatic lymph nodes, mean (SD)	5.1 (12.2)	3.5 (9.9)	.007
<b>Medication (days of therapy during 30 days before candidemia), mean (SD)</b>			
Steroid	8.6 (10.2)	5.8 (9.0)	<.001
Immunosuppressant	1.4 (5.7)	1.6 (5.1)	.43
Total peripheral nutrition	4.6 (7.8)	2.5 (5.8)	<.001
<b>Antibiotic use</b>	16.9 (9.3)	15.3 (9.8)	.002
Antianaerobic	10.1 (8.6)	5.5 (8.2)	<.001
Broad spectrum cephalosporine: 3rd generation	3.5 (4.8)	2.6 (4.5)	<.001
Carbapenem	4.8 (5.8)	2.5 (5.4)	<.001
Extended spectrum penicillin	4.8 (5.8)	2.4 (5.1)	<.001
Glycopeptide	3.8 (5.0)	1.8 (4.1)	<.001
<b>Vitals, mean (SD)</b>			
Systolic blood pressure (mmHg)	118.9 (23.4)	117.5 (20.8)	.21
Diastolic blood pressure (mmHg)	68.5 (14.0)	68.6 (13.4)	.89
Heart rate (bpm)	110.1 (21.3)	103.3 (20.6)	<.001
Respiratory rate (brpm)	21.1 (5.1)	20.1 (4.2)	<.001
Body temperature (°C)	37.2 (1.0)	37.6 (1.0)	<.001
Peripheral capillary oxygen saturation (%)	96.2 (6.1)	97.0 (4.2)	.03
<b>Laboratory work-up, mean (SD)</b>			
<b>Complete blood count</b>			
White blood cell count, blood ( $10^3/\mu\text{L}$ )	8.3 (11.2)	6.8 (18.0)	.02
Hemoglobin, blood (g/dL)	9.4 (1.4)	9.7 (1.8)	<.001

Characteristic	Candidemia events (n=501)	Control (n=2000)	P value
Platelet count, blood (10 <sup>3</sup> /μL)	90.1 (103.1)	104.7 (151.0)	.01
Segmented neutrophil (%)	62.3 (35.3)	54.0 (36.2)	<.001
Absolute neutrophil count (10 <sup>3</sup> /μL)	7.2 (10.3)	4.9 (7.5)	<.001
Absolute lymphocyte count (10 <sup>3</sup> /μL)	0.5 (0.7)	0.7 (1.6)	.009
<b>Acute phase reactants</b>			
Erythrocyte sedimentation rate (mm/h)	41.6 (36.8)	48.4 (34.3)	.001
C-reactive protein (mg/dL)	11.2 (9.0)	8.6 (7.9)	<.001
Procalcitonin, quantitative (ng/mL)	6.3 (16.3)	3.9 (14.4)	.07
<b>Coagulation</b>			
Prothrombin time (international normalized ratio)	1.4 (0.5)	1.3 (0.5)	<.001
<b>Chemistry</b>			
Total protein (g/dL)	5.3 (1.0)	5.8 (1.1)	<.001
Albumin (g/dL)	2.9 (0.5)	3.4 (0.6)	<.001
Globulin (g/dL)	2.3 (0.8)	2.5 (0.8)	.001
Cholesterol (mg/dL)	129.1 (64.8)	138.4 (51.3)	.006
Total bilirubin (mg/dL)	5.5 (9.1)	2.4 (4.7)	<.001
Alkaline phosphatase (U/L)	183.1 (178.2)	157.9 (217.1)	.007
Glucose fasting (mg/dL)	158.7 (71.2)	139.2 (59.2)	<.001
Blood urea nitrogen (mg/dL)	34.1 (25.3)	22.8 (17.4)	<.001
Creatinine (mg/dL)	1.1 (1.0)	1.0 (1.0)	.01
Uric acid (mg/dL)	3.5 (2.3)	3.6 (2.4)	.59
Calcium (mg/dL)	8.3 (0.8)	8.5 (0.8)	<.001
Phosphorus (mg/dL)	3.1 (1.2)	3.2 (1.1)	.03
Lactic acid (mmol/L)	2.9 (2.8)	2.2 (2.3)	<.001
<b>Other risk factors, n (%)</b>			
Candidemia history (within 30 days)	26 (5.2)	0 (0.0)	<.001
Candida isolation from nonblood specimen (within 30 days)	110 (22.0)	132 (6.6)	<.001
Central line inserted	223 (44.5)	854 (42.7)	.50
Major operation (within 30 days)	78 (15.6)	219 (10.9)	.005

**Table 2.** Outcome characteristics of the study population.

Characteristic	Candida event (n=501), n (%)	Control (n=2000), n (%)	P value
<b>Prescription of antifungal agents</b>			
On the day of the blood culture	70 (14.0)	233 (11.6)	.18
Within 3 days after blood culture	277 (55.3)	770 (38.5)	<.001
30-Day all-cause mortality	264 (52.7)	303 (15.2)	<.001

## Model Evaluation

The best model was the random forest model trained using variable group 1, which was selected as the CanDETEC algorithm (Table 3; Multimedia Appendix 4); the lowest performing model was the logistic regression model trained using variable group 4. Receiver operating characteristic curves are shown in Figure 3. The tree-based pipeline optimization

tool algorithm returned an extra-tree model, which showed the highest performance at the specified cut-off. Among the 30 auto-extractable variables, 5 variables showed the highest significance ( $P<.001$ ) in the prediction of candidemia: blood urea nitrogen level, 7-day variance of respiratory rate, total bilirubin level, 7-day variance of systolic blood pressure, and body weight (Multimedia Appendix 5).

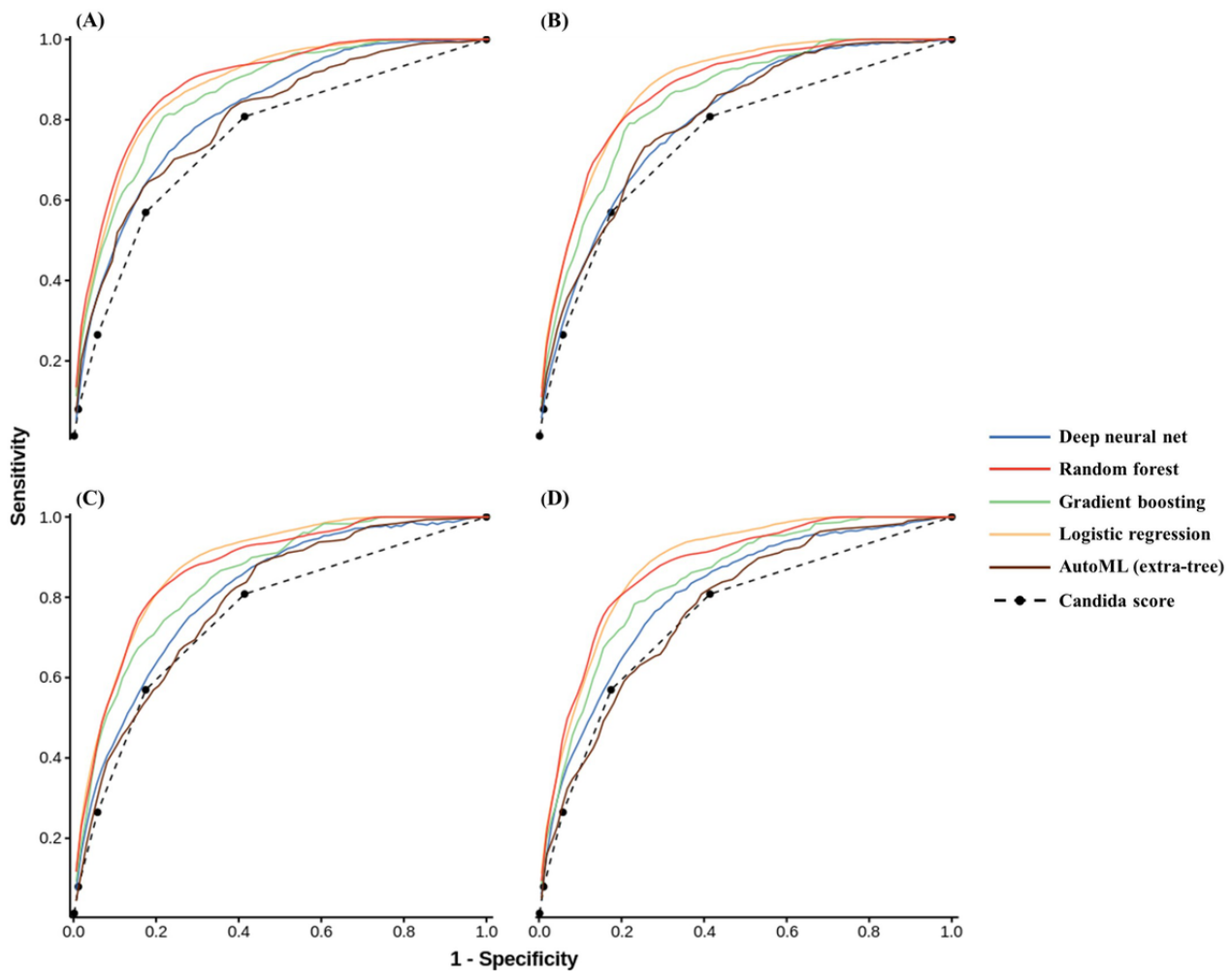


**Table 3.** Model performances by algorithm and variable group.

Algorithm and variable group	AUROC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Positive predictive value (95% CI)	Negative predictive value (95% CI)	F1 score (95% CI)
<b>Logistic regression</b>						
1	0.802 (0.802-0.802)	0.907 (0.907-0.907)	0.451 (0.451-0.451)	0.293 (0.293-0.293)	0.951 (0.951-0.951)	0.443 (0.443-0.443)
2	0.8 (0.8-0.8)	0.901 (0.901-0.901)	0.479 (0.479-0.479)	0.303 (0.303-0.303)	0.95 (0.95-0.95)	0.453 (0.453-0.453)
3	0.788 (0.788-0.788)	0.901 (0.901-0.901)	0.521 (0.521-0.521)	0.321 (0.321-0.321)	0.954 (0.954-0.954)	0.473 (0.473-0.473)
4	0.771 (0.771-0.771)	0.901 (0.901-0.901)	0.473 (0.473-0.473)	0.3 (0.3-0.3)	0.95 (0.95-0.95)	0.45 (0.45-0.45)
<b>Random forest</b>						
1 <sup>a</sup>	0.889 (0.888-0.889)	0.901 (0.901-0.902)	0.722 (0.719-0.724)	0.449 (0.447-0.451)	0.967 (0.967-0.967)	0.599 (0.597-0.601)
2	0.872 (0.872-0.873)	0.901 (0.901-0.902)	0.669 (0.667-0.672)	0.407 (0.405-0.409)	0.964 (0.964-0.964)	0.56 (0.559-0.562)
3	0.869 (0.869-0.87)	0.902 (0.901-0.902)	0.642 (0.639-0.645)	0.388 (0.386-0.39)	0.963 (0.963-0.963)	0.542 (0.54-0.544)
4	0.87 (0.87-0.871)	0.901 (0.901-0.901)	0.669 (0.666-0.672)	0.407 (0.404-0.409)	0.964 (0.964-0.964)	0.56 (0.558-0.562)
<b>Extra tree<sup>b</sup></b>						
1	0.881 (0.88-0.881)	0.901 (0.901-0.902)	0.67 (0.665-0.675)	0.408 (0.404-0.412)	0.964 (0.964-0.965)	0.561 (0.558-0.565)
2	0.882 (0.881-0.882)	0.902 (0.901-0.902)	0.715 (0.711-0.719)	0.443 (0.44-0.447)	0.967 (0.966-0.967)	0.594 (0.591-0.597)
3	0.879 (0.879-0.88)	0.901 (0.901-0.901)	0.708 (0.703-0.713)	0.438 (0.434-0.442)	0.966 (0.966-0.966)	0.589 (0.585-0.592)
4	0.879 (0.878-0.879)	0.902 (0.901-0.903)	0.717 (0.713-0.721)	0.445 (0.442-0.448)	0.967 (0.967-0.967)	0.596 (0.593-0.599)
<b>Gradient boosting</b>						
1	0.861 (0.861-0.862)	0.901 (0.901-0.901)	0.621 (0.621-0.621)	0.374 (0.374-0.374)	0.961 (0.961-0.961)	0.528 (0.528-0.528)
2	0.847 (0.847-0.847)	0.901 (0.901-0.901)	0.593 (0.592-0.593)	0.357 (0.357-0.357)	0.96 (0.96-0.96)	0.511 (0.511-0.512)
3	0.846 (0.846-0.846)	0.901 (0.901-0.901)	0.573 (0.573-0.573)	0.346 (0.346-0.346)	0.958 (0.958-0.958)	0.5 (0.5-0.5)
4	0.839 (0.839-0.839)	0.901 (0.901-0.901)	0.562 (0.562-0.563)	0.341 (0.341-0.341)	0.958 (0.957-0.958)	0.495 (0.494-0.495)
<b>Deep neural network</b>						
1	0.82 (0.818-0.821)	0.901 (0.901-0.902)	0.499 (0.494-0.504)	0.312 (0.309-0.314)	0.953 (0.952-0.953)	0.463 (0.461-0.466)
2	0.799 (0.797-0.801)	0.902 (0.901-0.902)	0.505 (0.5-0.51)	0.314 (0.312-0.317)	0.953 (0.953-0.954)	0.466 (0.464-0.469)
3	0.809 (0.807-0.811)	0.902 (0.901-0.902)	0.525 (0.516-0.534)	0.324 (0.32-0.327)	0.954 (0.953-0.956)	0.476 (0.472-0.48)
4	0.807 (0.804-0.81)	0.901 (0.901-0.902)	0.508 (0.499-0.517)	0.316 (0.313-0.32)	0.953 (0.952-0.954)	0.468 (0.464-0.472)

<sup>a</sup>CanDETEC model.<sup>b</sup>Automated machine learning.

**Figure 3.** Area under the receiver operating characteristic curve of the developed models by algorithm type and variables groups: (A) Group 1, (B) Group 2, (C) Group 3, and (D) Group 4. AutoML: automated machine learning.



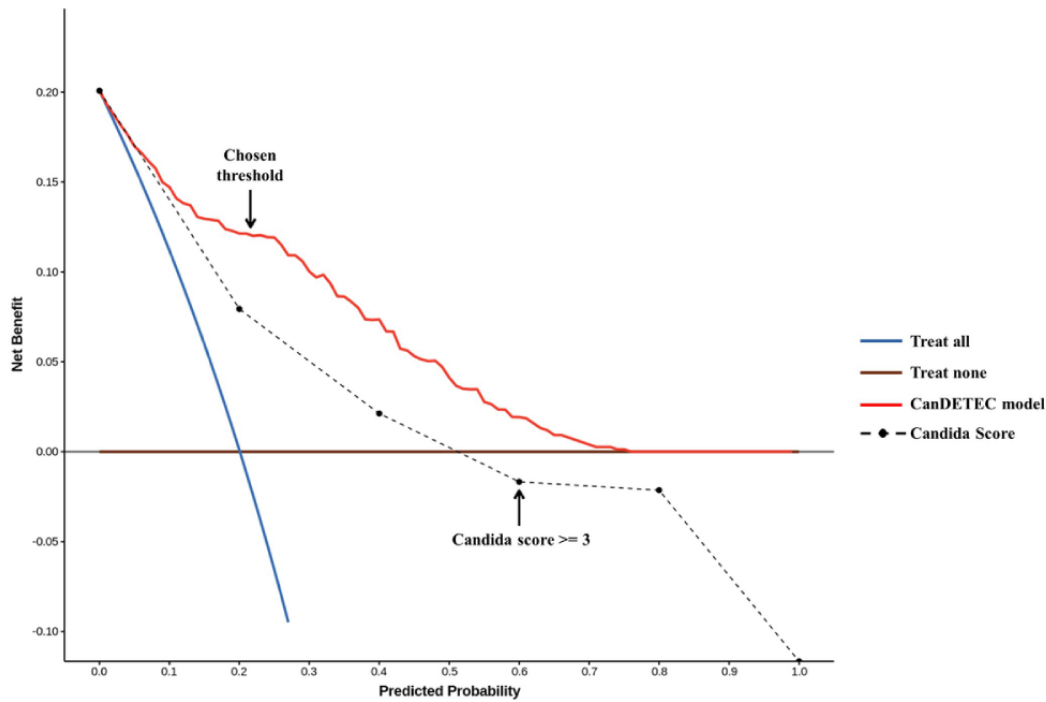
### CanDETEC Model

The threshold for categorizing the candidemia risk group was set at 0.216 by a predefined condition (sensitivity >0.90 and highest F1 score). At this cut-off point, the calculated net benefit was 0.121 (Figure 4). This threshold not only had a greater net benefit than that of treating all or none of the patients but was also close to the point showing the most significant net benefit. The coefficients, intercept, and  $R^2$  of the linear regression model

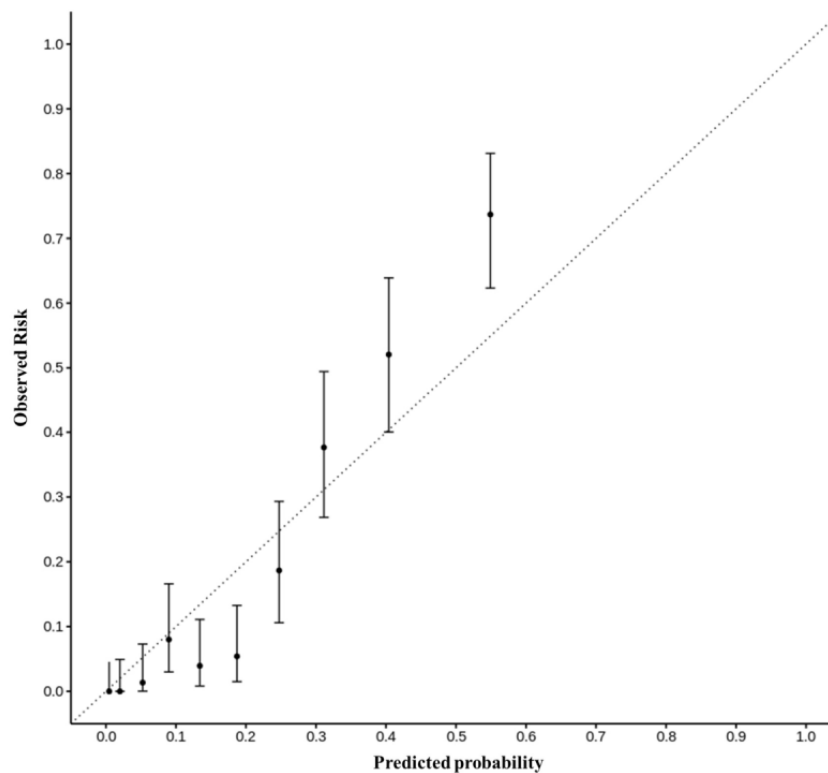
for evaluating calibration were 1.393, 0.078, and 0.93, respectively. These indexes show that the CanDETEC model was acceptably calibrated (Figure 5).

The diagnostic performance of the Candida Score, when the cut-off point was defined as  $\geq 3$ , was worse than that of all other models (AUROC 0.677, F1 score 0.354, sensitivity 0.265, specificity 0.942, positive predictive value 0.533, negative predictive value 0.836).

**Figure 4.** Decision curve of the CanDETEC model. Arrow indicates the threshold to determine candidemia high risk group.



**Figure 5.** Calibration plot of the CanDETEC model. A point represents mean decile grouped by predicted probability. Error bars represent 95% confidence interval.



## Discussion

### Principal Results

We developed a novel algorithm, called CanDETEC, to predict candidemia among patients with cancer with suspected sepsis. Given that this model only requires auto-extractable variables with low missing rates, it can be easily applied in clinical

settings to help clinicians with decision making in prescribing empirical antifungal agents.

Our CanDETEC model was developed using variables that reflect the dynamic status of patients with cancer. It seems that adopting these variables to the CanDETEC model will contribute to providing timely decision support when this algorithm applied to a real clinical setting. For example, the variable importance of variance of the respiration rate was second highest

([Multimedia Appendix 5](#)). Therefore, this model could also be used in real time.

### Comparison With Prior Work

The CanDETEC algorithm can be used to support clinicians' decision-making process in providing appropriate empirical treatment for candidemia in patients with cancer. Although candidemia contributes to the mortality of patients diagnosed with malignancies, the timely diagnosis of candidemia remains clinically challenging [8]. Given its low prevalence, busy clinicians often overlook candidemia as a possible cause of infection. Only 14.0% of patients (70/501) in our data set with candidemia had received empirical antifungal agents, and the 30-day all-cause mortality rate exceeded 50% (264/501, 52.7%), which is consistent with outcomes in previous studies [3,4,8,9]. Furthermore, infectious disease diagnostic processes that are currently used have a high cognitive burden as they require the collection and calculation of several complicated patient information variables [32]. Because the CanDETEC algorithm can be used without the clinician's subjective judgment, our model has the potential to support physicians' decision-making process with minimal additional workload.

Current tools for predicting candidemia have major limitations. Although the Candida Score is the most widely used tool to predict invasive candidemia, its sensitivity was reported to be only 37% in a recent external validation study [16], which was lower than the 81% in the original study [13]. This might have been a result of differences in the patient population. Static variables may also hinder the appropriate prediction of

candidemia because they cannot appropriately reflect changes in the clinical status of patients over time. Not surprisingly, the Candida Score showed relatively low diagnostic performance in our study (AUROC 0.677).

### Limitations

Our study has several limitations. First, CanDETEC was designed to predict candidemia when blood cultures were performed; however, the Candida Score was originally designed to predict invasive candidiasis, including candidemia. Thus, a comparison of diagnostic performance between our model and should be interpreted with caution. Second, this was a single-center retrospective study. Further multicenter prospective studies are required for external validation and to prove the clinical efficacy of the CanDETEC model. Third, although we developed a model with clinically acceptable performance, we only applied basic machine learning, such as random forest and gradient boosting. Recently, more complex ensemble models have been developed, and they have presented better performance compared to that of basic machine learning models in other medical domains [33,34]. Therefore, a follow-up study employing a state-of-the-art model should be conducted to examine whether the performance of the CanDETEC model could be improved.

### Conclusions

Our CanDETEC model, to predict candidemia in patients with cancer, is expected to reduce the mortality of patients with malignancy by helping the clinician with decision making for timely antifungal treatment.

---

### Acknowledgments

This study was supported by a grant from the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (grant number HI19C0275).

---

### Authors' Contributions

JY collected, coded, and analyzed data, developed the models, and co-wrote the manuscript. SK designed the study and co-wrote the manuscript. SH collected and coded the data and inspected the manuscript. KH selected clinically important variables and inspected the manuscript. JH developed the models and co-wrote the manuscript. WCC designed the study and oversaw data analysis and writing.

---

### Conflicts of Interest

None declared.

---

#### Multimedia Appendix 1

Variable group classification process.

[[XLSX File \(Microsoft Excel File\), 22 KB - medinform\\_v9i7e24651\\_app1.xlsx](#) ]

---

#### Multimedia Appendix 2

Detailed model parameters.

[[DOCX File, 21 KB - medinform\\_v9i7e24651\\_app2.docx](#) ]

---

#### Multimedia Appendix 3

Characteristics of candidemia events.

[[XLSX File \(Microsoft Excel File\), 9 KB - medinform\\_v9i7e24651\\_app3.xlsx](#) ]

## Multimedia Appendix 4

Bonferroni-corrected posthoc comparisons.

[\[XLSX File \(Microsoft Excel File\), 11 KB - medinform\\_v9i7e24651\\_app4.xlsx\]](#)

## Multimedia Appendix 5

Variable importance of the CanDETEC model.

[\[XLSX File \(Microsoft Excel File\), 12 KB - medinform\\_v9i7e24651\\_app5.xlsx\]](#)

## References

1. Clancy CJ, Nguyen MH. T2 magnetic resonance for the diagnosis of bloodstream infections: charting a path forward. *J Antimicrob Chemother* 2018 Mar 01;73(suppl\_4):iv2-iv5. [doi: [10.1093/jac/dky050](#)] [Medline: [29608754](#)]
2. Viscoli C, Girmenia C, Marinus A, Collette L, Martino P, Vandercam B, et al. Candidemia in cancer patients: a prospective, multicenter surveillance study by the Invasive Fungal Infection Group (IFIG) of the European Organization for Research and Treatment of Cancer (EORTC). *Clin Infect Dis* 1999 May;28(5):1071-1079. [doi: [10.1086/514731](#)] [Medline: [10452637](#)]
3. Tang H, Liu W, Lin H, Lai C. Epidemiology and prognostic factors of candidemia in cancer patients. *PLoS One* 2014 Jun 5;9(6):e99103 [FREE Full text] [doi: [10.1371/journal.pone.0099103](#)] [Medline: [24901336](#)]
4. Li D, Xia R, Zhang Q, Bai C, Li Z, Zhang P. Evaluation of candidemia in epidemiology and risk factors among cancer patients in a cancer center of China: an 8-year case-control study. *BMC Infect Dis* 2017 Aug 03;17(1):536 [FREE Full text] [doi: [10.1186/s12879-017-2636-x](#)] [Medline: [28768479](#)]
5. Anaissie EJ, Rex JH, Uzun O, Vartivarian S. Predictors of adverse outcome in cancer patients with candidemia. *Am J Med* 1998 Mar;104(3):238-245. [doi: [10.1016/s0002-9343\(98\)00030-8](#)] [Medline: [9552086](#)]
6. Uzun O, Anaissie E. Predictors of outcome in cancer patients with candidemia. *Ann Oncol* 2000 Dec;11(12):1517-1521 [FREE Full text] [doi: [10.1023/a:1008308923252](#)] [Medline: [11205457](#)]
7. Nucci M, Silveira MI, Spector N, Silveira F, Velasco E, Akiti T, et al. Risk factors for death among cancer patients with fungemia. *Clin Infect Dis* 1998 Jul;27(1):107-111. [doi: [10.1086/514609](#)] [Medline: [9675463](#)]
8. Morrell M, Fraser VJ, Kollef MH. Delaying the empiric treatment of candida bloodstream infection until positive blood culture results are obtained: a potential risk factor for hospital mortality. *Antimicrob Agents Chemother* 2005 Sep;49(9):3640-3645 [FREE Full text] [doi: [10.1128/AAC.49.9.3640-3645.2005](#)] [Medline: [16127033](#)]
9. Zilberberg MD, Kollef MH, Arnold H, Labelle A, Micek ST, Kothari S, et al. Inappropriate empiric antifungal therapy for candidemia in the ICU and hospital resource utilization: a retrospective cohort study. *BMC Infect Dis* 2010 Jun 03;10(1):150 [FREE Full text] [doi: [10.1186/1471-2334-10-150](#)] [Medline: [20525301](#)]
10. Horvath LL, Hospenthal DR, Murray CK, Dooley DP. Detection of simulated candidemia by the BACTEC 9240 system with plus aerobic/F and anaerobic/F blood culture bottles. *J Clin Microbiol* 2003 Oct;41(10):4714-4717 [FREE Full text] [doi: [10.1128/JCM.41.10.4714-4717.2003](#)] [Medline: [14532209](#)]
11. Gonzalez-Lara MF, Ostrosky-Zeichner L. Update on the diagnosis of candidemia and invasive candidiasis. *Curr Fungal Infect Rep* 2019 Nov 23;13(4):301-307. [doi: [10.1007/s12281-019-00367-1](#)]
12. Pittet D, Monod M, Suter PM, Frenk E, Auckenthaler R. Candida colonization and subsequent infections in critically ill surgical patients. *Ann Surg* 1994 Dec;220(6):751-758. [doi: [10.1097/0000658-199412000-00008](#)] [Medline: [7986142](#)]
13. León C, Ruiz-Santana S, Saavedra P, Galván B, Blanco A, Castro C, Cava Study Group. Usefulness of the "Candida score" for discriminating between Candida colonization and invasive candidiasis in non-neutropenic critically ill patients: a prospective multicenter study. *Crit Care Med* 2009 May;37(5):1624-1633. [doi: [10.1097/CCM.0b013e31819daa14](#)] [Medline: [19325481](#)]
14. Guillamet CV, Vazquez R, Micek ST, Ursu O, Kollef M. Development and validation of a clinical prediction rule for candidemia in hospitalized patients with severe sepsis and septic shock. *J Crit Care* 2015 Aug;30(4):715-720. [doi: [10.1016/j.jcrc.2015.03.010](#)] [Medline: [25813550](#)]
15. Atamna A, Eliakim-Raz N, Mohana J, Ben-Zvi H, Sorek N, Shochat T, et al. Predicting candidemia in the internal medicine wards: a comparison with gram-negative bacteremia-a retrospective study. *Diagn Microbiol Infect Dis* 2019 Sep;95(1):80-83. [doi: [10.1016/j.diagmicrobio.2019.04.007](#)] [Medline: [31129007](#)]
16. Laine ME, Flannery AH, Moody B, Thompson Bastin ML. Need for expanded Candida Score for empiric antifungal use in medically critically ill patients? *Crit Care* 2019 Jul 04;23(1):242 [FREE Full text] [doi: [10.1186/s13054-019-2525-3](#)] [Medline: [31272491](#)]
17. Altintop YA, Ergul AB, Koc AN, Atalay MA. Evaluation of Candida colonization and use of the Candida Colonization Index in a paediatric Intensive Care Unit: a prospective observational study. *Infez Med* 2019 Jun 01;27(2):159-167 [FREE Full text] [Medline: [31205039](#)]
18. Pfaller MA, Wolk DM, Lowery TJ. T2MR and T2Candida: novel technology for the rapid diagnosis of candidemia and invasive candidiasis. *Future Microbiol* 2016;11(1):103-117 [FREE Full text] [doi: [10.2217/fmb.15.111](#)] [Medline: [26371384](#)]



19. Arendrup MC, Andersen JS, Holten MK, Krarup KB, Reiter N, Schierbeck J, et al. Diagnostic performance of T2Candida among ICU patients with risk factors for invasive candidiasis. *Open Forum Infect Dis* 2019 May;6(5):ofz136 [FREE Full text] [doi: [10.1093/ofid/ofz136](https://doi.org/10.1093/ofid/ofz136)] [Medline: [31069244](https://pubmed.ncbi.nlm.nih.gov/31069244/)]
20. Tang D, Chen X, Zhu C, Li Z, Xia Y, Guo X. Pooled analysis of T2 Candida for rapid diagnosis of candidiasis. *BMC Infect Dis* 2019 Sep 11;19(1):798 [FREE Full text] [doi: [10.1186/s12879-019-4419-z](https://doi.org/10.1186/s12879-019-4419-z)] [Medline: [31510929](https://pubmed.ncbi.nlm.nih.gov/31510929/)]
21. Beaulieu-Jones B, Finlayson SG, Chivers C, Chen I, McDermott M, Kandola J, et al. Trends and focus of machine learning applications for health research. *JAMA Netw Open* 2019 Oct 02;2(10):e1914051 [FREE Full text] [doi: [10.1001/jamanetworkopen.2019.14051](https://doi.org/10.1001/jamanetworkopen.2019.14051)] [Medline: [31651969](https://pubmed.ncbi.nlm.nih.gov/31651969/)]
22. Johnson AEW, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine learning and decision support in critical care. *Proc IEEE Inst Electr Electron Eng* 2016 Feb;104(2):444-466 [FREE Full text] [doi: [10.1109/JPROC.2015.2501978](https://doi.org/10.1109/JPROC.2015.2501978)] [Medline: [27765959](https://pubmed.ncbi.nlm.nih.gov/27765959/)]
23. Haibo H, Garcia E. Learning from imbalanced data. *IEEE Trans Knowl Data Eng* 2009 Sep;21(9):1263-1284. [doi: [10.1109/TKDE.2008.239](https://doi.org/10.1109/TKDE.2008.239)]
24. Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A. Learning from imbalanced data in surveillance of nosocomial infection. *Artif Intell Med* 2006 May;37(1):7-18. [doi: [10.1016/j.artmed.2005.03.002](https://doi.org/10.1016/j.artmed.2005.03.002)] [Medline: [16233974](https://pubmed.ncbi.nlm.nih.gov/16233974/)]
25. Liu C, Huang L, Wang W, Chen T, Yen C, Yang M, et al. Candidemia in cancer patients: impact of early removal of non-tunneled central venous catheters on outcome. *J Infect* 2009 Feb;58(2):154-160. [doi: [10.1016/j.jinf.2008.12.008](https://doi.org/10.1016/j.jinf.2008.12.008)] [Medline: [19162330](https://pubmed.ncbi.nlm.nih.gov/19162330/)]
26. Karabinis A, Hill C, Leclercq B, Tancrede C, Baume D, Andremont A. Risk factors for candidemia in cancer patients: a case-control study. *J Clin Microbiol* 1988 Mar;26(3):429-432 [FREE Full text] [doi: [10.1128/jcm.26.3.429-432.1988](https://doi.org/10.1128/jcm.26.3.429-432.1988)] [Medline: [3356785](https://pubmed.ncbi.nlm.nih.gov/3356785/)]
27. León C, Ruiz-Santana S, Saavedra P, Almirante B, Nolla-Salas J, Alvarez-Lerma F, EPCAN Study Group. A bedside scoring system ("Candida score") for early antifungal treatment in nonneutropenic critically ill patients with Candida colonization. *Crit Care Med* 2006 Mar;34(3):730-737. [doi: [10.1097/01.CCM.0000202208.37364.7D](https://doi.org/10.1097/01.CCM.0000202208.37364.7D)] [Medline: [16505659](https://pubmed.ncbi.nlm.nih.gov/16505659/)]
28. Olson R, Moore J. TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In: Hutter F, Kotthoff L, Vanschoren J, editors. *Automated Machine Learning*. Heidelberg: Springer International Publishing; Jun 24, 2019:66-74.
29. Miller K, Mosby D, Capan M, Kowalski R, Ratwani R, Noaiseh Y, et al. Interface, information, interaction: a narrative review of design and functional requirements for clinical decision support. *J Am Med Inform Assoc* 2018 May 01;25(5):585-592. [doi: [10.1093/jamia/ocx118](https://doi.org/10.1093/jamia/ocx118)] [Medline: [29126196](https://pubmed.ncbi.nlm.nih.gov/29126196/)]
30. van Rossum G, Drake F. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace; 2009.
31. CanDETEC. GitHub. URL: <https://github.com/Smart-Health-Lab/CANDETEC> [accessed 2021-07-13]
32. Roosan D, Weir C, Samore M, Jones M, Rahman M, Stoddard GJ, et al. Identifying complexity in infectious diseases inpatient settings: an observation study. *J Biomed Inform* 2017 Jul;71S:S13-S21 [FREE Full text] [doi: [10.1016/j.jbi.2016.10.018](https://doi.org/10.1016/j.jbi.2016.10.018)] [Medline: [27818310](https://pubmed.ncbi.nlm.nih.gov/27818310/)]
33. Afzal M, Hussain M, Malik KM, Lee S. Impact of automatic query generation and quality recognition using deep learning to curate evidence from biomedical literature: empirical study. *JMIR Med Inform* 2019 Dec 09;7(4):e13430 [FREE Full text] [doi: [10.2196/13430](https://doi.org/10.2196/13430)] [Medline: [31815673](https://pubmed.ncbi.nlm.nih.gov/31815673/)]
34. Thongkam J, Xu G, Zhang Y. AdaBoost algorithm with random forests for predicting breast cancer survivability. 2008 Presented at: IEEE International Joint Conference on Neural Networks; June 1-8; Hong Kong. [doi: [10.1109/IJCNN.2008.4634231](https://doi.org/10.1109/IJCNN.2008.4634231)]

## Abbreviations

**AUROC:** area under the receiver operating characteristic curve

*Edited by C Lovis; submitted 29.09.20; peer-reviewed by M Afzal; comments to author 21.10.20; revised version received 09.11.20; accepted 17.06.21; published 26.07.21.*

### *Please cite as:*

Yoo J, Kim SH, Hur S, Ha J, Huh K, Cha WC

*Candidemia Risk Prediction (CanDETEC) Model for Patients With Malignancy: Model Development and Validation in a Single-Center Retrospective Study*

*JMIR Med Inform* 2021;9(7):e24651

URL: <https://medinform.jmir.org/2021/7/e24651>

doi: [10.2196/24651](https://doi.org/10.2196/24651)

PMID: [34309570](https://pubmed.ncbi.nlm.nih.gov/34309570/)

©Junsang Yoo, Si-Ho Kim, Sujeong Hur, Juhyung Ha, Kyungmin Huh, Won Chul Cha. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 26.07.2021. 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

# Prediction Model of Anastomotic Leakage Among Esophageal Cancer Patients After Receiving an Esophagectomy: Machine Learning Approach

Ziran Zhao<sup>1</sup>, MD; Xi Cheng<sup>2</sup>, MPH; Xiao Sun<sup>3</sup>, MPH; Shanrui Ma<sup>1</sup>, MD; Hao Feng<sup>1</sup>, MD; Liang Zhao<sup>1</sup>, MD

<sup>1</sup>Thoracic Surgery Department, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

<sup>2</sup>Department of Global Health Management, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, United States

<sup>3</sup>Department of Epidemiology, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, United States

**Corresponding Author:**

Liang Zhao, MD

Thoracic Surgery Department

National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital

Chinese Academy of Medical Sciences and Peking Union Medical College

Panjiayuan Nanli No 17

Beijing, 100021

China

Phone: 86 01087787150

Email: [drzhaoliang@126.com](mailto:drzhaoliang@126.com)

## Abstract

**Background:** Anastomotic leakage (AL) is one of the severe postoperative adverse events (5%-30%), and it is related to increased medical costs in cancer patients who undergo esophagectomies. Machine learning (ML) methods show good performance at predicting risk for AL. However, AL risk prediction based on ML models among the Chinese population is unavailable.

**Objective:** This study uses ML techniques to develop and validate a risk prediction model to screen patients with emerging AL risk factors.

**Methods:** Analyses were performed using medical records from 710 patients who underwent esophagectomies at the National Clinical Research Center for Cancer between January 2010 and May 2015. We randomly split (9:1) the data set into a training data set of 639 patients and a testing data set of 71 patients using a computer algorithm. We assessed multiple classification tools to create a multivariate risk prediction model. Our ML algorithms contained decision tree, random forest, naive Bayes, and logistic regression with least absolute shrinkage and selection operator. The optimal AL prediction model was selected based on model evaluation metrics.

**Results:** The final risk panel included 36 independent risk features. Of those, 10 features were significantly identified by the logistic model, including aortic calcification (OR 2.77, 95% CI 1.32-5.81), celiac trunk calcification (OR 2.79, 95% CI 1.20-6.48), forced expiratory volume 1% (OR 0.51, 95% CI 0.30-0.89); TLco (OR 0.56, 95% CI 0.27-1.18), peripheral vascular disease (OR 4.97, 95% CI 1.44-17.07), laparoscope (OR 3.92, 95% CI 1.23-12.51), postoperative length of hospital stay (OR 1.17, 95% CI 1.13-1.21), vascular permeability activity (OR 0.46, 95% CI 0.14-1.48), and fat liquefaction of incisions (OR 4.36, 95% CI 1.86-10.21). Logistic regression with least absolute shrinkage and selection operator offered the highest prediction quality with an area under the receiver operator characteristic of 72% in the training data set. The testing model also achieved similar high performance.

**Conclusions:** Our model offered a prediction of AL with high accuracy, assisting in AL prevention and treatment. A personalized ML prediction model with a purely data-driven selection of features is feasible and effective in predicting AL in patients who underwent esophagectomy.

(*JMIR Med Inform* 2021;9(7):e27110) doi:[10.2196/27110](https://doi.org/10.2196/27110)

**KEYWORDS**

anastomotic leakage; esophageal cancer; esophagectomy; machine learning; risk factors

## Introduction

Esophagectomies are important treatments for early-stage and locoregionally advanced esophageal cancer. However, esophagectomies are burdened with a high incidence of complications. Anastomotic leakage (AL), including cervical anastomotic leakage and intrathoracic anastomotic leakage, is a significant complication following an esophagectomy accounting for morbidity and mortality (5%-30%) [1]. Moreover, it is associated with prolonged intensive care unit stays, reduced quality of life, and higher hospital costs [2]. Accordingly, the prevention and optimal management of AL after an esophagectomy are of great importance. Investigations should be undertaken as soon as the risk factors of AL are recognized because any delay would substantially worsen the prognosis [3]. The timely detection of surgical and nonsurgical AL risk factors and the adoption of a proper approach are keys to the successful treatment of AL.

Previous conventional prediction models exploring AL risk factors have not validated their model's performance and provided the rationale for feature selection in their work. Analyses of several known predictive factors of AL have yielded poor statistical performance across studies [4]. Therefore, no consistent and clear predictive factors can be used to target patients with a risk of AL in clinical practice. Machine learning (ML) approaches are particularly suited to predictions based on real world evidence, which involves a computer algorithm learning important features of a data set and capturing complex relationships in the data to enable predictions about other unseen data. ML ensures a more accurate and robust prediction than conventional statistical models since it can capture nonlinear relationships among clinical features without human-biased intervention. It can predict AL for individual patients more accurately in terms of model performance and generalizability [5,6].

This study aims to use ML techniques to explore the risk factors that influence the occurrence of AL and the consequent clinical outcomes after an esophagectomy to inform the clinical management of AL. Various medical strategies are available to prevent AL after an esophagectomy, including patient screening and preparation, technical-surgical details, and postsurgical care management. Thus, the evidence generated from the prediction model can serve as a practical guide to the clinical management of patients undergoing esophagectomies with a particular focus on AL prevention.

## Methods

### Study Design and Participants

In this retrospective study, we collected data on 710 patients who underwent an esophagectomy for esophageal cancer at the Department of Thoracic Surgery in the National Clinical Research Center for Cancer in China between January 2010 and May 2015. Our data were collected by manual chart review. The protocols and guidelines for data abstraction from the medical record were developed prior to launching the medical record data collection effort in our hospital to ensure the reliability and accuracy of the data collection. Our medical data

investigator was trained carefully and followed strict protocols of data collection. Any discrepancies were reviewed jointly and discussed with our medical team to clarify any issues. The Independent Ethics Committee had approved this retrospective cohort study of the institution, and the requirement to obtain informed consent was waived.

We collected 76 features from patients' medical records, including patient-related information such as demography (age, gender, and BMI), smoking and alcohol consumption, surgical history, prescriptive medication, and the American Society of Anesthesiologists (ASA) classification. The comorbidities registered included diabetes mellitus, malnutrition, hypertension, other cardiovascular diseases (cardiac arrhythmia and coronary heart disease), chronic obstructive pulmonary disease (COPD), etc. Intra-operative features included timing of surgery, type of operation, incisions, and blood transfusions. The data set identified two kinds of esophagectomy, including open esophagectomy and minimally invasive esophagectomy (MIE). Besides total MIE, thoracoscopic or laparotomy assisted esophagectomy, or hybrid MIE were also included. The three most common techniques for thoracic esophageal cancer were: (1) the transhiatal approach, (2) Ivor Lewis esophagectomy, and (3) the McKeown technique. The postoperative features included hospital stay, complications, and mortality, etc.

### Outcome Definition

The primary outcome of this study is a diagnosis of AL which was ascertained through clinical symptoms and confirmed by endoscopy, radiological examination, clinical examination of the anastomosis, or reoperation [7]. The Esophagectomy Complications Consensus Group defined AL as a "full-thickness gastrointestinal defect involving the esophagus, anastomosis, staple line, or conduit irrespective of presentation or method of identification" [8]. In our clinical practice, an AL is first suspected if there were any (1) clinical signs, such as fever, abdominal pain, feculent drainage, purulent drainage, or signs of peritonism; (2) radiographic signs, such as fluid collection or gas containing collection; and (3) signs of anastomotic dehiscence during endoscopy. The definitive diagnostic tool for suspected AL is a computerized tomography scan with a contrast of the abdomen and pelvis, which will demonstrate the presence of any extraluminal contents. An additional assessment is urgent blood tests, including full blood count, a coagulation screen, etc [9].

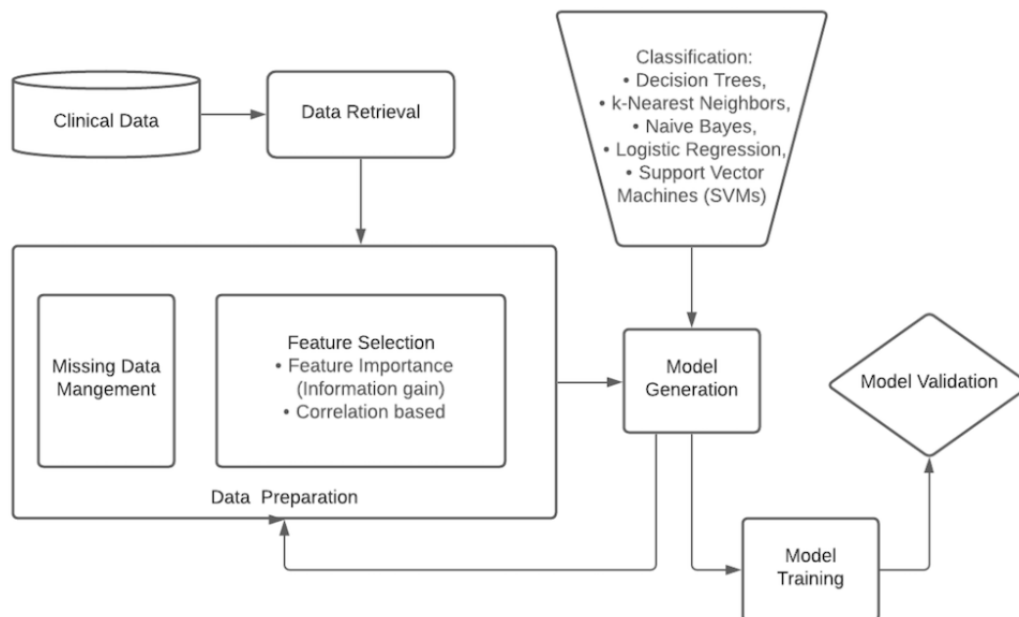
### Data Management and Machine Learning Approaches

Figure 1 illustrates the construction of the ML model and how the risk predictors (features) were handled. We first checked missing conditions and the balance of the data set; no variables were reported to have a missing percentage over 5%, meaning the completeness rate of every variable was above 95%. For this study, we randomly split the large data set (9:1) into a training data set (n=639) and a testing data set (n=71) using a computer algorithm. We split the data set using this ratio to allow sufficient training data to quantify the model's complexity while maintaining adequate data to validate the model [10]. The cross-validation process was iterated 9 times. Model parameters were optimized by a grid search, greedily tuning the model hyperparameters. The mean area under the receiver operating

characteristic curve (AUROC) was used to determine which model performed best and further tested with testing data set.

The sensitivity level of AUROC is set to 90%, which is considered clinically relevant.

**Figure 1.** Analysis workflow for data management and model development.



### Identification of Risk Factors (Feature Selection)

Identifying the most important features was based on the two most used feature selection filter methods in ML: (1) feature importance and (2) correlation-based feature selection. We used filter methods of feature selection because it is independent of the potential models [11]. Feature importance is a univariate filter that compares each feature's correlation with the outcome separately and removes features with zero importance according to a gradient boosting machine (GBM) learning model. Generally, importance provides a score indicating how useful or valuable each feature is in constructing the boosted decision trees within the model. The feature importance is averaged over 10 training runs of the GBM to reduce variance [12,13]. The correlation-based method is a multivariate filter that identifies the collinear features and removes the redundant features that are highly correlated with one another. These 2 feature selection methods have advantages in that they are more stable than the traditional statistical approaches, such as backward logistic regression, and they considerably minimized the model's over-fitting problem [13]. Similar to previous medical ML studies, we performed the 2 feature selection methods on all 76 features using our training data set and initially identified N features that have the least correlation to AL, then plotted the change in AUROC for the prediction of AL from 1 to N features [13]. The ML algorithm also plotted the change in cumulative importance and recognized the least number of features required, receiving above 99% of the cumulative importance. Thus, we included the smallest number of independent features into the final prediction model.

### Model Generation, Training, and Validation

Once our features were defined, we considered five different ML classification models (classifiers) to build our models: (1) logistic regression with regularizations, (2) a support vector

machine using Gaussian kernel, (3) a decision tree based on information gain, (4) a random forest including 9 decision tree classifiers based on Gin impurity, and (5) a naive Bayes classifier assuming a Gaussian distribution. These 5 algorithms were chosen for comparison because they are well-accepted ML methods typically used in medical applications [10]. Finally, models were validated with our testing data set, and the extended metrics (AUROC, accuracy, recall, F1 score, and precision) were reported.

### Statistical Analysis

This study aligns with TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) and TRIPOD-ML guidelines (see [Multimedia Appendices 1 and 2](#)) [14,15]. The complete set of patient medical data was utilized to maximize the power and generalizability of our results. All the analyses were performed using sklearn, pandas, numpy, and lightgbm packages in Python (version 3.6.1).

In the descriptive summary, categorical variables were presented as numbers and percentages and continuous variables as mean and standard deviation. The P values were also provided for the association of each factor with the presence and absence of AL using the Pearson chi-square method. The ranked risk features panel from the training data set provided the importance and regression coefficients of the association of each feature in the final prediction model. Finally, we presented the risk factors of AL associated with each feature using odds ratios (OR) and 95% CI.

### Ethics Approval and Consent to Participate

This retrospective cohort study was approved by the institutional review board of the National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital's Ethics



Committee, and the requirement to obtain informed consent was waived.

## Results

Demographic and symptom characteristics for training and testing data are depicted in [Table 1](#). The proportions were consistent between the two data sets, with 17% and 20% of patients indicating the presence of AL in the training and testing data sets, respectively. Patients with AL were generally male, with ages ranging from 50 to 70 years, who were more likely

to be smokers and heavy drinkers, and more likely to experience hypertension, peripheral vascular disease, and cardiac arrhythmia.

Feature selection using the training data set provided us the independent risk factors panel predicting AL. The plot of change of AUROC over the number of features indicated 34 features would yield a relatively high AUROC value (AUROC=0.78; [Figure 2](#)). The cumulative importance score plot identified at least 38 features required for our final model ([Figure 3](#); [Table 2](#)) [16,17].

**Table 1.** Demographic and symptom characteristics in the data sets defined by the presence or absence of AL.

	Training data set		<i>P</i> values <sup>a</sup>	Testing data set		<i>P</i> values
	No	Yes		No	Yes	
Presence of anastomotic leakage, n (%)	531 (83)	108 (17)		57 (80)	14 (20)	
<b>Age, n (%)</b>						
<50	64 (12)	80 (15)	.36	58 (11)	37 (7)	.06
50-59	196 (37)	207 (39)		271 (51)	112 (21)	
60-69	218 (41)	191 (36)		170 (32)	303 (57)	
>=70	53 (10)	53 (10)		37 (7)	74 (14)	
<b>Gender, n (%)</b>						
Male	419 (79)	451 (85)	.17	409 (77)	377 (71)	.66
Female	112 (21)	80 (15)		122 (23)	154 (29)	
<b>BMI, n (%)</b>						
15-19	27 (5)	48 (9)	.74	37 (7)	37 (7)	.73
20-25	340 (64)	287 (54)		372 (70)	340 (64)	
>=26	165 (31)	196 (37)		122 (23)	154 (29)	
<b>Ever smoked, n (%)</b>						
No	207 (39)	191 (36)	.53	234 (44)	228 (43)	.95
Yes	324 (61)	340 (64)		297 (56)	303 (57)	
<b>Ever alcohol heavy drinker, n (%)</b>						
No	212 (40)	159 (30)	.04	250 (47)	303 (57)	.52
Yes	319 (60)	372 (70)		281 (53)	228 (43)	
<b>Presence of aortic calcification, n (%)</b>						
No	419 (79)	292 (55)	<.001	457 (86)	303 (57)	.01
Yes	112 (21)	239 (45)		74 (14)	228 (43)	
<b>Presence of celiac trunk calcification, n (%)</b>						
No	473 (89)	366 (69)	<.001	457 (86)	457 (86)	.98
Yes	58 (11)	165 (31)		74 (14)	74 (14)	
<b>FEV1% category, n (%)</b>						
30-59	21 (4)	27 (5)	.13	37 (7)	74 (14)	<.001
60-79	122 (23)	165 (31)		101 (19)	377 (71)	
80-130	388 (73)	340 (64)		393 (74)	74 (14)	
<b>Abdominal surgery, n (%)</b>						
No	473 (89)	446 (84)	.18	446 (84)	494 (93)	.41
Yes	58 (11)	85 (16)		85 (16)	37 (7)	
<b>Presence of cardiac arrhythmia, n (%)</b>						
No	462 (87)	441 (83)	.31	457 (86)	457 (86)	.98
Yes	69 (13)	90 (17)		74 (14)	74 (14)	
<b>Presence of peripheral vascular disease, n (%)</b>						
No	515 (97)	473 (89)	<.001	531 (100)	531 (100)	–
Yes	16 (3)	58 (11)		0 (0)	0 (0)	
<b>Presence of hypertension, n (%)</b>						
No	409 (77)	340 (64)	<.001	398 (75)	377 (71)	.76

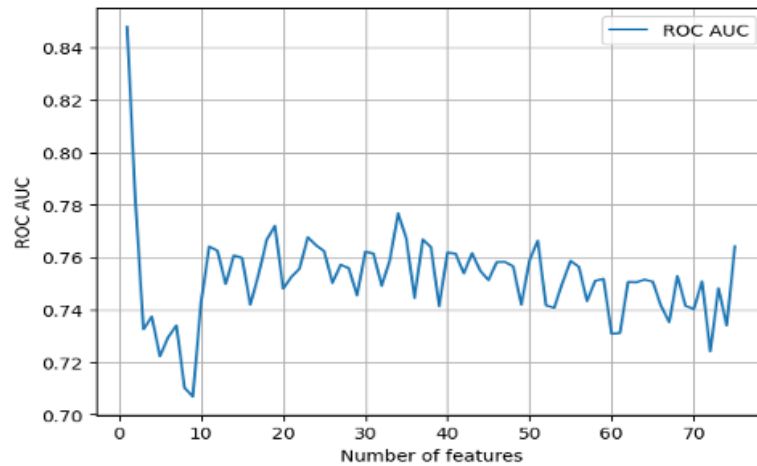
	Training data set			Testing data set		
	No	Yes	<i>P</i> values <sup>a</sup>	No	Yes	<i>P</i> values
Yes	122 (23)	191 (36)		133 (25)	154 (29)	
<b>Ever used hypotension drug, n (%)</b>						
No	457 (86)	425 (80)	.10	457 (86)	419 (79)	.50
Yes	74 (14)	106 (20)		74 (14)	112 (21)	
<b>Ever taken insulin, n (%)</b>						
No	504 (95)	489 (92)	.13	510 (96)	457 (86)	.12
Yes	27 (5)	42 (8)		21 (4)	74 (14)	
<b>Lesion length category, n (%)</b>						
1-3	117 (22)	74 (14)	.06	96 (18)	154 (29)	.25
4-5	218 (41)	234 (44)		297 (56)	303 (57)	
6-10	181 (34)	202 (38)		138 (26)	74 (14)	
>10	16 (3)	27 (5)		0 (0)	0 (0)	
<b>Position of lesion, n (%)</b>						
Upper esophagus	175 (33)	149 (28)	.90	186 (35)	303 (57)	.05
Middle esophagus	212 (40)	234 (44)		186 (35)	191 (36)	
Lower esophagus	69 (13)	96 (18)		74 (14)	37 (7)	
Upper-middle esophagus	37 (7)	16 (3)		37 (7)	0 (0)	
Lower-middle esophagus	27 (5)	37 (7)		27 (5)	0 (0)	
Multi-position	16 (3)	5 (1)		21 (4)	0 (0)	
<b>ASA<sup>b</sup> physical status classification, n (%)</b>						
1	117 (22)	80 (15)	<.001	122 (23)	37 (7)	.04
2	372 (70)	356 (67)		398 (75)	419 (79)	
3	42 (8)	101 (19)		11 (2)	74 (14)	
<b>Blood transfusion, n (%)</b>						
No	127 (24)	96 (18)	.97	196 (37)	154 (29)	.14
Yes	404 (76)	435 (82)		335 (63)	377 (71)	
<b>Type of anastomotic, n (%)</b>						
No	425 (80)	430 (81)	.79	435 (82)	340 (64)	.19
Yes	106 (20)	101 (19)		96 (18)	191 (36)	
<b>Tube stomach, n (%)</b>						
No	181 (34)	170 (32)	.07	101 (19)	191 (36)	.65
Yes	350 (66)	361 (68)		430 (81)	340 (64)	
<b>Surgical approach, n (%)</b>						
Nonthoraco-laparoscopy	181 (34)	133 (25)	.16	186 (35)	154 (29)	.68
Thoraco-laparoscopy	350 (66)	398 (75)		345 (65)	377 (71)	
<b>Laparoscope, n (%)</b>						
No	202 (38)	165 (31)	.03	234 (44)	266 (50)	.77
Yes	329 (62)	366 (69)		297 (56)	266 (50)	
<b>Histology grade, n (%)</b>						
0	276 (52)	218 (41)	.80	281 (53)	303 (57)	.95
1	255 (48)	313 (59)		250 (47)	228 (43)	

	Training data set			Testing data set		
	No	Yes	<i>P</i> values <sup>a</sup>	No	Yes	<i>P</i> values
<b>T classification, n (%)</b>						
0	42 (8)	27 (5)	.14	37 (7)	37 (7)	.31
1	53 (10)	58 (11)		48 (9)	74 (14)	
2	281 (53)	303 (57)		281 (53)	228 (43)	
3	159 (30)	143 (27)		170 (32)	191 (36)	
<b>Multiple primary, n (%)</b>						
No	499 (94)	478 (90)	.37	494 (93)	531 (100)	.36
Yes	32 (6)	53 (10)		37 (7)	0 (0)	
<b>N classification, n (%)</b>						
1	133 (25)	122 (23)	.26	175 (33)	154 (29)	.62
2	101 (19)	101 (19)		64 (12)	37 (7)	
3	239 (45)	234 (44)		212 (40)	191 (36)	
4	32 (6)	42 (8)		58 (11)	112 (21)	
5	27 (5)	32 (6)		21 (4)	37 (7)	
<b>Thyroglobulin level, n (%)</b>						
<1.7	441 (83)	473 (89)	.15	473 (89)	457 (86)	.70
>=1.7	90 (17)	58 (11)		58 (11)	74 (14)	
<b>Tumor vascular permeability, n (%)</b>						
<20	372 (70)	372 (70)	.92	324 (61)	419 (79)	.23
>=20	159 (30)	159 (30)		207 (39)	112 (21)	
<b>Postoperative ventilator-assisted breathing, n (%)</b>						
No	515 (97)	473 (89)	<.001	520 (98)	457 (86)	.04
Yes	16 (3)	58 (11)		11 (2)	74 (14)	
<b>Lung infection, n (%)</b>						
No	510 (96)	478 (90)	<.001	494 (93)	494 (93)	0.99
Yes	21 (4)	53 (10)		37 (7)	37 (7)	
<b>Pleural effusion or empyema, n (%)</b>						
No	515 (97)	446 (84)	<.001	520 (98)	419 (79)	<.001
Yes	16 (3)	85 (16)		11 (2)	112 (21)	
<b>Incision fat liquefaction and infection, n (%)</b>						
No	494 (93)	398 (75)	<.001	467 (88)	494 (93)	.59
Yes	37 (7)	133 (25)		64 (12)	37 (7)	
Hospital Length of Stay, mean (SD)	13.08 (7.2)	35.56 (23.2)	<.001	12.75 (4.13)	29.36 (16.72)	<.001

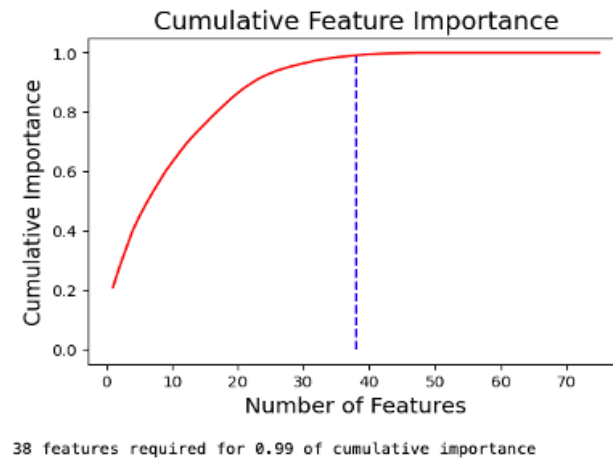
<sup>a</sup>*P* values are reported using the Pearson chi-square method.

<sup>b</sup>The American Society of Anesthesiologists Physical Status Classification System is a well-established assignment that assesses and communicates a patient's pre-anesthesia medical comorbidities.

**Figure 2.** Change of area under the receiver operating characteristic by the number of features.



**Figure 3.** Number of features and cumulative feature importance.



**Table 2.** Comparison of the model’s performance metrics with different machine learning classifiers.

	Logistic regression	Support vector classifier	Decision tree	Random forest	Gaussian naïve Bayes	Best score
Accuracy	0.91	0.90	0.84	0.90	0.82	Logistic regression
Precision	0.81	0.81	0.55	0.80	0.50	Logistic regression
Recall (sensitivity)	0.64	0.59	0.58	0.59	0.68	Gaussian naïve Bayes
F1 score	0.71	0.68	0.56	0.67	0.58	Logistic regression
AUROC	0.76	0.72	0.65	0.70	0.67	Logistic regression

After removing the features with only one unique value and one strong collinear feature, we identified the 36 most important risk factors to create our model, ensuring robustness and stability. Our panel of predictive risk features, listed and ranked by their importance, is shown in Table 3. The preoperative factors included patient’s age, gender, BMI, smoking and alcohol intake, malnutrition status, ASA index, cardiovascular disease (aortic calcification, celiac trunk calcification, peripheral vascular disease, cardiac arrhythmia, and hypertension), obstructive lung diseases test scores (forced vital capacity ratio [FEV1%], transfer factor for carbon monoxide [TLco]), surgical history (abdominal surgery), drug usage (insulin and hypertension drugs), and cancer staging (TNM classification of

malignant tumors). The intra-operative factors included operation time, lesion length and position of the lesion, availability of blood transfusion, and surgical approaches. The postoperative factors contained a prolonged hospital stay length and surgical complications such as arrhythmia, lung infection, pleural effusion, and fat liquefaction of incisions.

After the feature panel was determined, the model selection metrics showed the logistic regression with least absolute shrinkage and selection operator (LASSO) obtained the best median AUROC score, making it the most reliable ML classifier for this data set (Figure 4). In addition, it is more easily interpreted by the medical audience. To further improve the model’s performance and overcome the potential overfitting



risk caused by the large number of features, we added penalty items into the logistic model and used a grid search to find the optimal type of penalty (LASSO) and the hyperparameters used in the penalty term. The model’s performance was substantially improved using LASSO regularization.

Based on the final prediction model, multivariate logistic regression recognized the most significant risk factors as follows: aortic calcification (OR 2.77, 95% CI 1.32-5.81), celiac trunk calcification (OR 2.79, 95% CI 1.20-6.48), FEV1% (OR 0.51, 95% CI 0.30-0.89); TLco (OR 0.56, 95% CI 0.27-1.18), peripheral vascular disease (OR 4.97, 95% CI 1.44-.07), laparoscope (OR 3.92, 95% CI 1.23-12.51), postoperative

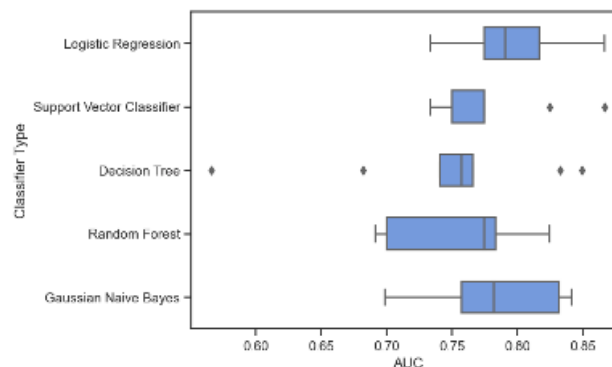
hospital length of stay (OR 1.17, 95% CI 1.13-1.21), vascular permeability activity (OR 0.46, 95% CI 0.14-1.48), and fat liquefaction of incisions (OR 4.36, 95% CI 1.86-10.21).

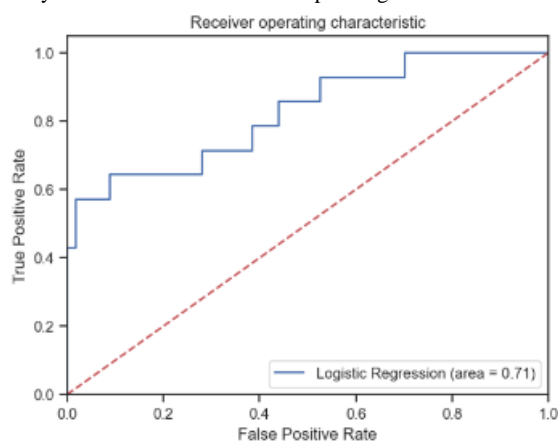
We used the testing data set to validate the model’s predictive ability. The AUROC curve was used to evaluate the model fitting. The logistic regression model with LASSO resulted in a clinically relevant AUROC of 71%, indicating good model performance (Figure 5). We also presented the AUROC, accuracy, recall, precision, and F1 score as extended metrics of both data sets. The AUROC accuracy and precision were consistent between the 2 data sets (Table 4).

**Table 3.** Panel of prediction factors selected in the training data set.

Features	Importance	Remain in model after correlation-based feature selection	Regression coefficients in final model to predict AL	P values in final model to predict AL	Odds ratio for AL (95% CI)
<b>Preoperative factors</b>					
Aortic calcification	29.8	Yes	1.0203	0.0069	2.77(1.32-5.81)
Celiac trunk calcification	32.4	Yes	1.0275	0.0167	2.79(1.20, 6.48)
Forced vital capacity ratio (FEV1%)	48.5	Yes	-0.6653	0.0177	0.51(0.30-0.89)
Transfer factor for carbon monoxide (TLCO) by single-breath (SB) method (%)	51.8	Yes	-0.5820	0.0111	0.56(0.27-1.18)
Peripheral vascular disease	17.6	Yes	1.6026	0.0109	4.97(1.44-17.07)
<b>Intra-operative and postoperative factors</b>					
Laparoscope	40.3	Yes	1.3665	0.0209	3.92(1.23-12.51)
Postoperative hospital stay	219.6	Yes	0.1571	0.0000	1.17(1.13-1.21)
Tumor vascular permeability (laboratory test)	20.1	Yes	-0.7663	0.0507	0.46(0.14-1.48)
Incision fat liquefaction and infection	23.5	Yes	1.4718	0.0007	4.36(1.86-10.21)

**Figure 4.** Comparison of model’s area under the receiver operating characteristic with different machine learning classifiers.



**Figure 5.** Final model performance presented by the area under the receiver operating characteristic.**Table 4.** Metrics for evaluating the machine learning application.

	Accuracy	Precision	Recall	F1 score	AUROC
Training data set	87%	88%	98%	93%	72%
Testing data set	87%	86%	43%	57%	71%

## Discussion

### Principal Findings

The study provided a panel with 36 features for predicting AL in patients who undergo esophagectomies, including detailed symptoms, surgical technical details, and complications. It can identify the presence of AL with high accuracy (87%) and precision (88%). In addition, our panel of risk factors is supported by the previous randomized controlled trials (RCTs), retrospective cohort studies, and meta-analyses [18-21].

Gaining insight into the risk factors of AL is crucial for designing an evidence-based treatment algorithm that will help guide clinical teams to perform timely AL management and support preoperative and postoperative optimization. Among the most important risk factors for AL development are 4 preoperative comorbidities, laparoscopic esophagectomies, and some postoperative complications. In general, the 4 preoperative comorbidities and most postoperative complications are modifiable factors that may guide patient-centered strategies. Full awareness of preoperative risk factors is essential for identifying high-risk patients and appropriately targeting them to mitigate the severe clinical consequences of AL. For example, if the patient is noticeably concerned about the postoperative AL complications, the clinical team might consider other nonsurgery treatments, such as chemotherapy. Likewise, the postoperative conditions help clinicians actively monitor patients' recovery after esophagectomies, allowing them to identify AL early.

### Preoperative Risk Factors (Patient Screening and Preparation)

The 4 preoperative comorbidities significantly linked to increased AL risk are peripheral vascular disease, aortic calcification, celiac trunk calcification, and COPD indicators (FEV1% and TLco results). In addition, hypertension, diabetes mellitus, and coronary artery disease are independent risk

factors. These preoperative comorbidities all have a negative impact on microvascular perfusion, and corresponding atherosclerosis might affect the etiology of AL [22]. Older esophageal cancer patients, whose nutrition status is often impaired, have a higher rate of atherosclerosis and new-onset atrial fibrillation, making them vulnerable to AL. Moreover, there might be an association between supra-aortic and coronary atherosclerosis and AL, implying that general atherosclerosis scores could predict AL risk. To optimize the preoperative screening of esophageal cancer patients, our study suggests a thorough investigation of atherosclerosis-related risks factors and continuous monitoring of perioperative hemodynamics is essential to prevent AL.

### Intra-operative Risk Factors (Surgical-Technical Aspects)

The prediction model indicates a laparoscopic esophagectomy alone increases the odds of AL by 3.92. However, the results require cautious interpretation. While esophagectomy surgical technique has evolved considerably, the scientific evidence regarding the superiority of specific esophagectomy techniques in reducing morbidity, such as AL, is not robust [23]. Our data collection started when the MIE was just initiated in our cancer center. The majority of our thoracic surgeons exclusively performed open esophagectomies, and they were at the early stages of learning MIE. The increased odds of AL linked to laparoscopic esophagectomies were primarily explained by the proficiency gain curve-associated morbidity since laparoscopic esophagectomies require extensive and adequate training for our thoracic surgeons.

### Postoperative Risk Factors (Medical Outcomes)

Accurate monitoring of postoperative complications has a significant impact on AL development [24]. The adverse medical outcomes associated with AL generally found in this study are the occurrence of fat liquefaction of incisions, reduced vascular permeability, and prolonged lengths of hospital stays. Recent

studies support the impact of our prediction on the outcomes. Kamarajah et al [18] summarized the meta-analysis results from previous studies in this field and confirmed the importance of the pulmonary complications (OR 4.54, 95% CI 2.99-6.89), cardiac complications (OR 2.44, 95% CI 1.77-3.37), and prolonged hospital stays (OR 5.91, 95% CI 1.41-24.97) [18]. Postoperative management should pay attention to any possible incision infections during follow-ups to prevent further development of anastomotic stricture.

### Comparison With Prior Work

One advantage of this study is the effective feature selection. Conventional statistical analyses identified various inconsistent risk factors which were cross-correlated. We approached this challenge by combining univariate and multivariable feature selection techniques to produce a stable panel. Important features were selected, internally cross-validated, and not connected to a specific learning algorithm; therefore, minimal human bias was involved [15].

### Limitations

This study had some limitations. First, because this is a retrospective study that includes consecutive patients, we could not determine the long-term sequelae of AL in the current

database, specifically pathological development. Second, the study was limited to a single center, and the results are therefore representative for the specific geographic region and cannot be generalized. Before extrapolating the model to other facilities, it is necessary to consider other risk factors such as geographical and treatment background. However, our hospital is one of China's top cancer research centers and can collect sufficient surgical esophageal cancer cases. Third, due to the complexity of ML models, substantial computing power is required for practical deployment. However, benefiting from the current development of electronic medical records and embedding automated ML algorithms can enable efficient and expedient risk calculations and substantially improve the convenience of utilizing ML models.

### Conclusions

The ML prediction model of AL provides insight into the important risk factors for designing evidence-based clinical management that will help guide physicians regarding AL prevention and treatment. However, additional prospective data collection is needed using a cohort study design or RCT design in multiple medical settings to confirm our findings' validity and establish a better risk prediction model.

### Acknowledgments

This study was supported by the Beijing Union Medical College's "Central University Basic Research Business Fees" Project (3332018079).

### Authors' Contributions

The authors thank the patients, their families, and the participating study teams for making this study possible. ZZ, XS, and LZ contributed to the study's conception and design. ZZ and XC contributed to the drafting of the manuscript. ZZ, XC, LZ, and XS contributed to the critical revision of the manuscript for intellectual content. SM, HF, and LZ contributed to the data acquisition and interpretation. ZZ contributed to the data analysis and interpretation. XC contributed to the model's development and validation. SM and HF contributed technical or material support. LZ obtained funding and provided study supervision. All authors read and approved the final manuscript.

### Conflicts of Interest

None declared.

Multimedia Appendix 1

Risk Factors Panel.

[[DOCX File, 17 KB](#) - [medinform\\_v9i7e27110\\_app1.docx](#) ]

Multimedia Appendix 2

TRIPOD Checklist.

[[PDF File \(Adobe PDF File\), 754 KB](#) - [medinform\\_v9i7e27110\\_app2.pdf](#) ]

### References

1. Verstegen MHP, Bouwense SAW, van Workum F, Ten Broek R, Siersema PD, Rovers M, et al. Management of intrathoracic and cervical anastomotic leakage after esophagectomy for esophageal cancer: a systematic review. *World J Emerg Surg* 2019;14:17-18 [[FREE Full text](#)] [doi: [10.1186/s13017-019-0235-4](https://doi.org/10.1186/s13017-019-0235-4)] [Medline: [30988695](#)]
2. Kassis ES, Kosinski AS, Ross P, Koppes KE, Donahue JM, Daniel VC. Predictors of anastomotic leak after esophagectomy: an analysis of the society of thoracic surgeons general thoracic database. *Ann Thorac Surg* 2013 Dec;96(6):1919-1926. [doi: [10.1016/j.athoracsur.2013.07.119](https://doi.org/10.1016/j.athoracsur.2013.07.119)] [Medline: [24075499](#)]

3. Li H, Wang D, Wei W, Ouyang L, Lou N. The predictive value of coefficient of PCT × BG for anastomotic leak in esophageal carcinoma patients with ARDS after esophagectomy. *J Intensive Care Med* 2019 Jul 10;34(7):572-577. [doi: [10.1177/0885066617705108](https://doi.org/10.1177/0885066617705108)] [Medline: [28486866](https://pubmed.ncbi.nlm.nih.gov/28486866/)]
4. Sun Z, Du H, Li J, Qin H. Constructing a risk prediction model for anastomotic leakage after esophageal cancer resection. *J Int Med Res* 2020 Apr 08;48(4):300060519896726 [FREE Full text] [doi: [10.1177/0300060519896726](https://doi.org/10.1177/0300060519896726)] [Medline: [32268818](https://pubmed.ncbi.nlm.nih.gov/32268818/)]
5. Chen JH, Asch SM. Machine learning and prediction in medicine - beyond the peak of inflated expectations. *N Engl J Med* 2017 Jun 29;376(26):2507-2509 [FREE Full text] [doi: [10.1056/NEJMp1702071](https://doi.org/10.1056/NEJMp1702071)] [Medline: [28657867](https://pubmed.ncbi.nlm.nih.gov/28657867/)]
6. Kruppa J, Ziegler A, König IR. Risk estimation and risk prediction using machine-learning methods. *Hum Genet* 2012 Oct 3;131(10):1639-1654 [FREE Full text] [doi: [10.1007/s00439-012-1194-y](https://doi.org/10.1007/s00439-012-1194-y)] [Medline: [22752090](https://pubmed.ncbi.nlm.nih.gov/22752090/)]
7. Rahbari NN, Weitz J, Hohenberger W, Heald RJ, Moran B, Ulrich A, et al. Definition and grading of anastomotic leakage following anterior resection of the rectum: a proposal by the International Study Group of Rectal Cancer. *Surgery* 2010 Mar;147(3):339-351. [doi: [10.1016/j.surg.2009.10.012](https://doi.org/10.1016/j.surg.2009.10.012)] [Medline: [20004450](https://pubmed.ncbi.nlm.nih.gov/20004450/)]
8. Low DE, Alderson D, Cecconello I, Chang AC, Darling GE, D Journo XB, et al. International consensus on standardization of data collection for complications associated with esophagectomy: Esophagectomy Complications Consensus Group (ECCG). *Ann Surg* 2015 Aug;262(2):286-294. [doi: [10.1097/SLA.0000000000001098](https://doi.org/10.1097/SLA.0000000000001098)] [Medline: [25607756](https://pubmed.ncbi.nlm.nih.gov/25607756/)]
9. Boone J, Rinkes IB, van Leeuwen M, van Hillegersberg R. Diagnostic value of routine aqueous contrast swallow examination after oesophagectomy for detecting leakage of the cervical oesophago-gastric anastomosis. *ANZ J Surg* 2008 Sep;78(9):784-790. [doi: [10.1111/j.1445-2197.2008.04650.x](https://doi.org/10.1111/j.1445-2197.2008.04650.x)] [Medline: [18844909](https://pubmed.ncbi.nlm.nih.gov/18844909/)]
10. Wei Q, Dunbrack RL. The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS One* 2013 Jul 9;8(7):e67863 [FREE Full text] [doi: [10.1371/journal.pone.0067863](https://doi.org/10.1371/journal.pone.0067863)] [Medline: [23874456](https://pubmed.ncbi.nlm.nih.gov/23874456/)]
11. Abdulrauf Sharifai G, Zainol Z. Feature selection for high-dimensional and imbalanced biomedical data based on robust correlation based redundancy and binary grasshopper optimization algorithm. *Genes (Basel)* 2020 Jun 27;11(7):717 [FREE Full text] [doi: [10.3390/genes11070717](https://doi.org/10.3390/genes11070717)] [Medline: [32605144](https://pubmed.ncbi.nlm.nih.gov/32605144/)]
12. Ebrahimi M, Mohammadi-Dehcheshmeh M, Ebrahimi E, Petrovski KR. Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Trees outperform other models. *Comput Biol Med* 2019 Nov;114:103456. [doi: [10.1016/j.combiomed.2019.103456](https://doi.org/10.1016/j.combiomed.2019.103456)] [Medline: [31605926](https://pubmed.ncbi.nlm.nih.gov/31605926/)]
13. Saeyns Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007 Oct 01;23(19):2507-2517. [doi: [10.1093/bioinformatics/btm344](https://doi.org/10.1093/bioinformatics/btm344)] [Medline: [17720704](https://pubmed.ncbi.nlm.nih.gov/17720704/)]
14. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Ann Intern Med* 2015 May 19;162(10):735. [doi: [10.7326/115-5093-2](https://doi.org/10.7326/115-5093-2)]
15. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *The Lancet* 2019 Apr;393(10181):1577-1579. [doi: [10.1016/s0140-6736\(19\)30037-6](https://doi.org/10.1016/s0140-6736(19)30037-6)]
16. Shang L. A Feature Selection Method Based on Information Gain and Genetic Algorithm. 2012 Apr 23 Presented at: 2012 International Conference on Computer Science and Electronics Engineering; 23-25 March 2012; Hangzhou, China p. 355-358 URL: <https://ieeexplore.ieee.org/document/6188038>
17. Rangarajan L. Effective feature selection for classification of promoter sequences. *PLoS One* 2016 Dec 15;11(12):e0167165 [FREE Full text] [doi: [10.1371/journal.pone.0167165](https://doi.org/10.1371/journal.pone.0167165)] [Medline: [27978541](https://pubmed.ncbi.nlm.nih.gov/27978541/)]
18. Kamarajah S, Lin A, Tharmaraja T, Bharwada Y, Bundred JR, Nepogodiev D, et al. Risk factors and outcomes associated with anastomotic leaks following esophagectomy: a systematic review and meta-analysis. *Dis Esophagus* 2020 Mar 16;33(3):14-17. [doi: [10.1093/dote/doz089](https://doi.org/10.1093/dote/doz089)] [Medline: [31957798](https://pubmed.ncbi.nlm.nih.gov/31957798/)]
19. Aoyama T, Atsumi Y, Hara K, Tamagawa H, Tamagawa A, Komori K, et al. Risk factors for postoperative anastomosis leak after esophagectomy for esophageal cancer. *In Vivo* 2020 Feb 28;34(2):857-862 [FREE Full text] [doi: [10.21873/invivo.11849](https://doi.org/10.21873/invivo.11849)] [Medline: [32111795](https://pubmed.ncbi.nlm.nih.gov/32111795/)]
20. de Mooij CM, Maassen van den Brink M, Merry A, Tweed T, Stoot J. Systematic review of the role of biomarkers in predicting anastomotic leakage following gastroesophageal cancer surgery. *J Clin Med* 2019 Nov 17;8(11):2005 [FREE Full text] [doi: [10.3390/jcm8112005](https://doi.org/10.3390/jcm8112005)] [Medline: [31744186](https://pubmed.ncbi.nlm.nih.gov/31744186/)]
21. Ubels S, Versteegen MHP, Rosman C, Reynolds JV. Anastomotic leakage after esophagectomy for esophageal cancer: risk factors and operative treatment. *Ann Esophagus* 2021 Mar;4:8-8. [doi: [10.21037/aoe-2020-18](https://doi.org/10.21037/aoe-2020-18)]
22. Zhao L, Zhao G, Li J, Qu B, Shi S, Feng X, et al. Calcification of arteries supplying the gastric tube increases the risk of anastomotic leakage after esophagectomy with cervical anastomosis. *J Thorac Dis* 2016 Dec;8(12):3551-3562 [FREE Full text] [doi: [10.21037/jtd.2016.12.62](https://doi.org/10.21037/jtd.2016.12.62)] [Medline: [28149549](https://pubmed.ncbi.nlm.nih.gov/28149549/)]
23. Zhou C, Ma G, Li X, Li J, Yan Y, Liu P, et al. Is minimally invasive esophagectomy effective for preventing anastomotic leakages after esophagectomy for cancer? A systematic review and meta-analysis. *World J Surg Oncol* 2015 Sep 04;13(1):269 [FREE Full text] [doi: [10.1186/s12957-015-0661-z](https://doi.org/10.1186/s12957-015-0661-z)] [Medline: [26338060](https://pubmed.ncbi.nlm.nih.gov/26338060/)]
24. Versteegen MHP, Bouwense SAW, van Workum F, Ten Broek R, Siersema PD, Rovers M, et al. Management of intrathoracic and cervical anastomotic leakage after esophagectomy for esophageal cancer: a systematic review. *World J Emerg Surg* 2019 Apr 4;14(1):17 [FREE Full text] [doi: [10.1186/s13017-019-0235-4](https://doi.org/10.1186/s13017-019-0235-4)] [Medline: [30988695](https://pubmed.ncbi.nlm.nih.gov/30988695/)]

## Abbreviations

**AL:** anastomotic leakage  
**ASA:** The American Society of Anesthesiologists  
**AUROC:** area under the receiver operating characteristic curve  
**COPD:** chronic obstructive pulmonary disease  
**FEV:** forced expiratory volume  
**GBM:** gradient boosting model  
**LASSO:** least absolute shrinkage and selection operator  
**ML:** machine learning  
**MIE:** minimally invasive esophagectomy  
**OR:** odds ratio  
**RCT:** randomized controlled trials  
**TLco:** transfer factor for carbon monoxide  
**TRIPOD:** Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

*Edited by G Eysenbach; submitted 14.01.21; peer-reviewed by Y Yoshida, B Zhao, T Tillmann; comments to author 05.02.21; revised version received 10.04.21; accepted 16.06.21; published 27.07.21.*

*Please cite as:*

*Zhao Z, Cheng X, Sun X, Ma S, Feng H, Zhao L*

*Prediction Model of Anastomotic Leakage Among Esophageal Cancer Patients After Receiving an Esophagectomy: Machine Learning Approach*

*JMIR Med Inform 2021;9(7):e27110*

*URL: <https://medinform.jmir.org/2021/7/e27110>*

*doi: [10.2196/27110](https://doi.org/10.2196/27110)*

*PMID: [34313597](https://pubmed.ncbi.nlm.nih.gov/34313597/)*

©Ziran Zhao, Xi Cheng, Xiao Sun, Shanrui Ma, Hao Feng, Liang Zhao. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 27.07.2021. 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

# Effects of Background Colors, Flashes, and Exposure Values on the Accuracy of a Smartphone-Based Pill Recognition System Using a Deep Convolutional Neural Network: Deep Learning and Experimental Approach

KyeongMin Cha<sup>1</sup>, MSc; Hyun-Ki Woo<sup>1,2</sup>, MSc; Dohyun Park<sup>1</sup>, MSc; Dong Kyung Chang<sup>1,3</sup>, MD, PhD; Mira Kang<sup>1,4</sup>, MD, PhD

<sup>1</sup>Department of Digital Health, Samsung Advanced Institute of Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea

<sup>2</sup>EvidNet Inc, Seongnam-si, Gyeonggi-do, Republic of Korea

<sup>3</sup>Division of Gastroenterology, Department of Internal Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

<sup>4</sup>Center for Health Promotion, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

**Corresponding Author:**

Mira Kang, MD, PhD

Department of Digital Health

Samsung Advanced Institute of Health Sciences & Technology

Sungkyunkwan University

81 Irwon-ro, Gangnam-gu

Seoul, 06351

Republic of Korea

Phone: 82 01099336838

Fax: 82 0234101000

Email: [kang.mirad@gmail.com](mailto:kang.mirad@gmail.com)

## Abstract

**Background:** Pill image recognition systems are difficult to develop due to differences in pill color, which are influenced by external factors such as the illumination from and the presence of a flash.

**Objective:** In this study, the differences in color between reference images and real-world images were measured to determine the accuracy of a pill recognition system under 12 real-world conditions (ie, different background colors, the presence and absence of a flash, and different exposure values [EVs]).

**Methods:** We analyzed 19 medications with different features (ie, different colors, shapes, and dosages). The average color difference was calculated based on the color distance between a reference image and a real-world image.

**Results:** For images with black backgrounds, as the EV decreased, the top-1 and top-5 accuracies increased independently of the presence of a flash. The top-5 accuracy for images with black backgrounds increased from 26.8% to 72.6% when the flash was on and increased from 29.5% to 76.8% when the flash was off as the EV decreased. However, the top-5 accuracy increased from 62.1% to 78.4% for images with white backgrounds when the flash was on. The best top-1 accuracy was 51.1% (white background; flash on; EV of +2.0). The best top-5 accuracy was 78.4% (white background; flash on; EV of 0).

**Conclusions:** The accuracy generally increased as the color difference decreased, except for images with black backgrounds and an EV of -2.0. This study revealed that background colors, the presence of a flash, and EVs in real-world conditions are important factors that affect the performance of a pill recognition model.

(*JMIR Med Inform* 2021;9(7):e26000) doi:[10.2196/26000](https://doi.org/10.2196/26000)

**KEYWORDS**

pill recognition; deep neural network; image processing; color space; color difference; pharmaceutical; imaging; photography; neural network; mobile phone

## Introduction

Recently, smartphone cameras have been used to not only take photos but also recognize objects via models with enhanced performance and artificial intelligence models [1,2]. The study of photo recognition is not only limited to a person or a thing, such as a car; it can even extend to analyzing a person's hair color or specifying the color of an object, such as a red car [3].

Many researchers are exploring new algorithms related to color in the field of image learning. For example, gray-scale images can be colored automatically by using a convolutional neural network (CNN) through a new method [4]. Additionally, Lunit—a well-known medical artificial intelligence

company—presented an algorithm that enhances the color of an image as if it was corrected by a professional [5].

Color is an important component that is used to recognize objects, especially pharmaceuticals. The United States Federal Drug Administration approves solid pharmaceuticals and pills, which have physical identifiers. Each pill should have its own unique physical features, that is, unique shapes, sizes, colors, and imprints (the letter or number carved onto the medicine), which need to be approved [6]. However, in some cases, all features of medicines, except for the color, can be the same [7]. For instance, Amaryl (glimepiride)—an oral pill for controlling the blood sugar levels of patients with diabetes—has identical physical features across all 1-mg, 2-mg, and 4-mg dosages except for their colors (Figure 1).

**Figure 1.** Examples of pills with the same physical features (except for color).



In 2016, the National Institutes of Health hosted a competition to promote the easy recognition of unknown medications. Even though the competition used reference images that were photographed in a professionally supervised setting, the accuracy of drug recognition was not very high. Since the quality of a picture taken by a smartphone can be greatly influenced by illumination (lighting), shading, and background color, it is difficult to develop a system for image recognition [8]. Pill colors are especially affected by lighting hues and fluorescent light (Figure 2). In addition, there are no quantitative analyses for determining how a pill recognition system can be affected by external factors [6,9]. The most recent work related to drug

recognition studies that involve deep learning has been conducted on wearable smart glasses developed for patients with visual impairment. Additionally, drug detection has been enhanced with feature pyramid networks and CNNs. However, despite recent improvements in pill recognition via a model approach, the effects of environmental factors have not been analyzed [10,11].

In this study, we sought to determine the accuracy of a pill recognition system under 12 different real-world conditions (ie, different background colors, the presence and absence of a flash, and different exposure values [EVs]).

**Figure 2.** Effects that external environments (fluorescent lighting) have on the colors of pills in images. A: Flash is on. B: Flash is off.



## Methods

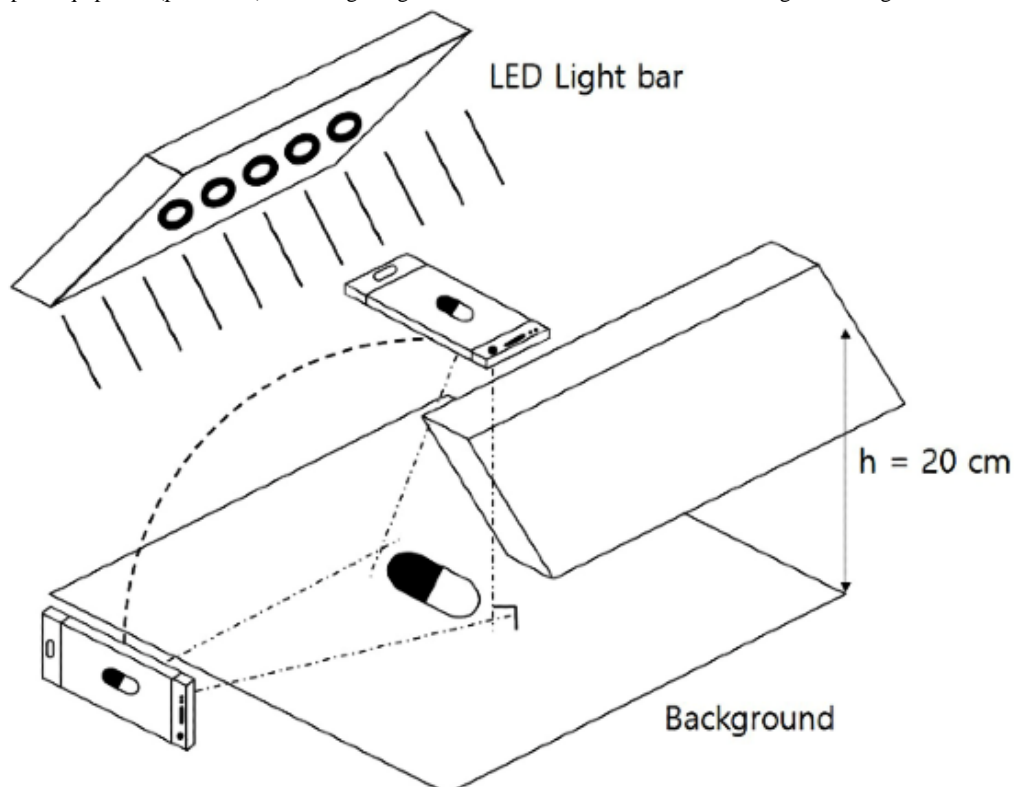
### Photo Shooting Equipment and Image Preprocessing

#### Data Acquisition Process for Reference Images

The smartphone used in this study was the Samsung Galaxy S7 Edge, which was equipped with a dual-pixel 12.0-megapixel

front camera with an aperture of  $f/1.7$ . An already intact camera app and the autofocus feature of the smartphone software were used. For lighting, 2 light-emitting diode panels were used. The flash was positioned at a height of 20 cm, and the intensity of illumination was set to 1145 lux. The background color was black, and the flash was turned off (Figure 3).

**Figure 3.** Photographic equipment (photo box) for taking images under the reference condition. LED: light-emitting diode.



### Data Acquisition Process for Real-World Images

The photos were taken under 12 conditions that involved

different background colors (black or white), the presence or absence of a flash, and 3 different EVs (+2.0, 0, and -2.0; [Table 1](#)).

**Table 1.** Real-world image sets for the 12 conditions.

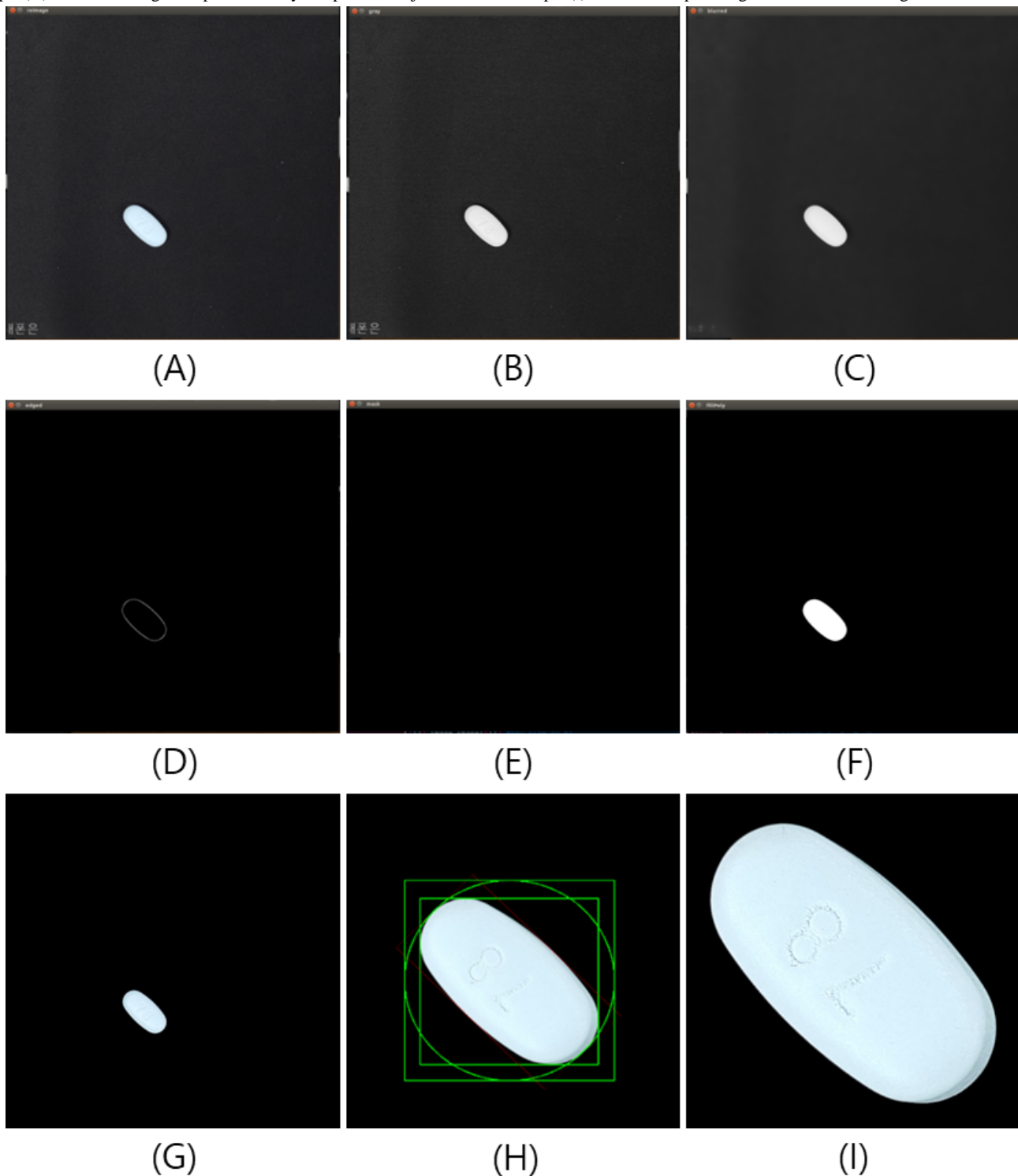
Image set name	Condition		
	Background color	Flash	Exposure value
B_O_EV-2.0	Black	On	-2.0
B_O_EV0	Black	On	0
B_O_EV+2.0	Black	On	+2.0
W_O_EV-2.0	White	On	-2.0
W_O_EV0	White	On	0
W_O_EV+2.0	White	On	+2.0
B_X_EV-2.0	Black	Off	-2.0
B_X_EV0	Black	Off	0
B_X_EV+2.0	Black	Off	+2.0
W_X_EV-2.0	White	Off	-2.0
W_X_EV0	White	Off	0
W_X_EV+2.0	White	Off	+2.0

### Image Preprocessing

[Figure 4](#) shows the 9 steps for processing images of the region of interest (ROI). This process was conducted to improve deep neural network-based image recognition accuracy by eliminating image noise and improving the quality of the picture [12]. Python 3.5.3 and the OpenCV 3.2 library were used to process each image [13]. The photos were converted to

gray-scale images and blurred to reduce image noise. Afterward, we experimented with applying the different threshold options of the OpenCV library to each pill image. The Canny edge detector algorithm was used to define the ROI (a drug's edge) [14,15]. Next, the processed picture was combined with the original picture, and all other areas except for the pill were omitted. Finally, the inner edge of the pill image was set within a square-shaped boundary, and this image was saved.

**Figure 4.** Image preprocessing algorithm for extracting an object. Step 1 (A): take pictures using a smartphone. Step 2 (B): convert image to a gray-scale image. Step 3 (C): use the blur and threshold options to process the image. Step 4 (D): Canny edge detection. Step 5 (E): create a black background. Step 6 (F): use the FillPoly function to process the image. Step 7 (G): use the bitwise operation to combine the original image with the processed image. Step 8 (H): draw a rectangle-shaped boundary and perform object extraction. Step 9 (I): use the final pill image as the reference image to train the model.










### Test Drug Type

A total of 19 different types of pills were used in this study. The different features of the pills (7 colors, 7 shapes, and 7 types)

can be seen in [Table 2](#). [Figure 5](#) shows all of the example images of the pills; the numbers on the upper left-hand corners were the labels used in the deep learning process.



**Table 2.** Characteristics of the reference set (shape, color, and dosage form).

Characteristic	Instances, n
<b>Shape</b>	
	6
	3
	4
	1
	2
	2
	1
<b>Color</b>	
Pink	1
Blue	5
White	5
Yellow	4
Green	1
Yellow-green	2
Orange	1
<b>Dosage form</b>	
Film-coated tablet	10
Sugar-coated tablet	2
Uncoated tablet	6
Hard capsule	1

**Figure 5.** Sample images of the pills used in the experiment. Yellow pills include pills 6, 12, 13, 14, and 16. Green pills include pills 5, 7, and 11. "Other" pills include the rest of the pills.

## Color Difference

### Factors Affecting the Color

Figure 6 shows the pill images that were taken under the

reference condition and under the 12 real-world conditions. The colors of the pills differed based on the background colors, the presence of a flash, and EVs.

**Figure 6.** Representative examples of pill images. A: reference condition. B: 12 real-world conditions (image sets: B\_O\_EV-2.0, B\_O\_EV0, B\_O\_EV+2.0, W\_O\_EV-2.0, W\_O\_EV0, W\_O\_EV+2.0, B\_X\_EV-2.0, B\_X\_EV0, B\_X\_EV+2.0, W\_X\_EV-2.0, W\_X\_EV0, and W\_X\_EV+2.0).



### Color Space

The spatial color concept, which is expressed as a 3D chart, was used to calculate the differences in color quantitatively. The Commission Internationale de l'Eclairage (CIE) L\*a\*b\* color space is a spatial color chart that is used worldwide to represent colors that can be detected by the human eye. After the red, green, and blue (RGB) color space is converted to a CIE XYZ color space, it is then converted to a CIE L\*a\*b\* color space that separates the lighting and the color [16]. The CIE and CIE 1976 L\*a\*b\* include some colors that human eyes cannot detect. L\* represents brightness with values that range from 0 to 100. Parameters a\* (green to red) and b\* (blue to yellow) range from -120 to 120 [17,18].

To quantify the color of the ROI, the process shown in Figure 7 was followed. By using the RGB analysis plugin of the ImageJ 1.52 program (National Institutes of Health), the RGB color space was changed to an XYZ color space [19] via the following equations:

$$X = 0.4303R + 0.3416G + 0.1784B \tag{1}$$

$$Y = 0.2219R + 0.7068G + 0.0713B \tag{2}$$

$$Z = 0.0202R + 0.1296G + 0.9393B \tag{3}$$

The XYZ color space was then converted to an L\*a\*b\* color space, as follows:

$$L^* = 116f(Y/Y_n) - 16 \tag{4}$$

$$a^* = 500(f[X/X_n] - f[Y/Y_n]) \tag{5}$$

$$b^* = 200(f[Y/Y_n] - f[Z/Z_n]) \tag{6}$$



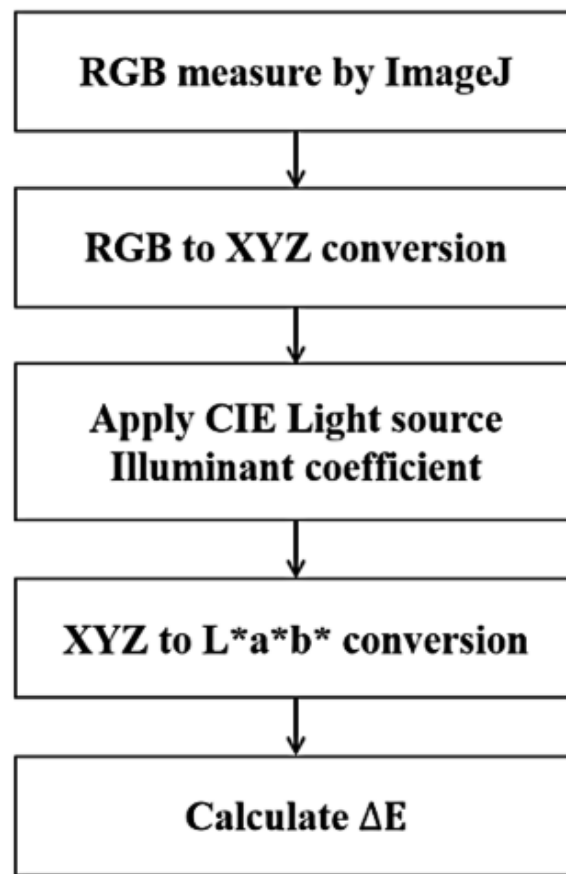
$$f(q) = 7.787q + (16/116) \quad (q \leq 0.008856) \tag{8}$$

After computing the values of L\*, a\*, and b\*, ΔE was calculated with the following equation, where ΔE is the color difference:



The color differences were calculated by subtracting the color distances in images taken under the real-world conditions from the color distances in images taken under the reference condition. The color distance of 19 medications was presented as means with SDs. A three-way repeated measures analysis of variance (ANOVA), which was followed by a Bonferroni posthoc test, was used to examine the effects that background color (black vs white), the presence of a flash (flash on vs flash off), and EV (+2.0, 0, and -2.0) had on color differences. A P value of <.05 was considered to be statistically significant. The statistical analysis was performed by using R software, version 3.6.2 (The R Foundation).

**Figure 7.** Color space conversion process. The conversion of RGB to CIE L\*a\*b\* involves equations 1-8.  $\Delta E$  is calculated by using equation 9. CIE: Commission Internationale de l'Éclairage; RGB: red, green, and blue.



### Model Learning Process

A total of 34,000 images were taken manually by using a smartphone. We used images without augmentation. The number of images in the training set was 19,000, and the number of images in the validation set was 5000. We used 5000 images for the tests conducted under the reference condition and 5000 images for the tests conducted under real-world conditions.

The model architecture used in this study was a CNN that used a deep learning algorithm (GoogLeNet) with 22 layers and 9 inception modules [20]. We used the NVIDIA Deep Learning Graphics Processing Unit Training System (DIGITS) for the learning framework [21]. In this framework, top-1 accuracy refers to the extent to which a model's answer exactly matches the expected answer. Top-5 accuracy refers to the extent to which the five highest model answers match the expected answer. Accuracy refers to the number of correct predictions divided by the total number of predictions. Loss refers to the penalty for a bad prediction. GoogLeNet has two auxiliary classifiers for combating the vanishing gradient problem. Loss1 is the first auxiliary classifier's output, and Loss2 is the second auxiliary classifier's output [20].

### Results

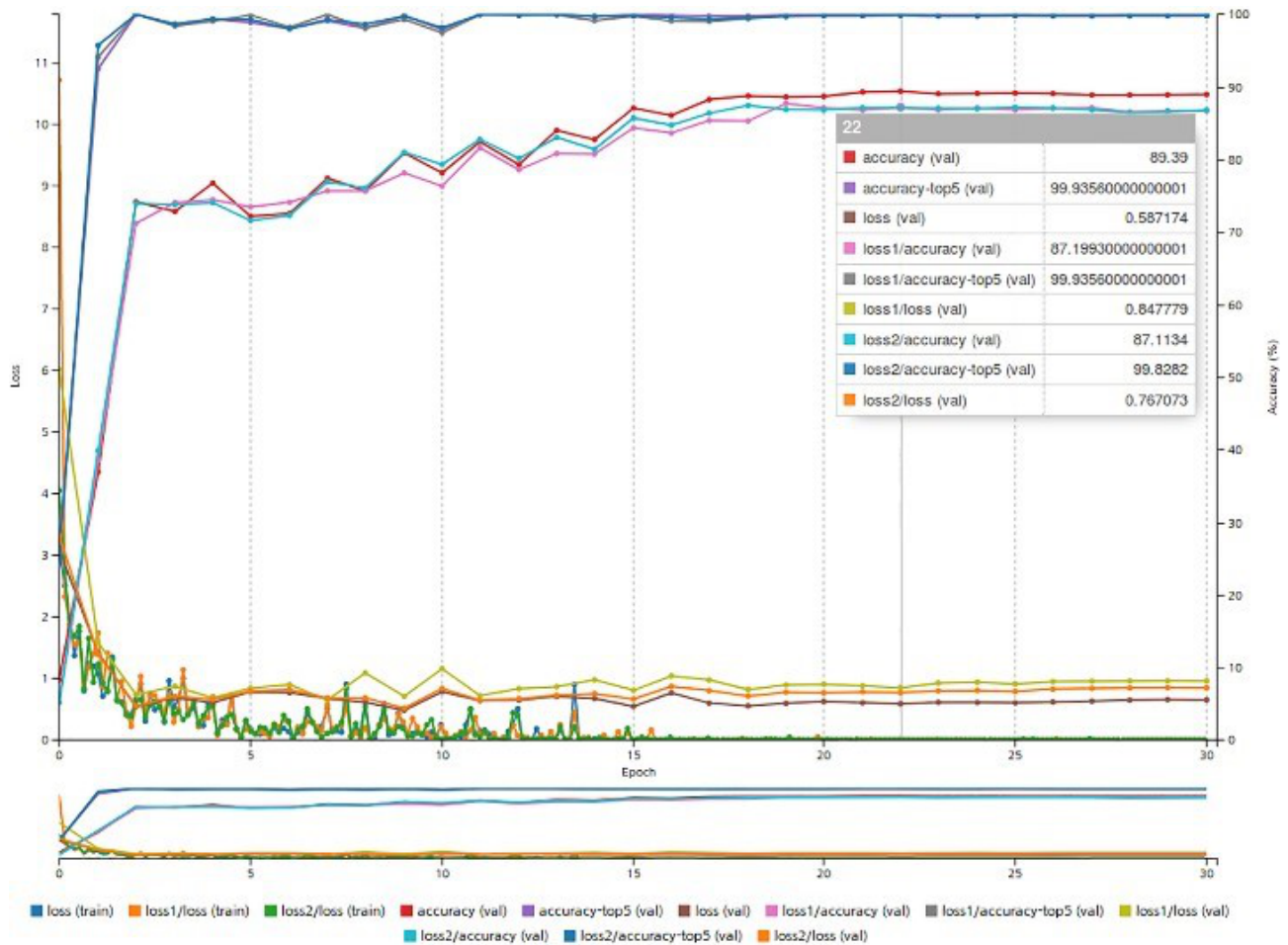
Figure 8 shows the results of model training via the DIGITS framework. Our model recognized the correct pill with a top-1 accuracy of 84.54% and a top-5 accuracy of 99.89% for the reference test image set. Figure 9 shows the top-1 and top-5

accuracies and the average color differences for images taken under the 12 real-world conditions. For images with black backgrounds, as the EV decreased, the top-1 and top-5 accuracies increased independently of the presence of a flash. The top-5 accuracy for images with black backgrounds increased from 26.8% to 72.6% when the flash was on and increased from 29.5% to 76.8% when the flash was off as the EV decreased. However, the top-5 accuracy increased from 62.1% to 78.4% for images with white backgrounds when the flash was on. The best top-1 accuracy was 51.1% (white background; flash on; EV of +2.0). The best top-5 accuracy was 78.4% (white background; flash on; EV of 0). The results of the repeated measures ANOVA and the Bonferroni posthoc test for over 19 medications, as displayed in Figure 9, were used to assess the variances in color differences. Color differences based on EV values varied significantly (all  $P$  values in the repeated measures ANOVA were  $<.05$ ). The results of the repeated measures ANOVA for color differences among 19 medications are as follows:  $P=.02$  (black background and flash on);  $P=.02$  (black background and flash off);  $P<.001$  (white background and flash on); and  $P<.001$  (white background and flash off). With regard to the Bonferroni posthoc test results, for images with white backgrounds that were taken with the flash turned on or off, all  $P$  values were  $<.001$  between the image groups with different EVs. Color differences among images with black backgrounds that were taken with the flash turned on were statistically different between the EV +2.0 and EV 0 groups ( $P=.004$ ) and between the EV 0 and EV -2.0 groups ( $P=.004$ ). Color differences among images with black backgrounds that were

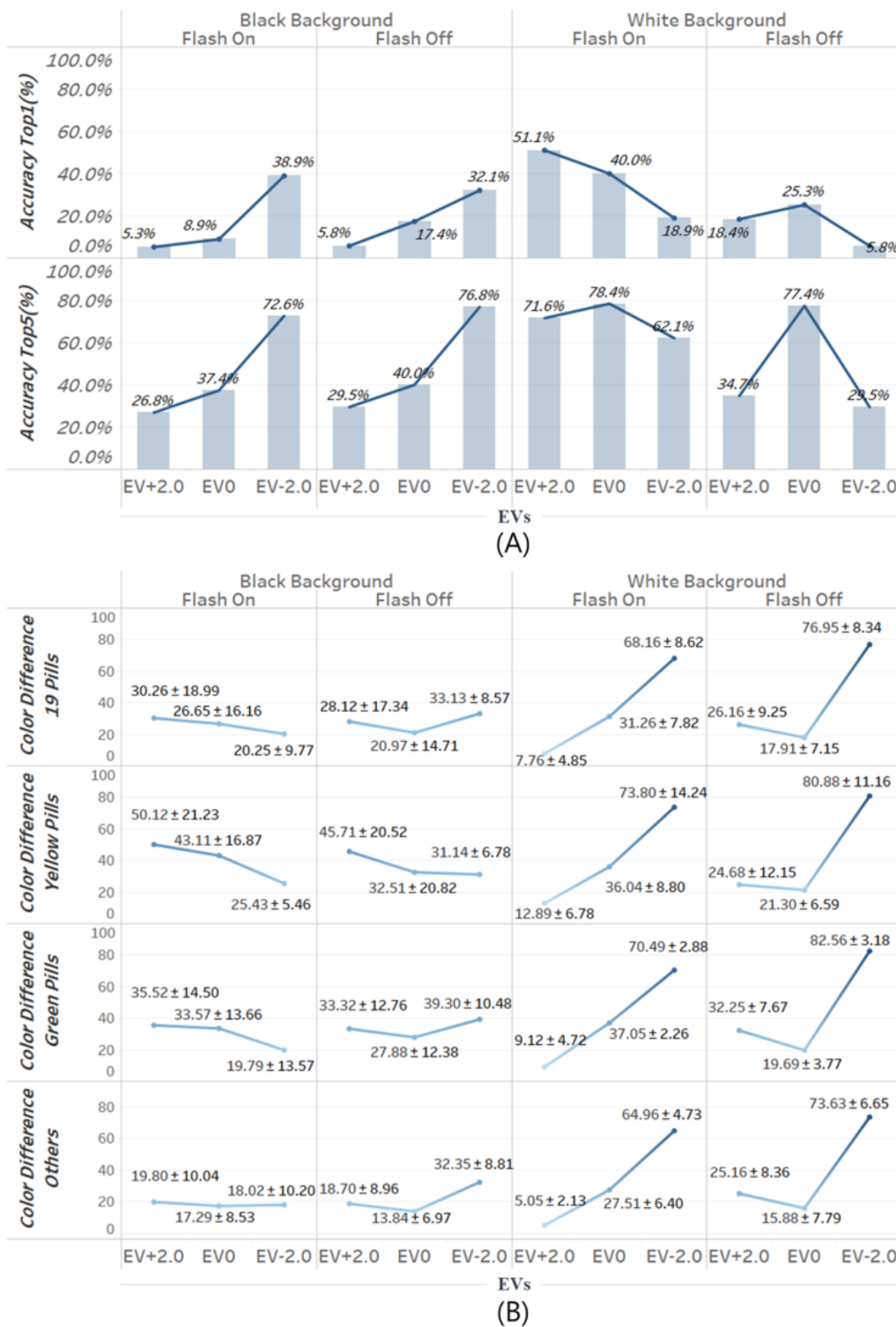
taken with the flash turned off were significantly different between the EV +2.0 and EV 0 groups ( $P=.005$ ) and between the EV 0 and EV -2.0 groups ( $P=.03$ ). When excluding the conditions of black backgrounds and an EV of -2.0, the accuracy generally increased as the color difference decreased. When the 19 medications were sorted into 3 groups by pill color (ie, yellow, green, and other), the color differences among the

color subgroups were not dependent on the colors of pills in images with white backgrounds. However, the color differences among the color subgroups were dependent on the colors of pills in images with black backgrounds. The pill color, as well as environmental factors such as the background color, the presence of a flash, and EVs, can affect the accuracy of a pill recognition system (Figure 9).

**Figure 8.** Model learning results. Top-1 accuracy refers to the extent to which a model’s answer exactly matches the expected answer. Top-5 accuracy refers to the extent to which the five highest model answers match the expected answer. Accuracy refers to the number of correct predictions divided by the total number of predictions. “(train)” refers to the training process and “(val)” refers to the validation process. Loss refers to the penalty for a bad prediction. Loss1 and Loss2 are two auxiliary classifiers of GoogLeNet.



**Figure 9.** A: Comparison of top-1 and top-5 accuracies. B: Comparison of color differences based on background color, the presence of a flash, and EV. Color differences are presented as means with SDs. EV: exposure value.



## Discussion

The National Library of Medicine Pill Image Recognition Challenge was hosted by the National Institutes of Health in 2016. The three winners obtained a mean average precision of 0.27, 0.09, and 0.08. Their top-5 accuracy values were 43%, 12%, and 11% for 5000 query and consumer images. Although the competition can be seen as a promising initial step for pill

identification, solid medication recognition systems are still in the difficult process of development. The reason for this seems to be that the quality of real-world images tends to be affected by illumination, shading, background color, or shooting direction, unlike reference images.

In our study, it was shown that smaller color differences yielded higher recognition accuracy except for images with black backgrounds and images with an EV of -2.0. In other words,



the accuracy of pill recognition is generally inversely proportional to color difference. These exceptions may have been due to the following: (1) it is believed that the Euclidean distance between two colors may not be proportional to the precise color difference; and (2) other factors, such as pill imprints, shapes, and colors, can influence the recognition rate.

Color differences are a crucial problem, especially for pill recognition systems. In previous studies, a few methods were suggested for enhancing pill recognition. MedSnap (MedSnap LLC) is a smartphone-based pill identification system that uses an adaptive color correction algorithm. However, despite the fact that it corrects for color differences, this system has a disadvantage; it has to use a controlled surface to improve its pill recognition rate [22]. In a study on a deep learning model for dermatology, the authors recommended retaking the photo if it is of poor quality due to brightness or noise levels. Thus,

adjusting the camera settings to match the optimized settings for a photo can yield better quality photos and improve the accuracy of medicine recognition systems [23,24]. Furthermore, the enhancement of drug detection via a model approach for minimizing color differences is warranted in the future.

This study reveals that background colors, the presence of a flash, and EVs in real-world conditions are important factors that affect the performance of pill recognition models. Depending on certain image conditions, pill colors can also affect pill recognition accuracy. However, this factor may not affect accuracy as much as environmental factors [25]. Further study is warranted on other factors, such as photography angles and heights, pill shapes, background colors, tablet and capsule conditions, and smartphone models that affect color differences and pill recognition accuracy [26-28].

## Acknowledgments

This work was supported by a National IT Industry Promotion Agency grant funded by the Ministry of Science and ICT and Ministry of Health and Welfare (project number: S1906-21-1001; Development Project of The Precision Medicine Hospital Information System). This work was also supported by the Technology Innovation Program (program 20005021: Establishment of Standardization and Anonymization Guidelines Based on a Common Data Model; program 20011642: common data model-based algorithm for treatment protocol service system development and spread), which was funded by the Ministry of Trade, Industry & Energy in Korea.

## Conflicts of Interest

None declared.

## References

1. Rivenson Y, Ceylan Koydemir H, Wang H, Wei Z, Ren Z, Günaydın H, et al. Deep learning enhanced mobile-phone microscopy. *ACS Photonics* 2018 Mar 15;5(6):2354-2364. [doi: [10.1021/acsphotonics.8b00146](https://doi.org/10.1021/acsphotonics.8b00146)]
2. Fan B, Zhu L, Du Y, Tang Y. A novel color based object detection and localization algorithm. 2010 Presented at: 2010 3rd International Congress on Image and Signal Processing; October 16-18, 2010; Yantai, China. [doi: [10.1109/CISP.2010.5646875](https://doi.org/10.1109/CISP.2010.5646875)]
3. van de Weijer J, Schmid C, Verbeek J, Larlus D. Learning color names for real-world applications. *IEEE Trans Image Process* 2009 Jul;18(7):1512-1523. [doi: [10.1109/TIP.2009.2019809](https://doi.org/10.1109/TIP.2009.2019809)] [Medline: [19482579](https://pubmed.ncbi.nlm.nih.gov/19482579/)]
4. Iizuka S, Simo-Serra E, Ishikawa H. Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Trans Graph* 2016 Jul 11;35(4):1-11. [doi: [10.1145/2897824.2925974](https://doi.org/10.1145/2897824.2925974)]
5. Park J, Lee JY, Yoo D, Kweon IS. Distort-and-recover: Color enhancement using deep reinforcement learning. arXiv. Preprint posted online on April 16, 2018 [FREE Full text]
6. Yaniv Z, Faruque J, Howe S, Dunn K, Sharlip D, Bond A, et al. The national library of medicine pill image recognition challenge: An initial report. 2017 Aug 17 Presented at: 2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR); October 18-20, 2016; Washington, DC, USA. [doi: [10.1109/AIPR.2016.8010584](https://doi.org/10.1109/AIPR.2016.8010584)]
7. NLM to retire Pillbox on January 29, 2021. National Library of Medicine. URL: <https://pillbox.nlm.nih.gov> [accessed 2021-07-06]
8. Zeng X, Cao K, Zhang M. MobileDeepPill: A small-footprint mobile deep learning system for recognizing unconstrained pill images. 2017 Jun Presented at: MobiSys'17: The 15th Annual International Conference on Mobile Systems, Applications, and Services; June 19-23, 2017; Niagara Falls, New York, USA p. 56-67. [doi: [10.1145/3081333.3081336](https://doi.org/10.1145/3081333.3081336)]
9. Larios Delgado N, Usuyama N, Hall AK, Hazen RJ, Ma M, Sahu S, et al. Fast and accurate medication identification. *NPJ Digit Med* 2019 Feb 28;2:10 [FREE Full text] [doi: [10.1038/s41746-019-0086-0](https://doi.org/10.1038/s41746-019-0086-0)] [Medline: [31304359](https://pubmed.ncbi.nlm.nih.gov/31304359/)]
10. Chang WJ, Chen LB, Hsu CH, Chen JH, Yang TC, Lin CP. MedGlasses: A wearable smart-glasses-based drug pill recognition system using deep learning for visually impaired chronic patients. *IEEE Access* 2020;8:17013-17024. [doi: [10.1109/access.2020.2967400](https://doi.org/10.1109/access.2020.2967400)]
11. Ou YY, Tsai AC, Zhou XP, Wang JF. Automatic drug pills detection based on enhanced feature pyramid network and convolution neural networks. *IET Computer Vision* 2020 Jan 20;14(1):9-17. [doi: [10.1049/iet-cvi.2019.0171](https://doi.org/10.1049/iet-cvi.2019.0171)]

12. Parker JR. Algorithms for Image Processing and Computer Vision. New Jersey, United States: John Wiley & Sons; 2010:1118021886.
13. Bradski G, Kaehler A. OpenCV. Dr Dobb's journal of software tools 2000;3:1-81 [FREE Full text]
14. Cunha A, Adão T, Trigueiros P. HelpmePills: A mobile pill recognition tool for elderly persons. Procedia Technology 2014;16:1523-1532. [doi: [10.1016/j.protcy.2014.10.174](https://doi.org/10.1016/j.protcy.2014.10.174)]
15. Canny J. A computational approach to edge detection. Readings in Computer Vision 1987:184-203. [doi: [10.1016/b978-0-08-051581-6.50024-6](https://doi.org/10.1016/b978-0-08-051581-6.50024-6)]
16. Rachmadi RF, Purnama IKE. Vehicle color recognition using convolutional neural network. arXiv. Preprint posted online on August 15, 2018 [FREE Full text]
17. León K, Mery D, Pedreschi F, León J. Color measurement in L\*a\*b\* units from RGB digital images. Food Res Int 2006 Dec;39(10):1084-1091. [doi: [10.1016/j.foodres.2006.03.006](https://doi.org/10.1016/j.foodres.2006.03.006)]
18. Robertson AR. The CIE 1976 Color-Difference Formulae. Color Res Appl 1977;2(1):7-11. [doi: [10.1002/j.1520-6378.1977.tb00104.x](https://doi.org/10.1002/j.1520-6378.1977.tb00104.x)]
19. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods 2012 Jul;9(7):671-675 [FREE Full text] [doi: [10.1038/nmeth.2089](https://doi.org/10.1038/nmeth.2089)] [Medline: [22930834](https://pubmed.ncbi.nlm.nih.gov/22930834/)]
20. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. 2015 Presented at: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 7-12, 2015; Boston, MA, USA. [doi: [10.1109/cvpr.2015.7298594](https://doi.org/10.1109/cvpr.2015.7298594)]
21. Yeager L, Bernauer J, Gray A, Houston M. DIGITS: the Deep learning GPU Training System. DIGITS; 2015 Presented at: ICML 2015 AutoML Workshop; July 11, 2015; Lille, France.
22. System and method of adaptive color correction for pill recognition in digital images. Google Patents. URL: <https://patentimages.storage.googleapis.com/68/38/45/c944a87f5be101/WO2014070871A1.pdf> [accessed 2021-07-09]
23. Han SS, Park GH, Lim W, Kim MS, Na JI, Park I, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network. PLoS One 2018 Jan 19;13(1):e0191493. [doi: [10.1371/journal.pone.0191493](https://doi.org/10.1371/journal.pone.0191493)] [Medline: [29352285](https://pubmed.ncbi.nlm.nih.gov/29352285/)]
24. Maron RC, Utikal JS, Hekler A, Hauschild A, Sattler E, Sondermann W, et al. Artificial intelligence and its effect on dermatologists' accuracy in dermoscopic melanoma image Classification: Web-Based survey study. J Med Internet Res 2020 Sep 11;22(9):e18091 [FREE Full text] [doi: [10.2196/18091](https://doi.org/10.2196/18091)] [Medline: [32915161](https://pubmed.ncbi.nlm.nih.gov/32915161/)]
25. Wang Z, Peng B, Huang Y, Sun G. Classification for plastic bottles recycling based on image recognition. Waste Manag 2019 Apr 01;88:170-181. [doi: [10.1016/j.wasman.2019.03.032](https://doi.org/10.1016/j.wasman.2019.03.032)] [Medline: [31079629](https://pubmed.ncbi.nlm.nih.gov/31079629/)]
26. Silva J, Rondon C, Cabrera D, Pineda Lezama OB. Influence of lighting and noise on visual color assessment in textiles. IOP Conf Ser Mater Sci Eng 2020 Jun 27;872:012033. [doi: [10.1088/1757-899x/872/1/012033](https://doi.org/10.1088/1757-899x/872/1/012033)]
27. Lee YB, Park U, Jain AK, Lee SW. Pill-ID: Matching and retrieval of drug pill images. Pattern Recognit Lett 2012 May;33(7):904-910. [doi: [10.1016/j.patrec.2011.08.022](https://doi.org/10.1016/j.patrec.2011.08.022)]
28. Chokchaitam S, Sukpornawan P, Pungpiboon N, Tharawut S. RGB compensation based on background shadow subtraction for low-luminance pill recognition. 2019 Presented at: 2019 4th International Conference on Control, Robotics and Cybernetics (CRC); September 27-30, 2019; Tokyo, Japan. [doi: [10.1109/crc.2019.00032](https://doi.org/10.1109/crc.2019.00032)]

## Abbreviations

- ANOVA:** analysis of variance
- CIE:** Commission Internationale de l'Eclairage
- CNN:** convolutional neural network
- DIGITS:** Deep Learning Graphics Processing Unit Training System
- EV:** exposure value
- RGB:** red, green, and blue
- ROI:** region of interest

*Edited by G Eysenbach; submitted 24.11.20; peer-reviewed by V Montmirail, R Krukowski; comments to author 02.12.20; revised version received 04.04.21; accepted 03.06.21; published 28.07.21.*

*Please cite as:*

*Cha K, Woo HK, Park D, Chang DK, Kang M*

*Effects of Background Colors, Flashes, and Exposure Values on the Accuracy of a Smartphone-Based Pill Recognition System Using a Deep Convolutional Neural Network: Deep Learning and Experimental Approach*

*JMIR Med Inform 2021;9(7):e26000*

URL: <https://medinform.jmir.org/2021/7/e26000>

doi: [10.2196/26000](https://doi.org/10.2196/26000)

PMID: [34319239](https://pubmed.ncbi.nlm.nih.gov/34319239/)

©KyeongMin Cha, Hyun-Ki Woo, Dohyun Park, Dong Kyung Chang, Mira Kang. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 28.07.2021. 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

# Experiences With Internet Triage of 9498 Outpatients Daily at the Largest Public Hospital in Taiwan During the COVID-19 Pandemic: Observational Study

Ding-Heng Lu<sup>1</sup>, MD; Chia-An Hsu<sup>1</sup>, MD; Eunice J Yuan<sup>1</sup>, MD; Jun-Jeng Fen<sup>2</sup>, MSc; Chung-Yuan Lee<sup>2</sup>, DPhil; Jin-Lain Ming<sup>3</sup>, DPhil; Tzeng-Ji Chen<sup>4,5,6,7</sup>, MD, PhD; Wui-Chiang Lee<sup>4,6</sup>, MD, DPhil; Shih-Ann Chen<sup>8,9,10</sup>, MD, DPhil

<sup>1</sup>Department of Medical Education, Taipei Veterans General Hospital, Taipei, Taiwan

<sup>2</sup>Information Management Office, Taipei Veterans General Hospital, Taipei, Taiwan

<sup>3</sup>Department of Nursing, Taipei Veterans General Hospital, Taipei, Taiwan

<sup>4</sup>Department of Medical Affairs and Planning, Taipei Veterans General Hospital, Taipei, Taiwan

<sup>5</sup>Big Data Center, Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan

<sup>6</sup>Institute of Hospital and Health Care Administration, School of Medicine, National Yang-Ming University, Taipei, Taiwan

<sup>7</sup>Department of Family Medicine, Taipei Veterans General Hospital, Taipei, Taiwan, ROC, Taipei, Taiwan

<sup>8</sup>Division of Cardiology, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

<sup>9</sup>Institute of Clinical Medicine, Cardiovascular Research Center, National Yang-Ming University, Taipei, Taiwan

<sup>10</sup>Cardiovascular Center, Taichung Veterans General Hospital, Taichung, Taiwan

**Corresponding Author:**

Jun-Jeng Fen, MSc

Information Management Office

Taipei Veterans General Hospital

No 201, Sec 2, Shih-Pai Road

Taipei, 11217

Taiwan

Phone: 886 2 2871 2121

Email: [fenjj@vghtpe.gov.tw](mailto:fenjj@vghtpe.gov.tw)

## Abstract

**Background:** During pandemics, acquiring outpatients' travel, occupation, contact, and cluster histories is one of the most important measures in assessing the disease risk among incoming patients. Previous means of acquiring this information in the examination room have been insufficient in preventing disease spread.

**Objective:** This study aimed to demonstrate the deployment of an automatic system to triage outpatients over the internet.

**Methods:** An automatic system was incorporated in the existing web-based appointment system of the hospital and deployed along with its on-site counterpart. Automatic queries to the virtual private network travel and contact history database with each patient's national ID number were made for each attempt to acquire the patient's travel and contact histories. Patients with relevant histories were denied registration or entry. Text messages were sent to patients without a relevant history for an expedited route of entry if applicable.

**Results:** A total of 127,857 visits were recorded. Among all visits, 91,195 were registered on the internet. In total, 71,816 of them generated text messages for an expedited route of entry. Furthermore, 65 patients had relevant histories, as revealed by the virtual private network database, and were denied registration or entry.

**Conclusions:** An automatic triage system to acquire outpatients' relevant travel and contact histories was deployed rapidly in one of the largest academic medical centers in Taiwan. The updated system successfully denied patients with relevant travel or contact histories entry to the hospital, thus preventing long lines outside the hospital. Further efforts could be made to integrate the system with the electronic medical record system.

(*JMIR Med Inform* 2021;9(7):e20994) doi:[10.2196/20994](https://doi.org/10.2196/20994)

**KEYWORDS**

COVID-19; hospital; information services; outpatients; patient; SARS-CoV-2; triage; virus

## *Introduction*

Since the outbreak of COVID-19, millions of individuals have been infected and tens of thousands of deaths have been reported worldwide [1]. This highly contagious and virulent disease has a higher case fatality rate than influenza. Moreover, sequelae such as pulmonary fibrosis have been observed in some recovered patients. As such, blocking the spread of the disease is a top priority during the pandemic [2]. Accordingly, governments worldwide have implemented various policies in hopes of decreasing the spread of COVID-19 [3], and the government of Taiwan is no exception [4].

Rigorous measures including travel restrictions and social distancing have been implemented to decrease the risk of community spread [5,6]. Various policies have also been enforced in hospitals to lower the risk of nosocomial infection. To contain the spread of the disease in inpatient departments, visitors to patients have been restricted to specific time slots, and the total number of visitors has been regulated [7,8]. However, while such regulations can be easily executed in an inpatient department, an outpatient department presents a greater challenge owing to the large number of daily outpatient visits and because the screening of ambulatory patients for relevant travel, occupation, contact, and cluster histories is necessary to minimize disease spread in an outpatient department.

Traditionally, screening of the travel, occupation, contact, and cluster histories of outpatients was conducted in examination rooms by physicians. However, given the current circumstances, allowing outpatients with a history of travel to high-risk countries to enter a hospital poses a risk of disease spread in the ambulatory service department [9]. A system to efficiently screen outpatients with relevant histories and deny them entry to the hospital is therefore of utmost importance. Thus, an automatic system that can acquire each patient's relevant travel and contact histories was deployed rapidly in Taipei Veterans General Hospital, one of the largest academic medical centers in Taiwan.

The aim of this study was to illustrate the deployment and utilization of the system. Further analysis focused on dynamic changes in the daily ambulatory services of the medical center with the implementation of this system. The experiences reported here will likely help hospitals worldwide in combating the COVID-19 pandemic and future pandemics.

## *Methods*

### **Methods Overview**

Taipei Veterans General Hospital is the largest public academic medical center in Taiwan. As of March 2020, this hospital had 2800 beds served by a staff of 6670 members. Until then, >8000 patient visits have been recorded on a daily basis. To contain the spread of COVID-19 in the outpatient department, the hospital deployed an automatic system to acquire patients' relevant travel and contact histories.

### **System Design**

A web-based appointment system for the hospital's ambulatory service was already in operation before the incorporation of the screening system. The system was implemented by querying the virtual private network (VPN) travel and contact history database maintained by the Ministry of Health and Welfare (Figure 1). Patients are required to enter their national ID number when attempting to book an appointment (Figure 2). The system then uses the national ID number to check for relevant travel or contact history through the VPN database. If the result turns out to be negative, an appointment is booked, and patients who entered their mobile phone numbers will receive text messages that provide them with access to an expedited entry route. However, if a patient has a potentially problematic travel or contact history, the attempt to book an appointment will fail. Patients can also book appointments on site. Their history will then be checked via the VPN database on inserting their health smart card before gaining entry through the regular route.



Figure 1. Design of the outpatient screening system. VPN: virtual private network.

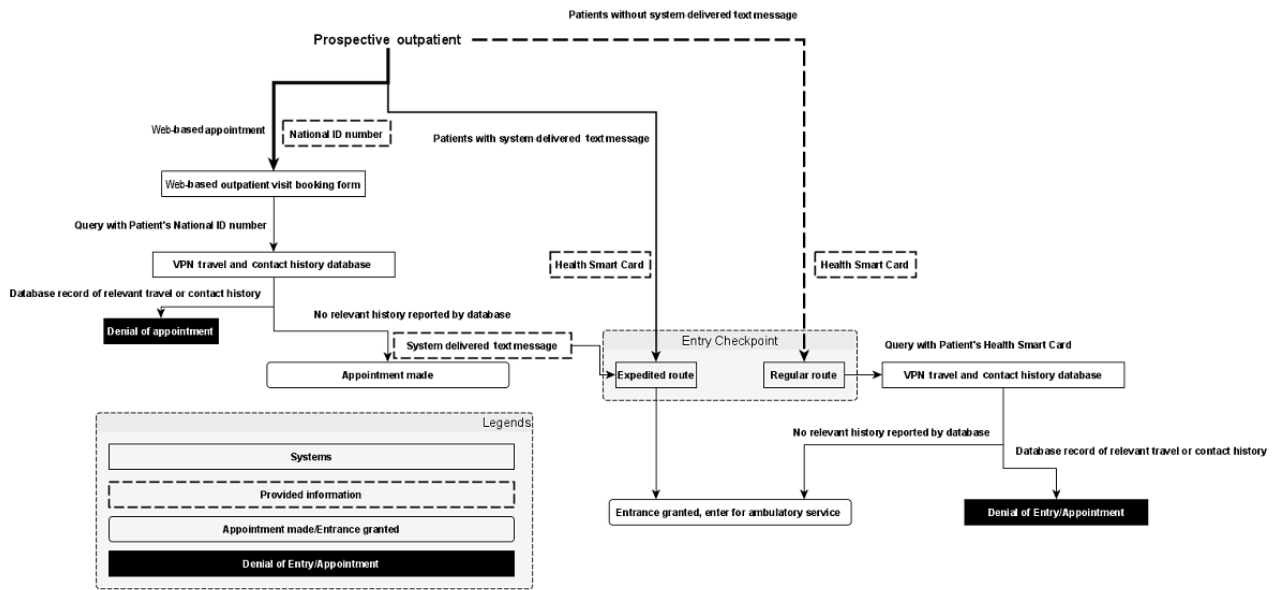


Figure 2. Screenshot of a web-based appointment form.

Upon arrival at the hospital, outpatients can enter the hospital through two routes: regular or expedited. The regular route is for patients who do not present text messages received from the screening system. These patients are asked to present their health smart card to the staff. The on-site VPN database query is then made, and a patient would be granted access after the confirmation of no potentially problematic histories provided by the VPN database. The patient will also receive a paper-based questionnaire that is to be submitted to the hospital in the examination room.

On the other hand, patients who successfully book an appointment on the internet and receive text messages from the hospital can enter the hospital through the expedited route. The staff will confirm the legitimacy of the text message and verify the patient's identity with the patient's health smart card. Once verified, the patient is allowed to enter the hospital and is provided the same aforementioned questionnaire.

**Data Processing**

The system described in this article was deployed on April 21, 2020. Data were retrieved from the registration database between April 21 and May 10, 2020. The data for each Sunday during that period were omitted since no regular outpatient service is

provided on Sundays. All web-based and on-site booked appointments were extracted. Each denied appointment was also extracted along with the reason for nonadmittance during the studied period.

**Statistical Analysis**

Descriptive statistics were assessed using Microsoft Excel 2016 (Microsoft Corp).

**Results**

During the study period, 127,857 visits were recorded. Among these visits, 91,195 were registered on the internet. In total, 71,816 of them generated text messages for the expedited route of entry (Table 1).

The average daily number of visits during the study period was 9498 on weekdays and 1463 on Saturdays. The highest number of visits (n=11,147) was recorded on May 6, 2020, and the lowest number (n=1351) was recorded on April 25, 2020.

In total, 71% of the visits were registered on the internet. Further, 78.76% of all patients who registered on the internet received text messages delivered from the system for the expedited route of entry.

Of all the visits registered on the internet, queries to the VPN database revealed relevant histories for 65 patients. Moreover, 66 patients entered invalid national ID numbers when attempting

to register for an outpatient visit. All of these attempts were blocked by the system. The denied attempts comprised approximately 0.14% of all web-based registration attempts.

**Table 1.** Total number of attempts and visits from April 21 to May 9, 2020.

Observations	April 21	April 22	April 23	April 24	April 25	April 27	April 28	April 29	May 1	May 2	May 4	May 5	May 6	May 7	May 8	May 9	Total
Attempts with an invalid ID, n	6	7	5	3	0	7	5	4	2	0	7	5	6	4	5	0	66
Attempts with relevant history, n	6	7	4	3	1	11	3	5	3	2	5	2	3	5	3	2	65
Text messages sent, n	4121	4961	5083	4434	848	5330	5358	6051	4693	908	5728	5699	6550	5816	5204	1032	71,816
Attempts to register, n	5468	6660	6363	5610	1010	6828	6781	7679	5960	1091	7171	7243	8163	7340	6622	1206	91,195
Total outpatient visits, n	8154	9294	8910	7866	1351	10,071	9247	10,405	8309	1437	10,650	10,017	11,147	10,128	9270	1601	127,857
Web-based registration rate of outpatients, %	67.06	71.66	71.41	71.32	74.76	67.80	73.33	73.80	71.73	75.92	67.33	72.31	73.23	72.47	71.43	75.33	71.33
Text message receipt rate on web-based attempts, %	75.37	74.49	79.88	79.04	83.96	78.06	79.01	78.80	78.74	83.23	79.88	78.68	80.24	79.24	78.59	85.57	78.75

## Discussion

### Principal Findings

To our knowledge, this is the first reported system to automatically screen for a patient's travel and contact histories prior to entry to a hospital. A similar check-in system for outpatient magnetic resonance imaging studies has been previously described [10], but there have been no reports of a hospital-wide service for the extraction of travel and contact histories. Previous studies on web-based outpatient resources of hospitals have focused on probing the features of official applications and appointment systems [11,12]. The purpose of the system reported herein is to deny access to patients with a relevant contact or travel history recorded in the VPN contact and travel history database. The goal was successfully achieved after the system's deployment on April 21, 2020.

### Challenges of Screening a Massive Number of Outpatients

Taipei Veterans General Hospital, being one of the largest public academic centers in Taiwan, is visited by over 10,000 patients each day. The traditional method of screening outpatients for relevant travel, occupation, cluster, and contact histories involves medical professionals or administrative staff taking patient histories and excluding those with relevant histories. This traditional method has allowed patients with relevant travel and contact histories to enter the hospital and could thus facilitate further disease spread. Therefore, a system to deny entry to patients with relevant histories was warranted [5].

However, with the large number of patients visiting the hospital's outpatient department, an efficient way to acquire

their histories was also required. Therefore, the automatic screening system utilizing the VPN travel and contact history database was an optimal solution for the obstacles encountered. The system acquires these histories by querying the database with the provided national ID numbers. With minimal training, staff could easily utilize the on-site screening system at the hospital entrance to stop patients with relevant histories from entering the outpatient department.

### Long Lines in the Early Phase of the Screening Policy

During the early phase of implementing the screening policy for outpatients, the hospital relied on on-site VPN database queries to deny entrance to patients with a relevant travel or contact history reported by the database. However, this soon proved time-consuming. Long lines were formed each morning during the opening hours of ambulatory services. Therefore, an expedited entry route was further designed to alleviate this problem.

The expedited route provided a faster track to patients with registered mobile phone numbers that passed the VPN database check before arriving at the hospital and was able to successfully mitigate the long lines at the entry checkpoint. Additionally, denying patients with a relevant history upon booking an appointment on the internet stops such patients from arriving at the hospital and further lowers the risk of disease spread.

### The Need for a Paper-Based Questionnaire

Though querying the VPN travel and contact database provided a fast and efficient means to deny hospital entry to individuals with a relevant travel or contact history, the information provided by the database was incomplete. Therefore, a separate

questionnaire was needed for recording each patient's occupational history and other pertinent information.

To ensure that every person visiting the outpatient department completed this questionnaire, a paper-based version of the questionnaire was distributed to each patient at the hospital entrance and then collected upon the patient's entry to the examination room before the patient was examined by the physician.

### Presence of Contact and Travel Histories Among Outpatients

The system revealed a total of 65 patients with relevant histories, including those with recent travel histories to high-risk countries within 2 weeks or those with positive contact histories, during the study period. This accounted for an extremely low percentage (0.05% of total visits during the study period) of the entire outpatient population.

However, one case of a nosocomial cluster at a medical center in Taiwan still indicated the importance of enforcing rigorous entrance screening policies at large medical facilities.

### Limitations

This study aimed to illustrate the design and utilization of an automatic system to acquire patients' relevant travel and contact histories. However, owing to the limited time spent on its development, the system was unable to generate records in the electronic medical record system. Moreover, the on-site screening system could not distinguish on-site-registered outpatients from family members and friends accompanying the patients. Therefore, no analysis could be performed for individuals who registered on site.

### Conclusions

This study demonstrated the successful deployment of an automatic system to acquire patients' relevant travel and contact histories. The utilization rate of the internet-based system is optimal (78.76% of all web-based registered visits) since it is incorporated into the already operating web-based registration system for outpatient visits. Further efforts could be made to integrate the system with the electronic medical record system.

### Acknowledgments

This study was supported by a grant (V109E-002-1) from Taipei Veterans General Hospital.

### Conflicts of Interest

None declared.

### References

1. Wu Y, Chen C, Chan Y. The outbreak of COVID-19: An overview. *J Chin Med Assoc* 2020 Mar;83(3):217-220 [FREE Full text] [doi: [10.1097/JCMA.0000000000000270](https://doi.org/10.1097/JCMA.0000000000000270)] [Medline: [32134861](https://pubmed.ncbi.nlm.nih.gov/32134861/)]
2. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA* 2020 Apr 07;323(13):1239-1242. [doi: [10.1001/jama.2020.2648](https://doi.org/10.1001/jama.2020.2648)] [Medline: [32091533](https://pubmed.ncbi.nlm.nih.gov/32091533/)]
3. Anderson RM, Heesterbeek H, Klinkenberg D, Hollingsworth TD. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *Lancet* 2020 Mar 21;395(10228):931-934 [FREE Full text] [doi: [10.1016/S0140-6736\(20\)30567-5](https://doi.org/10.1016/S0140-6736(20)30567-5)] [Medline: [32164834](https://pubmed.ncbi.nlm.nih.gov/32164834/)]
4. Wang CJ, Ng CY, Brook RH. Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing. *JAMA* 2020 Apr 14;323(14):1341-1342. [doi: [10.1001/jama.2020.3151](https://doi.org/10.1001/jama.2020.3151)] [Medline: [32125371](https://pubmed.ncbi.nlm.nih.gov/32125371/)]
5. Bai Y, Yao L, Wei T, Tian F, Jin D, Chen L, et al. Presumed Asymptomatic Carrier Transmission of COVID-19. *JAMA* 2020 May 14;323(14):1406-1407 [FREE Full text] [doi: [10.1001/jama.2020.2565](https://doi.org/10.1001/jama.2020.2565)] [Medline: [32083643](https://pubmed.ncbi.nlm.nih.gov/32083643/)]
6. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* 2020 Apr 24;368(6489):395-400 [FREE Full text] [doi: [10.1126/science.aba9757](https://doi.org/10.1126/science.aba9757)] [Medline: [32144116](https://pubmed.ncbi.nlm.nih.gov/32144116/)]
7. Adams JG, Walls RM. Supporting the Health Care Workforce During the COVID-19 Global Epidemic. *JAMA* 2020 May 21;323(15):1439-1440. [doi: [10.1001/jama.2020.3972](https://doi.org/10.1001/jama.2020.3972)] [Medline: [32163102](https://pubmed.ncbi.nlm.nih.gov/32163102/)]
8. Fauci AS, Lane HC, Redfield RR. Covid-19 - Navigating the Uncharted. *N Engl J Med* 2020 Mar 26;382(13):1268-1269 [FREE Full text] [doi: [10.1056/NEJMe2002387](https://doi.org/10.1056/NEJMe2002387)] [Medline: [32109011](https://pubmed.ncbi.nlm.nih.gov/32109011/)]
9. Wilder-Smith A, Chiew CJ, Lee VJ. Can we contain the COVID-19 outbreak with the same measures as for SARS? *Lancet Infect Dis* 2020 May;20(5):e102-e107 [FREE Full text] [doi: [10.1016/S1473-3099\(20\)30129-8](https://doi.org/10.1016/S1473-3099(20)30129-8)] [Medline: [32145768](https://pubmed.ncbi.nlm.nih.gov/32145768/)]
10. Pirasteh A, VanDyke M, Bolton-Ronacher J, Xi Y, Eastland R, Young D, et al. Implementation of an Online Screening and Check-In Process to Optimize Patient Workflow Before Outpatient MRI Studies. *J Am Coll Radiol* 2016 Aug;13(8):956-959.e5. [doi: [10.1016/j.jacr.2015.10.036](https://doi.org/10.1016/j.jacr.2015.10.036)] [Medline: [26786030](https://pubmed.ncbi.nlm.nih.gov/26786030/)]
11. Yang P, Chu F, Liu H, Shih M, Chen T, Chou L, et al. Features of Online Hospital Appointment Systems in Taiwan: A Nationwide Survey. *Int J Environ Res Public Health* 2019 Jan 09;16(2):171 [FREE Full text] [doi: [10.3390/ijerph16020171](https://doi.org/10.3390/ijerph16020171)] [Medline: [30634467](https://pubmed.ncbi.nlm.nih.gov/30634467/)]

12. Liu H, Lee W, Sun Y, Fen J, Chen T, Chou L, et al. Hospital-Owned Apps in Taiwan: Nationwide Survey. JMIR Mhealth Uhealth 2018 Jan 16;6(1):e22 [[FREE Full text](#)] [doi: [10.2196/mhealth.8636](https://doi.org/10.2196/mhealth.8636)] [Medline: [29339347](https://pubmed.ncbi.nlm.nih.gov/29339347/)]

## Abbreviations

**VPN:** virtual private network

*Edited by G Eysenbach; submitted 04.06.20; peer-reviewed by Q Ma, I Hochberg; comments to author 23.07.20; revised version received 05.05.21; accepted 23.05.21; published 27.07.21.*

*Please cite as:*

*Lu DH, Hsu CA, Yuan EJ, Fen JJ, Lee CY, Ming JL, Chen TJ, Lee WC, Chen SA*

*Experiences With Internet Triage of 9498 Outpatients Daily at the Largest Public Hospital in Taiwan During the COVID-19 Pandemic: Observational Study*

*JMIR Med Inform 2021;9(7):e20994*

*URL: <https://medinform.jmir.org/2021/7/e20994>*

*doi: [10.2196/20994](https://doi.org/10.2196/20994)*

*PMID: [34043524](https://pubmed.ncbi.nlm.nih.gov/34043524/)*

©Ding-Heng Lu, Chia-An Hsu, Eunice J Yuan, Jun-Jeng Fen, Chung-Yuan Lee, Jin-Lain Ming, Tzeng-Ji Chen, Wui-Chiang Lee, Shih-Ann Chen. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 27.07.2021. 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

# Social Media Opinions on Working From Home in the United States During the COVID-19 Pandemic: Observational Study

Ziyu Xiong<sup>1</sup>, BSc; Pin Li<sup>1</sup>, BSc, MSc, MSBA; Hanjia Lyu<sup>1</sup>, BEc, MSc; Jiebo Luo<sup>1</sup>, BEng, MSc, PhD

University of Rochester, Rochester, NY, United States

**Corresponding Author:**

Jiebo Luo, BEng, MSc, PhD

University of Rochester

3101 Wegmans Hall

Rochester, NY, 14627

United States

Phone: 1 585 276 5784

Email: [jluo@cs.rochester.edu](mailto:jluo@cs.rochester.edu)

## Abstract

**Background:** Since March 2020, companies nationwide have started work from home (WFH) owing to the rapid increase of confirmed COVID-19 cases in an attempt to help prevent the disease from spreading and to rescue the economy from the pandemic. Many organizations have conducted surveys to understand people's opinions toward WFH. However, the findings are limited owing to a small sample size and the dynamic topics over time.

**Objective:** This study aims to understand public opinions regarding WFH in the United States during the COVID-19 pandemic.

**Methods:** We conducted a large-scale social media study using Twitter data to portray different groups of individuals who have positive or negative opinions on WFH. We performed an ordinary least squares regression analysis to investigate the relationship between the sentiment about WFH and user characteristics including gender, age, ethnicity, median household income, and population density. To better understand the public opinion, we used latent Dirichlet allocation to extract topics and investigate how tweet contents are related to people's attitude.

**Results:** On performing ordinary least squares regression analysis using a large-scale data set of publicly available Twitter posts ( $n=28,579$ ) regarding WFH during April 10-22, 2020, we found that the sentiment on WFH varies across user characteristics. In particular, women tend to be more positive about WFH ( $P<.001$ ). People in their 40s are more positive toward WFH than those in other age groups ( $P<.001$ ). People from high-income areas are more likely to have positive opinions about WFH ( $P<.001$ ). These nuanced differences are supported by a more fine-grained topic analysis. At a higher level, we found that the most negative sentiment about WFH roughly corresponds to the discussion on government policy. However, people express a more positive sentiment when discussing topics on "remote work or study" and "encouragement." Furthermore, topic distributions vary across different user groups. Women pay more attention to family activities than men ( $P<.05$ ). Older people talk more about work and express a more positive sentiment regarding WFH.

**Conclusions:** This paper presents a large-scale social media-based study to understand the public opinion on WFH in the United States during the COVID-19 pandemic. We hope that this study can contribute to policymaking both at the national and institution or company levels to improve the overall population's experience with WFH.

(*JMIR Med Inform* 2021;9(7):e29195) doi:[10.2196/29195](https://doi.org/10.2196/29195)

**KEYWORDS**

characterization; COVID-19; social media; topic modeling; Twitter; work from home

## Introduction

**Background**

COVID-19 was first reported in China and then spread worldwide and has caused 22.3 million confirmed cases and more than 373,000 deaths in the United States as on January

11, 2021 [1]. To help prevent the virus from spreading and to salvage the economy, companies and schools nationwide have started work and study from home, respectively. According to a Gartner survey of 880 global human resources executives on March 17, 2020, almost 88% organizations have encouraged or required employees to work from home (WFH) [2]. Barrero et al [3] found that WFH might persist even after the pandemic.



Concerns may arise regarding productivity [4], willingness [5], and future trends [3,6] regarding work and study from home.

### Prior Studies

WFH has been a controversial issue that merits closer scrutiny. Palumbo [5] reported that WFH might incur side-effects such as a negative impact on work-life balance. This would lead to negative opinions on WFH when people tweet about it. Other studies have focused on specific categories. A survey of employees in Lithuania [7] reported that female employees appreciate WFH more than male employees because female employees can enjoy a healthier lifestyle, while male employees worry about career constraints. However, another survey conducted in the United States [4] shows “a gender gap in perceived work productivity”: before implementing WFH, female and male employees reported the same level of self-rated work productivity. After transitioning to WFH, male employees performed with higher productivity than female employees [4]. Regarding age, people in their 40s have more negative opinions on WFH because of their unfamiliarity with teleworking. People aged 30-39 years have the most positive opinions because they can enjoy time with their families and they are already accustomed with new technologies for teleworking [7]. Previous studies [3,7,8] also show that opinions concerning WFH vary across different socioeconomic groups. A similar social media study of public sentiments on WFH has been conducted in the United Kingdom [9] and reported that more than 70% of tweets concerning WFH expressed a positive sentiment, with the main topics including “traffic,” “drink,” and “e-commerce.”

Similar approaches have been implemented by researchers who mined Twitter posts on public attitudes toward face masks through natural language processing [10] using the Valence Aware Dictionary and Sentiment Reasoner (VADER) model [11] to perform sentiment analysis. Moreover, Twitter data have been used to study many different aspects of COVID-19, such as mining of overall public perception of COVID-19 [12], college students' attitudes toward the pandemic [13], people's attitude toward potential COVID-19 vaccines [14,15], sentiment analysis among pregnant women during quarantine [16], and monitoring of depression trends on Twitter during the COVID-19 pandemic [17]. These studies have used VADER for sentiment analysis, and most of them also include a time-series analysis. In addition, we followed the practice of using latent Dirichlet allocation (LDA) [18] to identify topics among a large text corpus. The M3-inference model [10] was used to portray different demographic groups.

### Study Objectives

In this study, we intend to understand public opinions on WFH, using large-scale social media data. Twitter has been a popular social media platform for people, especially in the United States, to express their opinions on what is happening around them. In contrast, the Boston Consulting Group used survey data to study employees' opinions regarding WFH owing to the COVID-19 pandemic [19]. However, social media data allow an opportunity for conducting a timelier study of many population-level issues on a larger scale [20]. We acquired data with an authorized Twitter developer account using Tweepy. This ensures reliability by acquiring first-hand and sufficient data for the study.

In this study, we also inferred user demographic information using Twitter user information. This is important, since we can carry out an in-depth assessment of the characteristics of those who are more pro-WFH. For example, when we consider gender, we understand that historically mothers have been mostly responsible for caring for children [21]. Therefore, we need information regarding the users' gender to determine whether there is any difference in sentiment toward WFH between women and men, as WFH would allow female employees to allocate more time to spend with their children.

Our goal is to understand the public opinions on WFH in the United States during the COVID-19 pandemic. In particular, we focused on the following research questions:

1. Who is more likely to tweet about WFH?
2. How does the sentiment of WFH vary across user demographics?
3. Regarding WFH, what do Twitter users mainly discuss? How does the content correlate with the sentiment of WFH?

To summarize, in a large-scale data set of publicly available Twitter posts concerning WFH during April 10-22, 2020, we found that women and older people are more likely to tweet about WFH. On performing ordinary least squares regression analysis, we confirm that sentiment of WFH varies across user characteristics. In particular, women tend to be more positive about WFH than men. People in their 40s are more positive toward WFH than those in other age groups. People from high-income areas are more likely to have positive opinions about WFH.

These nuanced differences are supported by a more fine-grained topic analysis. At a higher level, we found that the most negative sentiment about WFH roughly corresponds to discussions on government policy. However, people express more positive sentiment when discussing topics on “remote work and study” and “encouragement.” Furthermore, topic distributions vary across different user groups.

## Methods

### Methods Overview

In this section, we summarized the data collection process and the methods we used in the analyses. To address research questions 1 and 2, we discuss how we inferred user characteristics and the sentiment in the “Feature Inference” subsection. To address research question 3, we describe how we extracted the topics of tweets in the “Topic Modeling” subsection.

### Data Collection

We collected relevant English-language tweets through the Tweepy stream application programming interface (API) using keywords and hashtag-filtering. The filter keywords and hashtags are “WFH,” “workfromhome,” “work from home,” “#wfh,” and “#workingfromhome.” In total, 553,166 unique tweets with 23 attributes posted by 405,455 unique Twitter users during April 5-26, 2020, were sampled. We attempted to infer the gender, age, and ethnicity of these Twitter users, extract the population density of the area they resided in, and estimate the

sentiment of the tweets. There are 405,455 unique users in our data set, 313,815 (77.3%) of whom only tweeted once. After excluding duplicates and users with incomplete features, 28,579 unique Twitter users with all features were included in the data set.

## Feature Inference

### Sentiment

A normalized, weighted composite score was calculated for each tweet, using VADER [11] to measure the sentiment. The score ranged from  $-1$  (most negative) to  $+1$  (most positive). For validation, we randomly select 194 users' tweets within 1 month. By manually labeling the sentiment and comparing the sentiment scores with VADER scores, we found that the accuracy was 76%, suggesting that the automatic natural language processing methods we used provide adequate estimates of the sentiment of the tweets. The mean sentiment score was 0.242 (SD 0.448; range  $-0.967$  to  $0.984$ ; 25th percentile 0.000, 50th percentile 0.318, 75th percentile 0.617).

### Age and Gender

We applied the M3-inference model [22] to infer the gender and age of each Twitter user from their profile name, username (screen name), and profile description. Age is binned into four groups:  $\leq 18$  years, 19-29 years, 30-39 years, and  $\geq 40$  years. The gender distribution of Twitter users is biased toward men (71.8%) [23]. A similar pattern was also observed in our data set, where 57.9% of users are men and 42.1% are women. With respect to age, 37.08% of the users in our data set are older than 40 years, 37.6% are between 30 and 39 years old, 16.5% are between 19 and 29 years old, and the rest are younger than 19 years old. According to a report from the Pew Research Center [24], Twitter users are younger than the average US adult; 21% of adults are aged 18-29 years, 33% are aged 30-49 years, 26% are aged 50-64 years, and 20% are aged  $\geq 65$  years. The percentages of adults in the Twitter user population are 29%, 44%, 19%, and 8%, respectively. The pattern in our data set is more similar to that of the distribution of US adults.

### Ethnicity

To estimate the ethnicity of Twitter users, we applied the Ethnicolr API, which makes inferences on the basis of the last name and first name or just the last name of the Twitter user [25]. In our study, we removed emoji icons, hyphens, unrelated contents, and special characters to extract the last names and applied "census\_In" to infer the ethnicity, which included White, Black or African American, Asian or Pacific Islander, American Indian or Alaskan Native, and Hispanic.

In our data set, White (83.4%) was predominant over other ethnicities, while according to the US Census Bureau [26], White ethnicities accounted for 60.1% of the US population; 7.3% of Twitter users in this study were Asian or Pacific Islanders, while this ethnicity accounts for 6.1% of the US population; 6.5% of Twitter users in this study were Hispanic, while this ethnicity accounted for 18.5% of the US population; 2.5% of Twitter users in this study were Black or African American, while this ethnicity accounted for 13.4% of the US population; American Indian or Alaskan Natives constituted

0.26% of our study's Twitter user population. According to a report from the Pew Research Center [24], the proportion of individuals by race or ethnicity are almost the same between the US population and Twitter adult users. Interestingly, the proportions of White and Asian or Pacific Islander individuals are much higher than those in the general US population, which could be related to the labor force distributions of these 2 groups. In 2018, 54% of employed Asian and 41% of employed White individuals, compared with 31% of employed Black or African American and 22% of employed Hispanic individuals, worked in management, professional, and related occupations [27], which can most likely be managed from home [28]. Therefore, it is not surprising that there are more White and Asian or Pacific Islander individuals in our data set owing to the disparities in the occupations.

### Population Density

The USzipcode search engine was applied to extract the population density of each user's location that Twitter users self-report in their profile information. The population density is categorized as urban (greater than 3000), suburban (1000-3000), and rural (lower than 1000). Finally, 67.4% of users in this study were from urban areas, 14.6% were from suburban areas, and the rest were from rural areas. The majority of the users in our data set were from urban areas, which is consistent with the fact that 83% of the US population resides in urban areas [29]; however, there were proportionally fewer urban users in our data set than in the US population.

### Income

To investigate the relationship between people's attitude toward WFH and the gap between high- and low-income areas, we retrieved regional median income from the 2019 American Community Survey. Census API tools were used to extract the median income with an input of city-level user location. The median regional income was US \$33,538 (SD US \$10,298; range US \$3951-121,797; 25th percentile US \$28,072, 50th percentile US \$31,613, 75th percentile US \$36,336).

### Topic Modeling

We used LDA [18] to extract topics from the tweets. In our study, we used the stop words package of the Natural Language Toolkit library, extended with topic-related words (eg, "work" and "home"). To extract the most relevant topics, we only collected nouns, verbs, adjectives, and adverb lemmas. We use the spaCy package to screen all the words of the tweets and only includes those words whose postag is "NOUN," "ADJ," "VERB," or "ADV." We tuned the hyperparameters with nested looping topic numbers  $\alpha$  and  $\beta$ . Finally, we chose  $num\_topics=9$ ,  $\alpha=0.91$ ,  $\beta=0.31$ , and a coherence score  $C_v$  of 0.379.

## Results

### Sentiment Analysis

As indicated above, Twitter users' opinions of WFH were slightly positive. We attempted to investigate the relationship between user characteristics and the sentiment of discussions on WFH. We performed ordinary least squares regression analysis on our data set ( $n=28,579$ ). Descriptive statistics and

bivariate correlations are shown in Table 1. Table 2 summarizes the results of ordinary least squares regression analysis.

**Table 1.** Descriptive statistics and bivariate correlation coefficients for study variables<sup>a</sup>.

Variables	Mean (SD)	1	2	3	4	5	6	7	8	9	10
1	0.59 (0.49)										
2	0.09 (0.28)	-.04 <sup>b</sup>									
3	0.16 (0.37)	-.21 <sup>b</sup>	-.14 <sup>b</sup>								
4	0.38 (0.48)	.01	-.24 <sup>b</sup>	-.34 <sup>b</sup>							
5	0.03 (0.16)	.01 <sup>c</sup>	.02 <sup>b</sup>	-.00	-.03 <sup>b</sup>						
6	0.07 (0.26)	.01	.07 <sup>b</sup>	-.02 <sup>b</sup>	-.01	-.05 <sup>b</sup>					
7	0.01 (0.05)	.02 <sup>b</sup>	.03 <sup>b</sup>	-.01	-.00	-.01	-.01 <sup>c</sup>				
8	0.07 (0.25)	-.00	.02 <sup>b</sup>	.04 <sup>b</sup>	.03 <sup>b</sup>	-.04 <sup>b</sup>	.00	-.02 <sup>b</sup>			
9	0.25 (0.09)	.04 <sup>b</sup>	-.03 <sup>b</sup>	-.07 <sup>b</sup>	-.01	-.01	.04 <sup>b</sup>	.00	-.02 <sup>b</sup>		
10	0.67 (0.47)	-.02 <sup>b</sup>	.02 <sup>b</sup>	.03 <sup>b</sup>	.02 <sup>b</sup>	-.01	.05 <sup>b</sup>	-.00	.03 <sup>b</sup>	.05 <sup>b</sup>	
11	0.15 (0.35)	.02 <sup>b</sup>	-.02 <sup>b</sup>	-.02 <sup>b</sup>	.01 <sup>c</sup>	.01 <sup>c</sup>	-.02 <sup>b</sup>	.00	.00	.09 <sup>b</sup>	-.60

<sup>a</sup>Study variables: 1=gender (0=female, 1=male), 2=age ≤18 years (0=no, 1=yes), 3=age of 19-29 years (0=no, 1=yes), 4=age of 30-39 years (0=no, 1=yes), 5=Black or African American ethnicity (0=no, 1=yes), 6=Asian or Pacific Islander ethnicity (0=no, 1=yes), 7=American Indian or Alaskan Native ethnicity (0=no, 1=yes), 8=Hispanic ethnicity (0=no, 1=yes), 9=income (normalized using MinMaxScaler), 10=urban (0=no, 1=yes), and 11=suburban (0=no, 1=yes).

<sup>b</sup> $P < .001$ .

<sup>c</sup> $P < .05$ .

**Table 2.** Ordinary least squares regression outputs for public opinions (N=28,579) on working from home against demographics and other variables of interest.

Predictor	Sentiment score	
	$\beta$ (SE)	95% CI
Intercept	0.252 <sup>a</sup> (0.011)	0.231 to 0.273
Gender (0=female, 1=male)	-0.021 <sup>a</sup> (0.006)	-0.032 to 0.010
Age ≤18 years (0=no, 1=yes)	-0.084 <sup>a</sup> (0.010)	-0.103 to -0.064
Age of 19-29 years (0=no, 1=yes)	-0.076 <sup>a</sup> (0.008)	-0.092 to -0.060
Age of 30-39 years (0=no, 1=yes)	-0.022 <sup>a</sup> (0.005)	-0.034 to -0.010
Black or African American ethnicity (0=no, 1=yes)	0.023 (0.017)	-0.011 to 0.066
Asian or Pacific Islander ethnicity (0=no, 1=yes)	0.003 (0.010)	-0.020 to 0.020
Hispanic ethnicity (0=no, 1=yes)	-0.006 (0.011)	-0.027 to 0.016
American Indian or Alaskan Native ethnicity (0=no, 1=yes)	-0.013 (0.052)	-0.115 to 0.088
Income	0.143 <sup>a</sup> (0.031)	0.082 to 0.203
Urban (0=no, 1=yes)	-0.007 (0.007)	-0.021 to 0.007
Suburban (0=no, 1=yes)	-0.004 (0.009)	-0.023 to 0.014
F-statistics	15.25 <sup>a</sup>	
R <sup>2</sup>	0.006	
Adjusted R <sup>2</sup>	0.005	

<sup>a</sup> $P < .001$ .

### Women Tended to Be More Positive About WFH

Men were significantly more negative about WFH than women ( $P<.001$ ). This is consistent with the remote work survey report by Fast Company [30]. A more positive sentiment observed among women could be due to the change in working styles [30] and fewer work hours than those of men [31]. A previous survey [7] indicates that women favor WFH from the perspective of a healthier lifestyle.

### People in Their 40s Are More Positive Toward WFH Than Other Age Groups

Age is another perspective. Results of regression analysis revealed that as age increases, people are significantly more pro-WFH ( $P<.001$ ). This is consistent with the results of the survey conducted by Watkins [32] that generation Z individuals (people aged 8-23 years as of this writing) are more pro-office than millennials (aged 24-39 years). While assumptions exist among older employees, who might be unfamiliar with electronic devices and thus become more pro-office. However, an article in the Financial Times [33] reported contrasting observations. People aged  $\geq 40$  years are less likely to be re-employed; hence, they prefer to retain their current jobs while

avoiding the risk of being exposed to COVID-19, especially since this group is most vulnerable to COVID-19. Further details regarding these topics will be discussed in the following section. Furthermore, we observed the same pattern as that reported in a survey conducted among employees in Lithuania [7].

### People in Higher-Income Areas Are More Likely to Have Pro-WFH Opinions Than Those in Lower-Income Areas

Income was significantly correlated with the sentiment toward WFH ( $P<.001$ ). This is concurrent with our finding that people from urban areas would be more pro-WFH, since the regional median income would be higher in big cities [34]. This finding is also in line with that of Barrero et al [3] that high-income workers, in particular, enjoy the perks of WFH.

### Topic Analysis

Further, we attempted to investigate what Twitter users mainly discuss with reference to WFH. In particular, we investigated how the contents of the tweets correlate with the sentiment of WFH. Table 3 shows the 9 topics extracted by the LDA model. We assigned each topic a title on the basis of the top 10 keywords.

**Table 3.** Titles and the top 10 keywords of the topics extracted by the latent Dirichlet allocation model.

Topic#	Topic title	Topic keywords
1	Family activities	dog, try, today, wife, day, last, school, virtual, watch, look
2	Remote work or study	remote, new, time, covid, learn, great, help, many, support, join
3	Quarantine	pandemic, stay, safe, day, go, today, let, see, also, think
4	Dressing	dress, get, enough, adult, day, wear, time, zoom, right, thank
5	Government and policy	money, less, job, option, able, new, remotely, safely, force, take
6	COVID-19 side effects	people, still, do, job, good, say, first, place, probably, go
7	Encouragement	get, lot, world, honestly, fall, reveal, love, week, look
8	Back to office	go, back, know, office, time, feel, quarantine, hour, tip, covid
9	Leasing	office, well, year, instead, couple, permanently, think, lease, renew, similarly

Figure 1 shows the proportions of the topics. Topic 1 (family activities) contained the keywords “dog,” “wife,” and “watch” and accounts for 17.7% of the total tweets. Topic 2 (remote work or study) accounts for 16.7% of the total tweets, where people mostly tweeted about remote work and study. Topic 3 (quarantine) contained the keywords “pandemic,” “force,” and “stay.” Topic 4 (dressing) contained the keywords “dress” and “wear,” and most of the tweets based on this topic discuss what people wear when they WFH. Topic 5 (government and policy) accounts for 6.8% of the total tweets and contained the keywords “money,” “job,” and “force”; in this topic, many of the tweets mention the names of governors and express their concerns about WFH-related policies. An example of such a tweet is as follows:

*“@GovMurphy #CancelRentNJ If NJ doesn’t cancel rent, the consequences for everyone who can’t work from home will be catastrophic. The bailout money will go straight to landlords instead of feeding people.”*

Topic 6 (COVID-19 side effects) contained the keywords “still” and “job” and accounted for 8.5% of the total tweets. In this topic, people mostly complained about the influence of COVID-19, such as “job changing” and “staying home for so long.” An example of such a tweet is as follows:

*“I agree. I think the pandemic has shown the disparity of digital access in rural areas, and libraries in these kinds of communities should take note of what these people need so they can provide better services when all of this is over. @Sonaite @BakerChair #SLIS752 #752Diversity[QUOTE]@Sonaite @BakerChair @KaeliNWLlib I think this pandemic has shown a spotlight on the digital divide that still exists. Schools are scrambling to provide ‘continuity of education’ when children don’t have access to devices and/or reliable Internet.”*

Topic 7 (encouragement) accounted for 7.5% of the total tweets, where people express their support and inspire people to overcome difficulties together. One example of such a tweet is as follows:



*“If you are working out in this scary world today, my love to you. If you are working from home, my love to you. If you are out of work, my love to you. If you are lonely, my love to you. If you are sick, my love to you. If you are grieving, my love to you. My love. To you.”*

Topic 8 (back to office) contained the keywords “back” and “office” and accounted for 13.2% of the total tweets. Under this

topic, people mostly tweet about their opinions toward the office, including those inquiring if the return to “office” will finally materialize or when people will be able to go back to office. Topic 9 (leasing) contains keywords “year,” “lease,” “renew,” and “office,” where people argue that some companies might not renew their office lease for the next year because of how well-suited WFH is for these companies.

**Figure 1.** Topic distributions.

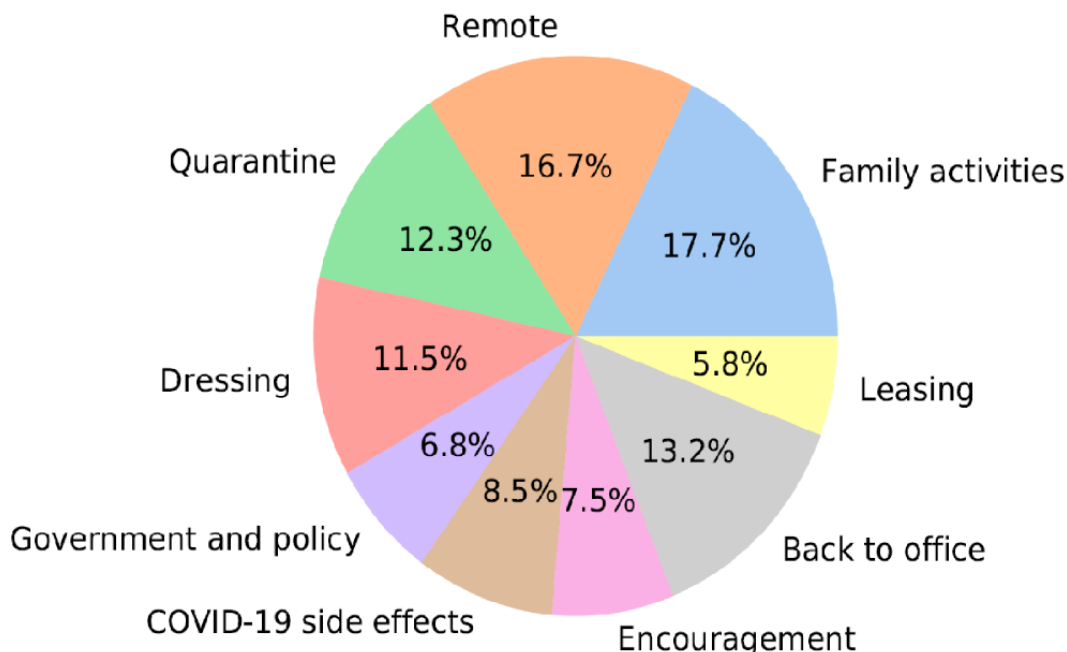
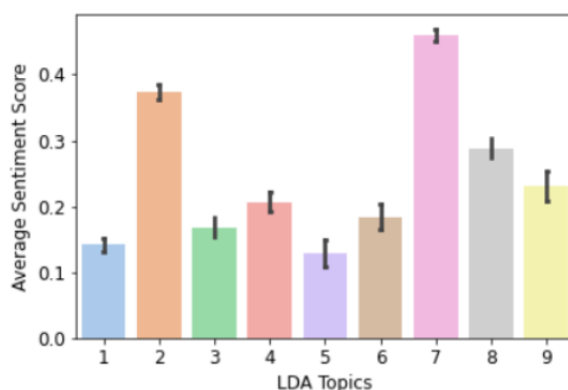


Figure 2 shows the average sentiment score of each topic. Topic 7 (encouragement) had the highest average sentiment score (0.460) and was considered the most positive topic; in contrast, topic 5 (government and policy) was the least positive topic (average sentiment score=0.129). As indicated above (“Feature Inference” subsection), the average sentiment score of all the users in our data set was 0.242. Among these 9 topics, all

expressed a positive sentiment toward WFH. Moreover, topics 2 (remote work or study), 7 (encouragement), and 8 (back to office) had a sentiment score above the average, while topics 1 (family activities), 3 (quarantine), 4 (dressing), 5 (government and policy), 6 (COVID-19 side effects), and 9 (leasing) had a sentiment score below the average.

**Figure 2.** Average sentiment score of each topic. Topics are as follows: 1=family activities, 2=remote work or study, 3=quarantine, 4=dressing, 5=government and policy, 6=COVID-19 side effects, 7=encouragement, 8=back to office, and 9=leasing. LDA: latent Dirichlet allocation.



**Money and Jobs are Discussed the Most When Government Accounts are Mentioned**

In topic 5 (government and policy), “money” and “job” were the most prominent keywords churned by the LDA model. On further exploring the tweets under this topic, we found that a

number of tweets mentioned government accounts and governor twitter accounts. An example of such a tweet is as follows:

*“@SenBobCasey @SenToomey @GovernorTomWolf supply chain workers discouraged. Work from home pieces of the chain are essential. TEMPORARY layoffs*



*making more than I am to wait to get called back to work, where's the incentive to work? Why aren't we included in stimulus 2.0?"*

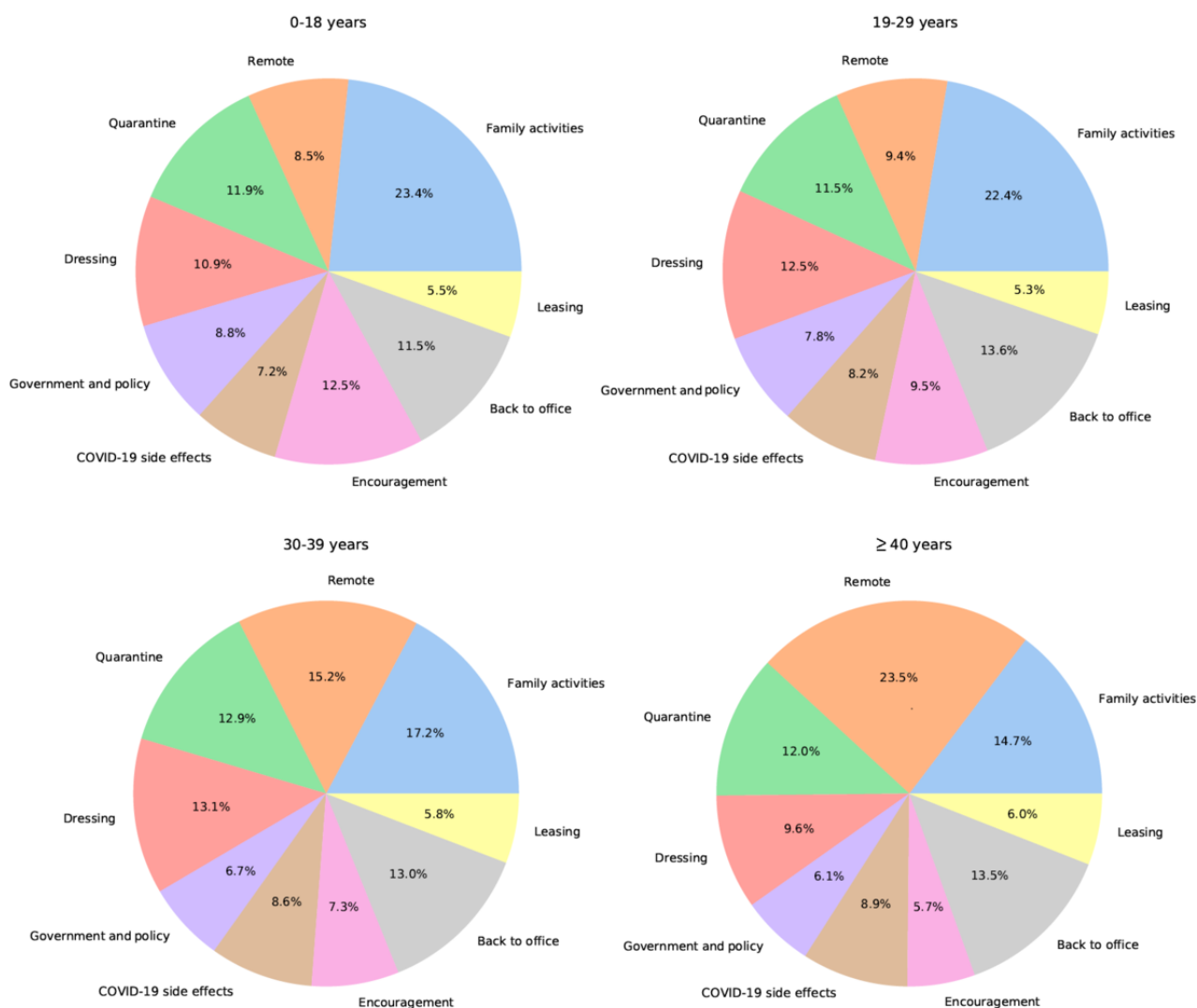
In addition, we also found some tweets about money, where people are somewhat worried about their financial status during the pandemic, such as losing money after being laid off from their jobs. Based on these findings, since April 2020 still marks the early stage of WFH, the economy would be a major priority for the government during this period.

### Family Activities and Remote Work or Study Conflicts Among Age Groups

As shown in Figure 3, as age increases, the proportion of people tweeting about remote work or study increased; moreover, lesser people tweeted about family activities. Among people aged 0-18 years, 23.4% of people tweeted about topic 1 (family activities) and 8.5% about topic 2 (remote work or study). In the age group of 19-29 years, 22.4% of people tweeted about

family activities and 9.4% tweeted about remote work or study. Among people in their 30s, 17.2% tweeted about family activities and 15.2% tweeted about remote work or study. Among people older than 40 years, only 14.7% tweeted about family activities and 23.5% tweeted about remote work or study. Overall, topic 2 (remote work or study) is largely a work-related topic, which highlights work-family conflicts. Frone et al [35] reported that family boundaries are more permeable than work boundaries. Interestingly, based on our findings, we conclude that in the WFH environment, family boundaries are becoming more permeable among older people. On average, people aged  $\geq 40$  years accounted for 37.08% of the study population. However, under topic 1 (family activities), only 30.8% of the tweets were from people older than 40 years; however, under topic 2 (remote work or study), 52.1% of the tweets were from people aged  $\geq 40$  years. These interesting patterns are consistent with our findings that family-work boundaries are becoming weaker among older people.

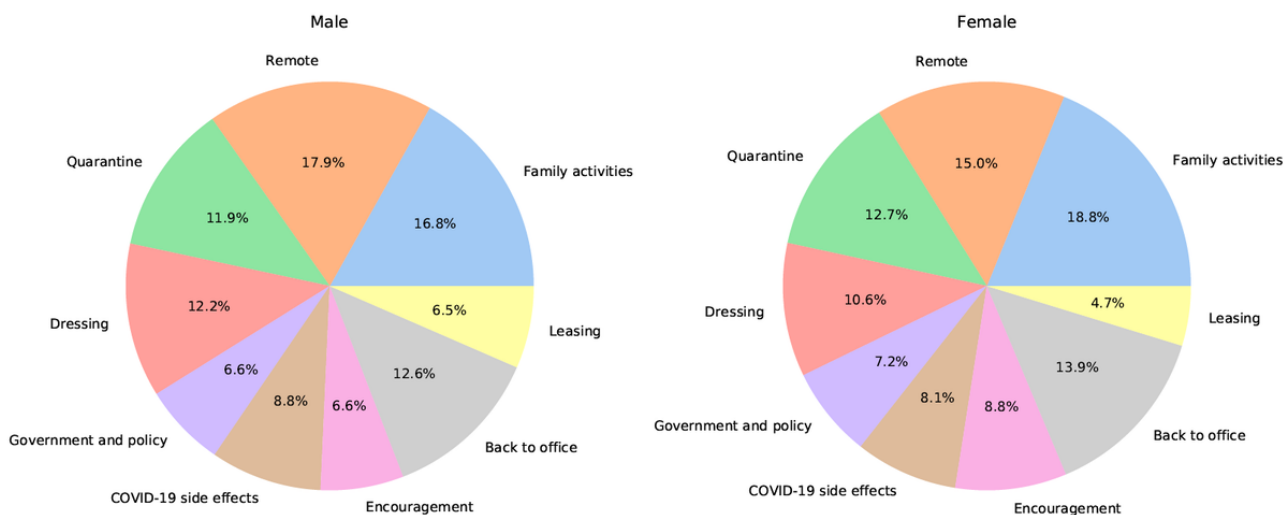
Figure 3. Topic distributions among different age groups.



## Superwomen in WFH

As indicated above, women expressed more positive attitudes to WFH than men. Collins et al [31] reported that this might be because women tend to have more reduced work hours, and we confirmed this finding from the perspective of thematic analysis. In the goodness-of-fit test, we found that the topic distributions among men and women (Figure 4) were significantly different ( $P < .001$ ). Based on the difference between topic distributions among men and women, we speculate that reduced work hours

Figure 4. Topic distribution by gender.



## Discussion

### Principal Findings

This study represents a large-scale quantitative analysis of public opinions on WFH in the United States during the COVID-19 pandemic. Through the lens of social media, we found that gender and age are the most influential features to public opinions about WFH. On performing ordinary least squares regression analysis, we found that the sentiment toward WFH varies across user characteristics. In particular, women are more positive about WFH, which could be related to the change in working styles [30] and reduced work hours compared to those of men [31]. People aged  $\geq 40$  years tend to be the most pro-WFH than other age groups. This could be owing to the fact that people of those ages are the most vulnerable to COVID-19, while also being the most difficult people to be re-employed upon losing their jobs. These people also need to work to mitigate the shrinkage of retirement savings that were invested in the rather inert stock market. People from high-income areas are more likely to have positive opinions about WFH, which echoes the findings of Barrero et al [3].

These nuanced differences are supported by a more fine-grained topic analysis. At a higher level, we found that all the topics expressed a positive sentiment about WFH. However, people expressed a more negative sentiment toward family activity and the government. Under the topic of family activity, we noticed that women pay more attention to family than men, and we identified superwomen in WFH. When people talk about government and policy, money and jobs are 2 major concerns.

allow women to spend more time with their children and take care of their families. Topic 1 (family activities) is 1 of the topics to which women pay more attention. On an in-depth analysis of the tweets, we found that many tweets were about spending time with children while working from home. An example of such a tweet is as follows:

*“That’d be me. I get to work from home and be with my kids. I’m loving every minute of this time with them!”*

Furthermore, based on our analysis by age groups, we noticed that the family-work boundary is another issue that varies among different age groups. As age increases, more people prefer to discuss work rather than family, which implies that family boundaries are becoming more permeable than work boundaries.

### Implications

Barrero et al [3] reported that WFH would persist even after the pandemic. It is critical to understand public opinions on WFH to help improve their experience and to design a more suitable and flexible work policy. Our study suggests that there are nuanced differences across user characteristics. Government and company policymakers could design a more customized work policy to not only increase work productivity but also improve work satisfaction among their employees. It is also important to address the WFH-related disparities that have been reported among different racial and socioeconomic groups [36,37].

### Limitations

Our study is focused on the relationship between user characteristics and the sentiment about WFH. However, user occupation can be included in future analyses. Since the ability to WFH varies among different jobs [28], 1 potential hypothesis could be that people of different occupations have different opinions about WFH; thus, occupations would have an impact on the sentiment of WFH. In addition, there are some limitations of only using the Ethnicolr API to infer ethnicity. The Ethnicolr API is trained on voter registration data from Florida. First, using data from a single state (albeit a representative state) may not be ideal, since the pattern of names can be different among

states. Another limitation is that our training data set was imbalanced (8,757,268 non-Hispanic White people, 1,853,690 non-Hispanic Black people, 2,179,106 Hispanic people, and 253,808 Asian people). Although these numbers are consistent with the population distribution in the United States, when training an inference model, we believe that the use of a more balanced data set could provide a better outcome.

## Conclusions

This is a large-scale social media-based study on people who are more likely to tweet about WFH. On performing ordinary

least squares regression analysis, our study shows how the sentiment of WFH varies across user characteristics. On conducting a content-based analysis, we carried out an in-depth analysis to determine what Twitter users mainly discuss and how the content of their tweets correlates with the sentiment of WFH. This paper contributes to a better understanding of public opinions on WFH in the United States during the COVID-19 pandemic and contributes to making policies both at national and institution or company levels to improve the overall population's experience of WFH.

## Authors' Contributions

All authors conceived and designed the study. HL collected the data. ZX and PL performed feature inference. PL conducted sentiment analysis. ZX applied the LDA models. ZX and PL analyzed the data and wrote the majority of the manuscript. All authors critically revised the manuscript.

## Conflicts of Interest

None declared.

## References

1. COVID Data Tracker. Centers for Disease Control and Prevention. URL: <https://covid.cdc.gov/covid-data-tracker> [accessed 2021-03-25]
2. Baker M. Gartner HR Survey Reveals 88% of Organizations Have Encouraged or Required Employees to Work From Home Due to Coronavirus. Gartner. 2020. URL: <https://www.gartner.com/en/newsroom/press-releases/2020-03-19-gartner-hr-survey-reveals-88--of-organizations-have-e> [accessed 2021-03-25]
3. Barrero J, Bloom N, Davis S. Why Working From Home Will Stick. SSRN Journal 2020. [doi: [10.2139/ssrn.3741644](https://doi.org/10.2139/ssrn.3741644)]
4. Feng Z, Savani K. Covid-19 created a gender gap in perceived work productivity and job satisfaction: implications for dual-career parents working from home. *GM* 2020 Sep 07;35(7/8):719-736. [doi: [10.1108/gm-07-2020-0202](https://doi.org/10.1108/gm-07-2020-0202)]
5. Palumbo R. Let me go to the office! An investigation into the side effects of working from home on work-life balance. *Int J Public Sect Manag* 2020 Oct 07;33(6/7):771-790. [doi: [10.1108/ijpsm-06-2020-0150](https://doi.org/10.1108/ijpsm-06-2020-0150)]
6. Chung H, Seo H, Forbes S, Birkett H. Working from home during the COVID-19 lockdown: Changing preferences and the future of work. University of Birmingham. URL: [https://kar.kent.ac.uk/83896/1/Working\\_from\\_home\\_COVID-19\\_lockdown.pdf](https://kar.kent.ac.uk/83896/1/Working_from_home_COVID-19_lockdown.pdf) [accessed 2021-07-21]
7. Raišienė AG, Rapuano V, Varkulevičiūtė K, Stachová K. Working from Home—Who Is Happy? A Survey of Lithuania's Employees during the COVID-19 Quarantine Period. *Sustainability* 2020 Jul 01;12(13):5332. [doi: [10.3390/su12135332](https://doi.org/10.3390/su12135332)]
8. Bick A, Blandin A, Mertens K. Work from Home Before and After the COVID-19 Outbreak. *SSRN Journal* 2021 [FREE Full text] [doi: [10.2139/ssrn.3786142](https://doi.org/10.2139/ssrn.3786142)]
9. Carroll F, Mostafa M, Thorne S. Working from home: Twitter reveals why we're embracing it. *The Conversation*. 2020. URL: <https://theconversation.com/working-from-home-twitter-reveals-why-were-embracing-it-136760> [accessed 2021-07-21]
10. Yeung N, Lai J, Luo J. Face Off: Polarized Public Opinions on Personal Face Mask Usage during the COVID-19 Pandemic. 2020 Presented at: 2020 IEEE International Conference on Big Data; December 10-13, 2020; Atlanta, GA. [doi: [10.1109/bigdata50022.2020.9378114](https://doi.org/10.1109/bigdata50022.2020.9378114)]
11. Hutto C, Gilbert E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. 2014 Presented at: Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14); June 1-4, 2014; Ann Arbor, MI URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>
12. 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/)]
13. Duong V, Luo J, Pham P, Yang T, Wang Y. The Ivory Tower Lost: How College Students Respond Differently than the General Public to the COVID-19 Pandemic. 2020 Presented at: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); December 7-10, 2020; The Hague. [doi: [10.1109/asonam49781.2020.9381379](https://doi.org/10.1109/asonam49781.2020.9381379)]
14. Lyu H, Wang J, Wu W, Duong V, Zhang X, Dye T, et al. Social Media Study of Public Opinions on Potential COVID-19 Vaccines: Informing Dissent, Disparities, and Dissemination. *arXiv*. Preprint posted online December 3, 2020 [FREE Full text] [doi: [10.1101/2020.12.12.20248070](https://doi.org/10.1101/2020.12.12.20248070)]
15. Wu W, Lyu H, Luo J. Characterizing Discourse about COVID-19 Vaccines: A Reddit Version of the Pandemic Story. *arXiv*. Preprint posted online January 15, 2021 [FREE Full text]

16. Talbot J, Charron V, Konkle A. Feeling the Void: Lack of Support for Isolation and Sleep Difficulties in Pregnant Women during the COVID-19 Pandemic Revealed by Twitter Data Analysis. *Int J Environ Res Public Health* 2021 Jan 06;18(2):393 [FREE Full text] [doi: [10.3390/ijerph18020393](https://doi.org/10.3390/ijerph18020393)] [Medline: [33419145](https://pubmed.ncbi.nlm.nih.gov/33419145/)]
17. Zhang Y, Lyu H, Liu Y, Zhang X, Wang Y, Luo J. Monitoring Depression Trend on Twitter during the COVID-19 Pandemic: Observational Study. *JMIR Infodemiol* 2021;1(1):e26769 [FREE Full text] [doi: [10.2196/26769](https://doi.org/10.2196/26769)]
18. Blei D, Ng A, Jordan M. Latent Dirichlet Allocation. *J Mach Learn Res* 2003;3:993-1022 [FREE Full text]
19. Dahik A, Lovich D, Kreafler C, Bailey A, Kilmann J, Kennedy D, et al. What 12,000 Employees Have to Say About the Future of Remote Work. Boston Consulting Group. 2020 Aug 11. URL: <https://www.bcg.com/publications/2020/valuable-productivity-gains-covid-19> [accessed 2021-03-25]
20. Jin X, Gallagher A, Cao L, Han J, Luo J. The wisdom of social multimedia: using flickr for prediction and forecast. 2010 Presented at: MM '10: ACM Multimedia Conference; October 25-29, 2010; Firenze. [doi: [10.1145/1873951.1874196](https://doi.org/10.1145/1873951.1874196)]
21. Thistle S. *From Marriage to the Market: The Transformation of Women's Lives and Work*. Berkeley, CA: University of California Press; 2006.
22. Wang Z, Hale S, Adelani D, Grabowicz P, Hartmann T, Flöck F, et al. Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. 2019 Presented at: WWW '19: The World Wide Web Conference; May 13-17, 2019; San Francisco, CA. [doi: [10.1145/3308558.3313684](https://doi.org/10.1145/3308558.3313684)]
23. Burger J, Henderson J, Kim G, Zarrella G, The MITRE Corporation. Discriminating Gender on Twitter. Defense Technical Information Center. 2011 Jan. URL: <https://apps.dtic.mil/sti/pdfs/AD1108485.pdf> [accessed 2021-07-21]
24. Wojcik S, Hughes A. Sizing Up Twitter Users. Pew Research Center. 2019. URL: <https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/> [accessed 2021-03-25]
25. Sood G, Laohaprapanon S. Predicting Race and Ethnicity From the Sequence of Characters in a Name. arXiv. Preprint posted online May 5, 2018 [FREE Full text]
26. Quick Facts. United States Census Bureau. URL: <https://www.census.gov/quickfacts/fact/table/US/PST045219> [accessed 2021-03-27]
27. Labor force characteristics by race and ethnicity, 2018. US Bureau of Labor Statistics. 2019. URL: <https://www.bls.gov/opub/reports/race-and-ethnicity/2018/home.htm> [accessed 2021-03-25]
28. Dingel JI, Neiman B. How many jobs can be done at home? *J Public Econ* 2020 Sep;189:104235 [FREE Full text] [doi: [10.1016/j.jpubeco.2020.104235](https://doi.org/10.1016/j.jpubeco.2020.104235)] [Medline: [32834177](https://pubmed.ncbi.nlm.nih.gov/32834177/)]
29. US Cities Factsheet. Center for Sustainable Systems, University of Michigan. URL: <http://css.umich.edu/factsheets/us-cities-factsheet> [accessed 2021-03-27]
30. Tennen-Zapier D. 4 ways remote work is better for women. Fast Company. 2020. URL: <https://www.fastcompany.com/90477102/4-ways-remote-work-is-better-for-women> [accessed 2021-03-24]
31. Collins C, Landivar L, Ruppner L, Scarborough W. COVID-19 and the Gender Gap in Work Hours. *Gen Work Organ* 2020 Jul 02 [FREE Full text] [doi: [10.1111/gwao.12506](https://doi.org/10.1111/gwao.12506)] [Medline: [32837019](https://pubmed.ncbi.nlm.nih.gov/32837019/)]
32. Watkins H. Gen Z and Millennials are Much More Pro-Office than Gen X and Baby Boomers. Hubble. 2021. URL: <https://hubblehq.com/blog/future-of-work-different-age-groups> [accessed 2021-03-25]
33. Older staff most likely to benefit from work from home'. *Financial Times*. URL: <https://www.ft.com/content/46ac277c-da5f-4b99-838d-b0ee228d8c57> [accessed 2021-06-04]
34. Median household income in the top 25 most populated cities in the United States in 2019 (in U.S. dollars). Statista. URL: <https://www.statista.com/statistics/205609/median-household-income-in-the-top-20-most-populated-cities-in-the-us/> [accessed 2021-03-25]
35. Frone M, Russell M, Cooper M. Prevalence of work-family conflict: Are work and family boundaries asymmetrically permeable? *J Organiz Behav* 1992 Dec;13(7):723-729 [FREE Full text] [doi: [10.1002/job.4030130708](https://doi.org/10.1002/job.4030130708)]
36. Chowkwanyun M, Reed AL. Racial Health Disparities and Covid-19 - Caution and Context. *N Engl J Med* 2020 Jul 16;383(3):201-203. [doi: [10.1056/NEJMp2012910](https://doi.org/10.1056/NEJMp2012910)] [Medline: [32374952](https://pubmed.ncbi.nlm.nih.gov/32374952/)]
37. Chang S, Pierson E, Koh PW, Gerardin J, Redbird B, Grusky D, et al. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 2021 Jan;589(7840):82-87. [doi: [10.1038/s41586-020-2923-3](https://doi.org/10.1038/s41586-020-2923-3)] [Medline: [33171481](https://pubmed.ncbi.nlm.nih.gov/33171481/)]

## Abbreviations

- API:** application programming interface
- LDA:** latent Dirichlet allocation
- VADER:** Valence Aware Dictionary and Sentiment Reasoner
- WFH:** work from home

*Edited by C Lovis; submitted 29.03.21; peer-reviewed by J Joo, R Lee; comments to author 19.04.21; revised version received 07.06.21; accepted 10.07.21; published 30.07.21.*

*Please cite as:*

*Xiong Z, Li P, Lyu H, Luo J*

*Social Media Opinions on Working From Home in the United States During the COVID-19 Pandemic: Observational Study*

*JMIR Med Inform 2021;9(7):e29195*

*URL: <https://medinform.jmir.org/2021/7/e29195>*

*doi: [10.2196/29195](https://doi.org/10.2196/29195)*

*PMID: [34254941](https://pubmed.ncbi.nlm.nih.gov/34254941/)*

©Ziyu Xiong, Pin Li, Hanjia Lyu, Jiebo Luo. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 30.07.2021. 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, Acceptance, and Concerns Surrounding App-Based Services to Overcome the COVID-19 Outbreak in South Korea: Web-Based Survey Study

Jihwan Park<sup>1</sup>, PhD; Jinhyun Han<sup>2</sup>, BA; Yerin Kim<sup>3</sup>; Mi Jung Rho<sup>2</sup>, PhD

<sup>1</sup>School of Software Convergence, College of Software Convergence, Dankook University, Yongin-si, Republic of Korea

<sup>2</sup>Department of Urology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

<sup>3</sup>Department of Korean Language and Literature, The Anyang University of Korea, Anyang-si, Republic of Korea

**Corresponding Author:**

Mi Jung Rho, PhD

Department of Urology, Seoul St. Mary's Hospital

College of Medicine

The Catholic University of Korea

222 Banpo-daero, Seocho-Gu

Seoul, 06591

Republic of Korea

Phone: 82 222585905

Email: [romy1018@naver.com](mailto:romy1018@naver.com)

## Abstract

**Background:** Since the COVID-19 outbreak, South Korea has been engaged in various efforts to overcome the pandemic. One of them is to provide app-based COVID-19-related services to the public. As the pandemic continues, a need for various apps has emerged, including COVID-19 apps that can support activities aimed at overcoming the COVID-19 pandemic.

**Objective:** We aimed to determine which apps were considered the most necessary according to users and evaluate the current status of the development of COVID-19-related apps in South Korea. We also aimed to determine users' acceptance and concerns related to using apps to support activities to combat COVID-19.

**Methods:** We collected data from 1148 users from a web-based survey conducted between November 11 and December 6, 2020. Basic statistical analysis, multiple response analysis, and the Wilcoxon rank sum test were performed using R software. We then manually classified the current status of the development of COVID-19-related apps.

**Results:** In total, 68.4% (785/1148) of the respondents showed high willingness to protect themselves from COVID-19 by using related apps. Users considered the epidemiological investigation app to be the most necessary app (709/1148, 61.8%) overall, followed by the self-management app for self-isolation (613/1148, 53.4%), self-route management app (605/1148, 52.7%), COVID-19 symptom management app (483/1148, 42.1%), COVID-19-related information provision app (339/1148, 29.5%), and mental health management app (270/1148, 23.5%). Despite the high intention to use these apps, users were also concerned about privacy issues and media exposure. Those who had an underlying disease and had experience using COVID-19-related apps showed significantly higher intentions to use those apps ( $P=.05$  and  $P=.01$ , respectively).

**Conclusions:** Targeting users is very important in order to design and develop the most necessary apps. Furthermore, to gain the public's trust and make the apps available to as many people as possible, it is vital to develop diverse apps in which privacy protection is maximized.

(*JMIR Med Inform* 2021;9(7):e29315) doi:[10.2196/29315](https://doi.org/10.2196/29315)

**KEYWORDS**

COVID-19; app-based services; acceptance; concerns; epidemiological investigation, self-route management app, privacy

## Introduction

### Background

Since the outbreak of COVID-19, countries worldwide have been engaging in various efforts to overcome the challenges associated with it. One of these efforts include providing app-based services, such as COVID-19 contact tracing apps, to support activities aimed at combating COVID-19 [1-4].

South Korea has been integrating digital technology to make it applicable to all fields [5,6], including surveillance, testing, contact tracing, and self-isolation, as well as apps providing COVID-19-related information. Several COVID-19-related apps have been developed and are currently being used, including the Self-Quarantine Safety Protection app [7] and apps for COVID-19 symptom management app and self-isolation. These apps have greatly helped South Korea in responding to the COVID-19 crisis. However, as the COVID-19 pandemic continues, the need for more diverse apps is emerging, such as COVID-19 vaccine apps [8], epidemiological investigation apps, self-route management apps, and mental health management apps.

To support activities aimed at overcoming COVID-19, diverse apps need to be developed for specific purposes. It is also vital to ensure that the majority of people can be assisted through these apps. Therefore, to ensure the effectiveness of COVID-19-related apps, we need to learn more about the apps that people need, as well as their acceptance and concerns regarding using these apps, for example, concerns regarding security issues. However, although security and information protection issues may arise while developing and using these apps [5,9-11], the use of technology during COVID-19 has focused on using larger amounts of personal data to contain the spread of COVID-19 [12], rather than reflecting on users' intentions or concerns.

For COVID-19-related technologies to be effective, most people need to be able to use them. To achieve this, we need to focus on users' intentions and concerns, rather than adopting a technical approach [13]. Therefore, in this study, we aimed to determine the most necessary apps as preferred by users and identify the current status of the development of COVID-19-related apps in South Korea. Furthermore, we aimed to determine users' acceptance and concerns related to using apps to overcome the COVID-19 crisis.

### Current Status of Development of COVID-19-Related Apps in South Korea

We organized various COVID-19-related apps developed in South Korea according to their release date (Table 1). Thus far, these apps can be classified according to the following main function types: (1) COVID-19-related information provision, (2) COVID-19 symptom management, (3) COVID-19 self-diagnosis, (4) self-route management, (5) mapping of COVID-19 cases, and (6) self-report of confirmed COVID-19 cases.

In early 2020, there were many apps providing COVID-19-related information, but over time, these evolved into COVID-19 symptom management and self-route management apps. Informational apps focus on providing information on the current status of COVID-19; subsequently, information relevant to the present state of COVID-19, such as information about masks and vaccines, is gradually updated and modified to remain relevant. For apps related to self-route management and mapping of COVID-19 cases, however, information is automatically saved using GPS or a QR code. Furthermore, these apps feature a function notifying users of the risk rate, such as mapping confirmed persons with COVID-19. Detailed information about the apps can be found in Table S1 of [Multimedia Appendix 1](#).

**Table 1.** COVID-19–related apps and app functions in South Korea.

No.	Release date	App name	Functions of COVID-19–related apps					OS
			Information provision	Symptom management	Self-diagnosis	Self-route management	Mapping of cases	
1	February 2020	CORNANOW	✓					Android
2	February 7, 2020	Corona Explorer (코로나 탐색기)	✓					Android
3	February 17, 2020	Corona App (코로나앱)	✓					Android
4	February 25, 2020	Corona contact test (코로나 접촉검사)					✓	Android
5	February 26, 2020	Corona 19 situation board (코로나19 상황판)	✓					Android
6	March 2, 2020	Corona 19 status board (코로나19 현황판)	✓					Android
7	March 6, 2020	Corona 19 Gyeongnam (코로나19 경남)	✓					Android
8	March 6, 2020	Corona compass (코로나침반)	✓				✓	Android
9	March 6, 2020	Corona Map (코로나맵)	✓					Android
10	March 9, 2020	Wear mask (웨어마스크)	✓					Android
11	March 9, 2020	Corona 19 news delivery (코로나19 소식전달)	✓					Android
12	March 10, 2020	Corona pin (코로나핀)	✓					Android
13	March 11, 2020	Coronaga (코로나가)					✓	Android
14	March 11, 2020	NEAR	✓					Android
15	March 12, 2020	Corona Map Wiki (코로나맵위키)	✓					Android
16	March 12, 2020	Carrot Mask	✓					Android
17	March 18, 2020	Mark (마크)	✓					Android
18	March 18, 2020	Coback Plus (코백플러스)	✓					Android
19	March 20, 2020	Where is the mask (마스크어디냐)	✓					Android
20	March 20, 2020	Mask time (마스크타임)	✓					Android
21	March 20, 2020	Let me know (알려줘)	✓					Android
22	March 30, 2020	Corona 19 self-diagnosis (코로나19 자가진단)			✓			Android
23	April 2020	BMC Corona 19 employee guardian BMC (코로나19 직원지킴이)		✓				Android, iOS
24	April 2020	Search for COVID-19 guidelines (코로나19 지침 검색)	✓					Android, iOS
25	April 6, 2020	Corona World (코로나월드)	✓					Android
26	May 22, 2020	JINOSYS	✓			✓		Android

No.	Release date	App name	Functions of COVID-19-related apps					OS
			Information provision	Symptom management	Self-diagnosis	Self-route management	Mapping of cases	
27	June 2020	Incheon Corona 19 freeze (인천 코로나19 꼼작마!)	✓	✓		✓		Android, iOS
28	June 2, 2020	School safety guard (학교 안전지킴이)		✓				Android
29	July 13, 2020	Corona Memo (코로나메모)				✓		Android
30	August 2020	FAMY 2.0				✓		Android
31	August 7, 2020	Corona index (코로나지수)	✓					Android
32	August 10, 2020	Corona Pass (코로나패스)				✓		Android
33	August 31, 2020	Corona detector (코로나 탐지기)					✓	Android
34	September 8, 2020	KFKOREA	✓					Android
35	September 11, 2020	Corona location tracking (코로나 위치추적)				✓		Android
36	October 6, 2020	Avoiding corona (코로나피하go)	✓			✓		Android
37	October 12, 2020	Koala	✓			✓		Android
38	December 1, 2020	Corona Alert (코로나알리미)	✓					Android
39	December 11, 2020	COVID SHIELD	✓					Android
40	December 23, 2020	Corona Safer	✓			✓	✓	Android
41	January 6, 2021	Hanyang Univ. Corona contact tracking app (코로나 접촉 추적앱)				✓	✓	Android
42	January 28, 2021	Corona traffic light (코로나 신호등)	✓					Android
43	February 3, 2021	Corona 19 vaccine reminder (코로나19 백신 알리미)	✓	✓				Android
44	February 8, 2021	All about the corona status (코현모)	✓					Android
45	February 9, 2021	Corona traffic safety (코로나 동선 안심이)				✓		Android, iOS
46	February 16, 2021	Corona magnifier (코로나돋보기)	✓					Android
47	March 3, 2021	Corona bored (코로나지겹다)	✓					Android
48	March 22, 2021	Corona vaccine reminder (코브리움)	✓					Android

## Methods

### Study Sample

We conducted a web-based survey between November 11 and December 6, 2020. The number of confirmed COVID-19 cases during the survey period ranged from 143 (on November 11) to 631 (on December 6). On November 1, 2020, the Korean government announced a plan to reorganize social distancing measures by subdividing social distancing into three to five stages; this came into effect on November 7, 2020. Thus, during the survey period, social distancing levels ranged from stage 1 to stage 2, based on the five stages of social distancing [14].

We had limitations in conducting a survey that included the total Korean population. Therefore, the survey was conducted keeping in mind the cost and time of distributing the questionnaire. In South Korea, as of December 6, 2020, Seoul, Gyeonggi-do, Incheon, and Daegu had the highest number of COVID-19 cases nationwide, accounting for 79% of all COVID-19 cases in South Korea [15].

We posted the survey recruitment notice on bulletin boards of online cafes, such as Korean portal online cafes (NAVER) [16], as well as university and college student community bulletin boards. In addition, a questionnaire was also distributed through referrals from cafe users. A total of 1170 people responded. After duplicate and incorrect responses were excluded, 1148 valid, completed questionnaires were obtained. The survey ended on December 6, where the proportion of survey respondents by region was similar to the proportion of COVID-19 cases by region as of December 6.

### Review of COVID-19–Related Apps and Functions Developed in South Korea

To determine the current status of apps developed in South Korea, we conducted a search on application software downloading services such as the Apple App Store, Google Play Store, and Naver One Store. We aimed to find all COVID-19–related apps developed after January 2020, that is, after the COVID-19 outbreak was reported. We used keywords such as “COVID,” “COVID-19,” “Corona,” “Corona 19,” and “infectious disease.” We excluded COVID-19–related apps developed by the Ministry of the Interior and Safety and the Ministry of Health and Welfare. Thus, we found a total of 54 apps. Among these, overseas apps and apps introduced before the COVID-19 outbreak were excluded. For the remaining 48 apps, two medical informatics professors (JP and MJR) and two researchers (JH and YK) manually organized the app features into categories (described below) over four meetings.

To categorize these apps, it was necessary to largely classify them by app features. However, there was no clear criteria for categorizing the app functions. Based on previous studies [12,17,18], we classified the apps developed in South Korea thus far into the following main function types to determine their current status: (1) COVID-19–related information provision, (2) COVID-19 symptom management, (3) COVID-19 self-diagnosis, (4) self-route management, (5) mapping of COVID-19 cases, and (6) self-report of COVID-19 confirmed cases.

### The Intention to Use COVID-19–Related Apps

We developed a questionnaire determining the intention to use COVID-19–related apps based on previous studies [19,20]. Intention to use is the most frequently used variable in research on technology acceptance and is widely used in the health care field [21,22]. Additionally, the questionnaire items were modified for this study. That is, “intention to use” was defined as the degree to which a user’s behavioral intention indicated their willingness to use COVID-19–related apps. Responses were given on a 5-point Likert scale ranging from 1 = “very unwilling” to 5 = “very willing.”

### Searches of App-Based Services to Support Activities to Combat COVID-19

COVID-19–related apps that were currently deemed as necessary were selected based on the six abovementioned functions. However, we added additional apps, namely the epidemiological investigation app, self-management app for self-isolation, and mental health management app.

Finally, we classified app-based services to support activities to overcome the COVID-19 crisis according to six app types: (1) epidemiological investigation apps, (2) self-management apps for self-isolation, (3) self-route management app, (4) COVID-19 symptom management app, (5) COVID-19–related information provision app, and (6) mental health management app.

### Statistical Analysis

The question asking participants which app services are needed to support activities aimed at overcoming the COVID-19 crisis was a multiple-response question; thus, multiple response analysis was used. The Wilcoxon rank sum test [23] was used to analyze people’s intention to use app-based services required to overcome COVID-19. Basic statistical analysis, multiple response analysis, and Wilcoxon rank sum test were conducted using R software (version 3.6.1). Furthermore, we manually classified the current status of the development of COVID-19–related apps.

### Ethics

The study procedures were carried out in accordance with the Declaration of Helsinki and were approved by the Institutional Review Board of Catholic University (MC20QISI0125). Participants’ data were anonymized to ensure confidentiality was maintained.

## Results

### Participants’ Characteristics

Of the total 1148 respondents, 675 (58.8%) were female and the majority (n=475, 41.4%) were in their 30s (Table 2). The proportion of married respondents was 50.6% (581/1148). Furthermore; 846 (73.7%) of the respondents had a university degree or higher; 592 (51.6%) were employed in professional, managerial, and white-collar jobs; and 128 (11.1%) were medical staff. Moreover, 883 (76.9%) respondents lived in Seoul, Gyeonggi-do, Incheon, and Daegu.



**Table 2.** Demographic characteristics (N=1148).

Characteristic	Participants, n (%)
<b>Gender</b>	
Male	473 (41.2)
Female	675 (58.8)
<b>Age</b>	
18 and 19	14 (1.2)
20-29	342 (29.8)
30-39	475 (41.4)
40-49	238 (20.7)
>50	79 (6.9)
<b>Marital status</b>	
Single	551 (48)
Married	581 (50.6)
Other (including divorced, separated, or widowed)	16 (1.4)
<b>Education</b>	
High school graduation or lower	93 (8.1)
College students	209 (18.2)
University graduation or higher	846 (73.7)
<b>Occupation</b>	
Other or unemployed	56 (4.9)
Service, sales, or production	96 (8.4)
Self-employed or freelancer	98 (8.5)
Office worker, professional, or administrative job	592 (51.6)
Housewife	133 (11.6)
Student	173 (15.1)
<b>Medical profession</b>	
No	1020 (88.9)
Yes	128 (11.1)
<b>Salary (US \$)<sup>a</sup></b>	
<1825.82	73 (6.4)
1825.82-3,651.63	426 (37.1)
3,651.63-5,477.45	330 (28.7)
>5,477.45	319 (27.8)
<b>Location</b>	
Seoul	420 (36.6)
Gyeonggi-do	299 (26)
Daegu Metropolitan City	103 (9)
Incheon Metropolitan City	61 (5.3)
Daejeon	57 (5)
Busan	51 (4.4)
Gyeongsangbuk-do	29 (2.5)
Chungcheongnam-do	25 (2.2)
Gwangju	21 (1.8)

Characteristic	Participants, n (%)
Ulsan Metropolitan City	17 (1.5)
Gyeongsangnam-do	15 (1.3)
Gangwon-do	13 (1.1)
Jeollabuk do	13 (1.1)
Chung-cheong bukdo	10 (0.9)
Sejong City	7 (0.6)
Jeju Special Self-Governing Province	4 (0.3)
Jeollanam-do	3 (0.3)

<sup>a</sup>A currency exchange rate of US \$1= ₩1095.40 is applicable (buy and sell base rate on January 13, 2021).

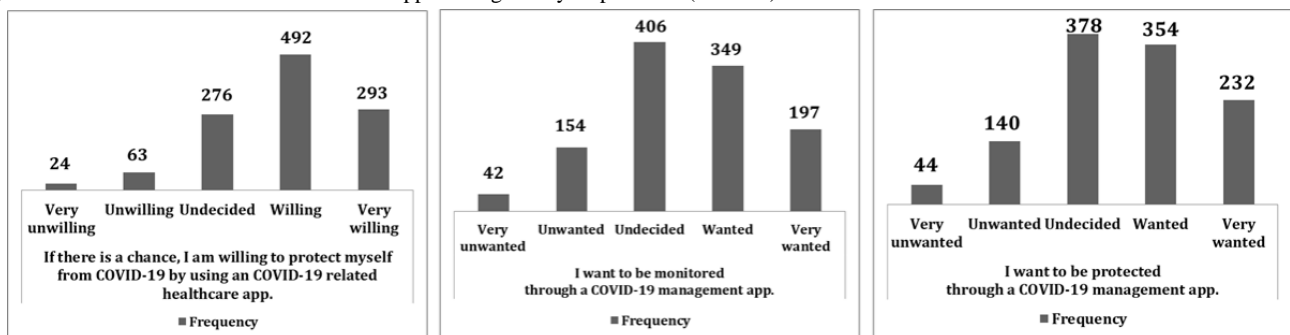
### COVID-19-Related Characteristics

Among the 1148 respondents, 95 (8.3%) had an underlying disease, such as high blood pressure, diabetes, asthma, kidney failure, or tuberculosis; 91 (7.9%) had experienced self-isolation due to COVID-19; 174 (15.2%) had experience with COVID-19 testing; 4 (0.3%) were confirmed COVID-19 cases; 78 (6.8%) had a family member or friend with COVID-19; 362 (31.5%) reported that they had jobs that were easily exposed to COVID-19; and 889 (77.4%) thought that their company was satisfactorily dealing with COVID-19 quarantine measures. Finally, 219 (19.1%) of the respondents had experience using COVID-19-related apps.

### Intention to Use COVID-19-Related Apps

This study assessed participants' willingness to use COVID-19-related apps as shown in Figure 1. The first question asked the 1148 respondents if they were willing to protect themselves from COVID-19 by using COVID-19-related health care apps (Table 3), to which 68.4% (n=785) reported that they were "willing" or "very willing." The second question asked the respondents whether they wanted to be monitored through a COVID-19 management app, to which 47.6% (n=546) of the respondents reporting that they wanted to be monitored. However, 35.4% (n=406) of the respondents had a neutral opinion about this. The last question asked respondents if they wanted to be protected through a COVID-19 management app; 51% (n=586) of the respondents wanted to be protected through a COVID-19 management app, whereas 32.9% (n=378) had a neutral opinion.

**Figure 1.** Intention to use COVID-19-related apps among survey respondents (N=1148).



**Table 3.** Respondents' (N=1148) intention to use COVID-19-related apps and the epidemiological investigation app.

Questions and intention to use the app	Participants, n (%)
<b>COVID-19-related apps</b>	
<b>If one is available, I am willing to protect myself from COVID-19 by using a COVID-19-related health care app.</b>	
Very unwilling	24 (2.1)
Unwilling	63 (5.5)
Undecided	276 (24.0)
Willing	492 (42.9)
Very willing	293 (25.5)
<b>I want to be monitored through a COVID-19 management app.</b>	
Very unwanted	42 (3.7)
Unwanted	154 (13.4)
Undecided	406 (35.4)
Wanted	349 (30.4)
Very wanted	197 (17.2)
<b>I want to be protected through a COVID-19 management app.</b>	
Very unwanted	44 (3.8)
Unwanted	140 (12.2)
Undecided	378 (32.9)
Wanted	354 (30.8)
Very wanted	232 (20.2)
<b>Epidemiological investigation app</b>	
<b>Are you willing to use the epidemiological investigation app?</b>	
Very unwilling	19 (1.7)
Unwilling	41 (3.6)
Undecided	238 (20.7)
Willing	554 (48.3)
Very willing	296 (25.8)

### App-Based Services to Support Activities to Overcome COVID-19

We surveyed which app-based services were needed to support activities aimed at overcoming COVID-19; this was a multiple-response question. Of the 1148 respondents, 709 (61.8%) reported that the epidemiological investigation app was the most necessary service (Table 4). In addition, respondents

stated that a self-management app for self-isolation (613/1148, 53.4%), preventive self-route management app (605/1148, 52.7%), COVID-19 symptom management app (483/1148, 42.1%), and COVID-19-related information provision app (339/1148, 29.5%) were needed. The lowest percentage of responses (270/1148, 23.5%) received were regarding the use of mental health management apps.

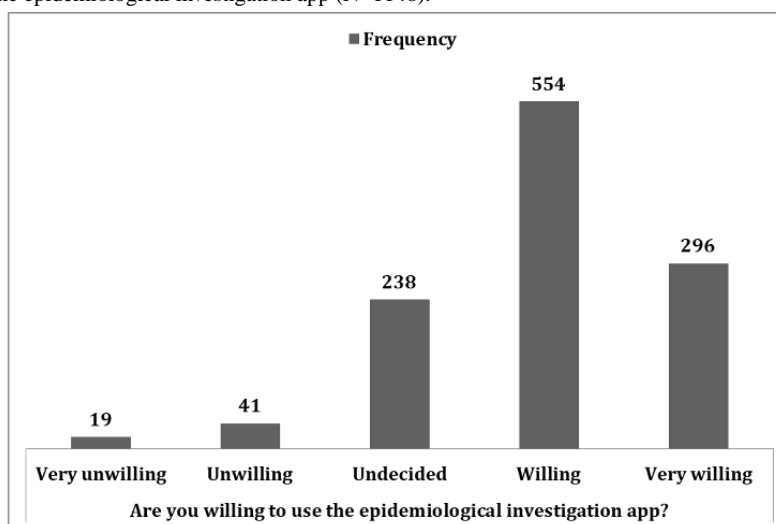
**Table 4.** App-based services needed to support activities to overcome COVID-19.

Question and responses	Value	
	Responses, n (%) (n=3019)	Participants, n (%) (N=1148)
<b>Which app-based services are needed to overcome COVID-19?</b>		
Epidemiological investigation app	709 (23.5)	709 (61.8)
Self-management app for self-isolation	613 (20.3)	613 (53.4)
Self-route management app	605 (20)	605 (52.7)
COVID-19 symptom management app	483 (16)	483 (42.1)
COVID-19-related information provision app	339 (11.2)	339 (29.5)
Mental health management apps	270 (8.9)	270 (23.5)

### Intention to Use and Reasons for Reluctance to Use the Epidemiological Investigations App

First, we inquired whether the respondents were willing to use the epidemiological investigation app, which was reported as the most necessary service. In total, 554 of the 1148 (48.3%) respondents reported that they would use this app, with 296 (25.8%) indicating that they were very willing (Figure 2 and Table 3).

Second, we inquired why they were reluctant to use the app. Regarding this, of the 1148 respondents, 480 (41.8%) of respondents cited privacy concerns, 449 (39.1%) expressed concerns about personal information exposure and media disclosure, and 202 (17.6%) did not have a reason (Table 5). The response rate of those who were not reluctant was very low (13/1148, 1.1%).

**Figure 2.** Willingness to use the epidemiological investigation app (N=1148).**Table 5.** Reasons for reluctance to use the epidemiological investigation app (N=1148).

Question and responses	Participants, n (%)
<b>If you are reluctant to use the epidemiological investigation app, why?</b>	
Privacy invasion problem	480 (41.8)
Personal information exposure and media exposure	449 (39.1)
Criticism and reproach of others	4 (0.3)
Not reluctant	13 (1.1)
No reason	202 (17.6)

### Intention to Use App-Based Services Required to Overcome the COVID-19 Crisis

The intention to use app-based services required to overcome COVID-19 were compared using the Wilcoxon rank sum test.

The various app-based services evaluated in this study were as follows: (1) epidemiological investigation app, (2) self-management app for self-isolation, (3) self-route management app, and (4) COVID-19 symptom management app. The results indicated whether there were any differences

in the intention to use these four apps according to the COVID-19-related characteristics of the respondents. These characteristics included the following: (1) presence of an underlying disease, (2) self-isolation experience, (3) COVID-19 test experience, (4) confirmed COVID-19 cases, (5) family members or friends with a confirmed COVID-19 case, (6) occupations that are easily exposed to COVID-19, (7) a company with good COVID-19 prevention strategies, and (8) experience with COVID-19-related apps (Table 6).

Regarding the presence of underlying disease and COVID-19-related app experience, there were significant differences in respondents' intention to use the epidemiological investigation app, self-management app for self-isolation, self-route management app, and COVID-19 symptom management app. Moreover, those who had an underlying disease and had experience using COVID-19-related apps showed significantly higher intention to use these four apps ( $P=.05$  and  $P=.01$ , respectively; Table 6).



**Table 6.** COVID-19–related characteristics and intention to use apps among survey participants (N=1148).

Variable and intention to use app	Participants, n (%)	Epidemiological investigation app		Self-management app for self-isolation		Self-route management app		COVID-19 symptom management app	
		Mean <sup>a</sup> (SD)	<i>P</i> value	Mean (SD)	<i>P</i> value	Mean (SD)	<i>P</i> value	Mean (SD)	<i>P</i> value
<b>Presence of underlying disease</b>									
No	1053 (91.7)	3.908 (0.871)	<i>.003<sup>b</sup></i>	3.944 (0.871)	<i>.03</i>	3.816 (0.900)	<i>.02</i>	3.885 (0.899)	<i>.04</i>
Yes	95 (8.3)	4.168 (0.794)		4.158 (0.719)		4.021 (0.850)		4.095 (0.787)	
<b>Self-isolation experience</b>									
No	1057 (92.1)	3.923 (0.868)	<i>.28</i>	3.953 (0.869)	<i>.29</i>	3.836 (0.898)	<i>.65</i>	3.900 (0.896)	<i>.68</i>
Yes	91 (7.9)	4.000 (0.869)		4.066 (0.757)		3.791 (0.901)		3.934 (0.854)	
<b>COVID-19 test experience</b>									
No	974 (84.8)	3.941 (0.842)	<i>.75</i>	3.977 (0.841)	<i>.41</i>	3.857 (0.857)	<i>.27</i>	3.912 (0.868)	<i>.97</i>
Yes	174 (15.2)	3.862 (0.999)		3.874 (0.965)		3.695 (1.088)		3.851 (1.020)	
<b>COVID-19 confirmed person</b>									
No	1144 (99.7)	3.929 (0.868)	<i>.96</i>	3.960 (0.860)	<i>.17</i>	3.836 (0.895)	<i>.12</i>	3.904 (0.892)	<i>.26</i>
Yes	4 (0.3)	4.000 (0.816)		4.500 (1.000)		3.000 (1.414)		3.500 (1.000)	
<b>Family or friends with confirmed COVID-19</b>									
No	1070 (93.2)	3.937 (0.851)	<i>.72</i>	3.959 (0.858)	<i>.49</i>	3.823 (0.900)	<i>.15</i>	3.901 (0.891)	<i>.71</i>
Yes	78 (6.8)	3.821 (1.066)		4.000 (0.912)		3.962 (0.860)		3.923 (0.908)	
<b>Occupations that are easily exposed to COVID-19</b>									
No	786 (68.5)	3.926 (0.852)	<i>.55</i>	3.961 (0.844)	<i>.64</i>	3.836 (0.867)	<i>.63</i>	3.912 (0.851)	<i>.91</i>
Yes	362 (31.5)	3.936 (0.902)		3.964 (0.897)		3.826 (0.962)		3.881 (0.976)	
<b>A company with good COVID-19 prevention</b>									
No	259 (22.6)	3.950 (0.841)	<i>.74</i>	4.042 (0.813)	<i>.11</i>	3.876 (0.797)	<i>.58</i>	3.969 (0.830)	<i>.23</i>
Yes	889 (77.4)	3.924 (0.876)		3.938 (0.874)		3.820 (0.925)		3.883 (0.909)	
<b>COVID-19–related app experience</b>									
No	929 (80.9)	3.896 (0.876)	<i>.004</i>	3.931 (0.877)	<i>.02</i>	3.799 (0.895)	<i>.002</i>	3.875 (0.896)	<i>.02</i>
Yes	219 (19.1)	4.073 (0.815)		4.091 (0.779)		3.977 (0.896)		4.018 (0.867)	

<sup>a</sup>Respondents' intention to use response values for each app, measured on a 5-point Likert scale, ranging from 1 = "very unwilling" to 5 = "very willing."

<sup>b</sup>Italicized values indicate statistical significance.

## Discussion

This study aimed to determine the most essential apps required to overcome COVID-19 and the current status of the development of COVID-19-related apps in South Korea. Furthermore, this study aimed to determine users' acceptance of and concerns related to the use of these apps.

First, respondents expressed a high level of willingness to use COVID-19-related apps. Many respondents indicated that they wanted to be protected and monitored by using COVID-19-related apps. However, many also had a neutral opinion. Thus, these apps need to be developed in a way to gain the trust of prospective users.

Second, the need to develop multiple apps emerged, which included epidemiological investigation apps, self-management apps for self-isolation, self-route management apps, COVID-19 symptom management apps (42%), and mental health management apps. Most of the respondents (61.8%) considered the epidemiological investigation app as the most needed app. In addition, the self-management app for self-isolation (53.4%), self-route management app (52.7%), COVID-19 symptom management app (42.1%), and mental health management app (23.5%) were marked as important, in that order.

In South Korea, there exists a self-management app for self-isolation, called "Self-quarantine Safety Protection App" [7,24]. However, based on the survey responses, it appears that various apps for self-management need to be developed. Regarding the self-route management app, apps using GPS and QR are increasingly being released; nevertheless, more apps are needed since their importance continues to increase. To illustrate, the symptom management app helps identify new symptoms of COVID-19 and estimates the predicted value of specific symptoms [25]. In addition, these apps appear to be helpful in developing reliable screening tools. Thus, it is important to develop and utilize symptom management apps that can be used by the general public. Moreover, it was confirmed that there is also a demand for an app that can manage fatigue, mental health, and symptoms such as depression and anxiety caused by working from home and COVID-19 itself. There is ongoing research about COVID-19 survivors [26] and mental illness issues such as depression and anxiety caused by COVID-19 [27-29]. It has been reported that some people experienced worsened mental health after the pandemic [30]. This problem also applies to the medical staff such as physicians and nurses [31-33]. Therefore, active participation from the private sector and government is required to overcome the challenges posed by COVID-19. Furthermore, mental health problems need to be urgently addressed for groups such as COVID-19 survivors, medical staff, and women [29].

In the event of the COVID-19 pandemic, the main purpose of epidemiological investigations is to prevent early spread [34]; therefore, if the epidemiological investigation is delayed, secondary and tertiary disease transmission can occur, and people may become infected without knowing when or how they were infected. However, when a pandemic such as COVID-19 occurs, difficulties in epidemiological investigations and lack of workforce to conduct epidemiological investigation

is often evident [35]. Therefore, there is also an urgent need to increase the number of epidemiological investigators, but this goal is difficult to achieve. To facilitate epidemiological investigations, a system that can actively cooperate with such investigations is needed. Consequently, if an epidemiological investigation app is developed, it can help actively provide basic information and medical records, including one's own movements, at the time of confirmation by the epidemiological investigator [36]. However, respondents expressed concern about infringement of personal information used by these services, such as COVID-19 contact tracing apps based on GPS or smartphone logs. Therefore, like the epidemiological investigation app, a self-route management app is also needed to reduce the fear of personal information infringement and increase the amount of information provided for epidemiological investigations.

Third, the importance of privacy invasion issues of COVID-19-related apps was emphasized in this study. Despite the high intention to use the epidemiological investigation app, people were very concerned about privacy invasion issues, personal information exposure, and media exposure. Thus, it is vital to consider how to resolve people's concerns about using these services, even after the necessary services are developed and available. Similarly, previous studies have found that people did not download and use contact tracing apps due to privacy concerns [37,38]. These findings suggest that it is important to design and develop apps deemed as necessary in order to overcome the COVID-19 crisis; however, to gain the public's trust and make such apps available to many people, minimum amounts of personal information should be used and seek the public interest based on this. In this regard, a service where users have authority over their information should be developed [39].

Fourth, we found that those who had an underlying disease and had experience using COVID-19-related apps showed a significantly higher intention to use apps such as the epidemiological investigation app, self-management app for self-isolation, self-route management app, and COVID-19 symptom management app. Interestingly, even if users reported self-isolation experiences, COVID-19 test experiences, COVID-19 confirmed experiences, and nearby confirmed cases, their willingness to use COVID-19-related apps was not higher. In addition, no differences were found between the intention to use the apps among respondents engaged in occupations that had a relatively high exposure to COVID-19 cases or those employed by companies that complied with COVID-19 quarantine regulations. This is a surprising result; that is, the spread of COVID-19 is prevalent, but this does not directly lead to the use of apps. Hence, understanding people's needs in the current situation is essential. The current findings suggest that a focus on promoting and distributing the service in view of the high intention of use by those with underlying diseases and those who have used COVID-19-related apps should be prioritized. In addition, the app should be promoted and distributed intensively in hospitals or health centers wherein people with underlying diseases may be easily accessible. It would also be beneficial to include a function that recommends other apps over existing apps so that various apps can be

exposed to active prospective users. Furthermore, it should be considered that many people reported that they do want to use apps to overcome the COVID-19 pandemic.

COVID-19-related apps developed and used in South Korea ranged from those providing information to those used for symptom management, COVID-19 self-diagnosis, self-route management, mapping of COVID-19 cases, and reporting of COVID-19 confirmed cases. Based on the information needed at present, the COVID-19-related information app was found to be faithful to its function. As the COVID-19 pandemic continues, the development and use of GPS- and QR-based self-movement management and mapping of COVID-19 case services are needed. However, as mentioned earlier, these apps still have privacy issues. In addition, according to existing studies, the development ratio of Android- and iOS-based apps are similarly developed and used for COVID-19-related apps [12], but in South Korea, COVID-19 apps are mostly based on Android. Thus, a need to develop iOS-based apps for various Korean smartphone users is evident.

Despite the meaningful results discussed thus far, this study has several limitations. First, a total of 1148 survey respondents were analyzed. However, there were only four patients with confirmed COVID-19 among these participants, which is a low rate of 0.3% of the total respondents. Thus, to obtain more

meaningful results, additional samples of COVID-19 confirmed cases should be collected. Second, although there are many COVID-19-related studies, there is limited published literature available. Third, 51.6% of the survey respondents were employed in white-collar jobs and managerial positions. Thus, occupational biases might have influenced the interpretation of results. Finally, to determine the current status of apps developed in South Korea, we conducted a search on application software downloading services. Two medical informatics professors and two researchers manually organized the app features in four meetings. To improve on this method, future research should apply tools to investigate the apps instead.

Despite these limitations, there are meaningful implications of the study's findings. It was found that the COVID-19 apps may support activities aimed at overcoming the COVID-19 pandemic. However, our findings emphasized how several actions and requirements are necessary to accomplish this aim. Our findings further identified the most essential apps, as well as provided future directions for app development to overcome COVID-19. This study also emphasized the need for information protection to guarantee maximum privacy for users, thus increasing the likelihood of more users. Overall, several insights into the development of apps related to COVID-19 were identified, which can be utilized in future developments and improvements of new and existing apps related to COVID-19.

---

## Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R111A1A01072400).

---

## Authors' Contributions

MJR is a principal investigator, and was involved in the research design, questionnaire development, interpretation of results, and writing of the manuscript. JP conducted analyses of the results, interpretation of results, and writing of the manuscript. JH and YK were involved in questionnaire development and interpretation of results, and review of COVID-19-related apps and functions developed in South Korea.

---

## Conflicts of Interest

The corresponding author (MJR) and the first author (JP) are a married couple. There are no competing financial interests.

---

## Multimedia Appendix 1

COVID-19-related apps in South Korea.

[[DOCX File, 2995 KB](#) - [medinform\\_v9i7e29315\\_app1.docx](#) ]

---

## References

1. Lee D, Lee J. Testing on the move: South Korea's rapid response to the COVID-19 pandemic. *Transp Res Interdiscip Perspect* 2020 May;5:100111 [[FREE Full text](#)] [doi: [10.1016/j.trip.2020.100111](https://doi.org/10.1016/j.trip.2020.100111)] [Medline: [34171015](#)]
2. Abeler J, Bäcker M, Buermeier U, Zillessen H. COVID-19 contact tracing and data protection can go together. *JMIR Mhealth Uhealth* 2020 Apr 20;8(4):e19359 [[FREE Full text](#)] [doi: [10.2196/19359](https://doi.org/10.2196/19359)] [Medline: [32294052](#)]
3. Wang S, Ding S, Xiong L. A new system for surveillance and digital contact tracing for COVID-19: spatiotemporal reporting over network and GPS. *JMIR Mhealth Uhealth* 2020 Jun 10;8(6):e19457 [[FREE Full text](#)] [doi: [10.2196/19457](https://doi.org/10.2196/19457)] [Medline: [32499212](#)]
4. Jonker M, de Bekker-Grob E, Veldwijk J, Goossens L, Bour S, Rutten-Van Mülken M. COVID-19 contact tracing apps: predicted uptake in the Netherlands based on a discrete choice experiment. *JMIR Mhealth Uhealth* 2020 Oct 09;8(10):e20741 [[FREE Full text](#)] [doi: [10.2196/20741](https://doi.org/10.2196/20741)] [Medline: [32795998](#)]
5. Park S, Choi GJ, Ko H. Information technology-based tracing strategy in response to COVID-19 in South Korea-privacy controversies. *JAMA* 2020 Jun 02;323(21):2129-2130. [doi: [10.1001/jama.2020.6602](https://doi.org/10.1001/jama.2020.6602)] [Medline: [32324202](#)]

6. Choi J, Lee S, Jamal T. Smart Korea: Governance for smart justice during a global pandemic. *Journal of Sustainable Tourism* 2020 Jun 10;29(2-3):541-550. [doi: [10.1080/09669582.2020.1777143](https://doi.org/10.1080/09669582.2020.1777143)]
7. Guide on the Installation of Self-quarantine Safety Protection App 2020. Central Disaster and Safety Countermeasures Headquarters, CDSCHQ. URL: [http://ncov.mohw.go.kr/upload/ncov/file/202004/1585732793827\\_20200401181953.pdf](http://ncov.mohw.go.kr/upload/ncov/file/202004/1585732793827_20200401181953.pdf) [accessed 2021-07-12]
8. Dasgupta N, Lazard A, Brownstein JS. Covid-19 vaccine apps should deliver more to patients. *The Lancet Digital Health* 2021 May;3(5):e278-e279. [doi: [10.1016/s2589-7500\(21\)00021-2](https://doi.org/10.1016/s2589-7500(21)00021-2)]
9. Ahmed N, Michelin RA, Xue W, Ruj S, Malaney R, Kanhere SS, et al. A survey of COVID-19 contact tracing apps. *IEEE Access* 2020;8:134577-134601. [doi: [10.1109/access.2020.3010226](https://doi.org/10.1109/access.2020.3010226)]
10. Morley J, Cowls J, Taddeo M, Floridi L. Ethical guidelines for COVID-19 tracing apps. *Nature* 2020 Jun;582(7810):29-31. [doi: [10.1038/d41586-020-01578-0](https://doi.org/10.1038/d41586-020-01578-0)] [Medline: [32467596](https://pubmed.ncbi.nlm.nih.gov/32467596/)]
11. Idrees SM, Nowostawski M, Jameel R. Blockchain-based digital contact tracing apps for covid-19 pandemic management: issues, challenges, solutions, and future directions. *JMIR Med Inform* 2021 Feb 09;9(2):e25245 [FREE Full text] [doi: [10.2196/25245](https://doi.org/10.2196/25245)] [Medline: [33400677](https://pubmed.ncbi.nlm.nih.gov/33400677/)]
12. Collado-Borrell R, Escudero-Vilaplana V, Villanueva-Bueno C, Herranz-Alonso A, Sanjurjo-Saez M. Features and functionalities of smartphone apps related to COVID-19: systematic search in app stores and content analysis. *J Med Internet Res* 2020 Aug 25;22(8):e20334 [FREE Full text] [doi: [10.2196/20334](https://doi.org/10.2196/20334)] [Medline: [32614777](https://pubmed.ncbi.nlm.nih.gov/32614777/)]
13. Luciano F. Mind the app-considerations on the ethical risks of COVID-19 apps. *Philos Technol* 2020 Jun 13;1-6 [FREE Full text] [doi: [10.1007/s13347-020-00408-5](https://doi.org/10.1007/s13347-020-00408-5)] [Medline: [32837867](https://pubmed.ncbi.nlm.nih.gov/32837867/)]
14. COVID-19 Outbreak Status. Webpage in Korean. KOSIS (Korean Statistical Information Service). URL: [https://kosis.kr/covid/covid\\_index.do](https://kosis.kr/covid/covid_index.do) [accessed 2021-07-12]
15. Coronavirus Infectious Disease-19 Outbreak in Korea (December 6) 2020. Agency KDCaP. URL: <http://ncov.mohw.go.kr/tcmBoardView.do?contSeq=361521#> [accessed 2021-07-12]
16. NAVER. Webpage in Korean. URL: <https://www.naver.com/> [accessed 2021-07-12]
17. Kondylakis H, Katehakis DG, Kouroubali A, Logothetidis F, Triantafyllidis A, Kalamaras I, et al. COVID-19 mobile apps: a systematic review of the literature. *J Med Internet Res* 2020 Dec 09;22(12):e23170 [FREE Full text] [doi: [10.2196/23170](https://doi.org/10.2196/23170)] [Medline: [33197234](https://pubmed.ncbi.nlm.nih.gov/33197234/)]
18. Ming LC, Untong N, Aliudin NA, Osili N, Kifli N, Tan CS, et al. Mobile health apps on covid-19 launched in the early days of the pandemic: content analysis and review. *JMIR Mhealth Uhealth* 2020 Sep 16;8(9):e19796 [FREE Full text] [doi: [10.2196/19796](https://doi.org/10.2196/19796)] [Medline: [32609622](https://pubmed.ncbi.nlm.nih.gov/32609622/)]
19. Venkatesh, Morris, Davis, Davis. User acceptance of information technology: toward a unified view. *MIS Quarterly* 2003;27(3):425. [doi: [10.2307/30036540](https://doi.org/10.2307/30036540)]
20. Rho MJ, Kim HS, Chung K, Choi IY. Factors influencing the acceptance of telemedicine for diabetes management. *Cluster Comput* 2014 Mar 12;18(1):321-331. [doi: [10.1007/s10586-014-0356-1](https://doi.org/10.1007/s10586-014-0356-1)]
21. Rho MJ, Kim H, Sun C, Wang G, Yoon K, Choi IY. Comparison of the acceptance of telemonitoring for glucose management between South Korea and China. *Telemed J E Health* 2017 Nov;23(11):881-890. [doi: [10.1089/tmj.2016.0217](https://doi.org/10.1089/tmj.2016.0217)] [Medline: [28598260](https://pubmed.ncbi.nlm.nih.gov/28598260/)]
22. Zhang Y, Liu C, Luo S, Xie Y, Liu F, Li X, et al. Factors influencing patients' intentions to use diabetes management apps based on an extended unified theory of acceptance and use of technology model: web-based survey. *J Med Internet Res* 2019 Aug 13;21(8):e15023 [FREE Full text] [doi: [10.2196/15023](https://doi.org/10.2196/15023)] [Medline: [31411146](https://pubmed.ncbi.nlm.nih.gov/31411146/)]
23. McKnight P, Najab J. Mann-Whitney U test. *The Corsini Encyclopedia of Psychology* 2010 Jan 30:1-1. [doi: [10.1002/9780470479216.corpsy0524](https://doi.org/10.1002/9780470479216.corpsy0524)]
24. Self-quarantine Safety Protection (App). Webpage in Korean. Ministry of the Interior and Safety. URL: <https://apps.apple.com/kr/app/%EC%9E%90%EA%B0%80%EA%B2%A9%EB%A6%AC%EC%9E%90-%EC%95%88%EC%A0%84%EB%B3%B4%ED%98%B8/id1502372537> [accessed 2021-07-12]
25. Zens M, Brammert A, Herpich J, Südkamp N, Hinterseer M. App-based tracking of self-reported COVID-19 symptoms: analysis of questionnaire data. *J Med Internet Res* 2020 Sep 09;22(9):e21956 [FREE Full text] [doi: [10.2196/21956](https://doi.org/10.2196/21956)] [Medline: [32791493](https://pubmed.ncbi.nlm.nih.gov/32791493/)]
26. Mazza MG, De Lorenzo R, Conte C, Poletti S, Vai B, Bollettini I, COVID-19 BioB Outpatient Clinic Study group, et al. Anxiety and depression in COVID-19 survivors: role of inflammatory and clinical predictors. *Brain Behav Immun* 2020 Oct;89:594-600 [FREE Full text] [doi: [10.1016/j.bbi.2020.07.037](https://doi.org/10.1016/j.bbi.2020.07.037)] [Medline: [32738287](https://pubmed.ncbi.nlm.nih.gov/32738287/)]
27. Hyland P, Shevlin M, McBride O, Murphy J, Karatzias T, Bentall RP, et al. Anxiety and depression in the Republic of Ireland during the COVID-19 pandemic. *Acta Psychiatr Scand* 2020 Sep;142(3):249-256. [doi: [10.1111/acps.13219](https://doi.org/10.1111/acps.13219)] [Medline: [32716520](https://pubmed.ncbi.nlm.nih.gov/32716520/)]
28. Lebel C, MacKinnon A, Bagshawe M, Tomfohr-Madsen L, Giesbrecht G. Elevated depression and anxiety symptoms among pregnant individuals during the COVID-19 pandemic. *J Affect Disord* 2020 Dec 01;277:5-13 [FREE Full text] [doi: [10.1016/j.jad.2020.07.126](https://doi.org/10.1016/j.jad.2020.07.126)] [Medline: [32777604](https://pubmed.ncbi.nlm.nih.gov/32777604/)]



29. Salari N, Hosseini-Far A, Jalali R, Vaisi-Raygani A, Rasoulpoor S, Mohammadi M, et al. Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: a systematic review and meta-analysis. *Global Health* 2020 Jul 06;16(1):57 [FREE Full text] [doi: [10.1186/s12992-020-00589-w](https://doi.org/10.1186/s12992-020-00589-w)] [Medline: [32631403](https://pubmed.ncbi.nlm.nih.gov/32631403/)]
30. Choi EPH, Hui BPH, Wan EYF. Depression and anxiety in Hong Kong during COVID-19. *Int J Environ Res Public Health* 2020 May 25;17(10):3740 [FREE Full text] [doi: [10.3390/ijerph17103740](https://doi.org/10.3390/ijerph17103740)] [Medline: [32466251](https://pubmed.ncbi.nlm.nih.gov/32466251/)]
31. Elbay RY, Kurtulmuş A, Arpacıoğlu S, Karadere E. Depression, anxiety, stress levels of physicians and associated factors in Covid-19 pandemics. *Psychiatry Res* 2020 Aug;290:113130 [FREE Full text] [doi: [10.1016/j.psychres.2020.113130](https://doi.org/10.1016/j.psychres.2020.113130)] [Medline: [32497969](https://pubmed.ncbi.nlm.nih.gov/32497969/)]
32. Labrague L, De Los Santos JAA. COVID-19 anxiety among front-line nurses: Predictive role of organisational support, personal resilience and social support. *J Nurs Manag* 2020 Oct;28(7):1653-1661 [FREE Full text] [doi: [10.1111/jonm.13121](https://doi.org/10.1111/jonm.13121)] [Medline: [32770780](https://pubmed.ncbi.nlm.nih.gov/32770780/)]
33. Liu Z, Han B, Jiang R, Huang Y, Ma C, Wen J, et al. Mental health status of doctors and nurses during COVID-19 epidemic in China. *SSRN Preprints*. Preprint posted online on March 18, 2020. [doi: [10.2139/ssrn.3551329](https://doi.org/10.2139/ssrn.3551329)]
34. Adhikari SP, Meng S, Wu Y, Mao Y, Ye R, Wang Q, et al. Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review. *Infect Dis Poverty* 2020 Mar 17;9(1):29 [FREE Full text] [doi: [10.1186/s40249-020-00646-x](https://doi.org/10.1186/s40249-020-00646-x)] [Medline: [32183901](https://pubmed.ncbi.nlm.nih.gov/32183901/)]
35. Yong SEF, Anderson DE, Wei WE, Pang J, Chia WN, Tan CW, et al. Connecting clusters of COVID-19: an epidemiological and serological investigation. *Lancet Infect Dis* 2020 Jul;20(7):809-815 [FREE Full text] [doi: [10.1016/S1473-3099\(20\)30273-5](https://doi.org/10.1016/S1473-3099(20)30273-5)] [Medline: [32330439](https://pubmed.ncbi.nlm.nih.gov/32330439/)]
36. Yamamoto K, Takahashi T, Urasaki M, Nagayasu Y, Shimamoto T, Tateyama Y, et al. Health observation app for COVID-19 symptom tracking integrated with personal health records: proof of concept and practical use study. *JMIR Mhealth Uhealth* 2020 Jul 06;8(7):e19902 [FREE Full text] [doi: [10.2196/19902](https://doi.org/10.2196/19902)] [Medline: [32568728](https://pubmed.ncbi.nlm.nih.gov/32568728/)]
37. Chan EY, Saqib NU. Privacy concerns can explain unwillingness to download and use contact tracing apps when COVID-19 concerns are high. *Comput Human Behav* 2021 Jun;119:106718 [FREE Full text] [doi: [10.1016/j.chb.2021.106718](https://doi.org/10.1016/j.chb.2021.106718)] [Medline: [33526957](https://pubmed.ncbi.nlm.nih.gov/33526957/)]
38. Ekong I, Chukwu E, Chukwu M. COVID-19 mobile positioning data contact tracing and patient privacy regulations: exploratory search of global response strategies and the use of digital tools in Nigeria. *JMIR Mhealth Uhealth* 2020 Apr 27;8(4):e19139 [FREE Full text] [doi: [10.2196/19139](https://doi.org/10.2196/19139)] [Medline: [32310817](https://pubmed.ncbi.nlm.nih.gov/32310817/)]
39. Okerefor K, Adebola O. Tackling the cybersecurity impacts of the coronavirus outbreak as a challenge to internet safety. *Int J IT Eng* 2020 Feb;8(2).

## Abbreviations

**NRF:** National Research Foundation of Korea

*Edited by C Lovis; submitted 01.04.21; peer-reviewed by S Rostam Niakan Kalhori, K Halttu; comments to author 25.04.21; revised version received 16.05.21; accepted 17.06.21; published 30.07.21.*

*Please cite as:*

*Park J, Han J, Kim Y, Rho MJ*

*Development, Acceptance, and Concerns Surrounding App-Based Services to Overcome the COVID-19 Outbreak in South Korea: Web-Based Survey Study*

*JMIR Med Inform* 2021;9(7):e29315

URL: <https://medinform.jmir.org/2021/7/e29315>

doi: [10.2196/29315](https://doi.org/10.2196/29315)

PMID: [34137726](https://pubmed.ncbi.nlm.nih.gov/34137726/)

©Jihwan Park, Jinhyun Han, Yerin Kim, Mi Jung Rho. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org/>), 30.07.2021. 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 Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities

Muhammad Ayaz<sup>1</sup>, MSc; Muhammad F Pasha<sup>1</sup>, PhD; Mohammed Y Alzahrani<sup>2</sup>, PhD; Rahmat Budiarto<sup>3</sup>, PhD; Deris Stiawan<sup>4</sup>, PhD

<sup>1</sup>Malaysia School of Information Technology, Monash University, Bandar Sunway, Malaysia

<sup>2</sup>Information Technology Department, College of Computer Science & Information Technology, Albaha University, Albaha, Saudi Arabia

<sup>3</sup>Informatics Department, Faculty of Science & Technology, Universitas Alazhar Indonesia, Jakarta, Indonesia

<sup>4</sup>Department of Computer Engineering, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia

**Corresponding Author:**

Muhammad Ayaz, MSc

Malaysia School of Information Technology

Monash University

Jalan Lagoon Selatan

Bandar Sunway, 47500

Malaysia

Phone: 60 0355146224

Email: [Muhammad.ayaz@monash.edu](mailto:Muhammad.ayaz@monash.edu)

**Related Article:**

This is a corrected version. See correction statement: <https://medinform.jmir.org/2021/8/e32869>

## Abstract

**Background:** Information technology has shifted paper-based documentation in the health care sector into a digital form, in which patient information is transferred electronically from one place to another. However, there remain challenges and issues to resolve in this domain owing to the lack of proper standards, the growth of new technologies (mobile devices, tablets, ubiquitous computing), and health care providers who are reluctant to share patient information. Therefore, a solid systematic literature review was performed to understand the use of this new technology in the health care sector. To the best of our knowledge, there is a lack of comprehensive systematic literature reviews that focus on Fast Health Interoperability Resources (FHIR)-based electronic health records (EHRs). In addition, FHIR is the latest standard, which is in an infancy stage of development. Therefore, this is a hot research topic with great potential for further research in this domain.

**Objective:** The main aim of this study was to explore and perform a systematic review of the literature related to FHIR, including the challenges, implementation, opportunities, and future FHIR applications.

**Methods:** In January 2020, we searched articles published from January 2012 to December 2019 via all major digital databases in the field of computer science and health care, including ACM, IEEE Explorer, Springer, Google Scholar, PubMed, and ScienceDirect. We identified 8181 scientific articles published in this field, 80 of which met our inclusion criteria for further consideration.

**Results:** The selected 80 scientific articles were reviewed systematically, and we identified open questions, challenges, implementation models, used resources, beneficiary applications, data migration approaches, and goals of FHIR.

**Conclusions:** The literature analysis performed in this systematic review highlights the important role of FHIR in the health care domain in the near future.

(*JMIR Med Inform* 2021;9(7):e21929) doi:[10.2196/21929](https://doi.org/10.2196/21929)

## KEYWORDS

Fast Health Interoperability Resources; FHIR; electronic health record; EHR; clinical document architecture; CDA; Substitutable Medical Applications Reusable Technologies; SMART; HL7; health standard; systematic literature review

## Introduction

### Background

In 2011, the proponent of Australian Health Level Seven (HL7) standards, Grahame Grieve, proposed an interoperability approach called Resources for Healthcare (RFH) as a new standard for better interoperability in digital health. Technically, RFH has been designed for web technology, and the resource is based on extensible markup language (XML) with an HTTP-based representational state transfer (REST)ful protocol and a distinct URL for each resource. The RFH standard was renamed Fast Health Interoperability Resources (FHIR) with extension of previous HL7 specifications (ie, HL7 version 2 and version 3) with consideration of modern web technologies [1].

The main idea behind FHIR was to build a set of resources and develop HTTP-based REST application programming interfaces (APIs) to access and use these resources. FHIR uses components called resources to access and perform operations on patient health data at the granular level. This feature makes FHIR a unique standard from all other standards because it was not available in all previous versions of HL7 (v2, v3) or the HL7 clinical document architecture (CDA).

The basic building blocks of FHIR are the so-called resources, a generic definition of common health care concepts (eg, patient, observation, practitioner, device, condition). FHIR uses JavaScript object notation and XML structures for data exchange and resources serialization. FHIR does not only support a RESTful to exchange resources but also manages and documents an interoperability paradigm.

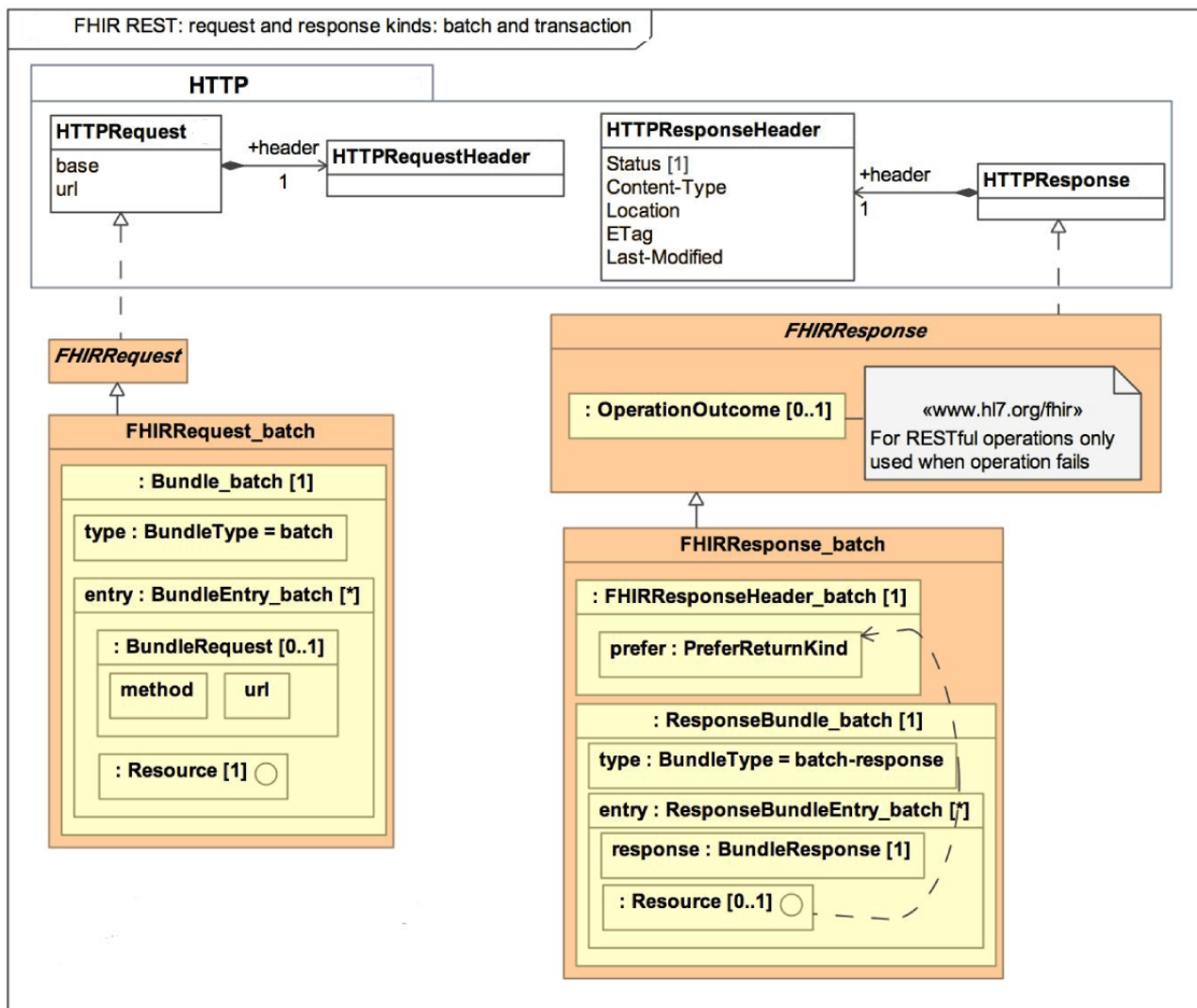
Since the first day of its introduction, FHIR has gained popularity and has been increasingly adopted by the health care industry. In 2018, six large technology companies, including

Microsoft, IBM, Amazon, and Google, pledged to remove barriers for health care interoperability and signed a letter that explicitly mentions FHIR as an emerging standard for the exchange of health data [2]. With incorporation of Substitutable Medical Applications Reusable Technologies (SMART), a platform for interoperable apps [3], FHIR can be expected to attract even more attraction in digital health in the future. Using FHIR for the exchange of medical data can provide potential benefits in a large number of domains, including mobile health apps, electronic health records (EHRs), precision medicine, wearable devices, big data analytics, and clinical decision support.

The main objective of FHIR is to reduce implementation complexity without losing information integrity. Moreover, this new standard combines the advantages of the previous HL7 (v2, v3, and CDA) standards and is expected to overcome their limitations. FHIR allows the developers to develop standardized browser applications that enable the user to access clinical data from any health care system regardless of the operating systems and devices that a health care system uses. For example, a user runs an application on the browser and will access data from a health care system using any device, whether it is running on a desktop, smartphone, Windows, Android, or Linux operating system. [Figure 1](#) represents the general architecture of FHIR [4].

The goal of this study was to gain a deeper understanding of the FHIR standard, and to review the use and adoption of the standard in current health care applications and organizations. This study can assist researchers and experts in understanding the FHIR architecture, design, implementation, resources, challenges, mapping, and adoption in health care informatics. Additionally, this systematic review identifies the key topics discussed in the context of FHIR in the literature.

Figure 1. General architecture of the Fast Health Interoperability Resources (FHIR) standard [4].



### FHIR Resource

A resource is the smallest discrete concept that can be maintained independently and is the smallest possible unit of a transaction [5]. Thus, a resource is a known identity providing meaningful data. Each resource has clear boundaries and differs from all other resources. A resource should be described in sufficient detail to define and support the medical data exchange that is involved in the process. According to the latest FHIR version (R4), the FHIR community has defined more than 150 resources to date [6]. These resources are divided into five major categories: (1) Administrative: location, organization, device, patient, group; (2) Clinical: CarePan, diagnostics, medication, allergy, family history; (3) Financial: billing, payment, support; (4) Infrastructure: conformance, document, message profile; and (5) Workflow: encounter, scheduling, order.

FHIR is the latest standard; however, to date, there has been no comprehensive systematic literature review performed in this area. Therefore, a systematic literature review was performed in this study to provide a broad view of FHIR, and to address various challenges, applications, and goals of FHIR highlighted in research in this field.

### Motivation and Objectives

Owing to its dynamic characteristics, FHIR is gaining popularity rapidly. It is expected that FHIR will soon become an icon for clinical information exchange in the health care sector. However, it also faces numerous challenges, which is the main motivation that inspired us to perform this systematic literature review. Despite its importance in health care research, there is no comprehensive review of the literature in the field.

There were five objectives of this study. The first objective was to profoundly investigate the literature related to FHIR and EHR to explore their multiple challenges in the health care domain and give a comprehensive summary of these issues. The second objective was to identify FHIR applications, goals, challenges, and their roles in the health care domain. The third aim was to address different models of FHIR implementation. Fourth, we addressed different existing and emerging challenges of electronic health implementation to provide the readers with up-to-date information about the different types of hurdles faced by health care project implementations. Finally, this review offers useful suggestions and recommendations about the solutions to these issues faced by health care stakeholders.

## Methods

### Design

This systematic literature review was conducted through the following steps: (1) establishing the research questions to be investigated; (2) identification of digital libraries to be explored and establishing the search strategy; (3) setting the criteria for selection of relevant articles; (4) setting the quality assessment criteria to select the best articles for this study; and (5) data extraction to address the research questions from the selected articles.

**Table 1.** Research questions and associated objectives.

Research questions	Objectives
SQ1: What are the types or models of FHIR <sup>a</sup> implementation?	To investigate various techniques, methods, or mechanisms used during the implementation of FHIR
SQ2: What are the common resources used in FHIR implementation?	To identify various resources used during the implementations of FHIR
SQ3: What are the applications that benefit from the use of FHIR?	To identify various types of applications that benefit from the FHIR standard (eg, mobile apps, SMART <sup>b</sup> on FHIR apps, research apps, HAPI <sup>c</sup> FHIR apps)
SQ4: What are approaches applied to map or migrate data from previous standards to FHIR?	To investigate various mechanisms on how to extract FHIR resources from HL7 <sup>d</sup> and other previous standards for mapping/migrating to the FHIR standard
SQ5: What are the goals of FHIR?	To identify or investigate the goals of the FHIR standard in the health care domain
SQ6: What are the challenges and open questions related to the FHIR domain?	To explore the challenges in the FHIR domain such as implementations (eg, FHIR API <sup>e</sup> , standard, interoperability)

<sup>a</sup>FHIR: Fast Health Interoperability Resources.

<sup>b</sup>SMART: Substitutable Medical Applications Reusable Technologies.

<sup>c</sup>HAPI: Health Level 7 application programming interface.

<sup>d</sup>HL7: Health Level 7.

<sup>e</sup>API: application programming interface.

### Search Strategy

After establishing the research questions, the next step was to search for articles to collect the required data. To perform a proper systematic literature review, an appropriate search is essential to define the scope and search keywords, which are the fundamental concepts of our research questions for retrieving accurate results.

There is a possibility that the search method may not identify some relevant studies. Therefore, to establish an optimized

**Textbox 1.** Search string for article retrieval.

```

{{{(Healthcare) or (eHealth) or (EHR)) and ((Standard) or (Protocols))} OR {(FHIR Approaches) or (FHIR Techniques) or (FHIR Methods)} OR
{(FHIR) and ((Implementation) or (Challenges) or (Barriers))} OR {(FHIR) and ((Resources) or (HL7 V2) or (HL7 CDA) or (HL7 CDA documents))}
OR {(FHIR and SMART) or (SMART on FHIR)} OR {(FHIR) and ((Mapping) or (Exchange))}

```

### Article Selection Process

#### Step 1

The following questions were defined for article selection: (1) What are the main domains/fields of the searched papers (eg, FHIR)? (2) Where are these papers published (conferences or

### Research Questions

According to Kitchenham et al [7], research questions are the most crucial part of any systematic literature review. Therefore, we have to set questions related to the focus fields, which are FHIR and EHR. We formulated specific research questions to identify the objectives in terms of problems, challenges, solutions, and goals. Our research questions identify the mentioned domain broadly and cover almost every aspect of the field, which is essential for the research purpose. [Table 1](#) summarizes the research questions and their corresponding objectives.

search string, Kitchenham et al [7] suggest breaking down the research questions into individual facets called research units, which include all of their associated acronyms, synonyms, abbreviations, related words, and alternative spellings combined using Boolean operators (AND, OR) for the construction of keyword phrases.

Finally, we obtained and used the search string shown in [Textbox 1](#) to retrieve the relevant articles.

journals)? (3) What should be the scope and credibility of these papers? (4) When were the papers published?

#### Step 2

To cover as many studies as possible, we selected the relevant articles from the literature by searching through well-known academic digital databases, including ACM, IEEE Xplore,

Springer, Google Scholar, PubMed, and ScienceDirect. These databases cover the most relevant conference and journal articles within the fields of health care and computer science. To limit the search, we set the range from January 2012 to December 2019. The search was performed during January 2020.

### Step 3

We selected the articles from all of the databases listed above on the basis of the search string (Textbox 1). We used the string and checked every article in chronological order, including title, abstract, keywords, introduction, background, methods, results, discussion, and conclusion. We then selected and downloaded the articles from the databases when the string or substring matched with any string in any of the above components of the article.

### Step 4

We removed duplicate articles retrieved from different databases, and manually filtered the collected articles using

**Table 2.** Inclusion and exclusion criteria.

Criteria	Inclusion	Exclusion
Subject	Full articles that deal with FHIR <sup>a</sup>	Articles that do not deal with FHIR or related acronyms, or do not address issues related to our research questions
Language	Articles published in English journals/proceedings	Articles published in non-English journals/proceedings
Access	Articles that provide access to the full text	Articles without access to the full text
Venue	Articles published in high-impact-factor conference proceedings or peer-reviewed journals	Articles from nonreputable journals/proceedings as well as books, notes, chapters, and press reports
Study type	Primary study	Nonprimary study, including literature review, informal literature surveys, theses, and articles that discuss aspects of health care without reference to FHIR
Publication history	Clear evidence of the article's print procedure and venue	Publication process with no proper scientific peer review or no clear evidence of the print venue
Keywords	Describe at least one part of our search string	Do not describe any part of the search string

<sup>a</sup>FHIR: Fast Health Interoperability Resources.

## Results

### Characteristics of Selected Articles

After performing the search queries, a total of 8144 articles from all five major digital databases were retrieved from the initial search. After thoroughly checking the web profiles of authors and their networks, 37 new articles were added with a snowballing procedure. From the 8181 retrieved articles, we first applied the duplication criteria, and then set the inclusion and exclusion criteria described in Table 2. Therefore, we first excluded all of the articles found in multiple databases. After removal of duplicates, 1514 articles remained. In the second phase, we discarded articles published in non-English journals/proceedings, resulting in 1442 articles for further screening. In the third phase, we excluded articles that were not primary studies such as reviews and survey papers. Finally, 892 articles remained for further screening.

In the fourth phase, we analyzed the remaining articles on the basis of their title, abstract, and keywords, and the number dropped to 278. In the final phase, after reading and analyzing

Endnote software to remove the articles included in multiple databases.

### Step 5

The inclusion criteria were full articles that deal with FHIR published in the English language in world-class conference proceedings or peer-reviewed journals between 2012 and 2019. The exclusion criteria were articles that address an FHIR-related issues but do not meet the inclusion criteria, such as books, theses (doctorate and masters), notes, chapters, press reports, informal literature surveys, literature surveys, papers without access to full text, and articles that discuss aspects outside of the scope of health care without reference to FHIR or EHR. All articles published in non-English journals/proceedings were removed. Table 2 provides further details of the inclusion and exclusion criteria used in this literature survey.

the full text of the articles, we selected 80 articles from the list to be included in the systematic review. Table 3 shows the results of the different phases of the selection process, Table 4 presents the articles chosen for our study, and Table 5 provides the geographic information of the publications.

As shown in Table 4, the distribution of the articles was 59% and 41% for journal articles and conference proceedings, respectively. The conferences represented are the main international conferences on health care or health care informatics, whereas the journals represent the world-class reputable journals in the field of computer science and health care. In terms of geography, as shown in Table 5, the number of publications related to FHIR published by researchers in the United States was the highest among represented countries. This indicates that the research in the field is quite active in the United States, which may become a factor that pushes the adoption of the standard in the country and in the rest of the world.

The 80 selected articles are arranged based in their primary subject categories in Table 6.



Once the articles were selected, we arranged them by ascending order of publication year. We then considered the attributes of the articles, including author names, article title, venue of publication (eg, journal article, conference proceeding), and publisher name. The complete list of the selected articles and their attributes is depicted in [Table 7](#).

**Table 3.** Phases of article selection and retrieval at each phase.

Phase	Description	Articles included for review, N
1	Total number of articles from all digital databases	8144
2	Snowball sampling	8181
3	Removal of duplicates	1514
4	Exclusion based on language	1442
5	Exclusion based on access and type of study (eg, reviews and survey papers)	892
6	Exclusion based on title, abstract, and keywords	278
7	Exclusion based on full text and nonprimary study	80

**Table 4.** Distribution of article types.

Publication year	Journal articles, N	Conference proceedings, N	Total, N
2012	0	0	0
2013	1	0	1
2014	0	1	1
2015	5	1	6
2016	6	4	10
2017	6	13	19
2018	10	7	17
2019	19	7	26
Total	47 (59%)	33 (41%)	80

**Table 5.** Geographic distribution of the selected articles.

Country	Articles, N	Year
Belgium	2	2018
Canada	4	2018, 2019
Czech Republic	2	2015
France	1	2017
Germany	7	2016, 2018, 2019
Ireland	3	2016, 2018, 2019
Netherlands	5	2016, 2017, 2019
Portugal	3	2017-2019
Switzerland	1	2019
Sweden	1	2017
United Arab Emirates	1	2018
United Kingdom	5	2016, 2017, 2019
United States	45	2013-2019

**Table 6.** Focus of the selected articles over time.

Category	2012, N	2013, N	2014, N	2015, N	2016, N	2017, N	2018, N	2019, N	Total, N
Apps	0	0	1	1	1	6	3	2	14
SMART <sup>a</sup>	0	0	0	0	3	1	1	0	5
FHIR <sup>b</sup> implementations models	0	0	0	1	1	3	6	1	12
FHIR resources	0	0	0	0	0	2	0	3	5
FHIR framework	0	0	0	0	0	0	0	11	11
Mapping framework/data model	0	1	0	4	4	2	4	4	19
Challenges	0	0	0	0	1	4	3	3	11
FHIR goals	0	0	0	0	0	1	0	2	3
Total	0	1	1	6	10	19	17	26	80

<sup>a</sup>SMART: Substitutable Medical Applications Reusable Technologies.

<sup>b</sup>FHIR: Fast Health Interoperability Resources.

**Table 7.** List of selected articles in ascending order of publication year.

Reference	Title	Year	Publisher	Venue
Bender and Sartipi [8]	HL7 FHIR: an agile and RESTful approach to healthcare information exchange	2013	IEEE <sup>a</sup>	Journal
Lamprinakos et al [9]	Using FHIR to develop a healthcare mobile application	2014	IEEE	Conference
Kasthurirathne et al [10]	Towards standardized patient data exchange: integrating a FHIR based API for the open medical record system	2015	IOS Press	Journal
Franz [11]	Applying FHIR in an integrated health monitoring system	2015	EuroMISE	Journal
Smits et al [12]	A comparison of two detailed clinical model representations: FHIR and CDA	2015	EuroMISE	Journal
Luz et al [13]	Providing full semantic interoperability for the Fast Healthcare Interoperability Resources schemas with resource description framework	2015	IEEE	Conference
Kasthurirathne et al [14]	Enabling better interoperability for healthcare: lessons in developing a standards based application programming interface for electronic medical record systems	2015	Springer	Journal
Jawaid et al [15]	Healthcare data validation and conformance testing approach using rule-based reasoning	2015	Springer	Journal
Rinner and Duftschmid [16]	Bridging the gap between HL7 CDA and HL7 FHIR: A JSON based mapping	2016	IOS Press	Journal
Ulrich et al [17]	Metadata repository for improved data sharing and reuse based on HL7 FHIR	2016	Elsevier	Journal
Ismail et al [18]	HL7 FHIR compliant data access model for maternal health information system	2016	IEEE	Conference
Mercorella et al [19]	An architectural model for extracting FHIR resources from CDA documents	2016	IEEE	Conference
Ruminski et al [20]	The data exchange between smart glasses and healthcare information systems using the HL7 FHIR standard	2016	IEEE	Conference
Doods et al [21]	Converting ODM metadata to FHIR questionnaire resources	2016	Springer	Journal
Bloomfield et al [22]	Opening the Duke electronic health record to apps: Implementing SMART on FHIR	2016	Elsevier	Journal
Lee et al [23]	Implementation of SMART APP Service Using HL7_FHIR	2016	IASER <sup>b</sup>	Journal
Mandel et al [3]	SMART on FHIR: a standards-based, interoperable apps platform for electronic health records	2016	Oxford	Journal
Andersen et al [24]	Point-of-care medical devices and systems interoperability: a mapping of ICE and FHIR	2016	IEEE	Conference
Minutolo et al [25]	Fuzzy on FHIR: a decision support service for healthcare applications	2017	Springer	Conference
Lee et al [26]	Profiling Fast Healthcare Interoperability Resources (FHIR) of family health history based on the clinical element models	2017	Elsevier	Journal
Abbas et al [27]	Mapping FHIR resources to ontology for DDI reasoning	2017	Linköping University	Conference
Yan et al [5]	Clinical decision support Based on FHIR data exchange standard	2017	Atlantis Press	Conference
Diomaiuta et al [28]	A FHIR-based system for the generation and retrieval of clinical documents	2017	Science and Technology Publications	Conference
Subhojeet et al [29]	Attribute based access control for healthcare resources	2017	ACM <sup>c</sup>	Conference
Saleh et al [30]	Using Fast Healthcare Interoperability Resources (FHIR) for the integration of risk minimization systems in hospitals	2017	IOS Press	Journal
Waghlikar et al [31]	SMART-on-FHIR implemented over i2b2	2017	Oxford	Journal
Li and Park [32]	Design and implementation of integration architecture of ISO 11073 DIM with FHIR resources using CoAp	2017	IEEE	Conference
Jiang et al [33]	A consensus-based approach for harmonizing the OHDSI common data model with HL7 FHIR	2017	IOS Press	Journal

Reference	Title	Year	Publisher	Venue
Shoumik et al [34]	Scalable micro-service based approach to FHIR server with Golang and No-SQL	2017	IEEE	Conference
Jánki et al [35]	Authorization solution for full stack FHIR HAPI access	2017	IEEE	Conference
Sanchez et al [36]	Achieving RBAC on RESTful APIs for mobile apps using FHIR	2017	IEEE	Conference
Clotet et al [37]	Differentiated synchronization plus FHIR a solution for EMR's ecosystem	2017	IEEE	Conference
Khalique and Khan [38]	An FHIR-based framework for consolidation of augmented EHR from hospitals for public health analysis	2017	IEEE	Conference
Aliakbarpoor et al [39]	Designing a HL7 compatible personal health record for mobile devices	2017	IEEE	Conference
Hong et al [40]	Shiny FHIR: an integrated framework leveraging Shiny R and HL7 FHIR to empower standards-based clinical data applications	2017	IOS Press	Journal
Leroux et al [41]	Towards achieving semantic interoperability of clinical study data with FHIR	2017	Springer	Journal
Jiang et al [42]	Developing a semantic web-based framework for executing the clinical quality language using FHIR	2017	Elsevier	Conference
Walinjkar and Woods [43]	FHIR tools for healthcare interoperability	2018	Biomedical Research Network	Journal
Kiourtis et al [44]	FHIR Ontology Mapper (FOM): aggregating structural and semantic similarities of ontologies towards their alignment to HL7 FHIR	2018	IEEE	Conference
Jeon et al [45]	Reactive server interface design for real-time data exchange in multiple data source and client	2018	IEEE	Conference
Gopinathan et al [46]	FHIR FLI: an open source platform for storing, sharing and analyzing lifestyle data	2018	Science and Technology Publications	Conference
Lackerbauer et al [47]	A model for implementing an interoperable electronic consent form for medical treatment using HL7 FHIR	2018	Elsevier	Journal
Stan and Miclea [48]	Local EHR management based on FHIR	2018	IEEE	Conference
Ahmad et al [49]	Implementation of SMART on FHIR in developing countries through SFPBRF	2018	ACM	Journal
Walonoski et al [50]	Validation and testing of Fast Healthcare Interoperability Resources standards compliance: data analysis	2018	JMIR	Journal
Urbauer et al [51]	Wearable activity trackers supporting elderly living independently: a standards based approach for data integration to health information systems	2018	ACM	Conference
Kamel and Nagy [52]	Patient-centered radiology with FHIR: an introduction to the use of FHIR to offer radiology a clinically integrated platform	2018	Springer	Journal
Borisov et al [53]	FHIR data model for intelligent multimodal interface	2018	IEEE	Conference
Hussain et al [54]	Learning HL7 FHIR using the HAPI FHIR server and its use in medical imaging with the SIIM dataset	2018	Springer	Journal
Crump et al [55]	Prototype of a standards-based EHR and genetic test reporting tool coupled with HL7-compliant infobuttons	2018	Elsevier	Journal
Peng et al [56]	Linking health web services as resource graph by semantic REST resource tagging	2018	Elsevier	Conference
Sharma and Aggarwal [57]	Mobile based application for predication of diabetes mellitus: FHIR standard	2018	Science Publisher Cooperation	Journal
Alves et al [58]	FHIRbox, a cloud integration system for clinical observations	2018	Elsevier	Journal
Kasparick et al [59]	IEEE 11073 SDC and HL7 FHIR – emerging standards for interoperability of medical system	2018	University of Rosstock	Journal
Zohner et al [60]	Challenges and opportunities in changing data structures of clinical document archives from HL7-V2 to FHIR-based archive solutions	2019	IOS Press	Journal
Maxhelaku and Kika [61]	Improving interoperability in healthcare using HL7 FHIR	2019	IDEAS	Conference

Reference	Title	Year	Publisher	Venue
Oemig [62]	HL7 version 2.x goes FHIR	2019	IOS Press	Journal
Kiourtis et al [63]	Structurally mapping healthcare data to HL7 FHIR through ontology alignment	2019	Springer	Journal
Metke-Jimenez and Hansen [64]	FHIRCap: transforming REDCap forms into FHIR resources	2019	Elsevier	Journal
Mukhiya et al [65]	A GraphQL approach to healthcare information exchange with HL7 FHIR	2019	Elsevier	Conference
Daumke et al [66]	Clinical text mining on FHIR	2019	Elsevier	Journal
Schleyer et al [67]	Preliminary evaluation of the Chest Pain Dashboard, a FHIR-based approach for integrating health information exchange information directly into the clinical workflow	2019	IOS Press	Journal
Kiourtis et al [68]	A string similarity evaluation for healthcare ontologies alignment to HL7 FHIR resources	2019	Springer	Journal
Kilintzis et al [69]	A sustainable HL7 FHIR based ontology for PHR data	2019	IEEE	Conference
Houta et al [70]	Use of HL7 FHIR to structure data in epilepsy self-management applications	2019	IEEE	Conference
Pfaff et al [71]	Fast Healthcare Interoperability Resources as a meta model to integrate common data models: development of a tool and quantitative validation study	2019	JMIR	Journal
Kondylakis et al [72]	Using XDS and FHIR to support mobile access to EHR information through personal health apps	2019	IEEE	Conference
Hong et al [73]	An interactive visualization tool for HL7 FHIR specification browsing and profiling	2019	Springer	Journal
Semenov et al [74]	Experience in developing an FHIR medical data management platform to provide clinical decision support	2019	MDPI <sup>d</sup>	Journal
Chapman et al [75]	A semi-autonomous approach to connecting proprietary EHR standards to FHIR	2019	Cornell University Library	Journal
Rivera Sánchez et al [76]	A service-based RBAC & MAC approach incorporated into the FHIR standard	2019	Elsevier	Journal
El-Sappagh et al [77]	A mobile health monitoring-and-treatment system based on integration of the SSN sensor ontology and the HL7 FHIR standard	2019	Springer	Journal
Eapen et al [78]	FHIRForm: an open-source framework for the management of electronic forms in healthcare	2019	IOS Press	Journal
Argüello-Casteleiro et al [79]	From SNOMED CT expressions to an FHIR RDF representation: exploring the benefits of an ontology-based approach	2019	RWTH Aachen University	Conference
Jenders [80]	Evaluation of the Fast Healthcare Interoperability Resources (FHIR) standard for representation of knowledge bases encoded in the Arden syntax	2019	Elsevier	Journal
Hong et al [81]	Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data	2019	Oxford	Journal
Eapen et al [82]	Drishti: A sense-plan-act extension to open mHealth framework using FHIR	2019	ACM	Conference
Mandl et al [83]	Beyond one-off integrations: a commercial, substitutable, reusable, standards-based, electronic health record-connected app	2019	JMIR	Journal
Sharma and Aggarwal [6]	HL-7 based middleware standard for healthcare information system: FHIR	2019	Springer	Journal
Baskaya et al [84]	mHealth4Afrika: implementing HL7 FHIR based interoperability	2019	Elsevier	Journal

<sup>a</sup>IEEE: Institute of Electrical and Electronics Engineers.

<sup>b</sup>IASER: Institute of Applied Social and Economic Research.

<sup>c</sup>ACM: Association for Computing Machinery.

<sup>d</sup>MDPI: Multidisciplinary Digital Publishing Institute.



## Quality Assessment

According to Kitchenham et al [7], it is essential to select and assess the best articles for every literature review, and the quality of selected studies must be verified before inclusion in the study. We evaluated the selected articles with regard to research

quality, related work, purposes of research, the obtained result, the methodology used, literature review, current and future objectives, conclusion, publication repository, and other factors. We evaluated the quality of each article according to the protocol defined by Roehrs [85] as displayed in [Textbox 2](#).

**Textbox 2.** Quality assessment criteria [85].

- Does the article state the purpose of the research?
- Does the article present the result related to objectives?
- Does the article have a research result?
- Does the article present a literature review and background?
- Does the article present an architecture proposal or research methodology?
- Does the article present a conclusion related to the research objectives?

The proposed quality criteria scores were assessed for each selected article. Although the majority of the selected articles did not fully satisfy all six criteria for evaluation, they complied with at least four out of the six quality assessment criteria listed in [Textbox 2](#).

All of the assessed articles clearly presented their research purpose, literature review, and were supported by a research methodology, bibliographical references, or models/architectural proposals. Based on this quality assessment, we did not exclude any articles from the corpus; this assessment only evaluates whether the articles have a satisfactory structure.

## Data Extraction and Addressing the Research Questions

### Process

In summary, the quality assessment of the selected articles was as follows. If an article identified by our search query criteria contained information related to our research questions, then the following three steps were applied. First, the title and abstract of the selected articles were carefully read to scrutinize whether the articles were relevant to our research questions. Second, we skimmed the entire article to assure that the required information was available. Finally, in the third round, we read through the entire article from start to end to ensure that this information was helpful for our study and could address the research questions.

To gather information from the selected articles corresponding to our research questions and criteria, we developed separate forms in Microsoft Word and Excel. We reviewed every section of the article from beginning to end and recorded details of the articles in these two forms whenever we found the answer to a corresponding research question. After compilation of the results, we placed these results in specific question-and-answer section tables and discarded the two temporary generated Word and Excel forms.

We collected the following types of data from each article: author name(s), affiliation and country name, venue (journal or conference), publication year. We then collected the answers to the set of research questions from these articles and recorded the details of selected articles for further processing.

## ***SQ1: What are the Types or Models of FHIR Implementation?***

To address this question, we reviewed the literature in the FHIR domain and investigated various techniques, methods, and mechanisms used in the implementation of FHIR in the health care sector.

At present, FHIR is the most attractive domain among health care researchers. Therefore, extensive efforts are being taken to implement FHIR with consideration of multiple aspects and diverse areas. We obtained 11 categories for FHIR implementation. From a platform point of view, we considered the implementation of mobile/tablet apps [3,9,28,35,39,51,57,60,65,70,72,77,82], standalone apps/servers [3,23,51,58,61], web services/API [3,11,14,26,34,36,40,48,53,56,71,78], and web-based tools/applications [18,26,37,40,42,46,55,69,78] categories. From a conceptual framework, we considered the categories of general FHIR implementation [48,59], using SMART on FHIR [3,14,22,31], HL7 API (HAPI)-FHIR server/library/applications [14,16,20,32,34-36,40,43,51,54,55,73,76,81,82], and FHIR general framework [14,40,78]. In consideration of compatibility, we chose FHIR data model/data exchange [5,11,26,45,55,66,67,73] and defining ontology to align with FHIR [44,68,69] as the main categories. In addition, we classified all implementation-related work under the miscellaneous category [21,25,30,32,47,60,71], such as FHIR implementation of the legacy clinical data repository system, FHIR implementation of the agent-based system, implementation of operational data model metadata [86] to FHIR questionnaire resource implementation, FHIR implementation of the electronic treatment form, Clinical Asset Mapping Program for FHIR, integration of the architecture of domain information model (ISO/IEEE 11073 DIM) [31] with FHIR, and FHIR-based decision support systems.

## ***SQ2: What are the Common Resources Used in FHIR Implementation?***

The FHIR community has defined more than 150 resources to date [6]. For this research question, we reviewed the literature in the FHIR domain to identify various FHIR resources used in implementation. We observed that approximately 82 different

types of resources have been used in FHIR implementation in various articles. The resource names and the articles mentioned in various resources are shown in [Table 8](#).

In the miscellaneous category, we list all of the articles that mention only one or two resources: (1) Activity Definition, (2) Adverse Reaction, (3) Adverse Event, (4) Address, (5) Billing, (6) Bundle, (7) Contraindication, (8) Conformance, (9) Consent, (10) Concept Map, (11) Claim, (12) Clinical, (13) Clinical Study Plan (14), Clinical Impression (15), Care Team, (16) Category, (17) Coverage, (18) Device Component, (19) Device

Observation Report, (20) Document Manifest (21), Document Reference (22), Dosage (23), Data Element (24), Diagnostic (25), Diagnostic Order (26), Drug Administration, (27) Element, (28) Element Definition (29), Equipment (30), Gender (31), Goal (32), Group (33), Intolerance (34), Imaging Study (35), Imaging Manifest (36), Medication Dispense, (37) Message Profile, (38) Nutrition Order, (39) Procedure Request, (40) Provenance (41), Provider, (42) Risk Assessment, (43) Research Definition, (44) Request Group, (45) Relative, (46) Related Person, (47) Schedule, (48) Specimen, (49) Staff (50), Structure Definition.

**Table 8.** List of resources used in Fast Health Interoperability Resources implementation.

Resource name	References
Allergy	[6,10,12,34,49]
Allergy Intolerance	[12,14,19,22,29,33,52,55,58,60,61,72,74,77]
Appointment	[13,55,70]
Condition	[3,10,12,16,18,22,23,26,31,36,40,52,55,60,64,66,71,73,74,76,77,79,81]
Composition	[14,16,70,81]
Care Plan	[6,34,36,39,41,49,61,74,76,77,82]
Device	[6,8,9,11,19,20,24,32,33,39,43,48,53,77]
Device Metric	[24,32,53]
Detected Issue	[25,74,77]
Document	[6,8,33]
Diagnostic Report	[28,44,52,60,63,74,84]
Encounter	[10,16,21,30,33,39,41,55,71,74,77]
Episode Of Care	[21,41,77,84]
Family Member History	[22,26,27,60,61,74,77,81]
Family History	[6,22,33]
Immunization	[22,60,74]
Location	[6,10,14,16,24,34,49,71,77]
Medication	[3,9,16,19,27,29,31,33,39,40,48,49,70,77,81]
Medication Administration	[60,70,71]
Medication Order	[16,22,23,70]
Medication Statement	[19,27,55,66,77,81]
Medication Prescription	[3,22,31]
Medication Request	[66,71,73,74,77]
Observation	[3,9-11,14,18-26,28-33,36,39-41,43,44,48,49,51-53,55,58,60,63,64,66,69-71,73-77,79,82,84]
Organization	[5,6,16,19,34,39,41,75,77,84]
Patient	[3,5,6,8-10,12,14,16-24,27-32,34,36,39-41,44,48-50,52,53,55,58,61,63,65,68,70-77,84]
Person	[14,46,69]
Plan Definition	[41,74,80]
Practitioner	[9,17-19,24,27-29,34,41,44,48,52,53,58,61,63,71,77]
Procedure	[3,40,60,66,71,72,74,77,81]
Questionnaire Response	[21,41,47,60,69,70,74,78,84]
Questionnaire	[17,21,41,47,64,65,70,74,84]
Miscellaneous	[3,5,6,10-12,14-17,21,23,24,28,29,32-34,41,44,46,47,49,52,55,60,66,71-75,77]

### ***SQ3: What are the Applications that Benefit from the use of FHIR?***

We attempted to thoroughly investigate the literature from various directions and provide the readers with a comprehensive summary of every aspect of FHIR. In this section, we consider the type of applications that can benefit from the FHIR standard, including health care systems/applications benefit in terms of interoperability/data exchange, rules, security/privacy,

conformance, health care process, and administration. Thus, we came up with eight categories based on how the applications make use of the FHIR standard (Table 9). In the miscellaneous category, we included all articles that address the type of applications that benefit from the FHIR standard but do not fall under any of the other categories mentioned above (eg, clinical applications for data exchange, testing applications). Table 9 shows the articles that address specific applications that benefit from the FHIR standard.

**Table 9.** Applications that benefit from the use of Fast Health Interoperability Resources (FHIR).

Applications types	References
Mobile apps	[5,9,25,32,34,35,39,40,42,49,57,59,67,72,73,76,77,82]
SMART <sup>a</sup> on FHIR	[3,14,22,49,67,72,73,83]
Research	[5,15,18,25,40,42,55,60,64,66,67,73,78]
Electronic records and medical practices	[25,39,46,52,55,67,73]
HAPI <sup>b</sup> FHIR	[14,26,34,40,43,73,81]
Graphic/images	[8,52,61,67]
Web-based	[34,42,55,59,67,73]
Miscellaneous	[26,50,60]

<sup>a</sup>SMART: Substitutable medical applications reusable technologies.

<sup>b</sup>HAPI: Health Level 7 application programming interface.

SMART on FHIR is mentioned under implementation in Table 8 as well as under applications in Table 9. Articles that mentioned SMART on FHIR implementation, either fully or partially, were grouped into one category, and the articles that mentioned any applications that benefit from the SMART on FHIR concept were considered as a different category. Thus, in Table 8, we list articles that mention SMART on FHIR in various implementations, whereas in Table 9, we list applications that benefit from SMART on FHIR platforms.

The SMART platform is a health data layer based on the FHIR API and resource definitions [87]. From the beginning, the SMART team selected platform components that emphasize web standards (eg, HTML, JavaScript, OAuth, and Resource Description Framework) [3]. This setup results in the HL7 legacy versions (ie, v2, v3, CDA) to be unable to implement SMART applications. All previous versions could not use a web API for data access and were unable to access data at the granular level. Additionally, the CDA is based on the reference information model and is lacking in sufficient detail, whereas version 2 suffers from inconsistencies across implementations and version 3 is complex, which leads to incompatible documents and systems [3].

In contrast, the FHIR standard uses web APIs for data access, which is capable of accessing the clinical data at the granular level. The SMART on FHIR concept does not exist without support of the FHIR standard. Therefore, the SMART concept is developed after introducing the FHIR standard, and SMART on FHIR when considered as a standard has some predecessors. All of these contribute in one way or another to the current standards (FHIR) [88]. Considering all of this evidence, we conclude that SMART on FHIR is the main beneficiary of the FHIR standard compared with the other standards.

During the literature review, we observed that mobile, research, and SMART on FHIR applications are the most common beneficiaries of the FHIR standard, followed by electronic records and medical practices, and web-based applications.

### ***SQ4: What Approaches are Applied to Map or Migrate Data from Other HL7-Based Legacy Systems to the FHIR-Based System?***

At present, HL7 (v2 and CDA) is the most popular data standard in the health care sector, with many countries still using this standard for medical data exchange. Specifically, more than 35 countries implement the HL7 v2 standard and 95% of US health care organizations are still using this standard for medical data sharing among various health care organizations [89].

Owing to its dynamic structure, FHIR provides numerous advantages such as flexibility to manage and retrieve granular clinical information from the whole document. Clinical practitioners and health care providers expect that the FHIR standard will soon occupy the health care market, and that it will replace all of the previous HL7 (eg, v2, v3, CDA) standards. Nevertheless, in this review, we found that FHIR is not likely to replace the previous HL7 (v2, v3, and CDA) standards within weeks or months, but might take years or decades. The rationale is related to the worldwide implementation of the earlier standards such as HL7 v2 and HL7 CDA. Furthermore, health care organizations argue that FHIR has not yet replaced the ubiquitous HL7 v2, and likely will not for several years, because many organizations have already recognized the value of adopting FHIR alongside legacy HL7 standards [27].

Therefore, for addressing this research question, we reviewed the literature in the FHIR domain and investigated various articles that address the mechanisms to extract FHIR resources

from HL7 or other previous standards, and to map or migrate them to the FHIR standard. We classified these mappings into six different categories (Table 10).

All of the included articles that address the mapping of any standards to the FHIR standard, but that are not in the six categories mentioned above, were categorized as

“miscellaneous/other standards to FHIR mapping.” Nevertheless, we found one study in which data were mapped from the FHIR standard to other standards, and the FHIR resource was mapped to the Web Ontology Language–based ontology [27]. Table 10 shows the list of articles categorized into different mapping categories.

**Table 10.** Approaches used to map or migrate data from other Health Level 7 (HL7)-based legacy systems to the Fast Healthcare Interoperability Resources (FHIR)-based system.

Techniques or methods	References
Map HL7 version 2 to FHIR	No relevant articles
Map HL7 CDA <sup>a</sup> documents, C-CDA <sup>b</sup> , or HL7 version 3 to FHIR	[5,8,12,16,19,60,79]
ODM <sup>c</sup> to FHIR	[21,41,64]
Map FHIR to other	[27]
i2b2 <sup>d</sup> to FHIR format	[31,71]
Health record data model to FHIR	[38,63,68,75,81,84]
Map other standards to FHIR	[13,17,24,33,64,71]

<sup>a</sup>CDA: clinical document architecture.

<sup>b</sup>C-CDA: consolidated clinical document architecture.

<sup>c</sup>ODM: operational data model.

<sup>d</sup>i2b2: Informatics for Integrating Biology and the Bedside.

### **SQ5: What are the Goals of FHIR?**

For this research question, we reviewed the literature in the FHIR domain to identify or investigate the goals of the FHIR standard in the health care domain. According to the objectives

of the reviewed articles, we divided the inquiries regarding the goals into seven different objectives (Table 11). Table 11 shows the articles that address various goals of the FHIR standard, demonstrating that most of these articles focus on the result rather than other goals and objectives.

**Table 11.** Goals of Fast Healthcare Interoperability Resources.

Goals	References
Simplify implementation without sacrificing information integrity	[5]
Patient satisfaction	[73]
Solve health problems (administrative and clinical)	[25,49,52]
Improve global health data interoperability	[49,59,67]
Enhance and maintain quality of data and accessibility	[60]
Result	[15,18,19,27,52,55,73]

### **SQ6: What are the Challenges and Open Questions Related to FHIR?**

As the latest standard in the health care domain, it is predictable that FHIR will face various challenges in terms of implementation, adoption, maintenance, data exchange, and other issues. In addition, numerous questions will be raised with respect to use of the FHIR standard. Therefore, for this research question, we reviewed the literature in the FHIR domain to identify various challenges faced by the FHIR standard. We

found 19 articles that discussed the implementation challenges, highlighting seven areas of challenge for the FHIR standard (Table 12).

Observations made during the literature review led us to conclude that implementing FHIR in any type of application is the most challenging task in the health care sector; 9 of the 19 related articles discussed this issue. Developers face various types of challenges during the development of any FHIR-based application. Table 12 lists the articles that mentioned these challenges.

**Table 12.** Challenges and open questions related to Fast Healthcare Interoperability Resources (FHIR).

Challenges	References
Implementations of FHIR in an application	[8,12,40,42,48,49,62,64,77]
Standards complexity	[8,27,62,81]
Adoptions	[40,56,83]
FHIR maintenance and specification	[41,42,62]
REST <sup>a</sup> ful approach	[56,59,65,73]
Mapping/migration challenging	[71,75,81]
Miscellaneous	[12,49]

<sup>a</sup>REST: representational state transfer.

## Discussion

### Principal Findings

This systematic literature review successfully identified both qualitative and quantitative sets of studies that enable obtaining a clear view of the FHIR standard in health care in the past 8 years, starting from the selected number of articles. Some of the most relevant studies in the field are highlighted according to systematic selection criteria. We identified the main topics associated with the use of FHIR in digital health. Many articles dealt with topics related to FHIR implementations and use resources, as well as data migration data models. As expected, various application categories such as mobile apps, SMART on FHIR applications, and research applications were the main topics associated with FHIR. Multiple challenges in FHIR adoption and implementation were also highlighted in the included articles. Interestingly, only a small number of relevant articles addressed FHIR goals.

At the beginning of this study, we planned to identify some common aspects in this field by answering some fundamental research questions. Hence, we established six research questions to address the objectives, goals, applications, and challenges of FHIR emerging in recent years. As a result, we can propose a taxonomy of the literature, and identify gaps to be further investigated on existing challenges and issues related to use of the FHIR standard in recent years. We also identify other common and related aspects with respect to interoperability, privacy, authorization (access control), data type, and testing and validation tools. For example, interoperability between the provider and hospital systems poses additional barriers to effective data sharing. In addition, various testing and validation tools are used to improve server compliance with the FHIR specification. Moreover, FHIR specifications define different data types to access and process the FHIR resources element.

Various FHIR-related studies aim to address FHIR implementation challenges such as data migration and cross-institutional sharing of clinical data in the clinical environment [60,71]. The main findings that are presented in these reviews and some other related studies include the importance of realizing EHR data interoperability via adoption of FHIR by health care providers. This adoption might be essential for the improvement of health care services with respect to health data sharing, integration, and availability.

Furthermore, use of the FHIR standard in the health care sector may enhance the chance of adoption of smart technologies in the health care domain, such as smartphones, mobile health apps, tablets, smart watches, fitness trackers, and any other future innovations [90]. Furthermore, use of artificial intelligence technologies and data sciences will also be dominant in implementing FHIR-based applications.

FHIR is viewed as the latest standard purely operating on resources, which are used for data storage, migration, and processing among multiple health care providers. The resources-based structure of FHIR is declared as distinct from other standards and is considered to be its main advantage. FHIR has several advantages that range from being a flexible standard, minimal implementation complexity, ability to display the patient history in a single document, granular data access, and avoiding message variability with the use of RESTful APIs.

FHIR is considered to be a unique pathway that can offer a solution to the interoperability issues of clinical data. Nevertheless, various studies indicate that FHIR also faces numerous challenges such as implementation, adoption, maintenance, mapping, and standard complexity. The RESTful API that makes FHIR unique from other standards also faces its own challenges in accessing sensitive health care data stored in the cloud environment [36].

Numerous studies have shown that several applications used in different domains are taking advantage of FHIR, including mobile apps, SMART on FHIR applications, research applications, electronic records and medical applications, graphic/image applications, HAPI FHIR application, and web-based applications.

To the best of our knowledge, this is the first comprehensive systematic literature review that focuses on FHIR-based EHR. There are some systematic literature reviews available in the FHIR domain; however, we found that the existing reviews are unable to explain the FHIR standard in detail. The FHIR standard is very rich, and therefore research in this domain is equally diverse with focus in various directions. Readers are interested in searching for articles that do not only introduce the FHIR standard but also explain various aspects of the standard in detail. For example, Lehne et al [91] only reviewed articles related to a general introduction of FHIR, without providing in-depth analysis on either FHIR or articles mentioning FHIR. In particular, the titles or identification of



included articles and more comprehensive details of the included articles are missing. Based on a thorough reading, we concluded that this previous review could not fully introduce and address various aspects of FHIR, such as challenges, applications, goals, mapping, and implementation models. Moreover, individual aspects of the reviewed articles were not explained adequately. Therefore, it is not possible to find a corresponding article when interested in a particular topic. Further, the authors focused on articles published between 2002 and 2018, although the FHIR concept was only introduced in 2011; thus, this was a mixed review of EHR and FHIR with little focus on FHIR itself. Lastly, the authors included only 15 references in the review, which is not sufficient for a systematic literature survey.

Similarly, another systematic literature review [90] only explained the general concept and current status of FHIR, whereas core issues such as challenges, goals, and application implementation were not discussed. Important articles that discuss the FHIR resources used in various application implementations were also not included in this review. Although this previous review analyzed some articles in the FHIR domain, the list of articles was not provided or explained properly. Therefore, it is quite difficult for readers to search for articles related to specific information of interest, such as FHIR applications, goals, challenges, and used resources. This information is the core requirement for readers interested in this field. Thus, we concluded that the existing reviews only introduced the FHIR standard without performing a comprehensive analysis of the current state of the field.

In this work, we deeply explored the literature and identified articles that not only mention the FHIR standard but also discuss its major aspects such as core challenges, applications, goals, mapping, used resources, and implementation models. In addition, we highlighted every article along with their references and addressed aspects such as those mentioned above. We searched existing databases for articles on the FHIR standard published between 2012 and 2019, and then included every article that discussed or mentioned even a single aspect of the FHIR standard. This approach provides a convenient resource for readers to easily search articles of interest in the literature.

FHIR is a new standard in the health care domain. It is still in the early stages of development and evaluation, and consequently faces numerous obstacles. We believe that these obstacles might eventually be overcome, thereby opening a new roadmap to solving the problem of data interoperability in the health care sector, which is in line with the findings of the literature review and remains the main objective of our research.

### Limitations

This study was limited to aspects related only to the FHIR standard rather than the general health care concept. In this

sense, the literature review focused exclusively on articles addressing FHIR concepts. This work sought to answer research questions proposed for providing an outline of the current literature related to FHIR without specifically assessing any computer system that refers to FHIR use. In addition, our search focused on articles published in various scientific journals related to health care and computer science within a limited time frame. This investigation was limited to articles selected from journals/conferences through implementations of standard steps of the systematic literature review methodology. We focused on scientific articles and did not address commercial or more technological approaches or solutions.

### Conclusion

This study provides a systematic literature review regarding the FHIR EHR standard, with the main objective of identifying and discussing the main issues, challenges, goals, and possible benefits from adoption of the FHIR standard in the health care sector. We have explored the FHIR-related literature and investigated articles associated with the FHIR standard in health care information systems. We identified various data models, methods/techniques used in FHIR implementation, FHIR beneficiary applications, and resources used in FHIR implementation. Various data mapping techniques/approaches, key challenges, and primary goals of FHIR were also explored. We observed that FHIR studies mainly focus on clinical data interoperability and portability issues between health care information systems.

The FHIR standard is capable of providing an optimized solution for medical data exchange between two systems and will establish data-sharing trust among health care providers. Furthermore, the FHIR standard is identical in terms of the support of smart technologies such as smartphones, tablets, mobile health apps, smart watches, and fitness trackers, which could solve numerous health care problems that were not possible for the previous standards (ie, HL7 v2, v3 and CDA). Based on this thorough investigation of the literature, we recommend the FHIR standard as a future suitable solution for addressing the health care interoperability problem. Nevertheless, FHIR itself faces some challenges such as implementation, standard complexity, and adoption, among others. Therefore, further research is required to address these challenges.

This review on the standard, purpose, and applications of FHIR will provide readers with a more comprehensive view and understanding of FHIR. This review should also help researchers and health care information technology professionals to access FHIR-associated information in the research community and to assess its impact on digital health. Lastly, this work can provide a roadmap, and suggest possible directions for future research and development in the FHIR domain.

---

### Conflicts of Interest

None declared.

---

### References

1. HL7 FHIR Release 4. URL: <https://www.hl7.org/fhir/?> [accessed 2014-01-02]
2. Tech industry looks to improve healthcare through cloud technology. Information Technology Industry Council (ITI). URL: <https://www.iti.org/news-events/news-releases/tech-industry-looks-to-improve-healthcare-through-cloud-technology> [accessed 2017-10-12]
3. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *J Am Med Inform Assoc* 2016 Sep;23(5):899-908 [FREE Full text] [doi: [10.1093/jamia/ocv189](https://doi.org/10.1093/jamia/ocv189)] [Medline: [26911829](https://pubmed.ncbi.nlm.nih.gov/26911829/)]
4. FHIR Overview - Architects. HL7 FHIR Release 4. URL: <https://www.hl7.org/fhir/overview-arch.html> [accessed 2011-01-05]
5. Yan H, Xiao L, Tian J. Clinical decision support based on FHIR data exchange standard. 2017 Presented at: 2nd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2017); May 13-14, 2017; Dalian, China. [doi: [10.2991/icmeit-17.2017.96](https://doi.org/10.2991/icmeit-17.2017.96)]
6. Sharma M, Aggarwal H. HL-7 based middleware standard for healthcare information system: FHIR. In: Proceedings of 2nd International Conference on Communication, Computing and Networking. Lecture Notes in Networks and Systems.: Springer; 2018 Sep 08 Presented at: ICCCN 2018; March 29-30, 2018; Chandigarh, India p. 889-899. [doi: [10.1007/978-981-13-1217-5\\_87](https://doi.org/10.1007/978-981-13-1217-5_87)]
7. Kitchenham BA, Brereton P, Turner M, Niazi MK, Linkman S, Pretorius R, et al. Refining the systematic literature review process—two participant-observer case studies. *Empir Software Eng* 2010 Jun 25;15(6):618-653. [doi: [10.1007/s10664-010-9134-8](https://doi.org/10.1007/s10664-010-9134-8)]
8. Bender D, Sartipi K. HL7 FHIR: an agile and RESTful approach to healthcare information exchange. : IEEE; 2013 Oct 10 Presented at: 26th IEEE International Symposium on Computer-Based Medical Systems; June 20-22, 2013; Porto, Portugal. [doi: [10.1109/cbms.2013.6627810](https://doi.org/10.1109/cbms.2013.6627810)]
9. Lamprinakos GC, Mousas AS, Kapsalis AP, Kaklamani DI, Venieris IS, Boufis AD, et al. Using FHIR to develop a healthcare mobile application. : IEEE; 2014 Dec 05 Presented at: 4th International Conference on Wireless Mobile Communication and Healthcare - "Transforming healthcare through innovations in mobile and wireless technologies"; November 3-5, 2014; Athens, Greece URL: <https://ieeexplore.ieee.org/document/7015927> [doi: [10.4108/icst.mobihealth.2014.257232](https://doi.org/10.4108/icst.mobihealth.2014.257232)]
10. Kasthurirathne SN, Mamlin B, Grieve G, Biondich P. Towards standardized patient data exchange: integrating a FHIR based API for the open medical record system. *Stud Health Technol Inform* 2015;216:932. [Medline: [26262234](https://pubmed.ncbi.nlm.nih.gov/26262234/)]
11. Franz B. Applying FHIR in an integrated health monitoring system. *Eur J Biomed Informatics* 2015 Oct 15;11(02):51-56. [doi: [10.24105/ejbi.2015.11.2.8](https://doi.org/10.24105/ejbi.2015.11.2.8)]
12. Smits M, Kramer E, Harthoorn M, Cornet R. A comparison of two Detailed Clinical Model representations: FHIR and CDA. *Eur J Biomed Informatics* 2015;11(02):1-15. [doi: [10.24105/ejbi.2015.11.2.3](https://doi.org/10.24105/ejbi.2015.11.2.3)]
13. Luz MP, Rocha de Matos Nogueira J, Cavalini LT, Cook TW. Providing Full Semantic Interoperability for the Fast Healthcare Interoperability Resources Schemas with Resource Description Framework. : IEEE; 2015 Oct 12 Presented at: International Conference on Healthcare Informatics; October 21-23, 2015; Dallas, TX. [doi: [10.1109/ichi.2015.74](https://doi.org/10.1109/ichi.2015.74)]
14. Kasthurirathne SN, Mamlin B, Kumara H, Grieve G, Biondich P. Enabling better interoperability for healthcare: lessons in developing a standards based application programming interface for electronic medical record systems. *J Med Syst* 2015 Nov;39(11):182. [doi: [10.1007/s10916-015-0356-6](https://doi.org/10.1007/s10916-015-0356-6)] [Medline: [26446013](https://pubmed.ncbi.nlm.nih.gov/26446013/)]
15. Jawaid H, Latif K, Mukhtar H, Ahmad F, Raza SA. Healthcare data validation and conformance testing approach using rule-based reasoning. In: Yin X, Ho K, Zeng D, Aickelin U, Zhou R, Wang H, editors. Health Information Science. HIS 2015. Lecture Notes in Computer Science, vol 9085. Cham: Springer; May 06, 2015:241-246.
16. Rinner C, Duftschmid G. Bridging the gap between HL7 CDA and HL7 FHIR: A JSON based mapping. *Stud Health Technol Inform* 2016;223:100-106. [Medline: [27139391](https://pubmed.ncbi.nlm.nih.gov/27139391/)]
17. Ulrich H, Kock AK, Duhm-Harbeck P, Habermann JK, Ingenerf J. Metadata repository for improved data sharing and reuse based on HL7 FHIR. *Stud Health Technol Inform* 2016;228:162-166. [Medline: [27577363](https://pubmed.ncbi.nlm.nih.gov/27577363/)]
18. Ismail S, Alshamari M, Qamar U, Butt WH, Latif K, Ahmad HF. HL7 FHIR compliant data access model for maternal health information system. : IEEE; 2016 Nov 02 Presented at: 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE); October 31-November 2, 2016; Taichung, Taiwan. [doi: [10.1109/bibe.2016.9](https://doi.org/10.1109/bibe.2016.9)]
19. Mercorella M, Ciampi M, Esposito M, Esposito A, De Pietro G. An architectural model for extracting FHIR resources from CDA documents. 2016 Dec 01 Presented at: 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS); November 28-December 1, 2016; Naples, Italy. [doi: [10.1109/sitis.2016.99](https://doi.org/10.1109/sitis.2016.99)]
20. Ruminski J, Bujnowski A, Kocejko T, Andrushevich A, Biallas M, Kistler R. The data exchange between smart glasses and healthcare information systems using the HL7 FHIR standard. 2016 Jun 08 Presented at: 9th International Conference on Human System Interactions (HSI); July 6-8, 2016; Portsmouth, UK. [doi: [10.1109/hsi.2016.7529684](https://doi.org/10.1109/hsi.2016.7529684)]
21. Doods J, Neuhaus P, Dugas M. Converting ODM metadata to FHIR questionnaire resources. *Stud Health Technol Inform* 2016;228:456-460. [Medline: [27577424](https://pubmed.ncbi.nlm.nih.gov/27577424/)]
22. Bloomfield RA, Polo-Wood F, Mandel JC, Mandl KD. Opening the Duke electronic health record to apps: Implementing SMART on FHIR. *Int J Med Inform* 2017 Mar;99:1-10. [doi: [10.1016/j.ijmedinf.2016.12.005](https://doi.org/10.1016/j.ijmedinf.2016.12.005)] [Medline: [28118917](https://pubmed.ncbi.nlm.nih.gov/28118917/)]

23. Lee CH, Kim YS, Lee YH. Implementation of SMART APP Service Using HL 7 \_ FHIR. 2016 Presented at: 2nd International Conference on Electronics, Electrical Engineering, Computer Science (EEECS); August 10-13, 2016; Qingdao, China URL: <https://www.semanticscholar.org/paper/Implementation-of-SMART-APP-Service-Using-HL-7--Lee-Kim/c116180f4ae8e255f5df934cbf2e11af827a0719>
24. Andersen B, Kasparick M, Ulrich H, Schlichting S, Golasowski F, Timmermann D, et al. Point-of-care medical devices and systems interoperability: A mapping of ICE and FHIR. 2016 Oct 31 Presented at: IEEE Conference on Standards for Communications and Networking (CSCN); October 31-November 2, 2016; Berlin, Germany. [doi: [10.1109/cscn.2016.7785165](https://doi.org/10.1109/cscn.2016.7785165)]
25. Minutolo A, Esposito M, De Pietro G. Fuzzy on FHIR: a Decision Support service for Healthcare Applications. In: Xhafa F, Barolli L, Amato F, editors. Advances on P2P, Parallel, Grid, Cloud and Internet Computing. 3PGCIC 2016. Lecture Notes on Data Engineering and Communications Technologies, vol 1. Cham: Springer; Nov 07, 2016.
26. Lee J, Hulse NC, Wood GM, Oniki TA, Huff SM. Profiling Fast Healthcare Interoperability Resources (FHIR) of family health history based on the clinical element models. AMIA Annu Symp Proc 2016;2016:753-762 [FREE Full text] [Medline: [28269871](https://pubmed.ncbi.nlm.nih.gov/28269871/)]
27. Abbas R, Al Khaldi IFHH, Ayele G, Nytnun JP. Mapping FHIR Resources to Ontology for DDI reasoning. 2017 Aug 10 Presented at: 15th Scandinavian Conference on Health Informatics; August 29-30, 2017; Kristiansand, Norway p. 30-40.
28. Diomaiuta C, Sicuranza M, Ciampi M, Pietro G. A FHIR-based system for the generation and retrieval of clinical documents. 2017 Mar 13 Presented at: 3rd International Conference on Information and Communication Technologies for Ageing Well and e-Health - ICT4AWE; April 28-29, 2017; Porto, Portugal. [doi: [10.5220/0006311301350142](https://doi.org/10.5220/0006311301350142)]
29. Subhojeet M, Ray I, Ray I, Shirazi H, Ong T, Kahn MG. Attribute based access control for healthcare resources. 2017 Mar 24 Presented at: ABAC '17: 2nd ACM Workshop on Attribute-Based Access Control; March 24, 2017; New York, NY. [doi: [10.1145/3041048.3041055](https://doi.org/10.1145/3041048.3041055)]
30. Saleh K, Stucke S, Uciteli A, Faulbrück-Röhr S, Neumann J, Tahar K, et al. Using Fast Healthcare Interoperability Resources (FHIR) for the integration of risk minimization systems in hospitals. Stud Health Technol Inform 2017;245:1378. [Medline: [29295457](https://pubmed.ncbi.nlm.nih.gov/29295457/)]
31. Wagholikar KB, Mandel JC, Klann JG, Wattanasin N, Mendis M, Chute CG, et al. SMART-on-FHIR implemented over i2b2. J Am Med Inform Assoc 2017 Mar 01;24(2):398-402 [FREE Full text] [doi: [10.1093/jamia/ocw079](https://doi.org/10.1093/jamia/ocw079)] [Medline: [27274012](https://pubmed.ncbi.nlm.nih.gov/27274012/)]
32. Li W, Park JT. Design and implementation of integration architecture of ISO 11073 DIM with FHIR resources using CoAP : IEEE; 2017 Jun 26 Presented at: International Conference on Information and Communications (ICIC); June 26-28, 2017; Hanoi, Vietnam. [doi: [10.1109/infoc.2017.8001674](https://doi.org/10.1109/infoc.2017.8001674)]
33. Jiang G, Kiefer RC, Sharma DK, Prud'hommeaux E, Solbrig HR. A consensus-based approach for harmonizing the OHDSI common data model with HL7 FHIR. Stud Health Technol Inform 2017;245:887-891 [FREE Full text] [Medline: [29295227](https://pubmed.ncbi.nlm.nih.gov/29295227/)]
34. Shoumik FS, Talukder MIMM, Jami AI, Protik NW, Hoque MM. Scalable micro-service based approach to FHIR server with golang and No-SQL. 2017 Presented at: 20th International Conference of Computer and Information Technology (ICCI); December 22-24, 2017; Dhaka, Bangladesh. [doi: [10.1109/iccitech.2017.8281846](https://doi.org/10.1109/iccitech.2017.8281846)]
35. Jánki ZR, Szabó Z, Bilicki V, Fidirich M. Authorization solution for full stack FHIR HAPI access. 2017 Nov 25 Presented at: IEEE 30th Neumann Colloquium (NC); November 24-25, 2017; Budapest, Hungary. [doi: [10.1109/nc.2017.8263266](https://doi.org/10.1109/nc.2017.8263266)]
36. Sánchez YKR, Demurjian SA, Baihan MS. Achieving RBAC on RESTful APIs for mobile apps using FHIR. 2017 Apr 06 Presented at: 5th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud); April 6-8, 2017; San Francisco, CA. [doi: [10.1109/mobilecloud.2017.22](https://doi.org/10.1109/mobilecloud.2017.22)]
37. Clotet R, Hernandez E, Huerta M, Rivas D. Differentiated synchronization plus FHIR a solution for EMR's ecosystem. 2017 Jun 07 Presented at: 2017 International Caribbean Conference on Devices, Circuits and Systems (ICCDCS); June 5-7, 2017; Cozumel, Mexico. [doi: [10.1109/iccdcs.2017.7959716](https://doi.org/10.1109/iccdcs.2017.7959716)]
38. Khaliq F, Khan SA. An FHIR-based framework for consolidation of augmented EHR from hospitals for public health analysis. 2019 Sep 22 Presented at: IEEE 11th International Conference on Application of Information and Communication Technologies (AICT); September 20-22, 2017; Moscow, Russia. [doi: [10.1109/icaict.2017.8687289](https://doi.org/10.1109/icaict.2017.8687289)]
39. Aliakbarpoor Y, Comai S, Pozzi G. Designing a HL7 compatible personal health record for mobile devices. 2017 Sep 13 Presented at: IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI); September 11-13, 2017; Modena, Italy. [doi: [10.1109/rtsi.2017.8065881](https://doi.org/10.1109/rtsi.2017.8065881)]
40. Hong N, Prodduturi N, Wang C, Jiang G. Shiny FHIR: an integrated framework leveraging Shiny R and HL7 FHIR to empower standards-based clinical data applications. Stud Health Technol Inform 2017;245:868-872 [FREE Full text] [Medline: [29295223](https://pubmed.ncbi.nlm.nih.gov/29295223/)]
41. Leroux H, Metke-Jimenez A, Lawley MJ. Towards achieving semantic interoperability of clinical study data with FHIR. J Biomed Semantics 2017 Sep 19;8(1):41 [FREE Full text] [doi: [10.1186/s13326-017-0148-7](https://doi.org/10.1186/s13326-017-0148-7)] [Medline: [28927443](https://pubmed.ncbi.nlm.nih.gov/28927443/)]
42. Jiang G, Prud'Hommeaux E, Solbrig HR. Developing a semantic web-based framework for executing the clinical quality language using FHIR. 2017 Sep 17 Presented at: CEUR Workshop Proceedings; 2017; Rome, Italy URL: [http://www.swat4ls.org/wp-content/uploads/2017/11/SWAT4LS-2017\\_paper\\_40.pdf](http://www.swat4ls.org/wp-content/uploads/2017/11/SWAT4LS-2017_paper_40.pdf)



43. Walinjkar A, Woods J. FHIR tools for healthcare interoperability. *Biomed J Sci Tech Res* 2018 Oct 10;9(5):1-15. [doi: [10.26717/bjstr.2018.09.001863](https://doi.org/10.26717/bjstr.2018.09.001863)]
44. Kiourtis A, Mavrogiorgou A, Kyriazis D. FHIR Ontology Mapper (FOM): aggregating structural and semantic similarities of ontologies towards their alignment to HL7 FHIR. 2018 Sep 17 Presented at: IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom); September 17-20, 2018; Ostrava, Czech Republic. [doi: [10.1109/healthcom.2018.8531149](https://doi.org/10.1109/healthcom.2018.8531149)]
45. Jeon DC, Lee DH, Hwang H. Reactive server interface design for real-time data exchange in multiple data source and clients. 2018 Oct 17 Presented at: International Conference on Information and Communication Technology Convergence (ICTC); October 17-19, 2018; Jeju, South Korea. [doi: [10.1109/ictc.2018.8539585](https://doi.org/10.1109/ictc.2018.8539585)]
46. Gopinathan K, Kaloumenos AN, Ajmera K, Matei A, Williams I, Davis A. FHIR FLI: an open source platform for storing, sharing and analysing lifestyle data. 2018 Dec 09 Presented at: 4th International Conference on Information and Communication Technologies for Ageing Well and e-Health - ICT4AWE, 227-233, 2018; 2018; Funchal, Madeira, Portugal. [doi: [10.5220/0006791302270233](https://doi.org/10.5220/0006791302270233)]
47. Lackerbauer AM, Lin AC, Krauss O, Hearn J, Helm E. A model for implementing an interoperable electronic consent form for medical treatment using HL7 FHIR. *Eur J Biomed Informatics* 2018 Feb 25;14(3):1-11. [doi: [10.24105/ejbi.2018.14.3.6](https://doi.org/10.24105/ejbi.2018.14.3.6)]
48. Stan O, Miclea L. Local EHR management based on FHIR. 2018 May 26 Presented at: IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR); May 24-26, 2018; Cluj-Napoca, Romania. [doi: [10.1109/aqtr.2018.8402719](https://doi.org/10.1109/aqtr.2018.8402719)]
49. Ahmad A, Azam F, Anwar MW. Implementation of SMART on FHIR in developing countries through SFPBRF. 2018 Nov 23 Presented at: ICBBE '18: Proceedings of the 2018 5th International Conference on Biomedical and Bioinformatics Engineering; November 2018; Okinawa, Japan. [doi: [10.1145/3301879.3301881](https://doi.org/10.1145/3301879.3301881)]
50. Walonoski J, Scanlon R, Dowling C, Hyland M, Ettema R, Posnack S. Validation and testing of Fast Healthcare Interoperability Resources standards compliance: data analysis. *JMIR Med Inform* 2018 Oct 23;6(4):e10870 [FREE Full text] [doi: [10.2196/10870](https://doi.org/10.2196/10870)] [Medline: [30355549](https://pubmed.ncbi.nlm.nih.gov/30355549/)]
51. Urbauer P, Frohner M, David V, Sauermann S. Wearable activity trackers supporting elderly living independently: a standards based approach for data integration to health information systems. 2018 Jun 22 Presented at: Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2018); June 20-22, 2018; Thessaloniki, Greece p. 302-309. [doi: [10.1145/3218585.3218679](https://doi.org/10.1145/3218585.3218679)]
52. Kamel PI, Nagy PG. Patient-centered radiology with FHIR: an introduction to the use of FHIR to offer radiology a clinically integrated platform. *J Digit Imaging* 2018 Jun 3;31(3):327-333 [FREE Full text] [doi: [10.1007/s10278-018-0087-6](https://doi.org/10.1007/s10278-018-0087-6)] [Medline: [29725963](https://pubmed.ncbi.nlm.nih.gov/29725963/)]
53. Borisov V, Minin A, Basko V, Syskov A. FHIR data model for intelligent multimodal interface. 2018 Nov 20 Presented at: 2018 26th Telecommunications Forum (TELFOR); November 20-21, 2018; Belgrade, Serbia. [doi: [10.1109/telfor.2018.8611918](https://doi.org/10.1109/telfor.2018.8611918)]
54. Hussain MA, Langer SG, Kohli M. Learning HL7 FHIR using the HAPI FHIR server and its use in medical imaging with the SIIM dataset. *J Digit Imaging* 2018 Jun 3;31(3):334-340 [FREE Full text] [doi: [10.1007/s10278-018-0090-y](https://doi.org/10.1007/s10278-018-0090-y)] [Medline: [29725959](https://pubmed.ncbi.nlm.nih.gov/29725959/)]
55. Crump JK, Del Fiol G, Williams MS, Freimuth RR. Prototype of a standards-based EHR and genetic test reporting tool coupled with HL7-compliant infobuttons. *AMIA Jt Summits Transl Sci Proc* 2018;2017:330-339 [FREE Full text] [Medline: [29888091](https://pubmed.ncbi.nlm.nih.gov/29888091/)]
56. Peng C, Goswami P, Bai G. Linking health web services as resource graph by semantic REST resource tagging. *Procedia Computer Science* 2018;141:319-326. [doi: [10.1016/j.procs.2018.10.194](https://doi.org/10.1016/j.procs.2018.10.194)]
57. Sharma M, Aggarwal H. Mobile based application for prediction of diabetes mellitus: FHIR Standard. *Int J Eng Technol* 2018 Mar 11;7(2.6):117. [doi: [10.14419/ijet.v7i2.6.10134](https://doi.org/10.14419/ijet.v7i2.6.10134)]
58. Alves N, Ferreira L, Lopes N, Varela M, Castro H, Ávila P, et al. FHIRbox, a cloud integration system for clinical observations. *Procedia Comput Sci* 2018;138:303-309. [doi: [10.1016/j.procs.2018.10.043](https://doi.org/10.1016/j.procs.2018.10.043)]
59. Kasparick M, Andersen B, Ulrich H, Franke S, Schreiber E, Rockstroh M, et al. IEEE 11073 SDC and HL7 FHIR – Emerging Standards for Interoperability of Medical Systems. 2018 Oct 12. URL: [https://www.amd.e-technik.uni-rostock.de/veroeff/2018\\_Kasparick\\_IEEE\\_11073\\_SDC\\_and\\_HL7\\_FHIR.pdf](https://www.amd.e-technik.uni-rostock.de/veroeff/2018_Kasparick_IEEE_11073_SDC_and_HL7_FHIR.pdf) [accessed 2021-07-24]
60. Zohner J, Marquardt K, Schneider H, Michel Backofen A. Challenges and opportunities in changing data structures of clinical document archives from HL7-V2 to FHIR-based archive solutions. *Stud Health Technol Inform* 2019 Aug 21;264:492-495. [doi: [10.3233/SHTI190270](https://doi.org/10.3233/SHTI190270)] [Medline: [31437972](https://pubmed.ncbi.nlm.nih.gov/31437972/)]
61. Maxhelaku S, Kika A. Improving interoperability in Healthcare using HI7 FHIR. 2019 Jan 12 Presented at: 47th International Academic Conference; 2019; Prague. [doi: [10.20472/iac.2019.047.012](https://doi.org/10.20472/iac.2019.047.012)]
62. Oemig F. HL7 Version 2.x Goes FHIR. *Stud Health Technol Inform* 2019 Sep 03;267:93-98. [doi: [10.3233/SHTI190811](https://doi.org/10.3233/SHTI190811)] [Medline: [31483260](https://pubmed.ncbi.nlm.nih.gov/31483260/)]
63. Kiourtis A, Mavrogiorgou A, Menychtas A, Maglogiannis I, Kyriazis D. Structurally mapping healthcare data to HL7 FHIR through ontology alignment. *J Med Syst* 2019 Feb 05;43(3):62. [doi: [10.1007/s10916-019-1183-y](https://doi.org/10.1007/s10916-019-1183-y)] [Medline: [30721349](https://pubmed.ncbi.nlm.nih.gov/30721349/)]

64. Metke-Jimenez A, Hansen D. FHIRCap: Transforming REDCap forms into FHIR resources. *AMIA Jt Summits Transl Sci Proc* 2019;2019:54-63 [FREE Full text] [Medline: 31258956]
65. Mukhiya SK, Rabbi F, I Pun VK, Rutle A, Lamo Y. A GraphQL approach to healthcare information exchange with HL7 FHIR. *Procedia Comput Sci* 2019;160:338-345. [doi: 10.1016/j.procs.2019.11.082]
66. Daumke P, Heitmann KU, Heckmann S, Martínez-Costa C, Schulz S. Clinical text mining on FHIR. *Stud Health Technol Inform* 2019 Aug 21;264:83-87. [doi: 10.3233/SHTI190188] [Medline: 31437890]
67. Schleyer TKL, Rahurkar S, Baublet AM, Kochmann M, Ning X, Martin DK, FHIR Development Team, et al. Preliminary evaluation of the Chest Pain Dashboard, a FHIR-based approach for integrating health information exchange information directly into the clinical workflow. *AMIA Jt Summits Transl Sci Proc* 2019;2019:656-664 [FREE Full text] [Medline: 31259021]
68. Kiourtis A, Mavrogiorgou A, Nifakos S, Kyriazis D. A string similarity evaluation for healthcare ontologies alignment to HL7 FHIR resources. In: Arai K, Bhatia R, Kapoor S, editors. *Advances in Intelligent Systems and Computing*. Cham: Springer; Dec 15, 2019:970-980.
69. Kilintzis V, Kosvyra A, Beredimas N, Natsiavas P, Maglaveras N, Chouvarda I. A sustainable HL7 FHIR based ontology for PHR data. 2019 Jul 27 Presented at: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 23-27, 2019; Berlin, Germany. [doi: 10.1109/embc.2019.8856415]
70. Houta S, Ameler T, Surges R. Use of HL7 FHIR to structure data in epilepsy self-management applications. 2019 Oct 23 Presented at: 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob); Oct 23, 2019; Barcelona, Spain. [doi: 10.1109/wimob.2019.8923179]
71. Pfaff ER, Champion J, Bradford RL, Clark M, Xu H, Fecho K, et al. Fast Healthcare Interoperability Resources (FHIR) as a meta model to integrate common data models: development of a tool and quantitative validation study. *JMIR Med Inform* 2019 Oct 16;7(4):e15199 [FREE Full text] [doi: 10.2196/15199] [Medline: 31621639]
72. Kondylakis H, Petrakis Y, Leivadaros S, Iatraki G, Katehakis D. Using XDS and FHIR to support mobile access to EHR information through personal health apps. 2019 Jun 05 Presented at: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS); June 5-7, 2019; Cordoba, Spain. [doi: 10.1109/cbms.2019.00058]
73. Hong N, Wang K, Wu S, Shen F, Yao L, Jiang G. An interactive visualization tool for HL7 FHIR specification browsing and profiling. *J Healthc Inform Res* 2019 Sep 10;3(3):329-344 [FREE Full text] [doi: 10.1007/s41666-018-0043-8] [Medline: 31598581]
74. Semenov I, Osenev R, Gerasimov S, Kopanitsa G, Denisov D, Andreychuk Y. Experience in developing an FHIR medical data management platform to provide clinical decision support. *Int J Environ Res Public Health* 2019 Dec 20;17(1):73 [FREE Full text] [doi: 10.3390/ijerph17010073] [Medline: 31861851]
75. Chapman M, Curcin V, Sklar EI. A semi-autonomous approach to connecting proprietary EHR standards to FHIR," pp. 1–20, 2019. arXiv. 2019 Nov 27. URL: <https://arxiv.org/abs/1911.12254> [accessed 2020-01-23]
76. Rivera Sánchez YK, Demurjian SA, Baihan MS. *Dig Commun Network* 2019 Nov;5(4):214-225. [doi: 10.1016/j.dcan.2019.10.004]
77. El-Sappagh S, Ali F, Hendawi A, Jang J, Kwak K. A mobile health monitoring-and-treatment system based on integration of the SSN sensor ontology and the HL7 FHIR standard. *BMC Med Inform Decis Mak* 2019 May 10;19(1):97 [FREE Full text] [doi: 10.1186/s12911-019-0806-z] [Medline: 31077222]
78. Eapen BR, Costa A, Archer N, Sartipi K. FHIRForm: An open-source framework for the management of electronic forms in healthcare. *Stud Health Technol Inform* 2019;257:80-85. [Medline: 30741177]
79. Argüello-Casteleiro M, Martínez-Costa C, Desdiz J, Maroto N, Prieto MJF, Stevens R. From SNOMED CT expressions to an FHIR RDF representation: Exploring the benefits of an ontology-based approach. 2019 Jan 19 Presented at: CEUR Workshop; 2019; Graz, Austria p. 1-13.
80. Jenders RA. Evaluation of the Fast Healthcare Interoperability Resources (FHIR) standard for representation of knowledge bases encoded in the Arden syntax. *Stud Health Technol Inform* 2019 Aug 21;264:1692-1693. [doi: 10.3233/SHTI190600] [Medline: 31438296]
81. Hong N, Wen A, Shen F, Sohn S, Wang C, Liu H, et al. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. *JAMIA Open* 2019 Dec;2(4):570-579 [FREE Full text] [doi: 10.1093/jamiaopen/ooz056] [Medline: 32025655]
82. Eapen BR, Archer N, Sartipi K, Yuan Y. Drishti: A sense-plan-act extension to open mHealth framework using FHIR. : IEEE; 2019 May 27 Presented at: 2019 IEEE/ACM 1st International Workshop on Software Engineering for Healthcare (SEH); May 27-29, 2019; Montreal, QC, Canada. [doi: 10.1109/seh.2019.00016]
83. Mandl KD, Gottlieb D, Ellis A. Beyond one-off integrations: a commercial, substitutable, reusable, standards-based, electronic health record-connected app. *J Med Internet Res* 2019 Feb 01;21(2):e12902-e12916 [FREE Full text] [doi: 10.2196/12902] [Medline: 30707097]
84. Baskaya M, Yuksel M, Erturkmen GBL, Cunningham M, Cunningham P. Health4Afrika - Implementing HL7 FHIR based interoperability. *Stud Health Technol Inform* 2019 Aug 21;264:20-24. [doi: 10.3233/SHTI190175] [Medline: 31437877]
85. Roehrs A, da Costa CA, Righi RDR, de Oliveira KSF. Personal health records: a systematic literature review. *J Med Internet Res* 2017 Jan 06;19(1):e13 [FREE Full text] [doi: 10.2196/jmir.5876] [Medline: 28062391]



86. Hume S, Aerts J, Sarnikar S, Huser V. Current applications and future directions for the CDISC Operational Data Model standard: A methodological review. *J Biomed Inform* 2016 Apr;60:352-362 [FREE Full text] [doi: [10.1016/j.jbi.2016.02.016](https://doi.org/10.1016/j.jbi.2016.02.016)] [Medline: [26944737](https://pubmed.ncbi.nlm.nih.gov/26944737/)]
87. SMART. Conduct Science. URL: <https://conductscience.com/digital-health/smart-on-fhir> [accessed 2021-07-24]
88. Sinhasane S. SMART on FHIR: a standard-based, interoperable apps platform to secure EHR records. Mobisoft Infotech. 2018 Oct 19. URL: <https://mobisoftinfotech.com/resources/blog/smart-on-fhir-to-secure-ehr-records/> [accessed 2021-07-24]
89. Clinical and Administrative Domains HL7 Version 2 Product Suite. HL7 International. URL: [https://www.hl7.org/implement/standards/product\\_brief.cfm?product\\_id=185](https://www.hl7.org/implement/standards/product_brief.cfm?product_id=185) [accessed 2021-07-24]
90. Saripalle RK. Fast Health Interoperability Resources (FHIR): current status in the healthcare system. *Int J E-Health Med Commun* 2019 Jan 10;10(1):76-93. [doi: [10.4018/IJEHMC.2019010105](https://doi.org/10.4018/IJEHMC.2019010105)]
91. Lehne M, Luijten S, Vom Felde Genannt Imbusch P, Thun S. The use of FHIR in digital health - a review of the scientific literature. *Stud Health Technol Inform* 2019 Sep 03;267:52-58. [doi: [10.3233/SHTI190805](https://doi.org/10.3233/SHTI190805)] [Medline: [31483254](https://pubmed.ncbi.nlm.nih.gov/31483254/)]

## Abbreviations

**API:** application programming interface  
**CDA:** clinical document architecture  
**EHR:** electronic health record  
**FHIR:** Fast Healthcare Interoperability Resources  
**HAPI:** Health Level 7 application programming interface  
**HL7:** Health Level 7  
**REST:** representational state transfer  
**RFH:** Resources for Healthcare  
**SMART:** substitutable medical applications reusable technologies  
**XML:** extensible markup language

*Edited by R Kukafka, G Eysenbach; submitted 29.06.20; peer-reviewed by R Cornet, G Grieve; comments to author 04.08.20; revised version received 22.09.20; accepted 31.05.21; published 30.07.21.*

*Please cite as:*

Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D

*The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities*

*JMIR Med Inform* 2021;9(7):e21929

URL: <https://medinform.jmir.org/2021/7/e21929>

doi: [10.2196/21929](https://doi.org/10.2196/21929)

PMID: [34328424](https://pubmed.ncbi.nlm.nih.gov/34328424/)

©Muhammad Ayaz, Muhammad F Pasha, Mohammed Y Alzahrani, Rahmat Budiarto, Deris Stiawan. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org>), 30.07.2021. 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

# Multifeature Fusion Attention Network for Suicide Risk Assessment Based on Social Media: Algorithm Development and Validation

Jiacheng Li<sup>1</sup>, BSc; Shaowu Zhang<sup>1</sup>, PhD; Yijia Zhang<sup>1</sup>, PhD; Hongfei Lin<sup>1</sup>, PhD; Jian Wang<sup>1</sup>, PhD

College of Computer Science and Technology, Dalian University of Technology, Dalian, China

**Corresponding Author:**

Yijia Zhang, PhD

College of Computer Science and Technology

Dalian University of Technology

No 2 Linggong Road

Ganjingzi District

Dalian, 116023

China

Phone: 86 13384118909

Email: [zhangyijia1979@gmail.com](mailto:zhangyijia1979@gmail.com)

## Abstract

**Background:** Suicide has become the fifth leading cause of death worldwide. With development of the internet, social media has become an imperative source for studying psychological illnesses such as depression and suicide. Many methods have been proposed for suicide risk assessment. However, most of the existing methods cannot grasp the key information of the text. To solve this problem, we propose an efficient method to extract the core information from social media posts for suicide risk assessment.

**Objective:** We developed a multifeature fusion recurrent attention model for suicide risk assessment.

**Methods:** We used the bidirectional long short-term memory network to create the text representation with context information from social media posts. We further introduced a self-attention mechanism to extract the core information. We then fused linguistic features to improve our model.

**Results:** We evaluated our model on the dataset delivered by the Computational Linguistics and Clinical Psychology 2019 shared task. The experimental results showed that our model improves the risk-F1, urgent-F1, and existence-F1 by 3.3%, 0.9%, and 3.7%, respectively.

**Conclusions:** We found that bidirectional long short-term memory performs well for long text representation, and the attention mechanism can identify the key information in the text. The external features can complete the semantic information lost by the neural network during feature extraction and further improve the performance of the model. The experimental results showed that our model performs better than the state-of-the-art method. Our work has theoretical and practical value for suicidal risk assessment.

(*JMIR Med Inform* 2021;9(7):e28227) doi:[10.2196/28227](https://doi.org/10.2196/28227)

**KEYWORDS**

suicide risk assessment; social media; infodemiology; attention mechanism; neural networks

## Introduction

The World Health Organization's statistical report showed that millions of people choose to commit suicide every year, and even more people are preparing to implement suicide. In 2016, 21.2 in 100,000 people chose to commit suicide worldwide. Moreover, approximately 300,000 people commit suicide in China every year, and the number of suicide attempts is close to 200,000. Suicide has become the fifth leading cause of death

worldwide [1]. The traditional suicide risk assessment method is only dependent on the diagnosis of psychologists, which has great deficiencies with respect to inefficiency and coverage. With development of the internet, social media platforms such as Twitter, Sina Weibo, and WeChat Moments have developed rapidly in recent years. Social media has gradually become an integral part of our lives. People communicate with each other through social media, and use it as a platform to express their emotions and share their opinions, including suicidal social media posters who use these platforms to express their feelings.

It is estimated that 68% of the people who use social media are 10 to 30 years old. Since the high-risk population for suicide is concentrated in the age group of 15 to 29 years, there is considerable overlap between these cohorts [2]. This means that social media is an important data source for studying psychological illnesses such as depression and suicide.

In recent years, text mining based on social media and its psychologically related submedia has become a hot topic in computational linguistics, which provides new research methods for social media-oriented suicide risk assessment. Many scholars have assessed suicide risk by extracting psychological features from texts. For example, Huang et al [3] proposed a method to detect the suicide risk of social media users by identifying mental vocabulary. Zhang et al [4] proposed a method of using linguistic features to assess suicide risk. However, this method has poor detection accuracy and generalization ability, leading to the development of machine learning-based approaches to tackle the task of suicide risk assessment. Kumar et al [5] analyzed the posting activities of posters on the SuicideWatch subreddit that followed celebrity suicide news. They proposed a suicide risk assessment method based on the Werther effect and latent Dirichlet allocation [6] model. De Choudhury et al [7] analyzed the transition process of user tweets from mental health content to suicide content. They proposed a statistical method based on propensity score matching to detect the user's suicidal intent. Bittar et al [8] proposed a method to detect suicide risk using machine learning for electronic health records. Ji et al [9] proposed a new data protection scheme and average difference reduction optimization strategy (AvgDiffLDP) to improve the machine learning model. In addition to machine learning-based methods, deep learning-based methods also have shown good performance in text classification. Shing et al [10] proposed a convolutional neural network (CNN) fused with external dictionary features to detect suicide risk. Mohammadi et al [11] proposed a multichannel classification model including a CNN and recurrent neural network (RNN).

It is necessary to judge the text from different angles when assessing the suicide risk of posts. However, it is difficult for a single model to fully capture the semantic information of the text. Therefore, inspired by previous work [10,11], we here propose a multifeature fusion recurrent attention model for the social media-oriented suicidal risk assessment task. The attention model is used to capture the semantic information in the text and merge it with other external features to better assess the effect.

The main contributions of this paper are divided into the following aspects. First, we propose a recurrent attention model. Using this model to represent the text can extract the core semantic information of the text. We further introduce a distribution loss function to reduce the impact of uneven data distribution.

Second, we fuse external features based on neural networks. These external features are valuable in suicide risk assessment and can further improve the performance of our model.

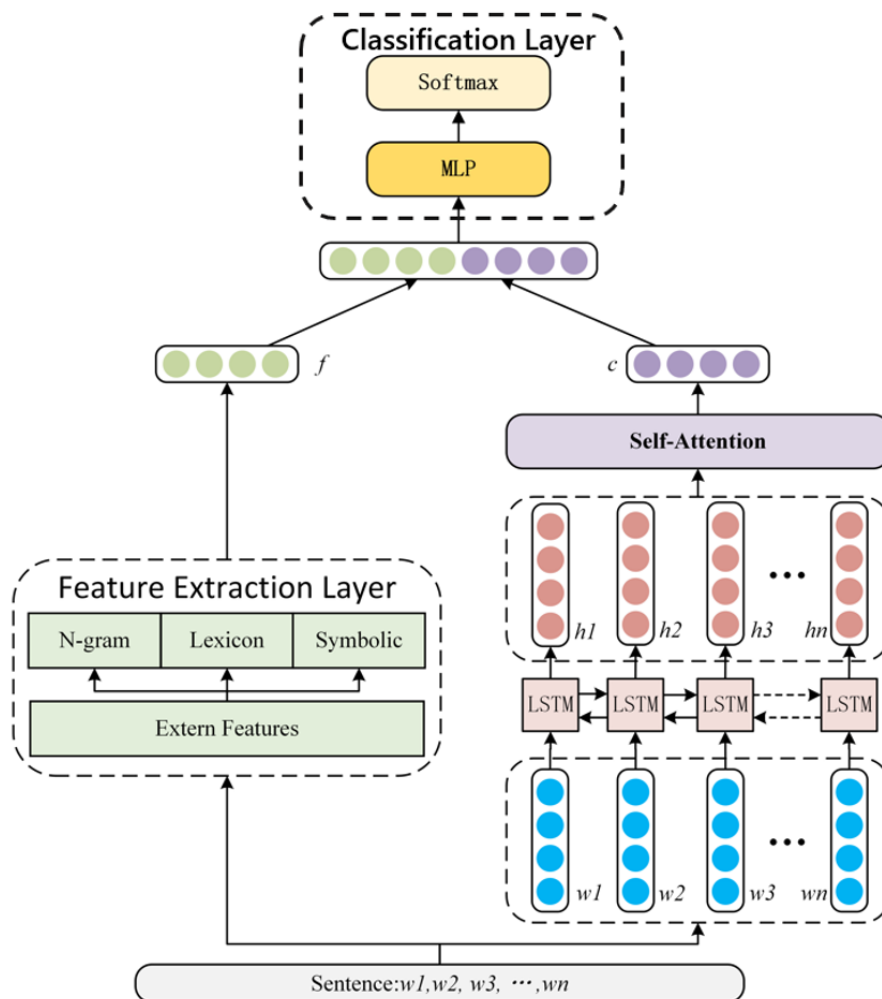
Finally, experimental results showed that our model achieved state-of-the-art performance on the suicide risk assessment dataset, demonstrating that the model has excellent performance and good practical value.

## Methods

### Multifeature Fusion Recurrent Attention Network

The multifeature fusion recurrent attention method proposed in this paper consists of four parts. The framework of our model is shown in Figure 1. The first part of the model uses a long short-term memory network (LSTM) to obtain the text representation  $T$ , which has an attention weight  $\alpha$  in the second part of the model-attention mechanism. The third part of the model is the feature extraction layer, which is used to capture features in the post that are difficult to be extracted by the neural network. The model then fuses the external feature vector with the attention vector to assess suicide risk.

**Figure 1.** Architecture of the multifeature fusion recurrent attention network. LSTM: long short-term memory; MLP: multilayer perceptron.



**LSTM Network**

The LSTM network was proposed by Hochreiter et al [12], which is a variant of the RNN. LSTM introduces a “gate layer” to control neurons to update information, increasing the ability to avoid long-distance dependency problems. LSTM further solves the gradient explosion and gradient disappearance of an RNN when training long text. Therefore, LSTM is the best choice for solving long text classification tasks. The algorithm process of LSTM is as follows:

$$f_k = \sigma(W^f x_k + V^f h_{k-1} + b^f) \quad (1)$$

$$i_k = \sigma(i^f x_k + V^i h_{k-1} + b^i) \quad (2)$$

$$o_k = \sigma(W^o x_k + V^o h_{k-1} + b^o) \quad (3)$$

$$c'_k = \tanh(W^c x_k + V^c h_{k-1} + b^c) \quad (4)$$

$$c'_k = f_k \odot c_{k-1} + i_k \odot c'_k \quad (5)$$

$$h_k = o_k \odot \tanh(c_k) \quad (6)$$

where  $\sigma$  represents the sigmoid function and  $\odot$  represents the element-wise multiplication of two vectors. If an input sequence is  $X=[x_1, x_2, x_3, \dots, x_N]$  for the input  $x_k(1 \leq k \leq N)$  of each position, LSTM needs three steps to output the hidden state  $h_k$ . In the first step, the forget gate *sigmoid* function decides whether the memory cell  $c_k$  needs to forget information based on the hidden

state  $h_{k-1}$  of the previous position and input  $x_k$ . The next step is to decide what information the memory cell needs to update, and this step can be divided into two parts. First, the input gate *sigmoid* function determines whether the memory cell needs to update information. Then, the *tanh* function will generate a new candidate value  $c'_k$ . The new state of the memory cell will be updated under the joint action of the forgetting gate and input gate. In the last step, the hidden state of this position is limited between 0 and 1 under the action of the *tanh* function, and the output gate *sigmoid* function decides whether the neuron needs to output.

LSTM can obtain the information of the current position through the above steps, but the text below is also essential. In the bidirectional LSTM (BiLSTM), the forward LSTM can extract the above information and the backward LSTM can extract the following information. The BiLSTM combines the above hidden state and the below hidden state in the same position to create a new hidden state, which can obtain more context information. The hidden state  $h_k$  of the BiLSTM is shown in Equation 9.



**Self-Attention Layer**

In a sentence, there are only a few words that can represent the semantic information of the entire sentence. If the model treats

every word the same way, the learning ability of the model will be wasted, which will reduce the efficiency of the model. Therefore, we introduce the attention mechanism to this process. This adds an attention weight to each word in the text so that the model will pay more attention to words with higher weights. The attention mechanism has achieved excellent performance in natural language processing tasks owing to its advantages of fewer parameters, faster model training, and stronger interpretability [11].

For the hidden state  $h_t$  from the BiLSTM, the calculation process to obtain the attention weight  $\alpha_t$  is as follows:

$$\alpha_t = \frac{\exp(\beta \cdot \text{score}(h_t, s))}{\sum_{s \in S} \exp(\beta \cdot \text{score}(h_t, s))}$$

where  $\beta$  are trainable parameters and  $\text{score}(h_t, s)$  is the attention score of the input hidden state  $h_t$ . Normalization of  $\alpha_t$  is the softmax function that can provide the attention weight of the input. The vector representation of the entire sentence can then be calculated by Equation 12:

$$H = \sum_{t=1}^T \alpha_t \cdot h_t$$

### Feature Extraction Layer

The neural network focuses on the semantic information of the text, but there are other linguistic features in the text that can help to assess suicide risk. We set up three sets of linguistic features: n-gram features, lexicon-based features, and symbolic features.

For n-gram features, we used bigram and trigram linguistic models as features, and we used term frequency-inverse document frequency (TF-IDF) weights to calculate the feature values. However, the feature matrix is very sparse, and therefore we used nonnegative matrix factorization [13] to reduce the dimension to 50.

For lexicon-based features, since a sentiment word represents the sentiment tendency of the entire text, we introduced the NRC [14] dictionary to capture the posters' emotions. We separately counted the number of emotional words representing

positive emotions, negative emotions, sadness, anger, despair, and fear in a post, and the length of the post. We combined these statistics as a lexicon-based feature vector.

For symbolic features, Stirman et al [15] proposed that suicidal people are self-oriented and they frequently use first-person pronouns. Yang et al [16] proposed that suicidal people frequently use rhetorical rhetoric to emphasize their emotions consciously. In social media posts, emojis are also used to express emotions. Therefore, we counted the number of first-person pronouns (eg, "I," "me," "mine," "myself"), question marks, and emojis in posts as symbolic features.

### Classification Layer

The classification layer used in this study consisted of two parts: a multilayer perceptron and softmax layer. The multilayer perceptron produces classification results and the classification probability is normalized by the softmax layer. We also used the distribution loss function to train the model. Owing to the small number of samples in the dataset, we introduced  $L_2$  regularization to reduce the overfitting problem of the model.

$$L = -\sum_{i=1}^N \sum_{j=1}^M y_j^i \log(\hat{y}_j^i) + \lambda \sum_{i=1}^N \sum_{j=1}^M \hat{y}_j^i$$

where  $N$  is the total number of training data,  $M$  is the number of categories, and  $q_i$  and  $\hat{y}_j^i$  represent the classification result and classification probability, respectively. In Equation 14,  $y_j^i$  is the ground truth,  $\lambda$  is the coefficient of the  $L_2$  regularization term, and  $\theta$  is a hyperparameter. In particular, we introduced the distribution weight  $\gamma$  in the loss function, which is a trainable parameter [17]. Categories with more training data have smaller weights. The distribution loss function can reduce the impact of an uneven data distribution.

### Experimental Settings

Before the experiment, we set the initial parameters based on previous modeling experience. We tuned the model parameters on the development set and achieved the best results. We used Adam to optimize the model. The parameters of the optimal model are shown in Table 1.

**Table 1.** Hyperparameter settings.

Hyperparameters	Optimal value
Word embedding dimension	300
BiLSTM <sup>a</sup> hidden units	200
Learning rate	0.2
Dropout rate	0.5
$L_2$ regularization weight	$10^{-5}$

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

## Results

### Dataset

The suicide risk assessment dataset was released by the Computational Linguistics and Clinical Psychology (CLPsych)

2019 shared task. The goal of CLPsych 2019 was to assess users' suicide risk based on their posts. The dataset constructed by Shin et al [10] in 2018 consists of posts published on the Reddit social media platform between 2005 and 2015. To protect users' privacy, their personal information was replaced by a user ID.



This paper is based on CLPsych-2019 task A (“From Keyboard to Clinic”). Texts used in our dataset were all derived from posts with varying degrees of suicide risk on the SuicideWatch subreddit. The CLPsych dataset was broken down to include 57,016 posts in the training set and 9611 posts in the test set, all from the SuicideWatch subreddit. Among them, the proportion of samples in each category was close to 1:1:1:1. The shortest sentence contained 14 words and the longest sentence contained 486 words. We defined the three following assessment methods to better assess the suicide risk and increase the practicality of the model: (1) suicide risk (risk), which has the same requirements of the CLPsych share task, divided into

four classes *a, b, c, d* from low to high; (2) suicide existence (existence), which is an indicator used to judge whether the poster has a suicidal intention so that the posts can be divided into two levels of exist versus not exist, with the latter indicating a shallow suicide risk (*class a*), and they are not likely to commit suicide in the near future; (3) suicide urgency (urgency), in which the post is divided into two levels of urgent versus not urgent according to the suicide risk, with the urgent level (classes *a, b*) indicating that the user needs psychological assistance urgently.

Table 2 shows the postassessment results obtained under the different suicide risk assessment methods.

**Table 2.** Example posts from the SuicideWatch subreddit.

Post	Risk	Existence	Urgency
A nihilist teetering on edge. Things were good before I came into being	a	Not exist	Not urgent
Has anyone attempted suicide and failed and then felt guilty for being incompetent?	b	Exist	Not urgent
Just sitting on a bench, waiting and thinking. I don't want to, but it feels like the best option.	c	Exist	Urgent
Tell me how to commit suicide painlessly.	d	Exist	Urgent

## Evaluation Metrics

In the experiments, the performance of our model was evaluated by the macroaverage  $F_1$  score. The verification method was as follows:

$$P = TP / (TP + FP) \quad (15)$$

$$R = TP / (TP + FN) \quad (16)$$

$$F_1 = 2 \times P \times R / (P + R) \quad (17)$$

where  $P$  and  $R$  are precision and recall, respectively. TP, FN, and FP represent the true positive, false negative, and false positive predictions, respectively. The  $F_1$  score is a harmonic average of precision and recall.

## Comparison With Baseline

To compare the performance of different models in the suicide assessment task, we tested different classification models on the training set. The experimental results are shown in Table 3.

**Table 3.** Experimental results of classification models.

Models	Risk- $F_1$	Existence- $F_1$	Urgency - $F_1$
SVM <sup>a</sup>	0.296	0.793	0.716
CNN <sup>b</sup>	0.336	0.834	0.742
LSTM <sup>c</sup>	0.397	0.862	0.766
BiLSTM <sup>d</sup>	0.404	0.863	0.774
BiLSTM+CNN	0.423	0.872	0.789
BiLSTM+Attention (proposed model)	0.448	0.887	0.796

<sup>a</sup>SVM: support vector machine.

<sup>b</sup>CNN: convolutional neural network.

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

<sup>d</sup>BiLSTM: bidirectional long short-term memory.

The inputs of the above models are all 300-dimensional Glove word embedding vectors. As shown in Table 3, the performance of the deep learning-based models was better than that of the machine learning-based models. The results of the LSTM and BiLSTM were also better than those of the CNN. In particular, LSTM was better than CNN for long text processing, and the performance of BiLSTM was better than that of LSTM. This shows that BiLSTM can capture more contextual semantic information. The results of the ensemble models were significantly better than those of the single models. In addition, different models showed different capabilities of semantic information extraction, and the combination of different models can supplement the missing semantic information of a single model. The result of the BiLSTM+Attention model was better than that of the BiLSTM+CNN model. This assessment demonstrated that our introduced attention mechanism is more suitable for this task.

### Comparison of Different Input Features

In addition to using the deep learning-based model, we also set up three sets of linguistic features: n-gram features, lexicon-based features, and symbolic features. To test the influence of different features on the suicide risk assessment task, we set up 6 sets of comparative experiments. We separately recorded the experimental results of a support vector machine (SVM) model. The experimental results are shown in [Table 4](#).

The *risk-F<sub>1</sub>* score using TF-IDF features was 0.257. The performance of the n-gram-based method was better than that of TF-IDF. The results of the trigram were better than those of the bigram. Using lexicon features had the most significant improvement on the results, whereas the symbolic features improved the performance to a lesser extent. Concatenating all feature vectors showed that using ensemble features was the best choice for our task, with a *risk-F<sub>1</sub>* score of 0.284.

**Table 4.** Experimental results of different features for support vector machine models.

Input	Risk-F <sub>1</sub>	Existence- F <sub>1</sub>	Urgency -F <sub>1</sub>
TF-IDF <sup>a</sup>	0.257	0.783	0.691
Bigram+TF-IDF	0.271	0.802	0.712
Trigram+TF-IDF	0.276	0.798	0.709
Lexicon+TF-IDF	0.282	0.826	0.721
Symbolic+TF-IDF	0.254	0.784	0.684
n-gram+lexicon+symbolic+TF-IDF	0.284	0.835	0.724

<sup>a</sup>TF-IDF: term frequency-inverse document frequency.

**Table 5.** Experimental results of deep learning-based models.

Models and input	Risk-F <sub>1</sub>	Existence- F <sub>1</sub>	Urgency -F <sub>1</sub>
BERT <sup>a</sup>	0.467	0.889	0.861
<b>BiLSTM<sup>b</sup></b>			
Word2vec	0.404	0.863	0.774
Glove	0.412	0.861	0.793
BERT	0.474	0.914	0.857
BERT+Features	0.481	0.923	0.863
<b>BiLSTM+Attention</b>			
Word2ve	0.448	0.887	0.796
Glove	0.456	0.891	0.787
BERT	0.507	0.915	0.863
BERT+Features	0.514	0.931	0.876

<sup>a</sup>BERT: bidirectional encoder representations from transformers.

<sup>b</sup>BiLSTM: bidirectional long short-term memory.

### Comparison With Other Existing Models

We compared our model with the methods of other teams in the CLPsych 2019 shared task, demonstrating that our model

We further compared the effects of embedding methods on the experimental results. The pretraining language model bidirectional encoder representations from transformers (BERT) can also be used for classification tasks alone. We compared the pretraining language model BERT with the BiLSTM and BiLSTM+Attention models, which showed excellent performance on our task. We used word2vec word embedding [18], Glove word embedding [19], and BERT embedding as the input of the model. The experimental results are shown in [Table 5](#).

The result improved slightly after adding LSTM. Using the pretrained language model BERT resulted in better performance than using the word embedding model. We also concatenated ensemble features at the classification layer, which further improved the performance of the model.

achieved the best results. The *risk-F<sub>1</sub>*, *urgent-F<sub>1</sub>*, and *existing-F<sub>1</sub>* all reached the highest levels with our proposed model ([Table 6](#)).

**Table 6.** Experimental results of existing methods.

Models	Risk- $F_1$	Existence- $F_1$	Urgency- $F_1$
Mohammadi et al [11]	0.481	0.922	0.776
Matero et al [20]	0.459	0.842	0.839
Bitew et al [21]	0.445	0.852	0.789
Iserman et al [22]	0.402	0.902	0.844
Allen et al [23]	0.373	0.876	0.773
González Hevia et al [24]	0.312	0.897	0.821
Multifeature fusion recurrent attention (this study)	0.514 (+0.033)	0.931 (+0.009)	0.876 (+0.037)

Mohammadi et al [11] proposed an ensemble method including 8 neural submodels to extract neural features. They then used the SVM classifier to classify the neural feature vector. They achieved a risk- $F_1$  score of 0.481 and an existence- $F_1$  score of 0.922 (the highest result in CLPsych 2019). González Hevia et al [24] also proposed an ensemble method combined with the result of the SVM classifier and a pretrained RNN. Marero et al [20] proposed multilevel dual-context language and BERT using the deep attention model to extract dual-context information. Their model was also fused with linguistic features and achieved the highest urgency- $F_1$  score of 0.839. Bitew et al [21] proposed a machine learning-based method, and

integrated the logistic regression classifier and the linear SVM classifier. Iserman et al [22] proposed a simple recursive partitioning model with lexicon features. Similarly, Allen et al [23] used CNN and Linguistic Inquiry and Word Count [25] features to assess suicide risk.

### Attention Visualization and Error Analysis

To analyze the effectiveness of the attention mechanism, we extracted the attention weight of the self-attention layer and visualized it with text. The attention visualization results are shown in Figure 2; a deeper color indicates a larger attention weight for the word.

**Figure 2.** Examples of attention visualization.

1. If I knew how to **kill** myself I will do it right now Its the best option right I've got
2. This is it guys I'am **tired of trying** and I cant keep **going**
3. perhaps not the right answer ask here but **how** many do you think have said **their last** words here
4. Been **having** the feeling for over a year

Among the four posts shown in Figure 2, the first two posts are classified into the right class by the model, whereas the last two posts are classified into the wrong category. As shown in the first post, “kill” has the largest weight, which is the core word of this post, and the model also pays attention to “knew” and “do it now.” The model then classified this post into *class d* (high suicide risk). In the second post, the model focused on “tired of trying” and “can’t keep going.” This shows that the model pays attention to words that represent the emotion of the poster. This post lacks the terms associated with high suicide risk, and therefore the model classified this post into *class c*.

In the third post (*class b*), the model focused on the terms “how” and “their last words.” However, the model did not learn that the subject of “last words” was “they” instead of the poster, and therefore mistakenly classified the post into *class d*. In the fourth post (*class a*), the model focused on “having,” “feeling,” and “for a year,” and mistakenly believed that this post reflects a high suicide risk. This is because we found that “feeling” is often associated with words that express negative emotions in the training set. Therefore, we believe that the accuracy can be improved by fusing external features.

## Discussion

### Principal Findings

The results of n-gram features based on TF-IDF weights were better than those obtained using TF-IDF features, which cannot capture the word order information in the text. However, the results of trigram features were inferior to those of bigram features. This shows that although n-gram features can capture the word order information, if multiple features are extracted, the feature vectors will be sparse and reduce the performance of the model. In the experiment, using dictionary features improved the model’s performance significantly. This demonstrates that the emotional tendency of a text can be represented by the limited number of emotional words in the text. The use of symbolic features showed only minor improvements on performance, indicating that punctuation in the text can also express part of the semantic information.

Our model uses the BERT pretraining model as input. The pretrain word vectors represent the semantic information of words, making up the missing information of word embedding models.

The experimental results further showed that BiLSTM performs well in extended text classification. BiLSTM can capture the semantic information of the context in the text and solve long-distance dependence in text processing. After adding the attention mechanism, the performance of the model was further improved. This shows that the attention mechanism can effectively make the model pay attention to the core semantic features of a text.

### Conclusions

This paper proposes a multifeature fusion recurrent attention network to assess the suicide risk of SuicideWatch subreddit posts. Our model uses the BERT pretrained language model as

input, which can create a more precise text representation than the word embedding model. The BiLSTM in the model can capture long-distance dependence and dual-content information. The self-attention mechanism can make the model focus on the core information of the post. The model achieved the best performance on the experimental dataset. Moreover, we introduced n-gram features, lexicon features, and symbolic features, which make up the missing information in the feature extraction of the recurrent attention network, thereby improving the accuracy of the model.

In our future work, we will introduce the personality characteristics of the posters and other social media attributes of the posters for further improving suicide risk assessment.

### Acknowledgments

The work is supported by grants from the National Natural Science Foundation of China (62072070).

### Authors' Contributions

JL designed the algorithm and experiments and wrote the paper. YZ provided theoretical guidance and the revision of this paper. SZ, YZ, HL, and JW contributed to the algorithm design. All authors read and approved the final manuscript.

### Conflicts of Interest

None declared.

### References

1. National Suicide Prevention Strategies: Progress, Examples and Indicators. World Health Organization. 2018. URL: <https://apps.who.int/iris/handle/10665/279765> [accessed 2021-06-30]
2. Lv M, Li A, Liu T, Zhu T. Creating a Chinese suicide dictionary for identifying suicide risk on social media. *PeerJ* 2015;3:e1455. [doi: [10.7717/peerj.1455](https://doi.org/10.7717/peerj.1455)] [Medline: [26713232](https://pubmed.ncbi.nlm.nih.gov/26713232/)]
3. Huang X, Lei L, Liu T. Detecting suicidal ideation in Chinese microblogs with psychological lexicons. 2014 Presented at: 2014 IEEE 11th International Conference on Ubiquitous Intelligence and Computing and 2014 IEEE International Conference on Autonomic and Trusted Computing and 2014 IEEE International Conference on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom); December 9-12, 2014; Bali, Indonesia p. 844-849. [doi: [10.1109/uic-atc-scalcom.2014.48](https://doi.org/10.1109/uic-atc-scalcom.2014.48)]
4. Zhang L, Huang X, Liu T, Li A, Chen Z, Zhu T. Using linguistic features to estimate suicide probability of Chinese microblog users. 2014 Nov 27 Presented at: International Conference on Human Centered Computing; 2014; Phnom Penh, Cambodia p. 549-559. [doi: [10.1007/978-3-319-15554-8\\_45](https://doi.org/10.1007/978-3-319-15554-8_45)]
5. Kumar M, Dredze M, Coppersmith G, De Choudhury M. Detecting changes in suicide content manifested in social media following celebrity suicides. *HT ACM Conf Hypertext Soc Media 2015 Sep*;2015:85-94 [FREE Full text] [doi: [10.1145/2700171.2791026](https://doi.org/10.1145/2700171.2791026)] [Medline: [28713876](https://pubmed.ncbi.nlm.nih.gov/28713876/)]
6. Blei DM, Ng A, Jordan MI. Latent dirichllocation. *J Machine Learn Res* 2003 Mar 4;3:993-1022. [doi: [10.1162/jmlr.2003.3.4-5.993](https://doi.org/10.1162/jmlr.2003.3.4-5.993)]
7. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. Discovering shifts to suicidal ideation from mental health content in social media. *Proc SIGCHI Conf Hum Factor Comput Syst* 2016 May;2016:2098-2110 [FREE Full text] [doi: [10.1145/2858036.2858207](https://doi.org/10.1145/2858036.2858207)] [Medline: [29082385](https://pubmed.ncbi.nlm.nih.gov/29082385/)]
8. Bittar A, Velupillai S, Roberts A, Dutta R. Text classification to inform suicide risk assessment in electronic health records. *Stud Health Technol Inform* 2019 Aug 21;264:40-44. [doi: [10.3233/SHTI190179](https://doi.org/10.3233/SHTI190179)] [Medline: [31437881](https://pubmed.ncbi.nlm.nih.gov/31437881/)]
9. Ji S, Long G, Pan S, Zhu T, Jiang J, Wang S. Detecting suicidal ideation with data protection in online communities. 2019 Presented at: International Conference on Database Systems for Advanced Applications; 2019; Thailand p. 225-229. [doi: [10.1007/978-3-030-18590-9\\_17](https://doi.org/10.1007/978-3-030-18590-9_17)]
10. Shing HC, Nair S, Zirikly A, Friedenber M, Daumé III H, Resnik P. Expert, crowdsourced, and machine assessment of suicide risk via online postings. 2018 Presented at: Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic; June 2018; New Orleans p. 25-36. [doi: [10.18653/v1/w18-0603](https://doi.org/10.18653/v1/w18-0603)]
11. Mohammadi E, Amini H, Kosseim L. ClaC at CLPsych 2019: Fusion of Neural Features and Predicted Class Probabilities for Suicide Risk Assessment Based on Online Posts. 2019 Presented at: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology; 2019; Minneapolis, MN p. 34-38. [doi: [10.18653/v1/W19-3004](https://doi.org/10.18653/v1/W19-3004)]

12. 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/)]
13. Févotte C, Idier J. Algorithms for nonnegative matrix factorization with the  $\beta$ -divergence. *Neur Comput* 2011 Sep;23(9):2421-2456. [doi: [10.1162/neco\\_a\\_00168](https://doi.org/10.1162/neco_a_00168)]
14. Mohammad SM. Word Affect Intensities. 2018 Presented at: Eleventh International Conference on Language Resources and Evaluation; May 2018; Miyazaki, Japan.
15. Stirman SW, Pennebaker JW. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosom Med* 2001;63(4):517-522. [doi: [10.1097/00006842-200107000-00001](https://doi.org/10.1097/00006842-200107000-00001)] [Medline: [11485104](https://pubmed.ncbi.nlm.nih.gov/11485104/)]
16. Yang Y, Zheng L, Zhang J. TI-CNN: Convolutional neural networks for fake news detection. arxiv. URL: <https://arxiv.org/abs/1806.00749> [accessed 2018-06-03]
17. Lin T, Goyal P, Girshick R. Focal loss for dense object detection. *IEEE Trans Patt Anal Machine Intell* 2017;42(2):318-327. [doi: [10.1109/iccv.2017.324](https://doi.org/10.1109/iccv.2017.324)]
18. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013 Presented at: 1st International Conference on Learning Representations, ICLR 2013; May 2-4, 2013; Scottsdale, AZ.
19. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25-29, 2015; Doha p. 1532-1543. [doi: [10.3115/v1/d14-1162](https://doi.org/10.3115/v1/d14-1162)]
20. Matero M, Idnani A, Son Y, Giorgi S, Vu H, Zamani M, et al. Suicide risk assessment with multi-level dual-context language and BERT. 2019 Presented at: Sixth Workshop on Computational Linguistics and Clinical Psychology; 2019; Minneapolis, MN p. 39-44. [doi: [10.18653/v1/w19-3005](https://doi.org/10.18653/v1/w19-3005)]
21. Bitew SK, Bekoulis G, Deleu J, Sterckx L, Zaparojets K, Demeester T, et al. Predicting suicide risk from online postings in Reddit The UGent-IDLab submission to the CLPsych 2019 shared Task A. 2019 Presented at: Sixth Workshop on Computational Linguistics and Clinical Psychology; 2019; Minneapolis, MN p. 158-161. [doi: [10.18653/v1/w19-3019](https://doi.org/10.18653/v1/w19-3019)]
22. Iserman M, Nalabandian T, Ireland M. Dictionaries and decision trees for the 2019 CLPsych Shared Task. 2019 Presented at: Sixth Workshop on Computational Linguistics and Clinical Psychology; 2019; Minneapolis, MN p. 188-194. [doi: [10.18653/v1/w19-3025](https://doi.org/10.18653/v1/w19-3025)]
23. Allen K, Bagroy S, Davis A, Krishnamurti T. ConvSent at CLPsych 2019 Task A: using post-level sentiment features for suicide risk prediction on Reddit. 2019 Presented at: Sixth Workshop on Computational Linguistics and Clinical Psychology; 2019; Minneapolis, MN p. 182-187.
24. González Hevia A, Cerezo Menéndez R, Gayo-Avello D. Analyzing the use of existing systems for the CLPsych 2019 Share Task. 2019 Presented at: Sixth Workshop on Computational Linguistics and Clinical Psychology; 2019; Minneapolis, MN p. 148-151. [doi: [10.18653/v1/w19-3017](https://doi.org/10.18653/v1/w19-3017)]
25. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The Development and psychometric properties of LIWC. 2015 Sep 15. URL: <https://repositories.lib.utexas.edu/handle/2152/31333> [accessed 2015-09-15]

## Abbreviations

**BERT:** bidirectional encoder representations from transformers  
**BiLSTM:** bidirectional long short-term memory network  
**CLPsych:** Computational Linguistics and Clinical Psychology  
**CNN:** convolutional neural network  
**LSTM:** long short-term memory network  
**RNN:** recurrent neural network  
**SVM:** support vector machine  
**TF-IDF:** term frequency-inverse document frequencies

*Edited by T Hao; submitted 25.02.21; peer-reviewed by C Sun, S Wang; comments to author 19.04.21; revised version received 30.04.21; accepted 05.05.21; published 09.07.21.*

*Please cite as:*

Li J, Zhang S, Zhang Y, Lin H, Wang J

Multifeature Fusion Attention Network for Suicide Risk Assessment Based on Social Media: Algorithm Development and Validation  
*JMIR Med Inform* 2021;9(7):e28227

URL: <https://medinform.jmir.org/2021/7/e28227>

doi: [10.2196/28227](https://doi.org/10.2196/28227)

PMID: [34255687](https://pubmed.ncbi.nlm.nih.gov/34255687/)



©Jiacheng Li, Shaowu Zhang, Yijia Zhang, Hongfei Lin, Jian Wang. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 09.07.2021. 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

# Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation

Lu Ren<sup>1</sup>, PhD; Hongfei Lin<sup>1</sup>, PhD; Bo Xu<sup>1,2</sup>, PhD; Shaowu Zhang<sup>1</sup>, PhD; Liang Yang<sup>1</sup>, PhD; Shichang Sun<sup>3</sup>, PhD

<sup>1</sup>Dalian University of Technology, Dalian, China

<sup>2</sup>State Key Lab for Novel Software Technology, Nanjing University, Nanjing, China

<sup>3</sup>Dalian Minzu University, Dalian, China

**Corresponding Author:**

Liang Yang, PhD

Dalian University of Technology

No. 2 Linggong Road

Dalian,

China

Phone: 86 041184706009

Email: [liang@dlut.edu.cn](mailto:liang@dlut.edu.cn)

## Abstract

**Background:** As a common mental disease, depression seriously affects people's physical and mental health. According to the statistics of the World Health Organization, depression is one of the main reasons for suicide and self-harm events in the world. Therefore, strengthening depression detection can effectively reduce the occurrence of suicide or self-harm events so as to save more people and families. With the development of computer technology, some researchers are trying to apply natural language processing techniques to detect people who are depressed automatically. Many existing feature engineering methods for depression detection are based on emotional characteristics, but these methods do not consider high-level emotional semantic information. The current deep learning methods for depression detection cannot accurately extract effective emotional semantic information.

**Objective:** In this paper, we propose an emotion-based attention network, including a semantic understanding network and an emotion understanding network, which can capture the high-level emotional semantic information effectively to improve the depression detection task.

**Methods:** The semantic understanding network module is used to capture the contextual semantic information. The emotion understanding network module is used to capture the emotional semantic information. There are two units in the emotion understanding network module, including a positive emotion understanding unit and a negative emotion understanding unit, which are used to capture the positive emotional information and the negative emotional information, respectively. We further proposed a dynamic fusion strategy in the emotion understanding network module to fuse the positive emotional information and the negative emotional information.

**Results:** We evaluated our method on the Reddit data set. The experimental results showed that the proposed emotion-based attention network model achieved an accuracy, precision, recall, and F-measure of 91.30%, 91.91%, 96.15%, and 93.98%, respectively, which are comparable results compared with state-of-the-art methods.

**Conclusions:** The experimental results showed that our model is competitive with the state-of-the-art models. The semantic understanding network module, the emotion understanding network module, and the dynamic fusion strategy are effective modules for depression detection. In addition, the experimental results verified that the emotional semantic information was effective in depression detection.

(*JMIR Med Inform* 2021;9(7):e28754) doi:[10.2196/28754](https://doi.org/10.2196/28754)

**KEYWORDS**

depression detection; attention network; emotional semantic information; dynamic fusion strategy; natural language processing; social media; emotion; mental health; algorithm; deep learning

## Introduction

### Background

As defined in the free dictionary, depression refers to the act of depressing or state of being depressed. Depression is usually regarded as one type of mood disorder; the main clinical feature of depression is the significant and persistent mood depression. The depressed patients' emotion can range from gloomy to grief, low self-esteem, and even to pessimism, which may cause suicidal attempts or behaviors [1]. The World Psychiatric Association set October 10 as the World Mental Health Day in 1992 to strengthen the awareness of the public on mental disorders. The latest report released by the World Health Organization (WHO) pointed out that [2] there were approximately 322 million patients with depression in the world, and the prevalence rate was about 4.4%. The number of patients with depression is growing year by year. From 2005 to 2015, the number of patients with depression worldwide increased by 18.4%. According to the statistics of the WHO [2], depression is one of the 20 main reasons that can cause suicide in the world, accounting for about 1.5% of suicides. It also accounts for the highest proportion of disability among the global diseases and is the main factor of global nonfatal health loss.

With the development of the internet in people's daily life, people began to share their feelings and problems on social media [3,4] such as Reddit and Twitter. The research of Park et al [5] showed that people with depression tend to post information about depression and even treatment on social media. Thus, we can get a lot of valuable information from social media. If we can judge whether a person has depression based on the information from the internet, it can help the doctors intervene early and avoid the happening of self-injury or suicide. Many researchers, coming from different disciplines such as computer science and psychology, have paid much attention on this topic. In addition, some advanced methods are proposed for depression detection. However, the detection accuracy still needs to be improved.

The goal of depression detection is to classify a person or a post as depressed or not. The performance of depression detection

on social media can help with the clinical treatment of depression. This problem needs to be solved. The posts of patients with depression usually contain strong emotions. We give three examples of the textual posts left on Reddit, including two depression-indicative posts and one standard post as follows.

- Example 1: "Today, I feel so horrible, it makes me want to die I made a fool of myself at work, felt so stupid after the meeting so I left work, told the boss I'm sick. Spent the remaining afternoon in bed." Label: depression
- Example 2: "That feeling when you hate who you are as a person but can't get yourself to change because you are so used to being like this for the past several years. I've become a shitty person. The thought of change seems impossible to me at this point." Label: depression
- Example 3: "Looking for cool ways to tell parents my wife is pregnant." Label: nondepression

Examples 1 and 2 contain strong emotional information made by the patients with depression. From example 1, the words, including *horrible*, *die*, and *stupid*, express strong negative emotions of the author. The words *hate* and *shitty* in example 2 also express the author's strong negative emotions. Example 3 shows the post of a regular user. It does not contain strong negative emotions. As previously mentioned, emotional semantic information usually provides us useful clues for depression detection.

We also counted the proportion of the positive words and the negative words that appeared in the depression-indicative posts and the standard posts of the Reddit data set [6], respectively. The statistical results are shown in Table 1. The percentage of positive emotion words in the table is calculated by  $\frac{\text{positive words}}{\text{total words}}$ . The percentage of negative emotion words was similar. In addition, we calculated the percentages of emotion words in the depression-indicative posts and the standard posts. The depressed users used more negative words than the nondepressed users. At the same time, they used less positive words in their posts than the nondepressed users. It can be concluded from the statistical results that the emotional semantic information may play an effective role for the depression detection task.

**Table 1.** Percentage of emotion words in posts.

Categories	Depression-indicative posts (%)	Standard posts (%)
Positive emotion words	8.62	9.41
Negative emotion words	6.70	4.85

Detecting depression automatically has made some progress. Many existing models detect depression based on the feature engineering such as bag of words [7,8], latent Dirichlet allocation (LDA) [9,10], N-gram [11], Linguistic Inquiry and Word Count (LIWC) dictionary [12], or their combinations [4,13,14]. Bag of words, LDA, and N-gram have been widely used in natural language processing (NLP) for feature extraction and have achieved great progress. LIWC can carry out quantitative analysis on the word categories (especially psychological words) of the text content, including the sentiment, emotion, and so on. Emotion extracted by LIWC is

often used in the depression detection task. With the development of deep learning in NLP, more and more studies use deep learning models for depression detection. Orabi et al [15] proposed a method based on deep learning (convolutional neural network [CNN] and recurrent neural network [RNN]) to detect depression. Gui et al [16] proposed a reinforcement learning method based on RNN for depression detection. Although these advanced deep learning based models can extract higher-level semantic information and have achieved great progress, they still lack effective extraction of the emotional semantic information. This may limit the ability of their model

because the emotional information may bring effective clues for depression detection, as shown in examples 1 and 2.

Before introducing our model and to understand our paper more conveniently, we give several definitions of concepts, including high-level emotional semantic information, semantic understanding network (SUN), emotion understanding network (EUN), and dynamic fusion strategy.

- High-level emotional semantic information denotes the emotional semantic information that is captured by deep learning.
- SUN is a deep learning method that is used to capture the contextual semantic information in the text for depression detection.
- EUN is a deep learning method that is used to capture the emotional semantic information in the text for depression detection.
- Dynamic fusion strategy denotes a fusion strategy that can fuse positive emotional information and negative emotional information dynamically.

To extract the emotional information effectively, we propose an emotion-based attention network (EAN) for depression detection. Our EAN model mainly contains two modules, including a SUN and an EUN. The SUN module is used to capture the contextual semantic information, which has been widely used in NLP. The EUN module is used to capture the emotional information because the emotional information plays an important role for depression detection as previously mentioned. As shown in [Table 1](#), the depression-indicative posts contained more negative words and less positive words, and the standard posts contained less negative words and more positive words. Thus, we designed the EUN module. The EUN module contains two units, including a positive emotion understanding unit and a negative emotion understanding unit, which are used to extract the positive emotional information and the negative emotional information, respectively. Apart from it, we also propose a dynamic fusion strategy in the EUN module to fuse the positive emotion information and the negative emotion information.

The main contributions of this paper can be summarized as follows:

- We propose a new deep learning framework for depression detection. We also design a special module to explicitly extract the high-level emotion information for depression detection in our framework.
- We take into consideration the positive emotion information and the negative emotion information simultaneously. At the same time, we apply a dynamic fusion strategy to fuse the positive emotion information and the negative information.
- We conduct experiments on the Reddit data set for depression detection. The experiments show our model can get state-of-the-art or comparable performance. The ablation study also verifies the effectiveness of the components proposed in our model.

## Related Work

In this section, we review the related work about depression detection on social media.

In recent years, with the development of social media, more and more people are willing to post their thoughts, emotions, or life details on social media, including Reddit, Twitter, and so on. Park et al [5] showed that people with depression tend to post information about depression and even treatment on social media. Thus, we can get a lot of valuable information from social media. More and more researchers began to analyze the mental health of the users based on the information from social media. As a result, depression detection based on social media has attracted a lot of attention.

De Choudhury et al [17] collected data from Twitter about the users with depression and the regular user, and combined the difference between their behavior on social media (depressed users manifested as decreased social activities, increased negative emotions and self-concern, a high degree and increased expression of religious thoughts, etc) and established a characteristic model for depression detection. Park et al [18] tested for users with depression through social media and conducted semistructured face-to-face interviews with 14 active users. The study concluded that users with depression regarded social media as a platform for social awareness and emotional sharing, while users with nondepression regarded social media as a platform for sharing information. Thus, emotional information is important in the task of detecting depression in social media.

Most of the existing methods for depression detection are based on feature engineering. LIWC is usually used to extract individual psychological states, such as positive and negative emotions, pronouns, and so on. Therefore, LIWC was often used for the depression detection task [4,12-14]. Kang et al [19] proposed a multimodal method for depression detection including text analysis, a word-based emoticon analysis, and a support vector machine-based image classifier. The authors applied visual sentiment ontology [20] and SentiStrength dictionaries to build a mood lexicon for emoticon analysis to enhance the results of depression detection. Shen et al [21] extracted six depression-related feature groups (including social network feature, user profile feature, visual feature, emotional feature, topic-level feature, and domain-specific feature) for depression detection. Hiraga [22] extracted linguistic features for depression detection, including character n-grams, token n-grams, and lemmas and selected lemmas. Hussain et al [3] developed an application called the Socially Mediated Patient Portal. The application could generate a series of features for depression detection.

Shneidman [23] presented depression that tended to be closely related to suicide. De Choudhury et al [24] analyzed Reddit users' posts on the topic of mental health that later turned to the topic of suicidal thoughts. This turn could be predicted by traits such as self-focus, poor language style, reduced social engagement, and expressions of despair or anxiety. Yates et al [25] proposed a neural framework for depression detection, and they presented that self-harm was closely related to depression. The Conference and Labs of Evaluation Forum for Early Risk

Prediction (CLEF eRISK) is a public competition about different areas such as health and safety [26]. CLEF eRISK 2018 is about the early detection of depression and anorexia [8,27]. CLEF eRISK 2019 is about the severity of symptoms of depression, self-injury, and anorexia [28].

Different from traditional feature engineering-based methods, deep learning methods mostly apply end-to-end models. Yates et al [25] proposed a neural framework based on a CNN for depression detection. Orabi et al [15] proposed a neural method based on a CNN and RNN for depression detection. Song et al [29] proposed a neural network that was named the feature attention network for depression detection. Gui et al [16] proposed a reinforcement learning method based on long short-term memory (LSTM) for depression detection. Ray et al [30] proposed a multilevel attention network to fuse the features from the multimodal for depression detection.

According to previous research on depression detection, it can be concluded that the emotional information is important in the task of depression detection. In addition, deep learning can take high-level semantic information into account, but the current deep learning methods for depression detection still lack effective extraction of the emotional semantic information. Thus, we propose a deep learning model to consider the high-level emotional information that is captured by the deep learning method for depression detection, which is named the EAN.

The structure of this paper is organized as follows. The Introduction section introduced the background and related work. The Methods section shows the details of the proposed model. The Results section gives the experiments in this paper. The Discussion section shows the conclusions and future work.

## Methods

### Data Sets

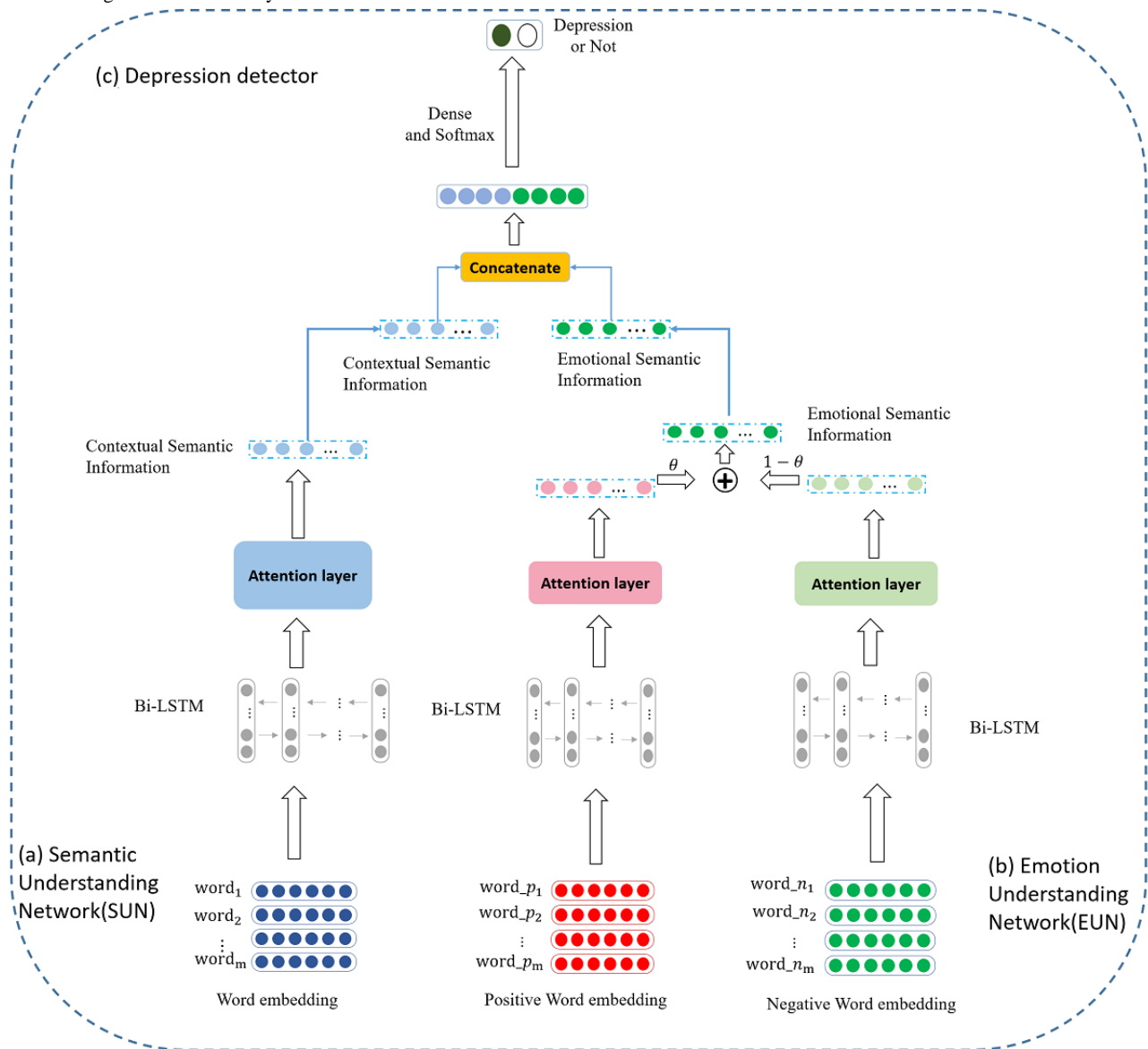
As a newly developed social media, Reddit has become a widely popular web-based discussion forum. Reddit users can discuss a variety of topics on this web-based platform anonymously. The topics discussed on the platform can be arranged in more than a million discussion groups. Due to the large amount of discussion text, Reddit attracts many researchers to conduct their studies with the data on the Reddit platform. Pirina and Çöltekin [6] built a data set for depression detection based on Reddit, which was named the Reddit data set. The samples in the Reddit data set [6] are collected from the Reddit platform. The Reddit data set [6] contains 1293 depression-indicative posts and 549 standard posts.

We preprocessed the Reddit data set, such as removing the stop words. We then counted the occurrence number of each word for the depression-indicative posts and the standard posts. We sorted the words according to the statistics and show the top of the word lists in Figure 1. We also counted the occurrence number of the positive emotion words and the negative emotion words for the depression-indicative posts and the standard posts. For all of the words, the positive emotion words and the negative emotion words with high frequency of occurrence are also shown in Textbox 1.

As shown in Textbox 1, from the most commonly used words of the depressed users, we can see many negatives words are also included in the most commonly used words such as depression or fucking. The most common words for nondepressed people are commonly used words in daily life. As can be seen from the list of negative words with high frequency of occurrence used by users with depression, the negative words used by users with depression are more intense than the negative words appearing in the posts of nondepressed users, such as suicide, die, kill, and hate.



**Figure 1.** The architecture of the emotion-based attention network model. There are two parts in our model, including a SUN and an EUN. bi-LSTM: bidirectional long short-term memory.



**Textbox 1.** Data analysis.**Depression-indicative posts**

- All text: i'm, like, feel, want, get, know, even, really, people, life, i've, one, time, think, would, never, depression, me, can't, go, going, things, don't, much, friends, make, good, it, still, could, back, anyone, years, anything, always, every, got, someone, fucking, help, day, see, something, work, ever, need, feeling, everything, talk, year
- Positive: friends, good, work, help, better, happy, job, love, hard, friend, family, care, wanted, best, sleep, sure, self, mind, understand, new, mental, hope, social, money, high, remember, working, reason, okay, close, together, great, normal, deal, believe, change, enjoy, birthday, honestly, nice, motivation, advice, loved, therapist, happiness, fun, boyfriend, saying, big
- Negative: depression, depressed, bad, fucking, nothing, alone, hate, shit, stop, lost, worse, anxiety, fuck, tired, sad, die, suicide, kill, relationship, wrong, pain, suicidal, problems, old, sorry, cry, lonely, therapy, hurt, stupid, constantly, issues, sick, crying, problem, afraid, weird, reddit, hospital, worst, hang, illness, dead, scared, dark, broken, shitty, broke, miserable, died

**Standard posts**

- All text: like, i'm, know, friend, would, feel, really, friends, want, time, get, one, even, said, always, never, told, got, family, go, things, me, think, best, make, mom, going, people, years, talk, also, still, back, something, much, see, say, could, i've, dad, tell, since, don't, started, us, me, it, made, help, parents
- Positive: friend, friends, family, best, sister, help, friendship, work, brother, good, new, sure, love, wanted, saying, together, advice, father, close, money, boyfriend, kids, care, hard, better, mad, understand, job, basically, happy, great, deal, child, high, moved, believe, fun, social, mind, baby, conversation, eventually, reason, married, big, change, spend, real, normal, nice
- Negative: bad, wrong, nothing, old, hang, problem, stop, hurt, upset, sorry, shit, issues, lost, alone, cut, angry, hate, problems, worse, depression, weird, sick, constantly, anxiety, sad, tired, annoyed, broke, bitch, scared, died, hell, afraid, crying, cancer, toxic, ignore, pregnant, lose, difficult, wait, fault, depressed, horrible, awkward, selfish, reply, fuck, confused, reddit

**Overview of the EAN Model**

In this section, we introduce the proposed model for depression detection briefly, which is called the EAN, as shown in Figure 1. The proposed EAN model mainly contains two parts, including a SUN and an EUN. The SUN module is used to capture the contextual semantic information in the depression-indicative posts. The EUN module is used to capture the emotional semantic information in the depression-indicative posts. Finally, we concatenated the features captured by the two parts and judged whether the text is depression-indicative or not by the depression detector. We give details on the SUN, the EUN, and the loss function next.

**Semantic Understanding Network**

The SUN was used to capture the contextual semantic information in the text for depression detection. There are three layers in the SUN module, including the word encoding layer, context encoding layer, and attention mechanism (Att) layer. We will introduce these three layers in more details.

**Word Encoding Layer**

We will introduce the word encoding layer in the SUN module briefly. The input of our task is text. The text can be denoted as  $w = \{w_1, w_2, \dots, w_n\}$ , where  $n$  denotes the length of the text, and  $w_i$  denotes the word in the text. In NLP tasks, words are usually mapped to the form of word vectors. Inspired by it, we also encoded every word into  $d$ -dimension word vector. We applied the pretrained Global Vectors for Word Representation (GloVe) [31] here. We then can get the textual representation  $S = R^{n \times d}$ , where  $n$  is the textual length and  $d$  is the dimension of the word.

**Context Encoding Layer**

The context encoding layer was used to obtain contextual information. Bidirectional long short-term memory (Bi-LSTM) [32] was widely used in NLP tasks to capture the contextual information. Inspired by this, we applied Bi-LSTM in the context encoding layer. Bi-LSTM contains a forward directional LSTM and a backward directional LSTM. The output Bi-LSTM contains two parts, including the forward LSTM output and the backward LSTM output.

LSTM was proposed by Hochreiter and Schmidhuber [33] and was used to capture the forward information in the text. LSTM cannot capture the backward information; therefore, Bi-LSTM was proposed. LSTM owns three gates and one cell, including an input gate  $i_t$ , a forget gate  $f_t$ , an output gate  $o_t$ , and a memory cell  $c_t$ . The operations of LSTM are as following.



Where  $x_t$  is the current input word vector,  $\otimes$  means the elementwise multiplication operation, and  $\sigma$  means the sigmoid function.  $W_f$ ,  $W_i$ ,  $W_c$ , and  $W_o$  represent the parameters that can be trained in the training processing.  $h_t$  is the hidden state vector.

$\otimes$  is the output of LSTM. More details on LSTM can be found in Hochreiter and Schmidhuber [33], and the output of Bi-LSTM is  $H = [H_1, H_2, \dots, H_n]$ .

**Attention Mechanism Layer**

The input of the Att layer is  $H = [H_1, H_2, \dots, H_n]$ . The Att is used to assign higher weights on the important words. We applied the Att to capture the important words in the depression-indicative posts for the depression detection task. The operations of the Att are based on the following equations:



Where  $H_i$  is the hidden state vector of Bi-LSTM,  $w$  and  $q_i$  are the weighted matrices, and  $h_{att}$  is the output of the Att.

### Emotion Understanding Network

Many research papers [19-21] and their experiments have proven the effectiveness of emotional feature in depression detection tasks. Inspired by this, we considered the high-level emotional semantic information in the depression-indicative posts based on the EUN. The EUN was used to capture the emotional semantic information in the text for depression detection. There are three layers in the EUN module, including the input layer, emotion encoding layer, and emotion fusion layer. We introduce these three layers in more detail in the following sections.

#### Input Layer

In this section, we introduce the inputs of the EUN module. The inputs include a positive emotion part and a negative emotion part. We applied the SenticNet application programming interface to divide the original texts into a positive emotional part and a negative emotional part. These two emotional parts are also mapped into a matrix of word vectors as in the word encoding layer in the SUN module, named  $R_{pos}$  and  $R_{neg}$ , respectively.

#### Emotion Encoding Layer

The emotion encoding layer is to encode the positive emotional information and the negative emotional information.  $R_{pos}$  and  $R_{neg}$  act as the inputs of the emotion encoding layer. There are two units in the emotion encoding layer, including the positive emotion understanding unit and the negative emotion understanding unit. These two units are used to capture positive emotional information and negative emotional information, respectively. We also applied Bi-LSTM to capture the contextual emotional information and the Att to capture the important emotions in the text in both units. The operations of Bi-LSTM and the Att are the same as the EUN module. We can get  $h_{pos}$  from the positive emotion understanding unit and  $h_{neg}$  from the negative emotion understanding unit.

#### Emotion Fusion Layer

The goal of the emotion fusion layer is to fuse the positive emotional information and the negative emotional information for depression detection. We get the positive emotional information  $h_{pos}$  and the negative emotional information  $h_{neg}$  from the emotion encoding layer, which can be learned in the training processing. Considering the difference of each text, we designed a dynamic fusion strategy that can dynamically fuse the positive emotional information  $h_{pos}$  and the negative emotional information  $h_{neg}$ . Inspired by the Att, we design a random floating point number  $\theta \in [0, 1]$ . It can be trained during the training. We can get the output  $h_{emo}$  of the EUN module with the following formula:

$$h_{emo} = \theta * h_{pos} + (1 - \theta) * h_{neg} \quad (10)$$

### Loss Function

As previously described, we get the contextual semantic information  $h_{att}$  from the SUN module and the emotional semantic information  $h_{emo}$  from the EUN module. In this section, we applied a concatenation operation to fuse the contextual semantic information  $h_{att}$  and the emotional semantic information  $h_{emo}$  as the final representation  $f_{final}$ :

$$f_{final} = \text{concatenate}[h_{att}; h_{emo}] \quad (11)$$

Accordingly, the final classification decision for depression detection is formulated by the softmax function:

$$y = \text{softmax}(W \cdot f_{final} + b) \quad (12)$$

The cross-entropy loss was used for depression detection in our model. The training goal was to minimize the loss.

## Results

### Implementation Details and Metrics

The unit size of Bi-LSTM in our experiments was 64. We applied the pretrained 300-dimension word embedding (GloVe) in the word encoding layer. In addition, the optimization function was Adam, and the batch size was 128. Following Tadesse et al [4], we also applied a 10-fold cross validation in our experiments; 90% of posts in the data sets were used as our training set, and the other 10% of posts were used as the testing set.

We applied the standard metrics, including accuracy, precision, recall, and F1-score, to evaluate the effectiveness of our model for depression detection. F1 is defined as follows:



### Comparison With Existing Methods

We compared the results of our model with many state-of-the-art methods on the Reddit data set. We compared it with the baselines, including LIWC, LDA, unigram, bigram, LIWC + LDA + unigram, LIWC + LDA + bigram [4], LSTM, Bi-LSTM, and Bi-LSTM + Att.

- LIWC: Tadesse et al [4] extracted the linguistic features and the psychological features based on LIWC [34] for depression detection.
- LDA: Tadesse et al [4] extracted 70 dimensional characteristics of the topic based on LDA. It can be helpful in discovering its underlying topic structures for depression detection.
- Unigram: Tadesse et al [4] extracted 3000 dimensional characteristics based on unigram in term frequency-inverse document frequency (TF-IDF) for depression detection.
- Bigram: Tadesse et al [4] extracted 2736 dimensional characteristics based on bigram in TF-IDF for depression detection.
- LIWC + LDA + unigram: The model is based on the aforementioned characteristics, including LIWC, LDA, and unigram, for depression detection.

- LIWC + LDA + bigram: The model is based on the aforementioned characteristics, including LIWC, LDA, and bigram, for depression detection.
- LSTM: LSTM was proposed by Hochreiter and Schmidhuber [33]. We applied the same word embedding in this paper, and the unit size was 128.
- Bi-LSTM: The Bi-LSTM was proposed by Graves et al [32]. We applied the same setting and the same word embedding in this paper.
- Bi-LSTM + Att: The model is based on Bi-LSTM and the Att.
- EAN: This model is proposed in this paper, which considers emotional semantic information based on deep learning.

As shown in [Table 2](#), the results based on deep learning are generally higher than the results based on feature engineering

methods. It is because deep learning can capture the higher semantic information of texts. In addition, we can also get the following conclusions.

The results based on bigram (bigram and LIWC + LDA + bigram) were higher than unigram (unigram and LIWC + LDA + unigram). It can be concluded that contextual information can improve the results of the model. The results based on Bi-LSTM were higher than LSTM. It can be concluded that considering bidirectional contextual semantic information is necessary. The results based on Bi-LSTM + Att were higher than Bi-LSTM; it can be proven that the Att is effective for the depression detection task. The proposed EAN model got the higher results because we took into consideration both the contextual semantic information and the emotional semantic information.

**Table 2.** Results compared with the existing models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
LIWC <sup>a,b</sup>	70	74	71	72
LDA <sup>b,c</sup>	75	75	72	74
Unigram <sup>b</sup>	70	71	95	81
Bigram <sup>b</sup>	79	80	76	78
LIWC + LDA + unigram <sup>b</sup>	78	84	79	81
LIWC + LDA + bigram <sup>b</sup>	91	90	92	91
LSTM <sup>d</sup>	87.03	90.30	91.67	90.98
Bi-LSTM <sup>e</sup>	86.46	88.08	95	91.41
Bi-LSTM + Att <sup>f</sup>	88.59	90.41	94.96	92.63
EAN <sup>g</sup> (our model)	91.3	91.91	96.15	93.98

<sup>a</sup>LIWC: Linguistic Inquiry and Word Count.

<sup>b</sup>Indicates that the results are shown in the literature [4].

<sup>c</sup>LDA: latent Dirichlet allocation.

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

<sup>e</sup>Bi-LSTM: bidirectional long short-term memory.

<sup>f</sup>Att: attention mechanism.

<sup>g</sup>EAN: emotion-based attention network.

## Detail Analysis

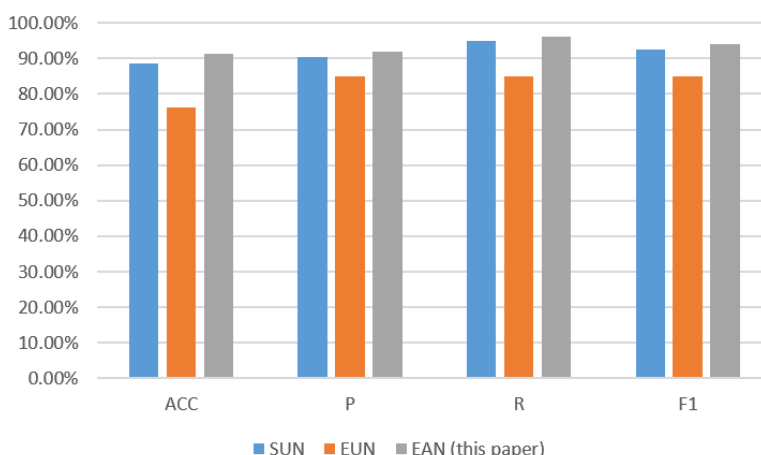
In this section, we analyze the effectiveness of the two modules (SUN and EUN), the effectiveness of different emotional semantic information, and the effectiveness of the dynamic fusion strategy.

### *The Effectiveness of SUN and EUN*

To verify the effectiveness of SUN and EUN, we designed a series of experiments. SUN means the proposed EAN model

without the EUN module. EUN means the proposed EAN model without the SUN module. As shown in [Figure 2](#), the EUN module obtained the worst results. This is because the model only considers the emotional semantic information without the complete semantic information. It verifies the effectiveness of our SUN module. The results of the EAN model were higher than the SUN module, which further verifies the effectiveness of our EUN module.

**Figure 2.** The effectiveness of the SUN and EUN. EAN: emotion-based attention network; EUN: emotion understanding network; SUN: semantic understanding network.

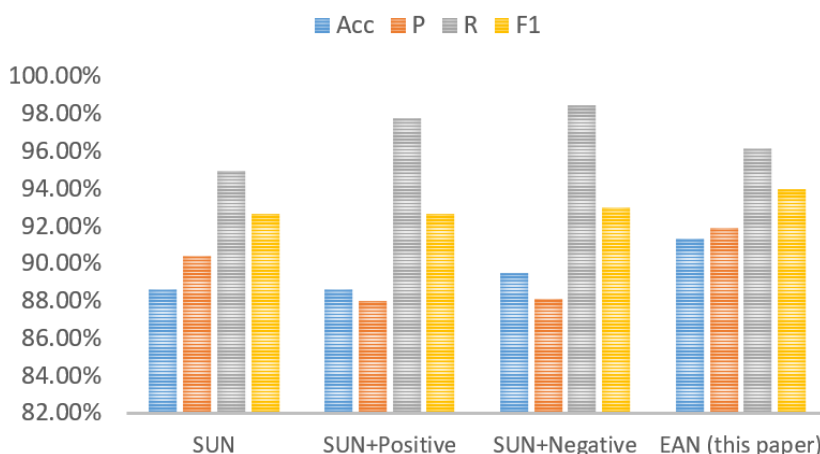


**The Effectiveness of Different Emotional Semantic Information**

To verify the effectiveness of different emotional semantic information, we designed a series of experiments, including without emotion (SUN), without positive emotion (SUN + negative), and without negative emotion (SUN + positive). As shown in Figure 3, the results of the SUN + positive model and

the SUN module were similar. It indicates that positive emotions have less effect on the model. Although the EAN model does not obtain the best recall value, it obtained the best P value, ACC value, and F1 value. From the experiments, our proposed EAN model obtained the best result compared to the three aforementioned baseline models. It also verified the effectiveness of each proposed module in our framework.

**Figure 3.** The effectiveness of different emotional semantic information. Acc: accuracy; EAN: emotion-based attention network; P: precision; R: recall; SUN: semantic understanding network.



**The Effectiveness of the Dynamic Fusion Strategy**

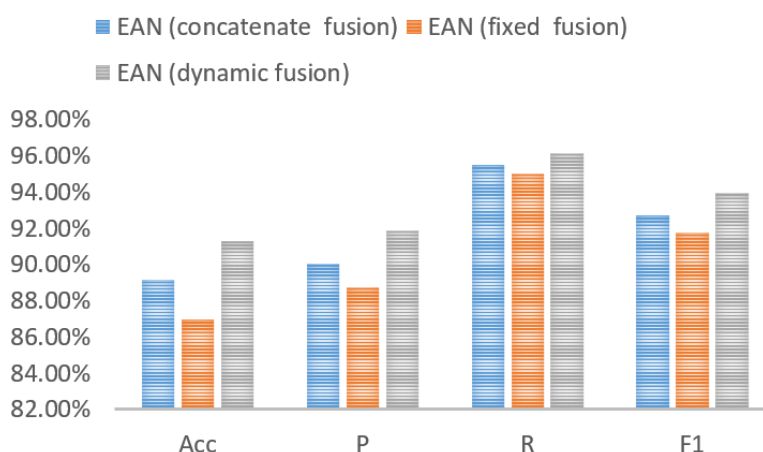
To verify the effectiveness of the dynamic fusion strategy, we designed a series of experiments including the EAN model with the concatenate fusion strategy, the EAN model with the fixed fusion strategy, and the EAN model with the dynamic fusion strategy. The EAN (concatenate fusion) model applies the concatenate operation in the emotion fusion strategy. The EAN (fixed fusion) model applies the fixed fusion operation in the emotion fusion layer. The  $\theta$  in equation 10 is fixed at 0.5. The EAN (dynamic fusion) model is the model proposed in this paper. As shown in Figure 4, the dynamic fusion method had the best results.

In this section, we designed a series of experiments to verify the effectiveness of the proposed EAN model, including the two modules in the EAN model, the different emotional semantic information, and the dynamic fusion method.

Some visualization results of the  $\theta$  to illustrate the effectiveness of the proposed dynamic fusion strategy intuitively are shown in Figure 5. As shown in Figure 5, the examples are both depression-indicative posts. The pie chart indicates the value of the  $\theta$  in the dynamic fusion strategy. We can see from the results that in the depression-indicative posts, the negative emotional information can be paid more attention.



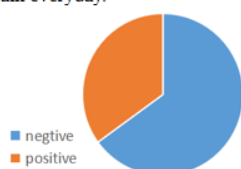
**Figure 4.** The effectiveness of the dynamic fusion strategy. Acc: accuracy; EAN: emotion-based attention network; P: precision; R: recall.



**Figure 5.** The visualization of the  $\theta$  in the dynamic fusion strategy. GF: girlfriend.

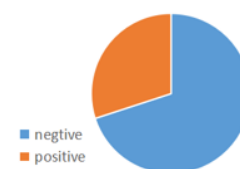
I attempted suicide and called my GF, who called the cops who then took me to the hospital. I kinda wish I had just died. We had just broken up and I got the urges to just end my life finally, I've felt like this for years upon years. I've lost so many people in my life recently due to deaths over the last few years, and this just brought me over the edge. I wish I could have ended my life so I don't have to wake up with so much pain everyday.

Label :Depression  
Predict label: Depression



Today, I feel so horrible, it makes me want to die I made a fool of myself at work, felt so stupid after the meeting so I left work, told the boss I'm sick. Spent the remaining afternoon in bed.

Label :Depression  
Predict label: Depression



## Discussion

### Conclusion

Depression attracts more and more attention from people and organizations now. With the development of computer technology, some researchers are trying to use computers to automatically identify people who are depressed. In this paper, we proposed an EAN model to explicitly extract the high-level emotion information for the depression detection task. The proposed EAN model consists of the SUN and the EUN. In the proposed model, we took into consideration the positive emotion information and the negative emotion information simultaneously. At the same time, we applied a dynamic fusion

strategy to fuse the positive emotion information and the negative information. The experimental results verified that the emotional semantic information is effective in depression detection.

### Future Work

According to WHO statistics, depression is one of the main causes of suicide in the world. We will focus on the relationship between depression and suicide. We will try to combine suicide detection with depression detection in our future work to improve the performance of both tasks by multitask learning. In addition, the future work will be combined with self-reported depressive symptoms or clinical diagnosis. Hopefully, our study can provide some technical supports in the field of health care.

### Acknowledgments

This study was partially supported by a grant from the Natural Science Foundation of China (No. 62076046, 61632011, 62006034, 61876031), the Ministry of Education Humanities and Social Science Project (No. 19YJCZH199), State Key Laboratory of Novel Software Technology (Nanjing University; No. KFKT2021B07), and the Fundamental Research Funds for the Central Universities (No. DUT21RC(3)015).

### Conflicts of Interest

None declared.

### References

1. Friedrich M. Depression is the leading cause of disability around the world. JAMA 2017 Apr 18;317(15):1517. [doi: [10.1001/jama.2017.3826](https://doi.org/10.1001/jama.2017.3826)] [Medline: [28418490](https://pubmed.ncbi.nlm.nih.gov/28418490/)]
2. Depression and other common mental disorders: global health estimates. World Health Organization. 2017. URL: <https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf> [accessed 2021-06-28]

3. Hussain J, Satti FA, Afzal M, Khan WA, Bilal HSM, Ansaar MZ, et al. Exploring the dominant features of social media for depression detection. *J Inf Sci* 2019 Aug 12;46(6):739-759. [doi: [10.1177/0165551519860469](https://doi.org/10.1177/0165551519860469)]
4. Tadesse MM, Lin H, Xu B, Yang L. Detection of depression-related posts in Reddit social media forum. *IEEE Access* 2019;7:44883-44893. [doi: [10.1109/access.2019.2909180](https://doi.org/10.1109/access.2019.2909180)]
5. Park M, Cha C, Cha M. Depressive moods of users portrayed in twitter. 2012 Presented at: ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD); August 12-16, 2012; Beijing, China.
6. Pirina I, Çöltekin Ç. Identifying depression on reddit: The effect of training data. 2018 Presented at: 018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task; October 2018; Brussels, Belgium p. 9-12. [doi: [10.18653/v1/w18-5903](https://doi.org/10.18653/v1/w18-5903)]
7. Nadeem M. Identifying depression on twitter. arXiv. Preprint posted online on July 25, 2016 [FREE Full text]
8. Paul S, Kalyani JS, Basu T. Early detection of signs of anorexia and depression over social media using effective machine learning frameworks. 2018 Presented at: CLEF 2018; September 10-14, 2018; Avignon, France p. 1-9.
9. Maupomé D, Meurs MJ. Using topic extraction on social media content for the early detection of depression. 2018 Presented at: CLEF 2018; September 10-14, 2018; Avignon, France p. 2125.
10. Resnik P, Armstrong W, Claudino L, Nguyen T, Nguyen VA, Boyd-Graber J. Beyond exploring supervised topic modeling for depression-related language in Twitter. 2015 Presented at: 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; June 5, 2015; Denver, Colorado p. e. [doi: [10.3115/v1/w15-1212](https://doi.org/10.3115/v1/w15-1212)]
11. Benton A, Mitchell M, Hovy D. Multi-task learning for mental health using social media text. arXiv. Preprint posted online on December 10, 2017 [FREE Full text]
12. Coppersmith G, Dredze M, Harman C, Hollingshead K. From ADHD to SAD: analyzing the language of mental health on twitter through self-reported diagnoses. 2015 Presented at: 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; June 5, 2015; Denver, Colorado p. 1-10. [doi: [10.3115/v1/w15-1201](https://doi.org/10.3115/v1/w15-1201)]
13. Wolohan JT, Hiraga M, Mukherjee A, Sayyed ZA. Detecting linguistic traces of depression in topic restricted text: attending to self-stigmatized depression with NLP. 2018 Presented at: The First International Workshop on Language Cognition and Computational Models; August 20, 2018; Santa Fe, New Mexico p. 11-21.
14. Tyschenko Y. Depression and anxiety detection from blog posts data. CORE. 2018. URL: <https://core.ac.uk/download/pdf/237085027.pdf> [accessed 2021-06-28]
15. Orabi AH, Buddhitha P, Orabi MH, Inkpen D. Deep learning for depression detection of Twitter users. 2018 Presented at: Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic; June 2018; New Orleans, LA p. 88-97. [doi: [10.18653/v1/w18-0609](https://doi.org/10.18653/v1/w18-0609)]
16. Gui T, Zhang Q, Zhu L, Zhou X, Peng M, Huang X. Depression detection on social media with reinforcement learning. In: Sun M, Huang X, Ji H, Liu Z, Liu Y, editors. Chinese Computational Linguistics 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings. Cham: Springer; 2019.
17. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. 2013 Presented at: Seventh International AAI Conference on Weblogs and Social Media; July 8-11, 2013; Cambridge, MA p. 1-10.
18. Park M, McDonald D, Cha M. Perception differences between the depressed and non-depressed users in Twitter. 2013 Presented at: Seventh International AAI Conference on Weblogs and Social Media; July 8-11, 2013; Cambridge, MA.
19. Kang K, Yoon C, Kim EY. Identifying depressive users in twitter using multimodal analysis. 2016 Presented at: International Conference on Big Data and Smart Computing (BigComp); January 18-20, 2016; Hong Kong, China p. 231-238. [doi: [10.1109/bigcomp.2016.7425918](https://doi.org/10.1109/bigcomp.2016.7425918)]
20. Borth D, Ji R, Chen T, Breuel T, Chang SF. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of the 21st ACM International Conference on Multimedia. 2013 Presented at: MM '13; October 21-25, 2013; Barcelona, Spain. [doi: [10.1145/2502081.2502282](https://doi.org/10.1145/2502081.2502282)]
21. Shen G, Jia J, Nie L, Feng F, Zhang C, Hu T, et al. Depression detection via harvesting social media: a multimodal dictionary learning solution. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017 Presented at: IJCAI-17; August 19-25, 2017; Melbourne, Australia p. 3838-3844. [doi: [10.24963/ijcai.2017/536](https://doi.org/10.24963/ijcai.2017/536)]
22. Hiraga M. Predicting depression for Japanese blog text. 2017 Presented at: ACL 2017, Student Research Workshop; July 2017; Vancouver, Canada p. 107-113. [doi: [10.18653/v1/p17-3018](https://doi.org/10.18653/v1/p17-3018)]
23. Shneidman ES. Suicide as psychache. *J Nerv Ment Dis* 1993 Mar;181(3):145-147. [doi: [10.1097/00005053-199303000-00001](https://doi.org/10.1097/00005053-199303000-00001)] [Medline: [8445372](https://pubmed.ncbi.nlm.nih.gov/8445372/)]
24. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. Discovering shifts to suicidal ideation from mental health content in social media. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 2016 Presented at: CHI '16; May 7-12, 2016; San Jose, CA p. 2098-2110. [doi: [10.1145/2858036.2858207](https://doi.org/10.1145/2858036.2858207)]
25. Yates A, Cohan A, Goharian N. Depression and self-harm risk assessment in online forums. arXiv. Preprint posted online on September 6, 2017 [FREE Full text]
26. Losada DE, Crestani F, Parapar J. eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. In: Jones GJF, Lawless S, Gonzalo J, Kelly L, Goeuriot L, Mandl T, et al, editors. Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings. Cham: Springer; 2017.

27. Trotzek M, Koitka S, Friedrich CM. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Trans Knowledge Data Eng* 2020 Mar 1;32(3):588-601. [doi: [10.1109/tkde.2018.2885515](https://doi.org/10.1109/tkde.2018.2885515)]
28. Losada DE, Crestani F, Parapar J. Overview of eRisk 2019 early risk prediction on the internet. In: Crestani F, Braschler M, Savoy J, Rauber A, Müller H, Losada DE, et al, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings*. Cham: Springer; 2019.
29. Song H, You J, Chung JW, Park JC. Feature attention network: interpretable depression detection from social media. 2018 Presented at: 32nd Pacific Asia Conference on Language, Information and Computation; December 1-3, 2018; Hong Kong, China.
30. Ray A, Kumar S, Reddy R, Mukherjee P, Garg R. Multi-level attention network using text, audio and video for depression prediction. In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 2019 Presented at: AVEC '19; October 21, 2019; Nice, France p. 81-88. [doi: [10.1145/3347320.3357697](https://doi.org/10.1145/3347320.3357697)]
31. Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 2014 Presented at: EMNLP '14; October 2014; Doha, Qatar p. 1532-1543. [doi: [10.3115/v1/d14-1162](https://doi.org/10.3115/v1/d14-1162)]
32. Graves A, Jaitly N, Mohamed AR. Hybrid speech recognition with Deep Bidirectional LSTM. 2013 Presented at: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding; December 8-12, 2013; Olomouc, Czech Republic p. 273-278. [doi: [10.1109/asru.2013.6707742](https://doi.org/10.1109/asru.2013.6707742)]
33. 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/)]
34. Pennebaker JW, Booth RJ, Boyd RL, Francis ME. *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates. 2001. URL: [http://downloads.liwc.net.s3.amazonaws.com/LIWC2015\\_OperatorManual.pdf](http://downloads.liwc.net.s3.amazonaws.com/LIWC2015_OperatorManual.pdf) [accessed 2021-06-28]

## Abbreviations

**Att:** attention mechanism

**Bi-LSTM:** bidirectional long short-term memory

**CLEF eRISK:** Conference and Labs of Evaluation Forum for Early Risk Prediction

**CNN:** convolutional neural network

**EAN:** emotion-based attention network

**EUN:** emotion understanding network

**GloVe:** Global Vectors for Word Representation

**LDA:** latent Dirichlet allocation

**LIWC:** Linguistic Inquiry and Word Count

**LSTM:** long short-term memory

**NLP:** natural language processing

**RNN:** recurrent neural network

**SUN:** semantic understanding network

**TF-IDF:** term frequency-inverse document frequency

**WHO:** World Health Organization

*Edited by T Hao, Z Huang, B Tang; submitted 13.03.21; peer-reviewed by T Qian, J Han; comments to author 05.05.21; revised version received 11.05.21; accepted 19.05.21; published 16.07.21.*

*Please cite as:*

Ren L, Lin H, Xu B, Zhang S, Yang L, Sun S

*Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation*

*JMIR Med Inform* 2021;9(7):e28754

URL: <https://medinform.jmir.org/2021/7/e28754>

doi: [10.2196/28754](https://doi.org/10.2196/28754)

PMID: [34269683](https://pubmed.ncbi.nlm.nih.gov/34269683/)

©Lu Ren, Hongfei Lin, Bo Xu, Shaowu Zhang, Liang Yang, Shichang Sun. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org>), 16.07.2021. 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

# Automatic Extraction of Lung Cancer Staging Information From Computed Tomography Reports: Deep Learning Approach

Danqing Hu<sup>1,2</sup>, MSc; Huanyao Zhang<sup>1,2</sup>, BSc; Shaolei Li<sup>3</sup>, MD; Yuhong Wang<sup>1,2</sup>, MSc; Nan Wu<sup>3</sup>, MD; Xudong Lu<sup>1,2</sup>, PhD

<sup>1</sup>College of Biomedical Engineering and Instrumental Science, Zhejiang University, Hangzhou, China

<sup>2</sup>Key Laboratory for Biomedical Engineering, Ministry of Education, Hangzhou, China

<sup>3</sup>Department of Thoracic Surgery II, Peking University Cancer Hospital & Institute, Beijing, China

**Corresponding Author:**

Xudong Lu, PhD

College of Biomedical Engineering and Instrumental Science

Zhejiang University

38 Zheda Road

Hangzhou, 310027

China

Phone: 86 13957118891

Email: [lvxd@zju.edu.cn](mailto:lvxd@zju.edu.cn)

## Abstract

**Background:** Lung cancer is the leading cause of cancer deaths worldwide. Clinical staging of lung cancer plays a crucial role in making treatment decisions and evaluating prognosis. However, in clinical practice, approximately one-half of the clinical stages of lung cancer patients are inconsistent with their pathological stages. As one of the most important diagnostic modalities for staging, chest computed tomography (CT) provides a wealth of information about cancer staging, but the free-text nature of the CT reports obstructs their computerization.

**Objective:** We aimed to automatically extract the staging-related information from CT reports to support accurate clinical staging of lung cancer.

**Methods:** In this study, we developed an information extraction (IE) system to extract the staging-related information from CT reports. The system consisted of the following three parts: named entity recognition (NER), relation classification (RC), and postprocessing (PP). We first summarized 22 questions about lung cancer staging based on the TNM staging guideline. Next, three state-of-the-art NER algorithms were implemented to recognize the entities of interest. Next, we designed a novel RC method using the relation sign constraint (RSC) to classify the relations between entities. Finally, a rule-based PP module was established to obtain the formatted answers using the results of NER and RC.

**Results:** We evaluated the developed IE system on a clinical data set containing 392 chest CT reports collected from the Department of Thoracic Surgery II in the Peking University Cancer Hospital. The experimental results showed that the bidirectional encoder representation from transformers (BERT) model outperformed the iterated dilated convolutional neural networks-conditional random field (ID-CNN-CRF) and bidirectional long short-term memory networks-conditional random field (Bi-LSTM-CRF) for NER tasks with macro-F1 scores of 80.97% and 90.06% under the exact and inexact matching schemes, respectively. For the RC task, the proposed RSC showed better performance than the baseline methods. Further, the BERT-RSC model achieved the best performance with a macro-F1 score of 97.13% and a micro-F1 score of 98.37%. Moreover, the rule-based PP module could correctly obtain the formatted results using the extractions of NER and RC, achieving a macro-F1 score of 94.57% and a micro-F1 score of 96.74% for all the 22 questions.

**Conclusions:** We conclude that the developed IE system can effectively and accurately extract information about lung cancer staging from CT reports. Experimental results show that the extracted results have significant potential for further use in stage verification and prediction to facilitate accurate clinical staging.

(*JMIR Med Inform* 2021;9(7):e27955) doi:[10.2196/27955](https://doi.org/10.2196/27955)

**KEYWORDS**

lung cancer; clinical staging; information extraction; named entity recognition; relation classification



## Introduction

### Background

Lung cancer is a group of diseases involving abnormal cell growth in the lung tissue with the potential to invade adjoining parts of the body and spread to other organs. It is the most commonly diagnosed cancer and the leading cause of cancer deaths worldwide [1], which has been a heavy burden on communities and a critical barrier to increasing life expectancy.

Clinical staging of lung cancer plays a critical role in making treatment decisions making and evaluating prognosis [2]. In current clinical practice, clinicians usually decide the clinical staging of lung cancer. Although various advanced diagnostic modalities with high sensitivity and specificity are used by clinical experts, clinical staging still disagrees with pathological staging in approximately one-half of patients, as reported in earlier studies [3,4]. Incorrect clinical staging of lung cancer may result in suboptimal treatment decisions, possibly leading to poor outcomes [3].

As an indispensable examination technique for lung cancer patients, chest computed tomography (CT) provides a large volume of valuable information about the primary tumor and lymph nodes, which is of paramount importance for clinical staging [2,5]. Besides, the reports record the inferences of radiologists about the findings from the images. Although this useful information in the form of natural language is effective and convenient for communication in medical clinical settings, its free-text nature poses difficulties when summarizing or analyzing this information for secondary purposes such as research and quality improvement. Moreover, manually extracting this information is time-consuming and expensive [6,7].

In this study, we aimed to develop an information extraction (IE) system to automatically extract valuable information from CT reports using natural language processing (NLP) techniques to support accurate clinical staging. We first summarized 22 questions about the diagnosis and staging of lung cancer based on the TNM stage guideline [8]. Subsequently, 14 types of entities and 4 types of relations were defined to represent the related information in the CT reports. Using the annotated reports, the following three state-of-the-art deep learning named entity recognition (NER) models were developed to label the entities: iterated dilated convolutional neural networks (ID-CNN) [9], bidirectional long short-term memory networks (Bi-LSTM) [10], and bidirectional encoder representation from transformers (BERT) [11]. Next, a novel relation classification (RC) approach using the relation sign constraint (RSC) was proposed to determine the relations between entities. Finally, a rule-based postprocessing (PP) module was developed to obtain the formatted results by analyzing the entities and relations extracted by NER and RC. We empirically evaluated our system using a real clinical data set. Experimental results showed that the system could extract entities and relations as well as obtain the answers to the questions correctly. Using these extracted results, we can verify the clinical staging accuracy and further develop staging prediction models to alleviate the problem of inaccurate clinical staging.

### Related Works

IE refers to the task of automatically extracting structured semantics (eg, entities, relations, and events) from unstructured text. Cancer information is often extracted from free-text clinical narratives, such as operation notes, radiology, and pathology reports, using rule-based, machine learning, or hybrid methods, which have been widely investigated [12]. In terms of staging information, most studies have extracted only the clinical or pathological stage statements (eg, Stage I, Stage II, and T3N2) but not detailed phenotypes [13-20]. Besides the stage statements, Savova et al [21] and Ping et al [22] extracted some tumor-related information such as the location and size. However, these extracted phenotypes are considerably limited in their ability to support staging, particularly for lymph nodes. To support diagnosis and staging, Yim et al [23] employed a hybrid method to recognize diverse entities and relations from radiology reports for hepatocellular cancer patients, but without further elaboration on how to exploit the extracted information. Chen et al [24] extracted information from various clinical notes including operation notes and CT reports to calculate the Cancer of Liver Italian Program (CLIP) score for hepatocellular cancer patients; however, they provide limited details about the radiology corpus extraction. Bozkurt et al first developed an IE pipeline to extract various types of information from mammography reports [25] and then used the extracted features as the inputs for Bayesian networks to predict malignancy of breast cancer [26].

These rule-based and conventional machine learning methods have extracted information about cancer successfully, and some of them have exploited the extracted results to provide further diagnosis and staging decision support. Nevertheless, the development of hand-craft features and usage of external resources like the Unified Machine Language System (UMLS) and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) are time-consuming and can even result in additional propagation errors [10,27,28]. Recently, with the rapid development of deep neural networks, advanced approaches exhibit excellent performance in many NLP tasks without tedious feature engineering [27-33]. Furthermore, some researchers began to adopt these advanced techniques to extract cancer information. Si et al [34] proposed a frame-based NLP method using Bi-LSTM-conditional random field (Bi-LSTM-CRF) to extract cancer-related information by a two-stage strategy. They first identified the keywords in the sentences to determine their frames and then employed models to label the entities in this frame. Using this strategy, they grouped the related entities by different frames. A limitation of this study is that they only evaluated each process in the pipeline using gold standard annotations separately but did not report the overall results of the pipeline. Gao et al [35] proposed a novel hierarchical attention network to predict the primary sites and histological grades of tumors in a text classification manner. Although this approach can directly provide the classification results and show the importance of each word in the text, the scope of the information extracted is considerably limited and insufficient to support cancer diagnosis and staging.

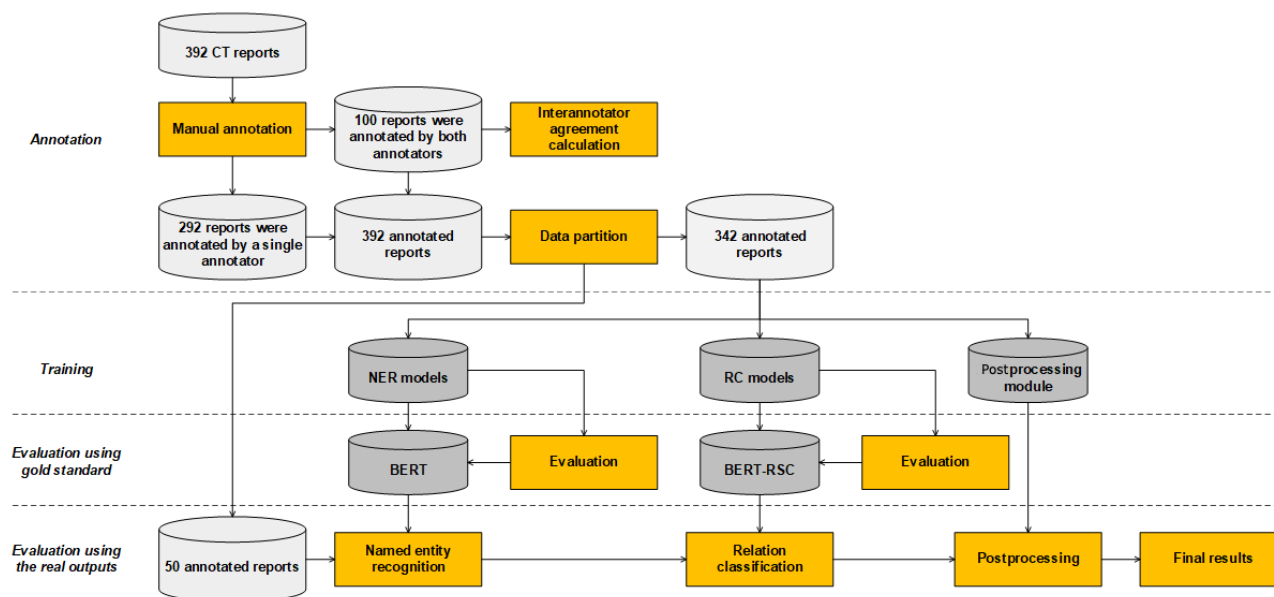
In this study, we aimed to develop an IE system using deep learning methods to extract information about lung cancer

staging from CT reports to better support the accurate clinical staging of lung cancer. Our specific contributions involve (1) defining a group of entity types and relation types to cover a wealth of information about lung cancer staging in CT reports, (2) applying advanced deep learning algorithms to develop the IE system, and (3) evaluating the performance of the IE system in a pipeline manner using real clinical CT reports.

## Methods

Figure 1 illustrates the development process of the IE system. First, we annotated the entities and relations in the collected CT reports as the gold standard. Next, the annotated CT reports were used to develop and evaluate the three core parts of the IE system. We also used 50 CT reports to verify the overall performance of the IE system in a pipeline manner. The details of each part are elaborated as follows.

**Figure 1.** Development process of the information extraction system. BERT: bidirectional encoder representation from transformers; BERT-RC: bidirectional encoder representation from transformers-relation classification; CT: computed tomography; NER: named entity recognition; RC: relation classification.



### Data Annotation

In clinical practice, clinicians usually follow the TNM staging guideline to stage the patients. Therefore, we first analyzed the eighth edition of the lung cancer TNM staging summary and parsed it into 41 questions to determine the scope of staging information (Multimedia Appendix 1). Note that the staging guideline covers three aspects of lung cancer (ie, tumor [T], lymph node [N], and metastases [M]), with detailed criteria. Chest CT can hardly provide all the information related to lung cancer staging. Clinicians also use other diagnostic modalities like positron emission tomography (PET), magnetic resonance imaging (MRI), and pathological biopsy to stage the patients. Thus, based on the content of the CT reports, 19 questions were identified under the clinician's guidance. Moreover, we also included 3 questions about the shape, density, and enhancement extent of the tumors. These 3 questions can facilitate the diagnosis of benign and malignant tumors. All 22 questions are listed in Table 1.

Based on the questions listed in Table 1, we defined 14 types of entities and 4 types of relations to represent the staging-related information in the CT reports. Table 2 shows

the defined entities. Figure 2 illustrates the entity–entity relation map.

Two medical informatics engineers were recruited to annotate the 392 CT reports by manually following the annotation guideline. The details of the annotation guideline are listed in Multimedia Appendix 2. Note that to obtain the annotation guideline, the annotators first independently annotated 10 reports and discussed the discrepancies until a consensus was reached in consultation with clinicians, resulting in a revised annotation guideline. Using the revised guideline, the annotators independently annotated 10 new reports and repeated the above process. In this manner, the guideline was refined by at least five iterations of annotation, discussion, consultation, and amendment, and then finalized. According to the final annotation guideline, we randomly selected 100 reports for annotation by both annotators to measure the interannotator agreement using the kappa statistic [36]. The remaining 292 reports were annotated only by either of the annotators. The BIO labeling scheme was employed to annotate the data. We employed brat [37] as the annotation tool. Figure 3 shows an example of the annotated CT reports.

**Table 1.** Questions about lung cancer diagnosis and staging<sup>a</sup>.

No.	Question	Type of answer	Stage
1	Whether the tumor can be visualized by imaging or bronchoscopy?	Yes/No	TX
2	What is the greatest dimension of the tumor?	Numerical	T1-4
3	Whether the tumor invades the lobar bronchus?	Yes/No	T1
4	Whether the tumor invades the visceral pleura?	Yes/No	T2
5	Whether there is an atelectasis or obstructive pneumonitis that extends to the hilar region, either involving part of the lung or the entire lung?	Yes/No	T2
6	Whether there is (are) associated separate tumor nodule (s) in the same lobe as the primary?	Yes/No	T3
7	Whether the tumor invades the great vessels?	Yes/No	T4
8	Whether the tumor invades the vertebral body?	Yes/No	T4
9	Whether there is (are) separate tumor nodule (s) in a different ipsilateral lobe to that of the primary?	Yes/No	T4
10	Whether there is regional lymph node metastasis?	Yes/No	N0
11	Whether there is metastasis in ipsilateral hilar lymph nodes, including involvement by direct extension?	Yes/No	N1
12	Whether there is metastasis in ipsilateral mediastinal lymph nodes?	Yes/No	N2
13	Whether there is metastasis in subcarinal lymph nodes?	Yes/No	N2
14	Whether there is metastasis in contralateral mediastinal lymph nodes?	Yes/No	N3
15	Whether there is metastasis in contralateral hilar lymph nodes?	Yes/No	N3
16	Whether there is metastasis in supraclavicular lymph nodes?	Yes/No	N3
17	Whether there is (are) separate tumor nodule (s) in a contralateral lobe?	Yes/No	M1a
18	Whether the tumor with pleural nodules?	Yes/No	M1a
19	Whether there is malignant pleural or pericardial effusion?	Yes/No	M1a
20 <sup>b</sup>	What is the shape of the tumor?	Text	NA
21 <sup>b</sup>	What is the density of the tumor?	Text	NA
22 <sup>b</sup>	What is the enhancement extent of the tumor?	Text	NA

<sup>a</sup>The stages are based on the eighth edition of the lung cancer TNM staging summary.

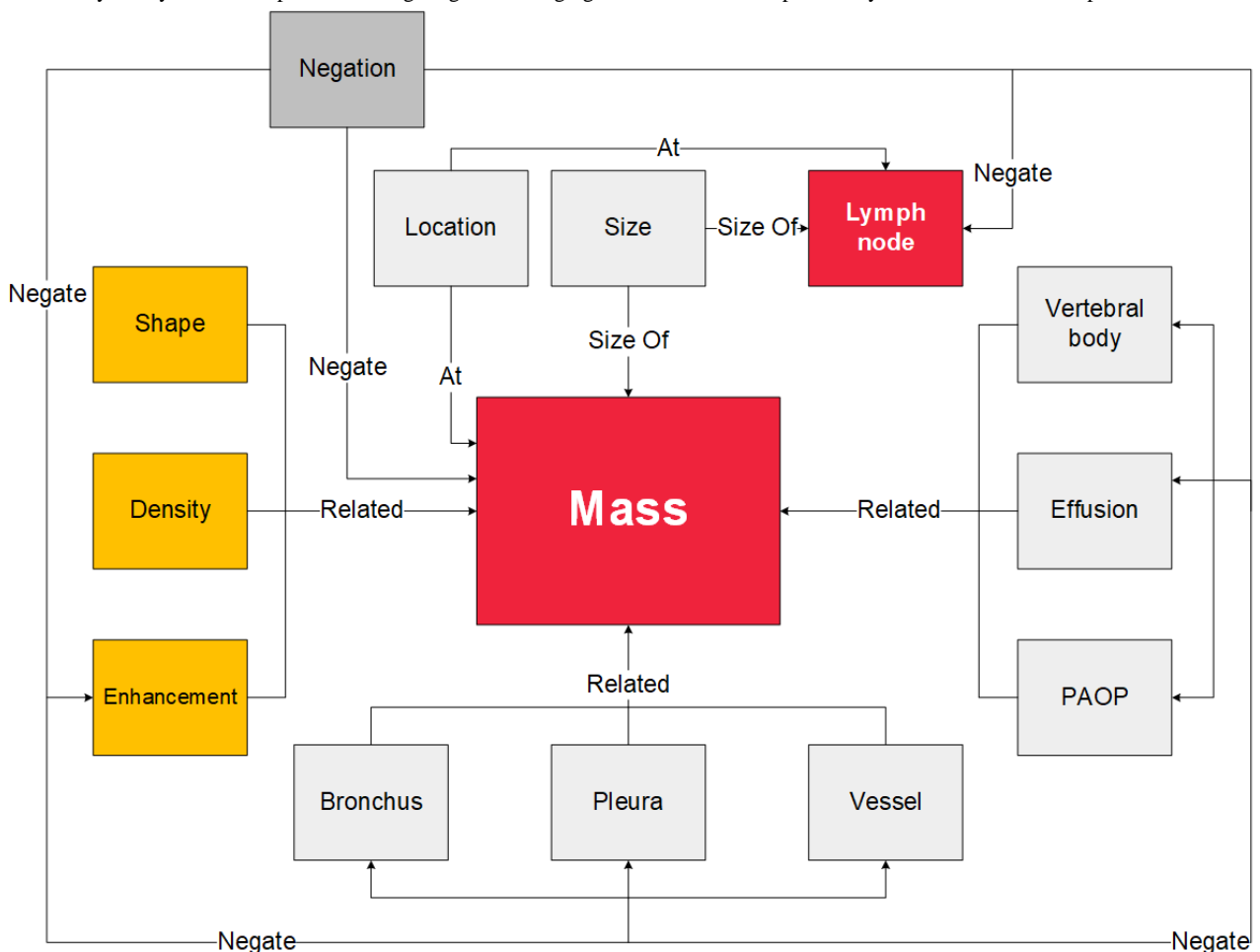
<sup>b</sup>The questions are not used for staging but are important for diagnosis of benign and malignant tumors.

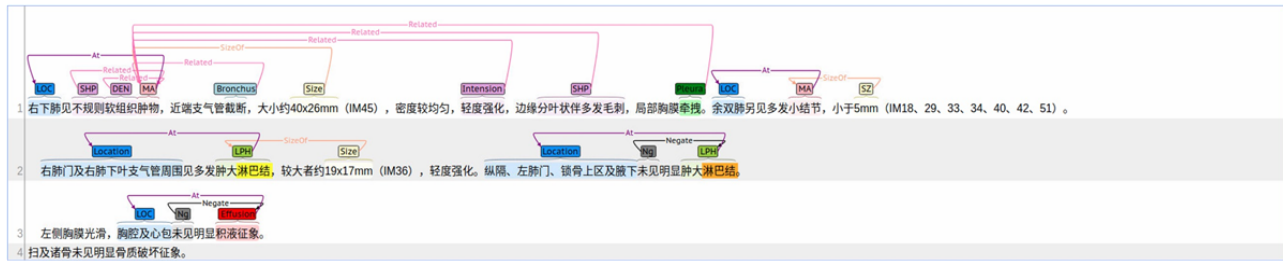
**Table 2.** Types of entities with descriptions and instances.

Entity type	Description	Instance
Mass	Suspected mass/nodule/lesion in the lung	肿物 (mass)
Lymph node	Suspected lymph node metastasis	肿大淋巴结 (enlarged lymph node)
Location	Location of mass or lymph node	左上肺右基底段 (right basal segment of the upper left lung)
Size	Size of mass or lymph node	25×22 cm
Negation	Negative words	未见 (unseen)
Density	Density of mass	磨玻璃密度 (ground glass density)
Enhancement	Enhancement extent of mass	强化明显 (significant enhancement)
Shape	Shape of mass	边缘见毛刺 (spiculate boundary)
Bronchus	Description of bronchial invasion	支气管狭窄 (bronchial stenosis)
Pleura	Description of pleural invasion or metastasis	胸膜凹陷 (pleural indentation)
Vessel	Description of great vessel invasion	包绕左肺动脉 (surrounds the right lower pulmonary artery)
Vertebral body	Description of vertebral body invasion	椎体见骨质破坏 (bone destruction seen in the vertebral body)
Effusion	Description of pleural or pericardial effusion	心包积液 (pericardial effusion)
PAOP <sup>a</sup>	Description of pulmonary atelectasis or obstructive pneumonitis	肺组织不张 (atelectasis)

<sup>a</sup>PAOP: pulmonary atelectasis/obstructive pneumonitis.

**Figure 2.** Entity–entity relation map for extracting lung cancer staging information. PAOP: pulmonary atelectasis/obstructive pneumonitis.



**Figure 3.** Annotated computed tomography report based on the annotation guideline.

## Word Embedding

As an unsupervised feature representation technique, word embedding maps the words to vectors of real values to capture the semantic and syntactic information from the corpus. In this study, we adopted the word embedding technique pretrained on the Chinese Wikipedia corpus using word2vec [38] for conventional CNN and recurrent neural network (RNN) models. Note that unlike English, Chinese words can be composed of multiple characters but with no space appearing between words. To incorporate the word segmentation information into the NER task, we first used jieba [39], a well-known Chinese text segmentation toolkit, to segment the sentence. Then, we used the randomly initialized real-value vectors to represent whether a character is the first, middle, or last character of the segmented word as in the segmentation embedding. For BERT, we used the default vocabulary to map the tokens to natural numbers.

## NER Process

NER is an essential technique to identify the types and boundaries of the entities of interest, which can drive other NLP tasks [40-43]. Recently developed deep learning NER methods exhibit more powerful performances than the traditional methods without tedious feature engineering [27,29,30,44]. In this study, we selected ID-CNN-CRF, Bi-LSTM-CRF, and BERT to recognize the entities.

ID-CNN is an advanced algorithm extending from the dilated CNN [45]. Instead of simply increasing the depth of a stacked dilated CNN, the ID-CNN applies the same small stack of dilated convolutions multiple times, with each iteration taking the result of the last application as the input to incorporate global information from a whole sentence and alleviate the overfitting problem. Bi-LSTM is another deep learning method using the recurrent neural network architecture that can capture the long-distance dependencies of context from both sides of the sequence and alleviate gradient vanishing or explosion during entity recognition from clinical text. A CRF layer was also employed on the ID-CNN and Bi-LSTM models, as it can exploit the relation constraints among different labels to find the optimal label path for sequence labeling tasks.

BERT is a novel language representation model pretrained on a large corpus using bidirectional transformers [46]. Unlike the traditional embedding methods that can only represent a word with polysemy using one fixed vector, BERT can dynamically adjust the representation depending on the context of the word.

It can also be easily fine-tuned to adapt to specific tasks, such as NER, RC, and question answering, and it has shown more powerful performance than conventional CNN and RNN models.

## RC Process

RC is the task of finding semantic relations between pairs of entities, which can group the relevant entities together to generate richer semantics [42,43]. Although traditional RC methods have achieved satisfactory performance [47,48], deep learning RC methods obtained better results and provided an effective way to alleviate the problem of hand-craft features [10,27,28]. In this study, we selected attention-based bidirectional long short-term memory networks (Attention-Bi-LSTM) [32] and BERT to classify the relations between entities.

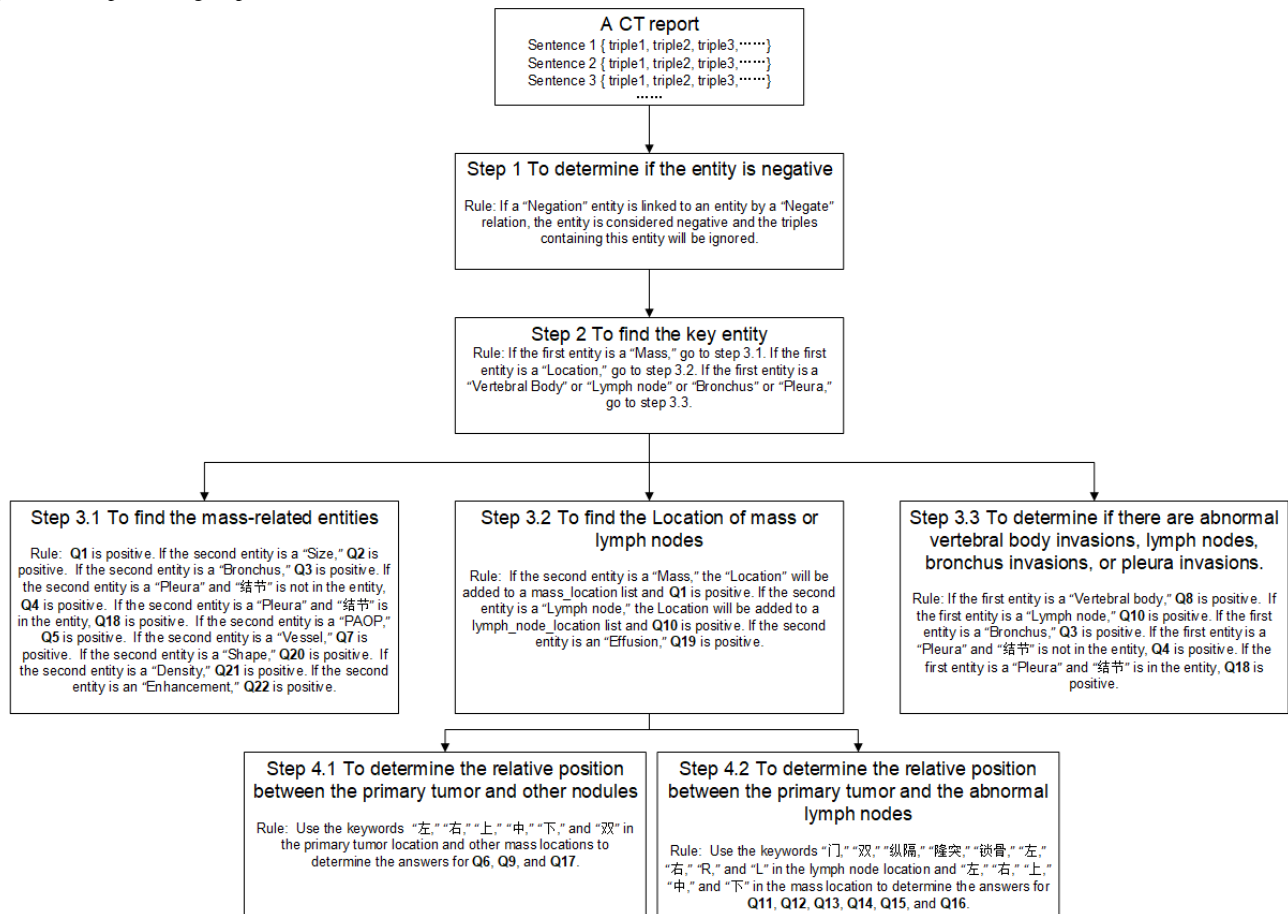
Note that in this study, two entities in a sentence can only have one type of relation or no relation depending on the definition in Figure 2. For instance, the relation between a lymph node entity and a location entity may be At or NoRelation, but definitely not a SizeOf relation. This information is useful for simplifying the multiclassification problem into a binary classification problem. We propose a novel approach, namely RSC, to use this extra information for RC. Before using the original sentence for relation classification, we first added the tags, namely At, SizeOf, Negate, Related, and NoRelation at the beginning of the sentence (eg, “At<e1>左肺门及纵隔4、5组</e1>见<e2>肿大淋巴结</e2>, 较大约14×12m.”). The added At tag is determined based on the two target entity types (location and lymph node). Then, the sentence with the tag can be input into the RC model. Using this method, we can simply incorporate the entity–entity relation constraints into the model to improve the prediction performance.

## PP Step

To obtain the answers to the questions listed in Table 1, it is not enough to directly use the extracted triples (entity 1–relation–entity 2), and further analysis is needed. For example, to answer the question on whether there is metastasis in ipsilateral mediastinal lymph nodes, we first need to know whether there exist a primary tumor and a mediastinal lymph node metastasis for this patient, and then determine the relative position of these two. In this study, we developed a rule-based PP module to process the extracted triples by the NER and RC models. The PP step is presented in Figure 4 and the rules are listed in Multimedia Appendix 3.



Figure 4. Postprocessing steps.



## Evaluation Metrics

To evaluate the performance of the models, we used the precision, recall, and F1 score as the evaluation metrics. Moreover, we also employed the microaverages and macroaverages for overall performance evaluation. The corresponding formulations are listed in [Multimedia Appendix 4](#).

## Results

### Data Annotation Results

A total of 392 chest CT reports of lung cancer patients were collected from the Department of Thoracic Surgery II in the

Peking University Cancer Hospital. Two medical informatics engineers were recruited to annotate the entities and relations based on the annotation guideline. The statistics of the annotations are summarized in [Tables 3 and 4](#). We had both the engineers annotate 100 CT reports to calculate the interannotator agreement, and the  $\kappa$  values were 0.937 for the entity annotation and 0.946 for the relation annotation, indicating the reliability of the annotation. Prior approval was obtained from the Ethics Committee of the Peking University Cancer Hospital to conduct this study.

**Table 3.** Statistics of annotated named entities.

Entity	Annotated entities, n
Mass	767
Lymph node	492
Location	1748
Size	699
Negation	808
Density	147
Enhancement	146
Shape	437
Bronchus	124
Pleura	262
Vessel	41
Vertebral body	25
Effusion	363
PAOP	78

**Table 4.** Statistics of annotated relations.

Relation	Annotated relations, n
At	1811
SizeOf	683
Related	988
Negate	803

## NER Results

To train and evaluate the NER models, we randomly separated 70% of the CT reports as the training set, 10% as the validation set, and 20% as the test set. The early stopping strategy was used on the validation set to avoid the overfitting problem. The hyperparameters used in this study are listed in [Multimedia Appendix 5](#). We repeated the entire training and evaluation process five times to reduce the possible bias that may be caused by data partitioning.

[Table 5](#) and [Figure 5](#) show the results of the NER models. As shown in [Table 5](#), the BERT model achieves the best overall performance with a macro-F1 score of 80.97% and a micro-F1 score of 88.5%. We can notice that the entities with several annotations or plain descriptions (eg, “Lymph Node,” “Negation,” “Size,” and “Effusion”) obtain satisfactory results with F1 scores greater than 90%. However, performances degraded for the entities with a small number of annotations or diverse descriptions (eg, “Shape,” “Pleura,” “Vessel,” “Vertebral Body,” and “PAOP”) [Figure 5](#) shows the results in a more intuitive manner with standard deviations.

By further analyzing the extractions, we found that most of the errors were due to an inexact match, where a predicted entity overlapped with the gold standard. For example, the predicted entity “余 (B-Location)肺 (I-Location)内 (O)” is an inexact match for the gold standard annotation “余 (B-Location)肺 (I-Location)内 (I-Location).” Although these extractions could not cover the gold standard exactly, the partially matched entities still contained useful information for RC and PP. We also calculated the inexact matching performances for each type of entity and have presented them in [Table 6](#) and [Figure 6](#).

As shown in [Table 6](#), the macro-F1 scores of ID-CNN-CRF, Bi-LSTM-CRF, and BERT using the inexact metrics are 89.6%, 89.96%, and 90.06%, which obtain improvements of 13.93%, 12.69%, and 9.09% compared with the exact metrics, respectively. Furthermore, the micro-F1 scores of the inexact metrics are all above 94%. Almost all the entities obtain better extraction results under the inexact matching scheme, especially those entities with diverse descriptions, which indicates that the extractions cover most of the annotations.

**Table 5.** Performance of the named entity recognition models.

Entity	ID-CNN-CRF <sup>a</sup>			Bi-LSTM-CRF <sup>b</sup>			BERT <sup>c</sup>		
	Precision (%)	Recall (%)	F1 score (%)	Precision (%)	Recall (%)	F1 score (%)	Precision (%)	Recall (%)	F1 score (%)
Mass	83.11	87.79	85.35	83.86	88.02	85.88	87.92	86.05	87.61
Lymph node	92.29	95.42	93.83	93.29	94.79	94.04	91.52	93.07	92.27
Location	84.85	87.40	86.1	86.99	89.3	88.12	87.93	86.99	87.44
Size	91.6	95	93.24	92.29	94.92	93.56	94.03	94.44	94.22
Negation	97.66	98.45	98.02	97.77	98.79	98.26	99.12	99.11	99.11
Density	64.16	69.66	66.61	68.4	71.47	69.73	75.55	68.49	71.75
Enhancement	74.48	81.04	77.47	74.33	78.4	76.14	81.39	75.03	77.69
Shape	82.65	83.85	83.21	78.95	83.38	81	82.72	81.8	82.2
Bronchus	66.45	67.96	67.11	62.57	69.55	65.66	74.17	76.88	75.1
Pleura	81.48	79.39	80.36	83.54	83.28	83.39	84.59	77.13	80.21
Vessel	37.52	41.59	39.05	44.5	43.13	43.27	68.09	54.53	58.51
Vertebral body	36.43	60.17	42.75	46.52	67.5	53.17	82	66.67	72.24
Effusion	97.02	97.25	97.11	95.77	97.3	96.51	98.32	97.25	97.78
PAOP <sup>d</sup>	47.67	51.33	49.11	50.28	57.25	53.04	65.86	53.2	57.46
Macroaverage	74.1	78.31	75.67	75.65	79.79	77.27	83.8	79.33	80.97
Microaverage	85.85	88.41	87.11	86.56	89.32	87.92	89.28	87.78	88.5

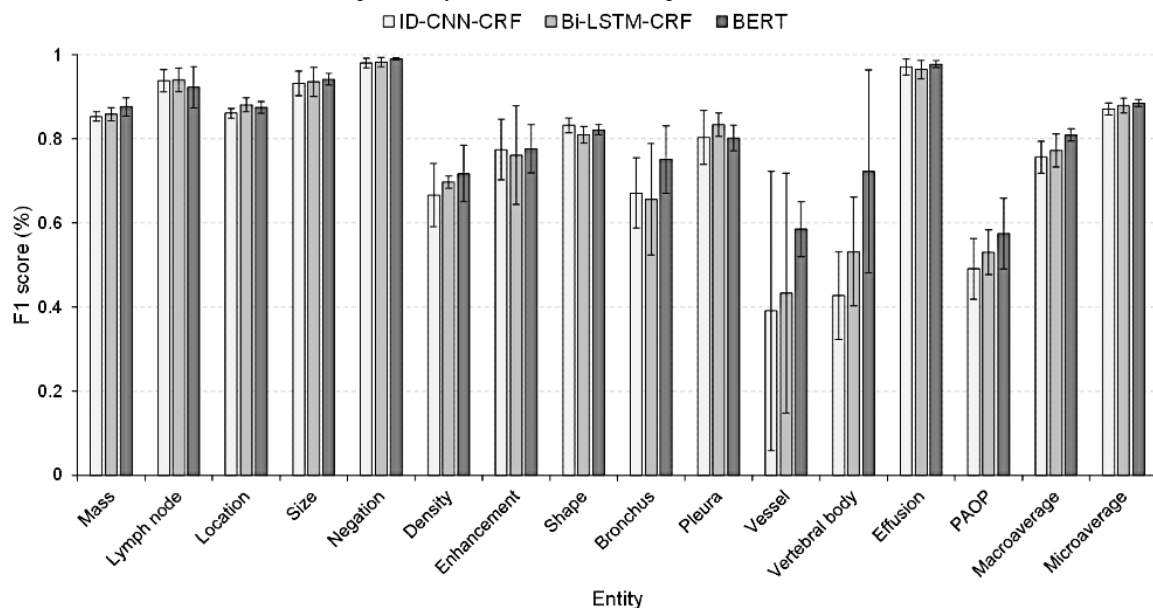
<sup>a</sup>ID-CNN-CRF: iterated dilated convolutional neural networks-conditional random field.

<sup>b</sup>Bi-LSTM-CRF: bidirectional long short-term memory networks- conditional random field.

<sup>c</sup>BERT: bidirectional encoder representation from transformers.

<sup>d</sup>PAOP: pulmonary atelectasis/obstructive pneumonitis.

**Figure 5.** F1 scores with bars showing the standard deviations of the named entity recognition models. Bi-LSTM-CRF: bidirectional long short-term memory networks-conditional random field; BERT: bidirectional encoder representation from transformers; ID-CNN-CRF: iterated dilated convolutional neural networks-conditional random field; PAOP: pulmonary atelectasis/obstructive pneumonitis.



**Table 6.** Performance of the named entity recognition models calculated using the inexact matching scheme.

Entity	ID-CNN-CRF <sup>a</sup>			Bi-LSTM-CRF <sup>b</sup>			BERT <sup>c</sup>		
	Precision (%)	Recall (%)	F1 score (%)	Precision (%)	Recall (%)	F1 score (%)	Precision (%)	Recall (%)	F1 score (%)
Mass	89.78	94.81	92.19	90.71	95.2	92.89	94.11	92.05	93.02
Lymph node	97.11	100.42	98.73	97.43	99	98.2	97.8	99.41	98.59
Location	91.88	94.66	93.24	92.73	95.2	93.95	95.01	93.99	94.47
Size	95.33	98.9	97.06	96.28	99.06	97.62	96.58	97.03	96.79
Negation	97.66	98.45	98.02	97.77	98.79	98.26	99.12	99.11	99.11
Density	84.09	90.53	86.95	82.39	86.13	84.01	94.48	85.49	89.64
Enhancement	85.64	93.26	89.11	86.79	92.3	89.25	91.53	85.03	87.73
Shape	91.59	92.92	92.22	88.76	93.94	91.16	92.23	91.12	91.6
Bronchus	83.20	85.03	84.01	79.4	89.19	83.73	84.76	87.76	85.75
Pleura	93.07	90.44	91.67	92.86	92.54	92.67	93.12	85.7	88.73
Vessel	81.29	79.18	79.09	84.66	73.82	77.52	89.03	67.5	75.58
Vertebral body	63.81	92.5	71.76	65.76	86.67	72.28	92	73.33	80.24
Effusion	98.4	98.64	98.5	98.18	99.74	98.93	100	98.91	99.45
PAOP <sup>d</sup>	80.1	84.64	81.84	84.23	96.39	89.03	90.23	74.99	80.13
Macroaverage	88.07	92.46	89.6	88.43	92.71	89.96	93.57	87.96	90.06
Microaverage	92.66	95.42	94.01	92.87	95.84	94.32	95.39	93.81	94.57

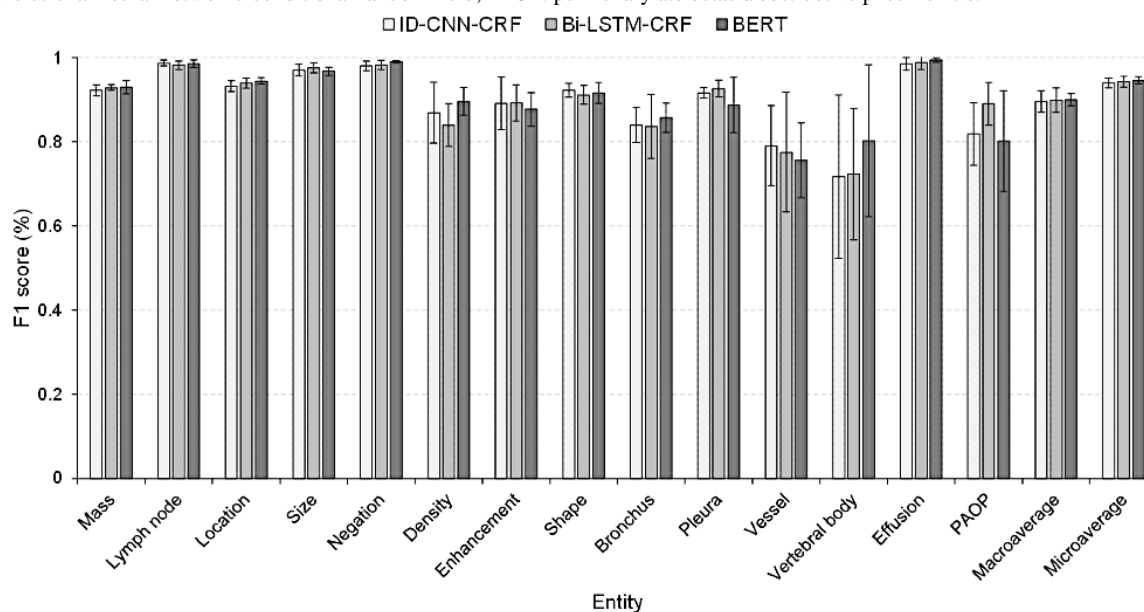
<sup>a</sup>ID-CNN-CRF: iterated dilated convolutional neural networks-conditional random field.

<sup>b</sup>Bi-LSTM-CRF: bidirectional long short-term memory networks- conditional random field.

<sup>c</sup>BERT: bidirectional encoder representation from transformers.

<sup>d</sup>PAOP: pulmonary atelectasis/obstructive pneumonitis.

**Figure 6.** Inexactly matching F1 scores with bar showing the standard deviations of the named entity recognition models. Bi-LSTM-CRF: bidirectional long short-term memory networks-conditional random field; BERT: bidirectional encoder representation from transformers; ID-CNN-CRF: iterated dilated convolutional neural networks-conditional random field; PAOP: pulmonary atelectasis/obstructive pneumonitis.



## RC Results

To evaluate the proposed RC method, the data set was randomly separated such that 70%, 10%, and 20% of the CT reports were

used as the training, validation, and test sets, respectively. Attention-Bi-LSTM and BERT were selected as the baselines. The annotated entities were provided in this step for evaluating the performance of the RC models. The hyperparameters of the

RC models are listed in [Multimedia Appendix 5](#). We also repeated the entire training and evaluation process five times with different random seeds to alleviate the possible bias caused by data partitioning.

[Table 7](#) and [Figure 7](#) show the experimental results of the RC models. As depicted in [Table 7](#), all the four models achieve

excellent performances with macro-F1 values above 95% and micro-F1 values above 97%. Comparing the baseline and proposed methods indicates that the RSC improves the performances of both the baseline models, especially for the Related RC. Moreover, the BERT-RSC achieves the best performance among all the models.

**Table 7.** Performance of the proposed and baseline relation classification models.

Relation	Baseline						Proposed					
	Attention-Bi-LSTM <sup>a</sup>			BERT <sup>b</sup>			Attention-Bi-LSTM-RSC <sup>c</sup>			BERT-RSC <sup>d</sup>		
	Precision (%)	Recall (%)	F1 score (%)	Precision (%)	Recall (%)	F1 score (%)	Precision (%)	Recall (%)	F1 score (%)	Precision (%)	Recall (%)	F1 score (%)
At	96.02	94.47	95.23	96.3	95.39	95.83	96.25	94.79	95.5	96.95	95.55	96.23
SizeOf	97.05	97.51	97.27	98.13	97.35	97.73	97.19	98.1	97.61	98.11	98.42	98.25
Related	88.17	91.47	89.65	85.22	94.7	89.64	88.95	92.31	90.55	89.17	96.27	92.56
Negate	98.7	97.07	97.87	99.38	99.63	99.5	99.33	97.82	98.56	99.38	99.74	99.56
NoRelation	98.7	98.67	98.68	99.11	98.57	98.84	98.83	98.77	98.8	99.22	98.87	99.05
Macroaverage	95.73	95.84	95.74	95.63	97.13	96.31	96.11	96.36	96.2	96.57	97.77	97.13
Microaverage	97.88	97.88	97.88	98.11	98.11	98.11	98.08	98.08	98.08	98.48	98.48	98.37

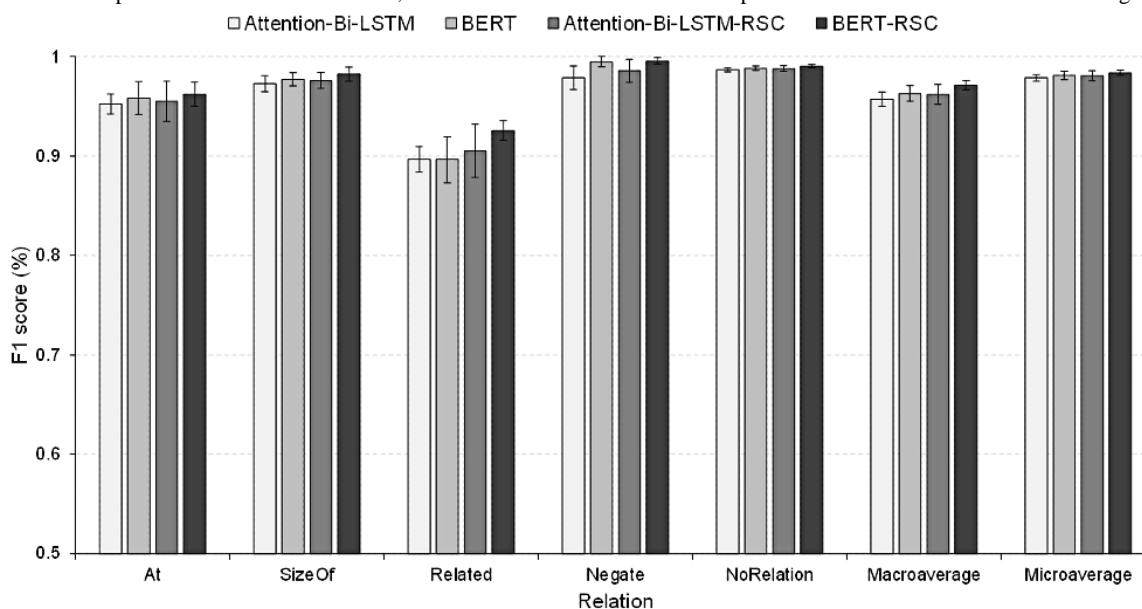
<sup>a</sup>ID-CNN-CRF: iterated dilated convolutional neural networks-conditional random field.

<sup>b</sup>Bi-LSTM-CRF: bidirectional long short-term memory networks- conditional random field.

<sup>c</sup>BERT: bidirectional encoder representation from transformers.

<sup>d</sup>BERT-RSC: bidirectional encoder representation from transformers-relation sign constraint.

**Figure 7.** F1 scores with bars showing the standard deviations of the relation classification models. Attention-Bi-LSTM: attention-based bidirectional long short-term memory networks; Attention-Bi-LSTM-RSC: attention-based bidirectional long short-term memory-relation sign constraint; BERT: bidirectional encoder representation from transformers; BERT-RSC: bidirectional encoder representation from transformers-relation sign constraint.



### PP Results

Based on the experimental results presented above, we selected the BERT model for NER and RC. Note that instead of using the annotated data, we directly used the output of the NER model as the input for RC and employed the PP module to analyze the triples extracted by NER and RC to verify the performance of the IE system. We randomly selected 50 reports,

for which both the annotators manually answered the 22 questions. [Table 8](#) shows the number of positive answers annotated to each question in the 50 reports and the experimental results of the IE system for each question. The experimental results prove that the IE system achieves a macro-F1 score of 94.57% and a micro-F1 score of 96.74%, indicating that the system can effectively extract information related to lung cancer staging from CT reports.



By analyzing the incorrect answers, we found that the main reason for inaccurate extraction was that some entities or relations were not recognized by the system. For example, missing “Mass” or “At” relations made it impossible to determine the relative position between the primary tumor and

other nodules, resulting in low recall values of Q6, Q9, and Q18. Besides, missing “Bronchus,” “PAOP,” “Vessel,” “Density,” and “Enhancement” entities led to low recall values of Q3, Q5, Q7, Q21, and Q22 when relevant descriptions were inherently scarce.

**Table 8.** Experimental results of the developed information extraction system.

No.	Number of positive answers annotated	Precision (%)	Recall (%)	F1 score (%)
1	50	100	100	100
2	47	97.83	95.74	96.77
3	16	100	87.50	93.33
4	27	100	96.3	98.11
5	6	83.33	83.33	83.33
6	17	100	82.35	90.32
7	5	100	80	88.89
8	2	100	100	100
9	14	100	85.71	92.31
10	28	100	100	100
11	18	100	100	100
12	22	95.45	95.45	95.45
13	1	100	100	100
14	19	95	100	97.44
15	6	100	100	100
16	5	100	100	100
17	20	95	95	95
18	5	80	80	80
19	2	100	100	100
20	28	96.3	92.86	94.55
21	14	100	85.71	92.31
22	16	85.71	80	82.76
Macroaverage		96.76	92	94.57
Microaverage		97.49	95.99	96.74

## Discussion

### Principal Findings

In this study, we developed an IE system to extract information related to lung cancer staging from CT reports automatically. The experimental results indicate that the IE system can effectively extract the useful entities and relations using the NER and RC models and accurately obtain the answers to the questions about lung cancer staging using the PP module. The extracted information shows significant potential to support further research about accurate lung cancer clinical staging.

Although the macro-F1 score of NER is only 80.97%, which seems insufficient to support RC and PP, the IE system still achieves satisfactory results. The main reason is that the PP module exploits the key characters in the extracted entities or

only the presence of the entities to obtain the answers but does not need the complete entities. For example, the annotation of the sentence “右肺下叶基底段见软组织密度肿块” is [Location\_B, Location\_I, Location\_I, Location\_I, Location\_I, Location\_I, Location\_I, O, Mass\_B, Mass\_I, Mass\_I, Mass\_I, Mass\_I, Mass\_I, Mass\_I, Mass\_I], but the NER result is [Location\_B, Location\_I, Location\_I, Location\_I, Location\_I, O, Location\_I, O, Mass\_B, Mass\_I, Mass\_I, Mass\_I, Mass\_I, Mass\_I, Mass\_I, Mass\_I], which means the Location entity extracted is merely “右肺下叶基.” However, this partial Location entity is correctly linked to the Mass entity “软组织密度肿块” with an “At” relation by the RC model, and the key characters “右” and “下” in the Location entity can support the following PP step. The high macro-F1 and micro-F1 of the inexact matching scheme indicate that most of the entities can be extracted completely or partially by the NER model. Furthermore, the extractions cover most of the key characters needed during the PP step.

For the RC task, all the four models achieve satisfactory performances. This is because the descriptions are similar in many sentences so that the models can easily learn these patterns. However, for the Related relation, none of the models obtain the perfect performance. The main reason is that some types of entities like “Vertebral Body” and “Vessel” are rare and have diverse descriptions, making it difficult for these models to learn the corresponding patterns. The addition of RSC may make the descriptions more uniform so that the models may learn the patterns more easily.

For the NER and RC tasks, the advanced pretrained BERT model achieves better performance compared to the conventional CNN and RNN methods, thus verifying the superiority of large language representation models for various NLP tasks.

### Limitations

Although the rule-based PP module can accurately obtain the answers to the defined questions by analyzing the extracted entities and relations, these hard-coded rules are difficult to maintain and update. Furthermore, for better use of clinical knowledge (eg, enlarged lymph nodes with a minimum diameter greater than 10 mm are often considered metastatic), we need to establish a more comprehensive knowledge base to analyze the extracted information. Ontology, as a formal representation of medical knowledge, has become the standard method to develop knowledge bases [49,50]. In future, we can use the Web Ontology Language (OWL) [51] to construct the knowledge graph and employ the Semantic Web Rule Language (SWRL) [52] to develop the reasoning rules for lung cancer staging.

In this study, we explored the feasibility of extracting information related to lung cancer staging from CT reports using an NER+RC+PP pipeline in a single hospital. When generalizing this approach to other hospitals, the entity and relation definitions as well as the annotation strategy can be important

references for the same application, and the developed pipeline can also be reused. However, if researchers want to customize the entity types or relation types to suit their purpose or if the writing style of CT reports is significantly different from that in the reports that we used, fine-tuning of BERT using the newly annotated reports may be a possible way to obtain satisfactory generalization.

### Future Research

In the current study, pathological staging was not applied as the gold standard to evaluate the correctness of the extracted results. This is mainly because in clinical practice, clinicians use not only CT but also PET, MRI, and other diagnostic modalities to stage patients. Therefore, it is insufficient to use only the information extracted from the CT report to stage the patients. In future, we plan to extract staging information from other examination reports and use this multisource information to verify the staging correctness from a more comprehensive perspective. Moreover, by combining various details such as laboratory tests, disease history, and radiomics data, we can employ advanced machine learning algorithms to develop clinical staging prediction models to further alleviate the large number of disagreements between clinical and pathological stages.

### Conclusions

In this study, we developed an IE system to extract lung cancer staging information from CT reports automatically using NLP techniques. Experimental results obtained using real clinical data demonstrated that the IE system could effectively extract the relevant entities and relations using the NER and RC models. It could also accurately answer the staging questions using the rule-based PP module, thus proving the potential of this system for lung cancer staging verification and clinical staging prediction.

---

### Acknowledgments

This work was supported by the National Key R&D Program of China (grant 2018YFC0910700).

---

### Authors' Contributions

DH, SL, XL, and NW conceptualized the study. SL acquired the clinical data. SL, YW, and HZ annotated the data. HZ, DH, and YW designed and implemented the algorithms and conducted the experiments. DH, HZ, YW, and SL analyzed the experimental results. DH wrote the manuscript with revision assistance from SL, XL, and NW. All authors have read and approved the manuscript.

---

### Conflicts of Interest

None declared.

---

#### Multimedia Appendix 1

Parsed questions about lung cancer staging.

[[PDF File \(Adobe PDF File\), 123 KB - medinform\\_v9i7e27955\\_app1.pdf](#)]

---

#### Multimedia Appendix 2

Annotation guideline.

[[PDF File \(Adobe PDF File\), 167 KB - medinform\\_v9i7e27955\\_app2.pdf](#)]

## Multimedia Appendix 3

Postprocessing rules.

[\[PDF File \(Adobe PDF File\), 101 KB - medinform\\_v9i7e27955\\_app3.pdf \]](#)

## Multimedia Appendix 4

Evaluation metrics.

[\[PDF File \(Adobe PDF File\), 52 KB - medinform\\_v9i7e27955\\_app4.pdf \]](#)

## Multimedia Appendix 5

Hyperparameters of the named entity recognition and relation classification models. Attention-Bi-LSTM-RSC: attention-based bidirectional long short-term memory-relation sign constraint; Bi-LSTM-CRF: bidirectional long short-term memory networks-conditional random field; BERT: bidirectional encoder representation from transformers; BERT-RSC: bidirectional encoder representation from transformers-relation sign constraint; ID-CNN-CRF: iterated dilated convolutional neural networks-conditional random field.

[\[PDF File \(Adobe PDF File\), 115 KB - medinform\\_v9i7e27955\\_app5.pdf \]](#)**References**

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018 Nov;68(6):394-424 [[FREE Full text](#)] [doi: [10.3322/caac.21492](https://doi.org/10.3322/caac.21492)] [Medline: [30207593](https://pubmed.ncbi.nlm.nih.gov/30207593/)]
2. Ettinger D, Wood D, Aggarwal C, Aisner D, Akerley W, Bauman J, et al. NCCN clinical practice guidelines in oncology. Non-Small Cell Lung Cancer. URL: <https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1450> [accessed 2019-09-16]
3. Navani N, Fisher DJ, Tierney JF, Stephens RJ, Burdett S, NSCLC Meta-analysis Collaborative Group. The accuracy of clinical staging of stage I-IIIa non-small cell lung cancer: an analysis based on individual participant data. *Chest* 2019 Mar;155(3):502-509 [[FREE Full text](#)] [doi: [10.1016/j.chest.2018.10.020](https://doi.org/10.1016/j.chest.2018.10.020)] [Medline: [30391190](https://pubmed.ncbi.nlm.nih.gov/30391190/)]
4. Heineman DJ, Ten Berge MG, Daniels JM, Versteegh MI, Marang-van de Mheen PJ, Wouters MW, et al. The quality of staging non-small cell lung cancer in the Netherlands: data from the Dutch lung surgery audit. *Ann Thorac Surg* 2016 Nov;102(5):1622-1629. [doi: [10.1016/j.athoracsur.2016.06.071](https://doi.org/10.1016/j.athoracsur.2016.06.071)] [Medline: [27665479](https://pubmed.ncbi.nlm.nih.gov/27665479/)]
5. Wood D, Kazerooni E, Baum S, Eapen G, Ettinger D, Ferguson J, et al. NCCN clinical practice guidelines in oncology. Lung Cancer Screening. URL: <https://www.nccn.org/guidelines/guidelines-detail?category=2&id=1441> [accessed 2019-09-16]
6. Yim W, Yetisgen M, Harris WP, Kwan SW. Natural language processing in oncology: a review. *JAMA Oncol* 2016 Jun 01;2(6):797-804. [doi: [10.1001/jamaoncol.2016.0213](https://doi.org/10.1001/jamaoncol.2016.0213)] [Medline: [27124593](https://pubmed.ncbi.nlm.nih.gov/27124593/)]
7. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform* 2019 Apr 27;7(2):e12239 [[FREE Full text](#)] [doi: [10.2196/12239](https://doi.org/10.2196/12239)] [Medline: [31066697](https://pubmed.ncbi.nlm.nih.gov/31066697/)]
8. Detterbeck FC, Boffa DJ, Kim AW, Tanoue LT. The eighth edition lung cancer stage classification. *Chest* 2017 Jan;151(1):193-203. [doi: [10.1016/j.chest.2016.10.010](https://doi.org/10.1016/j.chest.2016.10.010)] [Medline: [27780786](https://pubmed.ncbi.nlm.nih.gov/27780786/)]
9. Strubell E, Verga P, Belanger D, McCallum A. Fast and accurate entity recognition with iterated dilated convolutions. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.: Association for Computational Linguistics; 2017 Presented at: Conference on Empirical Methods in Natural Language Processing; Sept 07-11; Copenhagen, Denmark p. 2670-2680. [doi: [10.18653/v1/D17-1283](https://doi.org/10.18653/v1/D17-1283)]
10. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. ArXiv. Preprint posted online on Aug 9, 2015 [[FREE Full text](#)]
11. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv. Preprint posted online on Oct 11, 2018.
12. Datta S, Bernstam EV, Roberts K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *J Biomed Inform* 2019 Oct 04;100:103301 [[FREE Full text](#)] [doi: [10.1016/j.jbi.2019.103301](https://doi.org/10.1016/j.jbi.2019.103301)] [Medline: [31589927](https://pubmed.ncbi.nlm.nih.gov/31589927/)]
13. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *J Am Med Inform Assoc* 2010 Jul;17(4):440-445 [[FREE Full text](#)] [doi: [10.1136/jamia.2010.003707](https://doi.org/10.1136/jamia.2010.003707)] [Medline: [20595312](https://pubmed.ncbi.nlm.nih.gov/20595312/)]
14. Warner JL, Levy MA, Neuss MN, Warner JL, Levy MA, Neuss MN. ReCAP: feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. *J Oncol Pract* 2016 Feb;12(2):157-158 [[FREE Full text](#)] [doi: [10.1200/JOP.2015.004622](https://doi.org/10.1200/JOP.2015.004622)] [Medline: [26306621](https://pubmed.ncbi.nlm.nih.gov/26306621/)]
15. Schroeck FR, Lynch KE, Chang JW, MacKenzie TA, Seigne JD, Robertson DJ, et al. Extent of risk-aligned surveillance for cancer recurrence among patients with early-stage bladder cancer. *JAMA Netw Open* 2018 Sep;1(5):e183442 [[FREE Full text](#)] [doi: [10.1001/jamanetworkopen.2018.3442](https://doi.org/10.1001/jamanetworkopen.2018.3442)] [Medline: [30465041](https://pubmed.ncbi.nlm.nih.gov/30465041/)]

16. Cary C, Roberts A, Church AK, Eckert G, Ouyang F, He J, et al. Development of a novel algorithm to identify staging and lines of therapy for bladder cancer. *J Clin Oncol* 2017 May 20;35(15\_suppl):e18235 [FREE Full text] [doi: [10.1200/jco.2017.35.15\\_suppl.e18235](https://doi.org/10.1200/jco.2017.35.15_suppl.e18235)]
17. Schroeck F, Lynch K, Chang JW, Robertson D, Seigne J, Goodney P, et al. MP44-01 a national study of risk-aligned surveillance practice for non-muscle invasive bladder cancer. *J Urol* 2018 Apr;199(4S):e587. [doi: [10.1016/j.juro.2018.02.1420](https://doi.org/10.1016/j.juro.2018.02.1420)]
18. AAlAbdulsalam AK, Garvin JH, Redd A, Carter ME, Sweeny C, Meystre SM. Automated extraction and classification of cancer stage mentions from unstructured text fields in a central cancer registry. *AMIA Jt Summits Transl Sci Proc* 2018 May;2017:16-25 [FREE Full text] [Medline: [29888032](https://pubmed.ncbi.nlm.nih.gov/29888032/)]
19. Nunes A, Green E, Dalvi T, Lewis J, Jones N, Seeger J. Abstract P5-08-20: a real-world evidence study to define the prevalence of endocrine therapy-naïve hormone receptor-positive locally advanced or metastatic breast cancer in the US. *Cancer Res* 2017 Feb;77(4 Supplement):P5-08-20 [FREE Full text] [doi: [10.1158/1538-7445.SABCS16-P5-08-20](https://doi.org/10.1158/1538-7445.SABCS16-P5-08-20)]
20. Giri A, Levinson R, Keene S, Holman G, Smith S, Clayton L, et al. Abstract 4229: preliminary results from the pharmacogenetics ovarian cancer knowledge to individualize treatment (POCKIT) study. *Cancer Res* 2018 Jul;78(13):4229 [FREE Full text] [doi: [10.1158/1538-7445.AM2018-4229](https://doi.org/10.1158/1538-7445.AM2018-4229)]
21. Savova GK, Tseytlin E, Finan S, Castine M, Miller T, Medvedeva O, et al. DeepPhe: a natural language processing system for extracting cancer phenotypes from clinical records. *Cancer Res* 2017 Nov 01;77(21):e115-e118 [FREE Full text] [doi: [10.1158/0008-5472.CAN-17-0615](https://doi.org/10.1158/0008-5472.CAN-17-0615)] [Medline: [29092954](https://pubmed.ncbi.nlm.nih.gov/29092954/)]
22. Ping X, Tseng Y, Chung Y, Wu Y, Hsu C, Yang P, et al. Information extraction for tracking liver cancer patients' statuses: from mixture of clinical narrative report types. *Telemed J E Health* 2013 Sep;19(9):704-710. [doi: [10.1089/tmj.2012.0241](https://doi.org/10.1089/tmj.2012.0241)] [Medline: [23869395](https://pubmed.ncbi.nlm.nih.gov/23869395/)]
23. Yim W, Denman T, Kwan SW, Yetisgen M. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. *AMIA Jt Summits Transl Sci Proc* 2016 Jul;2016:455-464 [FREE Full text] [Medline: [27570686](https://pubmed.ncbi.nlm.nih.gov/27570686/)]
24. Chen L, Song L, Shao Y, Li D, Ding K. Using natural language processing to extract clinically useful information from Chinese electronic medical records. *Int J Med Inform* 2019 Apr;124:6-12. [doi: [10.1016/j.ijmedinf.2019.01.004](https://doi.org/10.1016/j.ijmedinf.2019.01.004)] [Medline: [30784428](https://pubmed.ncbi.nlm.nih.gov/30784428/)]
25. Bozkurt S, Lipson JA, Senol U, Rubin DL. Automatic abstraction of imaging observations with their characteristics from mammography reports. *J Am Med Inform Assoc* 2015 Apr;22(e1):e81-e92. [doi: [10.1136/amiainf-2014-003009](https://doi.org/10.1136/amiainf-2014-003009)] [Medline: [25352567](https://pubmed.ncbi.nlm.nih.gov/25352567/)]
26. Bozkurt S, Gimenez F, Burnside ES, Gulkesen KH, Rubin DL. Using automatically extracted information from mammography reports for decision-support. *J Biomed Inform* 2016 Aug;62:224-231 [FREE Full text] [doi: [10.1016/j.jbi.2016.07.001](https://doi.org/10.1016/j.jbi.2016.07.001)] [Medline: [27388877](https://pubmed.ncbi.nlm.nih.gov/27388877/)]
27. Jauregi Unanue I, Zare Borzeshi E, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *J Biomed Inform* 2017 Dec;76:102-109 [FREE Full text] [doi: [10.1016/j.jbi.2017.11.007](https://doi.org/10.1016/j.jbi.2017.11.007)] [Medline: [29146561](https://pubmed.ncbi.nlm.nih.gov/29146561/)]
28. Zhang D, Wang D. Relation classification via recurrent neural network. ArXiv. Preprint posted online on Aug 5, 2015 [FREE Full text]
29. Zhang Y, Wang X, Hou Z, Li J. Clinical named entity recognition from Chinese electronic health records via machine learning methods. *JMIR Med Inform* 2018 Dec 17;6(4):e50 [FREE Full text] [doi: [10.2196/medinform.9965](https://doi.org/10.2196/medinform.9965)] [Medline: [30559093](https://pubmed.ncbi.nlm.nih.gov/30559093/)]
30. 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](https://pubmed.ncbi.nlm.nih.gov/28699566/)]
31. Zeng D, Liu K, Lai S, Zhou G, Zhao J. Relation classification via convolutional deep neural network. Dublin, Ireland: Dublin City University and Association for Computational Linguistics; 2014 Presented at: The 25th International Conference on Computational Linguistics; August 23-29, 2014; Dublin, Ireland p. 2335-2344.
32. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, et al. Attention-based bidirectional long short-term memory networks for relation classification. Berlin, Germany: Association for Computational Linguistics; 2016 Presented at: The 54th Annual Meeting of the Association for Computational Linguistics; August 7-12, 2016; Berlin, Germany p. A URL: <https://aclanthology.org/P16-2034/> [doi: [10.18653/v1/p16-2034](https://doi.org/10.18653/v1/p16-2034)]
33. Luo Y. Recurrent neural networks for classifying relations in clinical notes. *J Biomed Inform* 2017 Aug;72:85-95 [FREE Full text] [doi: [10.1016/j.jbi.2017.07.006](https://doi.org/10.1016/j.jbi.2017.07.006)] [Medline: [28694119](https://pubmed.ncbi.nlm.nih.gov/28694119/)]
34. Si Y, Roberts K. A frame-based NLP system for cancer-related information extraction. *AMIA Annu Symp Proc* 2018;2018:1524-1533 [FREE Full text] [Medline: [30815198](https://pubmed.ncbi.nlm.nih.gov/30815198/)]
35. Gao S, Young MT, Qiu JX, Yoon H, Christian JB, Fearn PA, et al. Hierarchical attention networks for information extraction from cancer pathology reports. *J Am Med Inform Assoc* 2018 Mar 01;25(3):321-330 [FREE Full text] [doi: [10.1093/jamia/ocx131](https://doi.org/10.1093/jamia/ocx131)] [Medline: [29155996](https://pubmed.ncbi.nlm.nih.gov/29155996/)]
36. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. *J Am Med Inform Assoc* 2005 May;12(3):296-298 [FREE Full text] [doi: [10.1197/jamia.M1733](https://doi.org/10.1197/jamia.M1733)] [Medline: [15684123](https://pubmed.ncbi.nlm.nih.gov/15684123/)]



37. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a web-based tool for NLP-assisted text annotation. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.: Association for Computational Linguistics; 2012 Presented at: The 13th Conference of the European Chapter of the Association for Computational Linguistics; April 23-27, 2012; Avignon, France p. 102-107.
38. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. ArXiv. Preprint posted online on Jan 16, 2013 [[FREE Full text](#)]
39. Sun J. jieba. Jieba Chinese word segmentation module. URL: <https://github.com/fxsjy/jieba> [accessed 2021-07-03]
40. Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. A comprehensive study of named entity recognition in Chinese clinical text. J Am Med Inform Assoc 2014 Sep;21(5):808-814 [[FREE Full text](#)] [doi: [10.1136/amiajnl-2013-002381](https://doi.org/10.1136/amiajnl-2013-002381)] [Medline: [24347408](https://pubmed.ncbi.nlm.nih.gov/24347408/)]
41. Wang H, Zhang W, Zeng Q, Li Z, Feng K, Liu L. Extracting important information from Chinese Operation Notes with natural language processing methods. J Biomed Inform 2014 Apr;48:130-136 [[FREE Full text](#)] [doi: [10.1016/j.jbi.2013.12.017](https://doi.org/10.1016/j.jbi.2013.12.017)] [Medline: [24486562](https://pubmed.ncbi.nlm.nih.gov/24486562/)]
42. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49 [[FREE Full text](#)] [doi: [10.1016/j.jbi.2017.11.011](https://doi.org/10.1016/j.jbi.2017.11.011)] [Medline: [29162496](https://pubmed.ncbi.nlm.nih.gov/29162496/)]
43. Kang T, Zhang S, Tang Y, Hrubby GW, Rusanov A, Elhadad N, et al. EliIE: an open-source information extraction system for clinical trial eligibility criteria. J Am Med Inform Assoc 2017 Nov 01;24(6):1062-1071 [[FREE Full text](#)] [doi: [10.1093/jamia/ocx019](https://doi.org/10.1093/jamia/ocx019)] [Medline: [28379377](https://pubmed.ncbi.nlm.nih.gov/28379377/)]
44. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 2010 Jan;17(1):19-24 [[FREE Full text](#)] [doi: [10.1197/jamia.M3378](https://doi.org/10.1197/jamia.M3378)] [Medline: [20064797](https://pubmed.ncbi.nlm.nih.gov/20064797/)]
45. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv e-prints 2015 Nov 23:07122 [[FREE Full text](#)]
46. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. ArXiv. Preprint posted online on June 12, 2017.
47. GuoDong Z, Jian S, Jie Z, Min Z. Exploring various knowledge in relation extraction. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.: Association for Computational Linguistics; 2005 Jun Presented at: The 43rd Annual Meeting of the Association for Computational Linguistics; June 25-30, 2005; Ann Arbor, Michigan p. 427-434. [doi: [10.3115/1219840.1219893](https://doi.org/10.3115/1219840.1219893)]
48. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.: Association for Computational Linguistics; 2009 Aug Presented at: The Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP; Aug 7-12, 2009; Suntec, Singapore p. 1003-1011.
49. Chen R, Huang Y, Bau C, Chen S. A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection. Expert Syst Appl 2012 Mar;39(4):3995-4006 [[FREE Full text](#)] [doi: [10.1016/j.eswa.2011.09.061](https://doi.org/10.1016/j.eswa.2011.09.061)]
50. Zhang Y, Gou L, Tian Y, Li T, Zhang M, Li J. Design and development of a sharable clinical decision support system based on a semantic web service framework. J Med Syst 2016 May;40(5):118. [doi: [10.1007/s10916-016-0472-y](https://doi.org/10.1007/s10916-016-0472-y)] [Medline: [27002818](https://pubmed.ncbi.nlm.nih.gov/27002818/)]
51. OWL 2 Web Ontology Language Document Overview (Second Edition). URL: <https://www.w3.org/TR/owl2-overview/> [accessed 2021-07-03]
52. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. URL: <https://www.w3.org/Submission/SWRL/> [accessed 2021-07-03]

## Abbreviations

- Attention-Bi-LSTM:** attention-based bidirectional long short-term memory networks  
**Bi-LSTM:** bidirectional long short-term memory networks  
**Bi-LSTM-CRF:** bidirectional long short-term memory networks- conditional random field  
**BERT:** bidirectional encoder representation from transformers  
**CLIP:** Cancer of Liver Italian Program  
**CRF:** conditional random field  
**CT:** computed tomography  
**ID-CNN:** iterated dilated convolutional neural networks  
**ID-CNN-CRF:** iterated dilated convolutional neural networks-conditional random field  
**IE:** information extraction  
**MRI:** magnetic resonance imaging  
**NER:** named entity recognition  
**NLP:** natural language processing  
**OWL:** Web Ontology Language



**PAOP:** pulmonary atelectasis/obstructive pneumonitis  
**PET:** positron emission tomography  
**PP:** postprocessing  
**RC:** relation classification  
**RNN:** recurrent neural network  
**RSC:** relation sign constraint  
**SNOMED CT:** Systematized Nomenclature of Medicine Clinical Terms  
**SWRL:** Semantic Web Rule Language  
**UMLS:** Unified Machine Language System

*Edited by T Hao, Z Huang, B Tang; submitted 15.02.21; peer-reviewed by Z Su, H Chen, K Roberts; comments to author 29.04.21; revised version received 27.05.21; accepted 07.06.21; published 21.07.21.*

*Please cite as:*

*Hu D, Zhang H, Li S, Wang Y, Wu N, Lu X*

*Automatic Extraction of Lung Cancer Staging Information From Computed Tomography Reports: Deep Learning Approach*

*JMIR Med Inform 2021;9(7):e27955*

*URL: <https://medinform.jmir.org/2021/7/e27955>*

*doi: [10.2196/27955](https://doi.org/10.2196/27955)*

*PMID: [34287213](https://pubmed.ncbi.nlm.nih.gov/34287213/)*

©Danqing Hu, Huanyao Zhang, Shaolei Li, Yuhong Wang, Nan Wu, Xudong Lu. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 21.07.2021. 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

# Head and Tail Entity Fusion Model in Medical Knowledge Graph Construction: Case Study for Pituitary Adenoma

An Fang<sup>1,2</sup>, MS; Pei Lou<sup>2</sup>, MS; Jiahui Hu<sup>2</sup>, PhD; Wanqing Zhao<sup>2</sup>, MSc; Ming Feng<sup>3</sup>, MD; Huiling Ren<sup>2</sup>, MSL; Xianlai Chen<sup>4,5</sup>, PhD

<sup>1</sup>Life Science College, Central South University, Changsha, China

<sup>2</sup>Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China

<sup>3</sup>Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China

<sup>4</sup>Big Data Institute, Central South University, Changsha, China

<sup>5</sup>National Engineering Lab for Medical Big Data Application Technology, Central South University, Changsha, China

**Corresponding Author:**

Xianlai Chen, PhD

Big Data Institute

Central South University

932 South Lushan Road

Changsha, 410083

China

Phone: 86 731 88879583

Email: [chenxianlai@csu.edu.cn](mailto:chenxianlai@csu.edu.cn)

## Abstract

**Background:** Pituitary adenoma is one of the most common central nervous system tumors. The diagnosis and treatment of pituitary adenoma remain very difficult. Misdiagnosis and recurrence often occur, and experienced neurosurgeons are in serious shortage. A knowledge graph can help interns quickly understand the medical knowledge related to pituitary tumor.

**Objective:** The aim of this study was to develop a data fusion method suitable for medical data using data of pituitary adenomas integrated from different sources. The overall goal was to construct a knowledge graph for pituitary adenoma (KGPA) to be used for knowledge discovery.

**Methods:** A complete framework suitable for the construction of a medical knowledge graph was developed, which was used to build the KGPA. The schema of the KGPA was manually constructed. Information of pituitary adenoma was automatically extracted from Chinese electronic medical records (CEMRs) and medical websites through a conditional random field model and newly designed web wrappers. An entity fusion method is proposed based on the head-and-tail entity fusion model to fuse the data from heterogeneous sources.

**Results:** Data were extracted from 300 CEMRs of pituitary adenoma and 4 health portals. Entity fusion was carried out using the proposed data fusion model. The F1 scores of the head and tail entity fusions were 97.32% and 98.57%, respectively. Triples from the constructed KGPA were selected for evaluation, demonstrating 95.4% accuracy.

**Conclusions:** This paper introduces an approach to fuse triples extracted from heterogeneous data sources, which can be used to build a knowledge graph. The evaluation results showed that the data in the KGPA are of high quality. The constructed KGPA can help physicians in clinical practice.

(*JMIR Med Inform* 2021;9(7):e28218) doi:[10.2196/28218](https://doi.org/10.2196/28218)

## KEYWORDS

knowledge graph; pituitary adenoma; entity fusion; similarity calculation

## Introduction

Pituitary adenoma is one of the most common central nervous system tumors. Most of the benign adenomas are characterized by swelling growth, which can be cured by surgery or medicine

[1]. However, a small number of pituitary adenomas are not sensitive to surgery, radiotherapy, and drug therapy, and metastasis will lead to pituitary adenocarcinoma [2]. At present, there are difficulties in the diagnosis and treatment of pituitary adenoma [3]. In some cases, pituitary adenocarcinoma can even

be life-threatening [4] and the prognosis is extremely poor. Therefore, pituitary adenoma has become a hot topic in life science research, and an open knowledgebase of pituitary adenoma is needed.

A knowledge graph is a general framework for formal description of knowledge, which can describe knowledge in the form of triples as a “head entity-relation-tail entity,” one of the most popular knowledge representation methods currently adopted [5]. Well-known open-domain knowledge graphs include Freebase, DBpedia, YAGO, and NELL, among others [6]. Knowledge graphs are also widely used in the medical field. Gong et al [7] proposed a method to build a diabetes knowledgebase by mining the web; they extracted knowledge from the semistructured content of the vertical portal and then mapped the information onto a unified knowledge graph. Ernst et al [8] constructed a biomedical science knowledge graph in which they extracted data using distant supervision methods and used logical reasoning for consistency checks. Rotmensch et al [9] designed an automatic extraction framework to directly extract diseases and symptoms from electronic medical records (EMRs), and automatically constructed a knowledge graph.

Data fusion is an important step of the integration of heterogeneous data in the construction of knowledge graphs. Entity fusion includes methods based on character similarity, clustering, deep learning, and others. Zhang et al [10] proposed a novel multisource medical data integration and mining solution for better health care services, which can search for similar medical records in a time-efficient and privacy-preserving manner. Wang et al [11] extracted different semantic words using multimodal trees and performed multigranularity feature fusion on the data. Li et al [12] proposed a novel fusion-embedding learning model, G2SKGE, which aims to learn the subgraph structure information of the entity in a

knowledge graph. Li et al [13] proposed an approach to build a knowledge graph for hepatocellular carcinoma, and applied a biomedical information extraction system to filter and fuse the data.

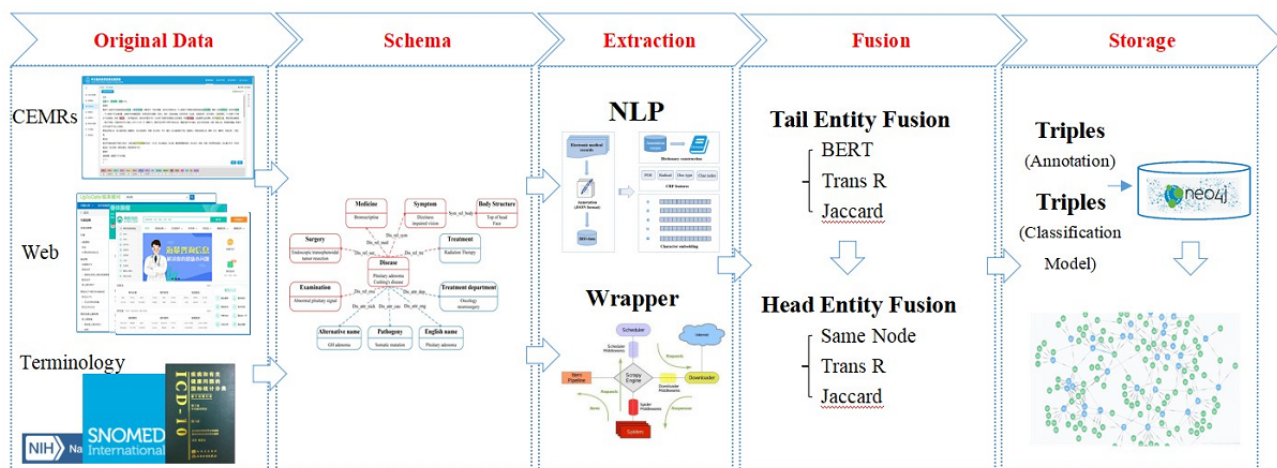
In this study, we extracted data from patient EMRs and medical websites, fused the entities using our proposed head-and-tail entity fusion model, and constructed a medical knowledge graph for pituitary adenoma (KGPA). The main contributions of this study are as follows. First, there is currently no Chinese knowledgebase for pituitary adenoma. Therefore, this study presents the complete process of knowledge graph construction, which was used to construct the KGPA. Second, to integrate the data extracted from different sources, we propose a fusion method suitable for medical data that was used in the process of KGPA construction. The method includes two steps: tail entity fusion and head entity fusion. Finally, knowledge of pituitary adenoma, such as the typical symptoms of different pituitary adenoma-related diseases, can be clearly revealed by searching the KGPA. According to doctors' feedback on use of the KGPA, the content displayed in the KGPA was considered to be consistent with the actual clinical situation.

## Methods

### Overview

According to the characteristics of pituitary adenoma diseases combined with the characteristics of Chinese electronic medical records (CEMRs) and Chinese health websites, we designed the construction framework of the KGPA, as shown in Figure 1, which includes 5 steps: raw data collection, schema design, data extraction, data fusion, and data storage and visualization. Each step is introduced in detail below, with emphasis on the proposed data fusion model.

**Figure 1.** Process for construction of the knowledge graph for pituitary adenoma. CEMR: Chinese electronic medical record; NLP: natural language processing; BERT: bidirectional encoder representations from transformer.



### Data Schema

The knowledge graph includes a data layer and a schema layer [14]. Entities, relations, and attributes in the data layer are regulated and restricted by the schema. The schema was based on several open-access authoritative terminologies and ontologies, including the UMLS Semantic Network [15], the

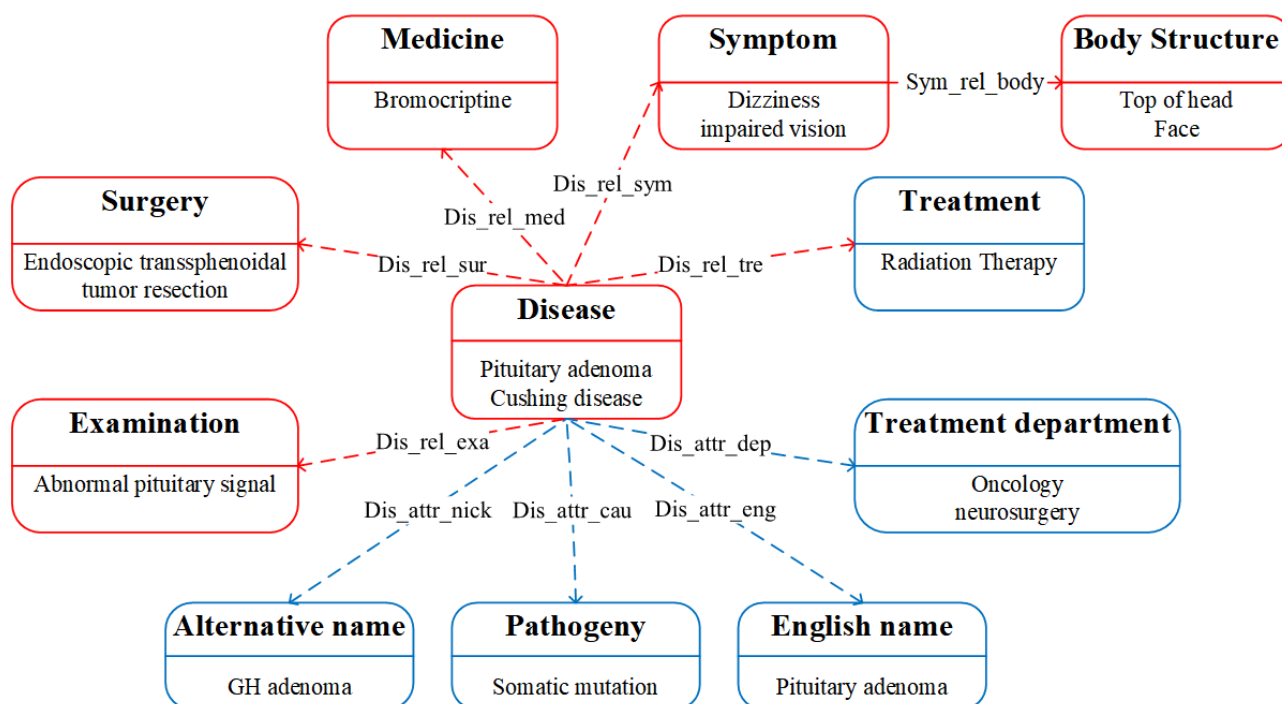
concept definitions in SNOMED-CT [16], and the International Statistical Classification of Diseases and Related Health Problems (ICD-10). In addition, the natural language processing datasets defined by the Informatics for Integrating Biology & the Bedside [17] and CEMRs Entity and Relations Annotation Specifications defined by Harbin Institute of Technology [18] were also referenced for this task. With the help of clinical

experts, a combination of top-down and bottom-up approaches was used to construct the KGPA schema.

In our previous study of CEMRs data extraction, we found that the medical diagnosis and treatment activities could be summarized based on symptoms (symptom) and abnormal results (examination) [19]. The doctor will give a comprehensive diagnosis conclusion (disease) and corresponding treatment measures (surgery, medicine). Therefore, the mentioned entities and the relations between them were abstracted for design of the schema. The CEMRs are detailed but contain a limited number of concepts; therefore, we extracted data from medical

websites to expand the concepts. Through analyzing the data types of the websites, six types of concepts were added to the schema: pathogeny, treatment, examination, treatment department, English name, and alternative name. The most frequently used disease term in websites was selected as the concept of the disease, and then treatment and examination were defined as related entities. Pathogeny, treatment department, English name, and alternative name were defined as the attributes of the disease. Attributes can be used to describe the internal characteristics of the disease entities; the more attributes there are, the more complete the information of the entity will be [20]. The KGPA schema is shown in Figure 2.

**Figure 2.** Schema of the knowledge graph for pituitary adenoma (KGPA). Concepts extracted from Chinese electronic medical records are in red. Concepts extracted from health websites are in blue. GH: growth hormone.



**Data Extraction**

**Process**

In the process of data extraction, entities and relations were first extracted from unstructured information in CEMRs. For website data, specific HTML wrappers were constructed to directly extract the triples (eg, Cushing syndrome, Symptom, Lethargy). The details are described below.

**EMR Data Extraction**

CEMRs include information on admission, discharge summary, disease course, and a medical record summary, among other details. Since the history of present illness (HPI) in the admission record contains a large amount of detailed patient

symptoms and preliminary examination information, the HPI was selected as the main data source in our study.

The Chinese Clinical Natural Language Processing System (CCNLP) [21] developed by our team was used to annotate entities and relations in CEMRs, as shown in Figure 3. The CCNLP allows user to customize the entities and relations. According to the definition of the schema, we defined 6 types of entities and 5 types of relations in the CCNLP. Two clinicians were invited to perform annotation. The conditional random field model is embedded in the system, which can train the annotated corpus and assist in annotation. The results of the two annotators were evaluated by the consistency evaluation function of the CCNLP [22].

Figure 3. Medical text annotation using the Chinese Clinical Natural Language Processing System (CCNLP) system.



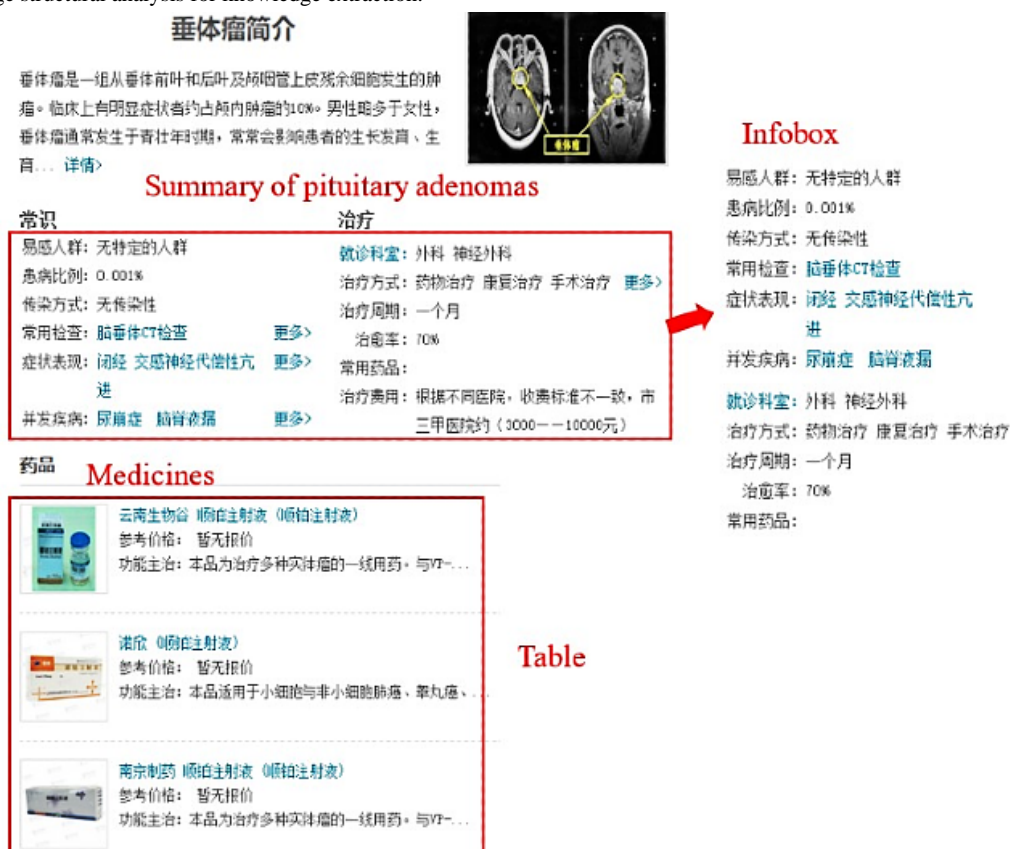
Web Data Extraction

The web data were mainly collected from medical websites and high-quality encyclopedia websites. The extracted disease entities in the CEMRs were used as search terms on the medical websites. Since single medical website retrieval is not comprehensive, four websites with higher data quality were used: xywy [23], UpToDate [24], Baidu Encyclopedia [25], and chunyuisheng [26]. All of these websites provide HTML pages of diseases, symptoms, treatments, and other relevant details.

This enabled obtaining sufficient medical knowledge to construct the knowledge graph.

Since the websites shared similar structures, xywy was selected as an example to illustrate the details of pages and its structures used for data extraction. As shown in Figure 4, the information in “Infobox” can be directly extracted and stored as triples. The “Medicines” data in the website are stored in a tabular format. We extracted the title and first lines of the tables, which were combined as triples. Different wrappers were designed to extract information from different web pages.

Figure 4. Web page structural analysis for knowledge extraction.





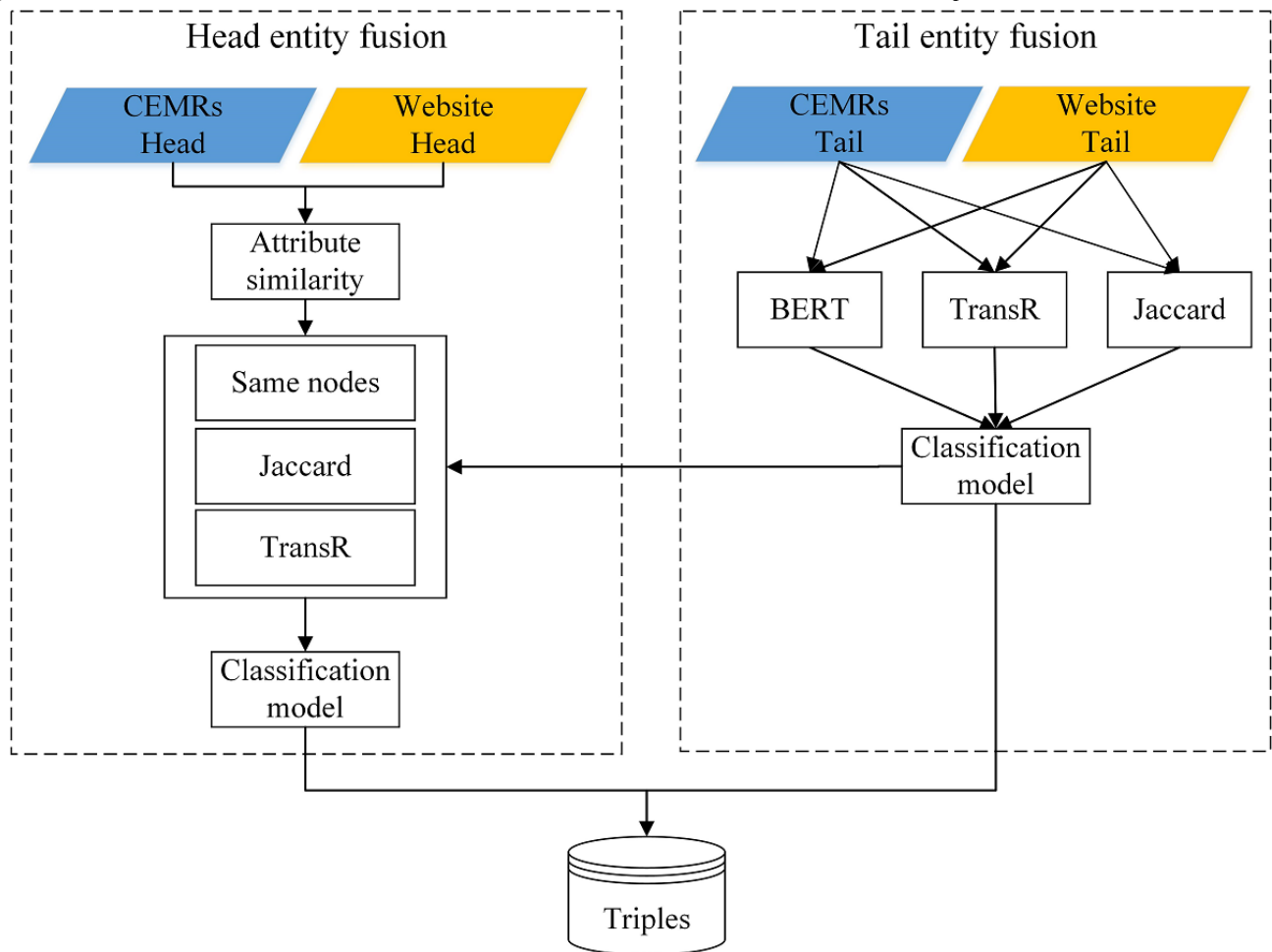
**Data Fusion**

**Framework**

Triples from different sources may have complements, redundancies, or even conflicts among each other. To ensure accuracy of the data in the knowledge graph, a data fusion method was proposed as shown in Figure 5. The data were fused by calculating the similarity of head entities and tail entities. The purpose of similarity calculation is to find the optimal alignment between the website entities and CEMR entities. The

fusion methods were carried out in two steps. First, the similarity of tail entities (symptoms and examinations contained in both data sources) were calculated based on bidirectional encoder representations from transformer (BERT), the TransR model, and the Jaccard coefficient. Tail entity fusion enabled obtaining a more consistent entity expression. Second, the structural information of the graph was used to merge the head entities (diseases) through the TransR model, Jaccard coefficient, and the count of same nodes.

**Figure 5.** Data fusion framework. CEMR: Chinese electronic medical record; BERT: bidirectional encoder representations from transformer.



**Tail Entity Fusion Model**

**Features**

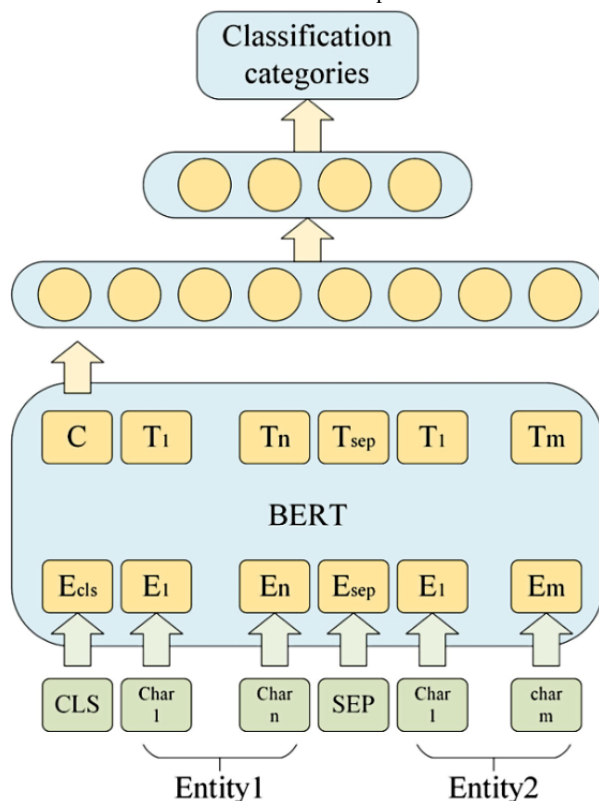
In the entity fusion task, there are only two types of training results (positive and negative); therefore, this can be converted into a binary classification problem. In the tail entity fusion experiment, three different features were constructed as model inputs: semantic similarity, TransR similarity, and Jaccard similarity.

**Semantic Similarity Calculation Based on BERT**

A semantic model is widely used in the similarity calculation of textual data. In this study, the semantic classification model

was trained with labeled data. BERT-Base, Chinese [27] was used to construct the embedding of the tail entities in CEMRs and website data, as shown in Figure 6. Tail entities can be regarded as short sentences, and the matching problem of entity pairs can be modeled as a classification task. The first output vector of the coding layer “C” is taken as the semantic representation of the entity pair. “[CLS]” represents the beginning of a sentence and “[SEP]” separates the two sentences. “E” represents the word embedding of the input character and “T” represents the contextual representation of the input character. The semantic categories are then calculated using two full connection layers: full connection layer 1 uses a tanh activation function and full connection layer 2 normalizes the probability of each class with the softmax function.

Figure 6. Semantic similarity calculation model based on bidirectional encoder representations from transformer (BERT).



**Knowledge Representation Learning**

Knowledge representation learning methods do not rely on textual information but rather obtain the depth characteristics of the data by mapping the entities to low-dimensional space vectors. A total of 4684 pituitary adenoma triples were used to test the data representation ability of the Trans models [28]. We evaluated the performance of the models using hits@10 (ie, the proportion of correctly aligned entities ranked in the top 10 predictions); a higher hits@10 value indicates better performance. The evaluation results were 0.27 for TransE, 0.37 for TransH, and 0.39 for TransR. Therefore, TransR was selected for knowledge representation learning. The extracted triples were used as positive examples (head [h], relation [r], tail [t]). For each positive triple, we randomly replaced its head entity (h', r, t) or tail entity (h, r, t') to generate a negative triple. A mapping matrix  $M_r$  was used to describe the relational space of relation r. Using the gradient descent method to update the parameters, we obtained the vector of the tail entities  $trans\_vec$ . The cosine similarity cos was used to calculate the tail entity similarity of the two data sources, as shown in Equation 1:

$$Sim_{teal\_trans}(m_i, n_i) = \text{argmax}(\cos[trans\_vec_{m_i}], \cos[trans\_vec_{n_i}])$$

(1)

**Jaccard Coefficient**

The Jaccard coefficient was selected as the third feature of tail entity fusion. The Jaccard coefficient refers to the ratio of the number of intersection elements to the union elements in two sets; the higher the Jaccard value, the higher the similarity. We assigned each tail entity in the CEMRs and websites to sets  $t_1$  and  $t_2$ , respectively. The Jaccard coefficient represents the ratio

of the same number of Chinese characters in the two words to the total number of characters, as shown in Equation 2:

$$Jaccard(t_1, t_2) = \frac{|t_1 \cap t_2|}{|t_1 \cup t_2|} = \frac{|t_1 \cap t_2|}{|t_1| + |t_2| - |t_1 \cap t_2|}$$

(2)

**Head Entity Fusion Model**

**Features**




When merging head entities (diseases), the similarity of the two attributes and their structures were mainly considered. That is, if two head entities are the same, their neighboring entities should also be similar.

**Attribute Similarity**

Entity alignment can be performed using the alternative name attribute or the English name attribute of the disease. If the head entities in the two data sources have the same alternative name or English name, the two entities can be considered the same. For example, “垂体生长激素腺瘤” (growth hormone-secreting pituitary adenoma) has alternative names of “pituitary growth hormone secreting adenoma” and “GH adenoma.” Therefore, we can align “pituitary growth hormone secreting adenoma” and “GH adenoma” to “growth hormone-secreting pituitary adenoma.”

**Structural Similarity Fusion Model**

When the head entities cannot be aligned by the attribute, we propose using the structural similarity model to fuse entities. Three different features were chosen as the classifier model’s inputs: the number of identical tail nodes, Jaccard similarity, and TransR similarity, as shown in Equation 3.

The head entity and the tail entity have a 1-N relationship. Taking two disease sets  from two data sources as an example,  represents the number of identical tail nodes in different sets and  represents the ratio of the same number of characters to the total number of characters of two sets. The order of words in the set is not considered. For the attribute similarity, the vector representation of entities was trained using the TransR model, whereas in this case, we calculated the vector of the head entity using the TransR model.



After the head entities of two heterogeneous data sources were fused, the triples containing all of the disease information were obtained. Finally, to standardize the disease names in the knowledge graph, we mapped them to the ICD codes.

**Table 1.** Number of relations before and after data fusion.

Relation	Head entity	Tail entity	Prefusion	After fusion
Diseases_rel_Symptom	disease	symptom	3154	1940
Diseases_rel_Surgery	disease	surgery	55	45
Diseases_rel_Medicines	disease	medicine	245	182
Diseases_rel_Examination	disease	examination	437	274
Symptoms_rel_Body structure	symptom	body	396	281
Diseases_rel_Treatment	disease	treatment	110	109
Diseases_attr_Pathogeny	disease	pathogeny	122	104
Diseases_attr_Department	disease	department	71	44
Diseases_attr_English name	disease	English name	71	42
Diseases_attr_Alternative name	disease	alternative name	23	20

## Data Fusion

Two hundred medical records were randomly selected for the fusion experiment. The ratio of the training set and test set was 8:2. The experiment was trained under Windows 10, and the model based on the TensorFlow framework was used.

The proposed tail entity fusion model was used to perform entity fusion for symptoms and examinations. Before the fusion began, different entities with the same conceptual semantics extracted from different websites were merged to reduce duplication and computation. A vector representation of 768 dimensions was constructed through the Chinese BERT model, and then the similarity results were obtained by full connection layers. A 50-dimensional vector was obtained by the TransR model and the cosine similarity was used to calculate the entity pair similarity values. The Jaccard coefficient was used as a numerical feature. These three results were taken as features into the classification model. Three different classification models were adopted for training: logistic regression, decision tree, and neural network. The results are shown in [Table 2](#). The neural network showed the best performance.

## Results

### Data Extraction

Three hundred clinical medical records and 4 portal websites were selected as data sources to construct the KGPA. Although these are all Chinese resources, our proposed approach is not dependent on a particular language and can be applied to data resources in other language in the same way. The data in CEMRs were annotated by two doctors using the CCNLP system [21]. With the consistency test function of the system, the consistency of the annotations reached 95.2%. Website data were extracted according to the wrapper defined in this study. [Table 1](#) shows the number of all entities extracted from the two types of data sources. The concepts are abundant in websites, whereas the CEMRs included more symptom entities, which can help to expand more data types for the KGPA. The “Prefusion” column of [Table 1](#) shows the number of all relations extracted from the two types of data sources.

Subsequently, the triples completed by the tail entity fusion model were used for the head entity fusion experiment. A total of 65 head entities were fused between CEMRs and websites. Among them, 17 entities could be directly mapped by disease name, 6 entities could be fused by attribute (eg, growth hormone-secreting pituitary adenoma, pituitary microadenoma, Cushing syndrome, hypothyroidism), and 42 head entities were fused based on the proposed structural similarity fusion model. The three classification models above were used for training. As shown in [Table 2](#), the decision trees performed better when fusing head entities because the data inputs to the model were smaller than the fusing tail entities. With the increase of data volume, the advantages of the neural network were reflected in the fusion of tail entities.

Additionally, we divided the features into four variants for an ablation study. We selected logistic regression as the classification model to explore the contribution of different features to the model, and these results are also shown in [Table 2](#). These three features had nearly the same contributions to the model in the head entity fusion. For a specific disease knowledge graph, the Jaccard similarity feature played a major role in the

tail entity ablation experiment, and the features based on BERT and TransR simply contributed by fine-tuning the model.

Table 3 shows that our proposed model has higher accuracy than previous models. Compared with previous models, we

divided the entities into head entities and tail entities and fused them according to different characteristics. Different concepts were considered separately in the step-by-step fusion process, which improved the precision of the fusion.

**Table 2.** Head and tail fusion model performance.

Fusion model	Precision (%)	Recall (%)	F-score (%)
<b>Head entity fusion</b>			
<b>Linear regression models</b>			
Ja <sup>a</sup> +TransR	83.37	84.06	83.71
Sa <sup>b</sup> +TransR	83.37	84.55	83.95
Ja+Sa	83.85	84.55	84.19
Ja+Sa+TransR	83.92	84.61	84.26
Neural network	97.29	97.03	97.16
Decision tree	97.47	97.18	97.32
<b>Tail entity fusion</b>			
<b>Linear regression models</b>			
BERT <sup>c</sup> +TransR	61.73	61.74	61.73
Ja+BERT	95.76	95.83	95.79
Ja+TransR	95.89	95.93	95.90
Ja+BERT+TransR	95.92	95.94	95.93
Neural network	98.43	98.72	98.57
Decision tree	98.18	98.05	98.11

<sup>a</sup>Ja : Jaccard similarity.

<sup>b</sup>Sa: identical tail nodes in different sets: .

<sup>c</sup>BERT: bidirectional encoder representations from transformer.

**Table 3.** Model comparison.

Model	Research field	Method	F1-score
Ruan et al [29]	Symptom	Align entities according to the string similarities of the entity names and attribute values	— <sup>a</sup>
Yang et al [30]	Disease, medicine	Align entities according to the entity's attribute types (attr <sub>bool</sub> , attr <sub>numeric</sub> , attr <sub>string</sub> , attr <sub>time</sub> )	0.60
Sun et al [31]	Disease, medicine, symptom	Character similarity of entity pairs and degree centrality of entities in the graph	0.76
Liu et al [32]	Disease, medicine, examination	Semantic classification model based on pretrained BERT <sup>b</sup>	0.83
Our model	Symptom, examination, disease	Multifeature learning based on head-and-tail entities	0.97

<sup>a</sup>Not provided.

<sup>b</sup>BERT: bidirectional encoder representations from transformer.

The triples obtained after data fusion were stored and visualized in Neo4j [33]. The KGPA contained 1789 entities and 3041 pairs of relations of 73 pituitary adenoma-related diseases. For a knowledge graph, accuracy is of great importance. However, there is currently no gold standard for pituitary adenoma knowledge graph validation. To evaluate the quality of the knowledge graph, the accuracy of triples was used as an indicator. Three hundred triples were randomly sampled and each triple was manually evaluated by two physicians; the accuracy reached 95.4%.

## Discussion

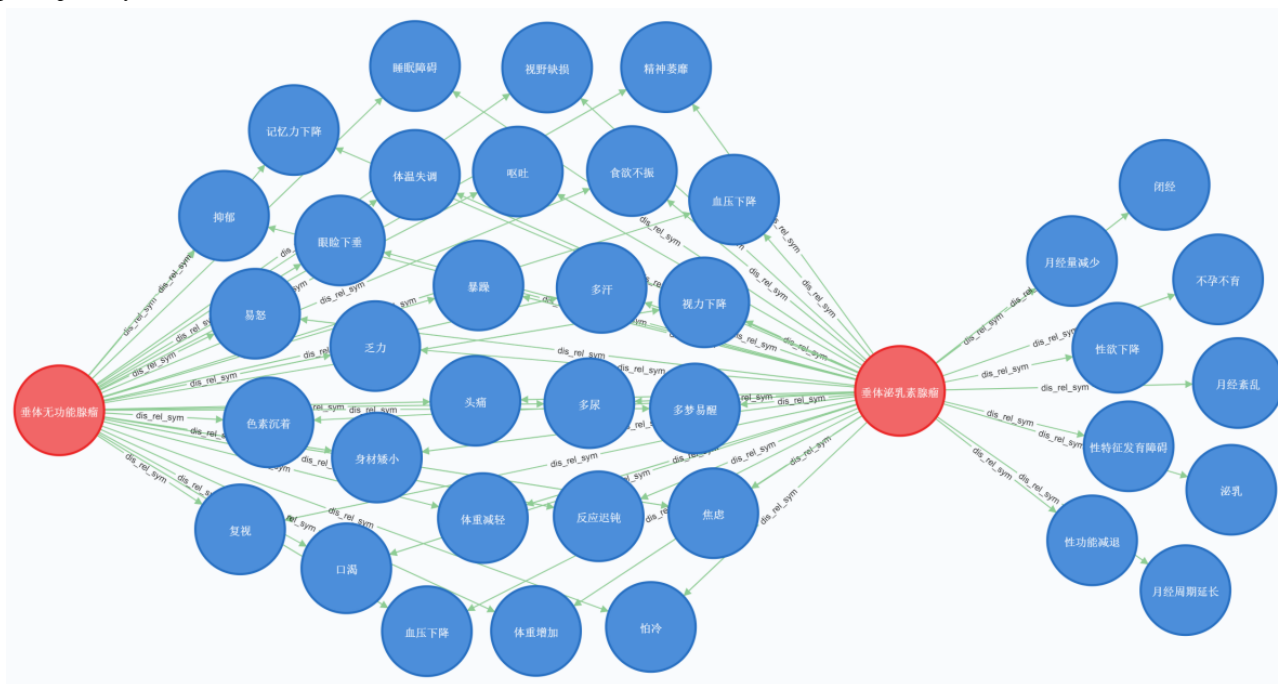
### Principal Findings

A knowledge graph was constructed by mining CEMRs and web resources. In the process of KGPA construction, to solve the problem of knowledge duplication between heterogeneous data sources, we proposed a head-and-tail entity fusion model. The model showed good performance on the fusion of medical data.

The KGPA was proven to be effective when displaying the typical symptoms of pituitary adenoma-related diseases. For example, the query for symptoms of disease “prolactin (PRL)-secreting pituitary adenomas” differed from the query for the disease “nonfunctioning pituitary adenoma” using the following query in Cypher: “MATCH (p:dis{disease: 垂体泌

乳素腺瘤})-[:dis\_rel\_sym]->(n), (m)<-[:dis\_rel\_sym]-(q:dis{disease:垂体无功能腺瘤}), WHERE (m)<-(n), RETURN p,n,q.” As shown in Figure 7, the entities in the middle of the graph are symptoms of both diseases and the entities on the right are typical symptoms unique to the disease “PRL-secreting pituitary adenomas.”

**Figure 7.** Differences of typical symptoms between “prolactin-secreting pituitary adenomas” and “nonfunctioning pituitary adenoma” in the knowledge graph for pituitary adenoma.



Searching for the KGPA by Cypher, we found that most pituitary adenoma-related diseases have the following basic symptoms: headache, vision problems, fatigue, slow reaction, mood problems, changes in height and weight, changes in appetite, and changes in sleep. Nonfunctioning pituitary adenoma has all of these basic symptoms listed above. In addition to the basic symptoms, pituitary thyroid-stimulating hormone adenoma is also associated with symptoms of goiter, palpitation, and exophthalmos. The typical symptoms of PRL-secreting pituitary adenomas are associated with the reproductive system, decreased libido, and menstrual changes in women. The typical symptoms of pituitary growth hormone adenoma are altered facial features, enlarged hands and feet, snoring, and metabolic disorders. Cushing syndrome is characterized by obesity, altered skin color, increased hair, and edema. Based on clinicians’ feedback on the use of the KGPA, the knowledge in the KGPA was consistent with the actual clinical situation. The KGPA will be useful for clinical interns in diagnosis and treatment, and may also be helpful for medical students to quickly master knowledge of pituitary adenoma-related diseases.

**Limitations**

The KGPA was constructed by integrating CEMRs and web data related to pituitary adenoma. However, since we only focused on pituitary tumors, the data volume was relatively small. In the next step, we plan to try to extend the method proposed in this study to the entire neurosurgery field or even larger fields and apply the knowledge graph to clinical practice.

**Conclusion**

This study shows that entities and relations extracted from heterogeneous data sources such as CEMRs and health websites can be used to construct a knowledge graph after entity fusion. The head-and-tail entity fusion model proposed in this paper achieved 97% in accuracy, which is higher than that reported for previous models. The KGPA constructed in this study can be used to discover the knowledge hidden in the source text, such as typical symptoms unique to the disease “PRL-secreting pituitary adenomas.” Based on clinicians’ feedback, the knowledge in the KGPA was consistent with the actual clinical situation. The knowledge graph constructed will be useful and helpful for patients, medical students, and interns to assist in obtaining information for symptoms, diagnosis, treatment, and disease pathogenesis.

**Acknowledgments**

This research has been funded by the Science and Technology Innovation 2030-Major Project (2020AAA0104902), the Chinese Academy of Medical Sciences Initiative for Innovative Medicine (2017-I2M-3-014), the Chinese Academy of Medical Sciences



and Peking Union Medical College Fundamental Scientific Research Funds Project of the Central Public Welfare Research Institution (2018PT33005), and the Hunan Provincial Key Research and Development Program (2020SK2089).

### Authors' Contributions

AF designed the methods, analyzed the results of experiments, and drafted the paper. PL, JH, and WZ extracted the data and performed the data fusion. MF collected the electronic medical records and annotated the dataset. MF and HR evaluated the pituitary adenoma knowledge graph. XC supervised the research and revised the paper. All authors read and approved the final manuscript.

### Conflicts of Interest

None declared.

### References

1. Osamura R, Egashira N, Miyai S, Yamazaki M, Takekoshi S, Sanno N, et al. Molecular pathology of the pituitary. Development and functional differentiation of pituitary adenomas. *Front Horm Res* 2004;32:20-33. [doi: [10.1159/000079036](https://doi.org/10.1159/000079036)] [Medline: [15281338](https://pubmed.ncbi.nlm.nih.gov/15281338/)]
2. Kinoshita Y, Tominaga A, Usui S, Arita K, Sugiyama K, Kurisu K. Impact of subclinical haemorrhage on the pituitary gland in patients with pituitary adenomas. *Clin Endocrinol (Oxf)* 2014 May;80(5):720-725. [doi: [10.1111/cen.12349](https://doi.org/10.1111/cen.12349)] [Medline: [24125536](https://pubmed.ncbi.nlm.nih.gov/24125536/)]
3. Kim JP, Park BJ, Kim SB, Lim YJ. Pituitary apoplexy due to pituitary adenoma infarction. *J Korean Neurosurg Soc* 2008 May;43(5):246-249 [FREE Full text] [doi: [10.3340/jkns.2008.43.5.246](https://doi.org/10.3340/jkns.2008.43.5.246)] [Medline: [19096606](https://pubmed.ncbi.nlm.nih.gov/19096606/)]
4. Kaushik C, Ramakrishnaiah R, Angtuaco EJ. Ectopic pituitary adenoma in persistent craniopharyngeal canal. *J Comput Assist Tomogr* 2010;34(4):612-614. [doi: [10.1097/rct.0b013e3181d8e5d1](https://doi.org/10.1097/rct.0b013e3181d8e5d1)]
5. Byambasuren O, Yang Y, Sui Z, Dai D, Chang B, Li S, et al. Preliminary study on the construction of Chinese medical knowledge graph. *J Chinese Inf Process* 2019;33(10):1-9 [FREE Full text]
6. Li L, Wang P, Yan J, Wang Y, Li S, Jiang J, et al. Real-world data medical knowledge graph: construction and applications. *Artif Intell Med* 2020 Mar;103:101817. [doi: [10.1016/j.artmed.2020.101817](https://doi.org/10.1016/j.artmed.2020.101817)] [Medline: [32143785](https://pubmed.ncbi.nlm.nih.gov/32143785/)]
7. Gong F, Chen Y, Wang H, Lu H. On building a diabetes centric knowledge base via mining the web. *BMC Med Inform Decis Mak* 2019 Apr 09;19(Suppl 2):49 [FREE Full text] [doi: [10.1186/s12911-019-0771-6](https://doi.org/10.1186/s12911-019-0771-6)] [Medline: [30961582](https://pubmed.ncbi.nlm.nih.gov/30961582/)]
8. Ernst P, Siu A, Weikum G. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC Bioinformatics* 2015 May 14;16:157 [FREE Full text] [doi: [10.1186/s12859-015-0549-5](https://doi.org/10.1186/s12859-015-0549-5)] [Medline: [25971816](https://pubmed.ncbi.nlm.nih.gov/25971816/)]
9. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. *Sci Rep* 2017 Jul 20;7(1):5994. [doi: [10.1038/s41598-017-05778-z](https://doi.org/10.1038/s41598-017-05778-z)] [Medline: [28729710](https://pubmed.ncbi.nlm.nih.gov/28729710/)]
10. Zhang Q, Lian B, Cao P, Sang Y, Huang W, Qi L. Multi-source medical data integration and mining for healthcare services. *IEEE Access* 2020;8:165010-165017. [doi: [10.1109/access.2020.3023332](https://doi.org/10.1109/access.2020.3023332)]
11. Wang C, Wang H, Zhuang H, Li W, Han S, Zhang H, et al. Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree. *J Biomed Inform* 2020 Nov;111:103583. [doi: [10.1016/j.jbi.2020.103583](https://doi.org/10.1016/j.jbi.2020.103583)] [Medline: [33010427](https://pubmed.ncbi.nlm.nih.gov/33010427/)]
12. Li W, Zhang X, Wang Y, Yan Z, Peng R. Graph2Seq: fusion embedding learning for knowledge graph completion. *IEEE Access* 2019;7:157960-157971. [doi: [10.1109/access.2019.2950230](https://doi.org/10.1109/access.2019.2950230)]
13. Li N, Yang Z, Luo L, Wang L, Zhang Y, Lin H, et al. KGHC: a knowledge graph for hepatocellular carcinoma. *BMC Med Inform Decis Mak* 2020 Jul 09;20(Suppl 3):135 [FREE Full text] [doi: [10.1186/s12911-020-1112-5](https://doi.org/10.1186/s12911-020-1112-5)] [Medline: [32646496](https://pubmed.ncbi.nlm.nih.gov/32646496/)]
14. Nickel M, Murphy K, Tresp V, Gabrilovich E. A review of relational machine learning for knowledge graphs. *Proc IEEE* 2016 Jan;104(1):11-33. [doi: [10.1109/jproc.2015.2483592](https://doi.org/10.1109/jproc.2015.2483592)]
15. Unified Medical Language System (UMLS). National Library of Medicine. URL: <https://www.nlm.nih.gov/research/umls/index.html> [accessed 2021-02-09]
16. SNOMED International. URL: <https://www.snomed.org/> [accessed 2021-02-09]
17. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011 Sep 01;18(5):552-556 [FREE Full text] [doi: [10.1136/amiajnl-2011-000203](https://doi.org/10.1136/amiajnl-2011-000203)] [Medline: [21685143](https://pubmed.ncbi.nlm.nih.gov/21685143/)]
18. Yang J, Guan Y, He B, Qu CY, Yu QB, Liu YX, et al. Corpus construction for named entities and entity relations on Chinese electronic medical records. *J Softw* 2016;2725-2746. [doi: [10.13328/j.cnki.jos.004880](https://doi.org/10.13328/j.cnki.jos.004880)]
19. Yang C, Hu J, Fang A. Study on the building of clinical text natural language processing system—taking cTAKES as an example. *J Med Inform* 2018;39(12):48-53.
20. Su X, Fan K, Shi W. Privacy-preserving distributed data fusion based on attribute protection. *IEEE Trans Ind Inf* 2019 Oct;15(10):5765-5777. [doi: [10.1109/tii.2019.2912175](https://doi.org/10.1109/tii.2019.2912175)]

21. Chinese Clinical Natural Language Processing System (CCNLP). 2021. URL: <http://ccnlp.imicams.ac.cn/> [accessed 2021-02-09]
22. Hu J, Fang A, Zhao W, Yang C, Ren H. Annotating Chinese e-medical record for knowledge discovery. *Data Anal Knowl Discov* 2019;3(7):123-132.
23. xywy. URL: <http://www.xywy.com/> [accessed 2020-12-20]
24. UpToDate. URL: <https://www.uptodate.cn/home/> [accessed 2020-12-20]
25. Baidu Encyclopedia. URL: <https://baike.baidu.com/> [accessed 2020-12-20]
26. chunyuisheng. URL: <https://www.chunyuisheng.com/> [accessed 2020-12-20]
27. Devlin J, Chang MW, Lee K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. 2018. URL: <https://arxiv.org/abs/1810.04805> [accessed 2021-07-07]
28. Lin H, Liu Y, Wang W, Yue Y, Lin Z. Learning entity and relation embeddings for knowledge resolution. *Procedia Comput Sci* 2017;108:345-354. [doi: [10.1016/j.procs.2017.05.045](https://doi.org/10.1016/j.procs.2017.05.045)]
29. Ruan T, Wang M, Sun J, Wang T, Zeng L, Yin Y, et al. An automatic approach for constructing a knowledge base of symptoms in Chinese. *J Biomed Semantics* 2017 Sep 20;8(Suppl 1):33 [FREE Full text] [doi: [10.1186/s13326-017-0145-x](https://doi.org/10.1186/s13326-017-0145-x)] [Medline: [29297414](https://pubmed.ncbi.nlm.nih.gov/29297414/)]
30. Fu Y, Liu M, Qiao R. Construction of Chinese knowledge graph of heart disease. *J Wuhan Univ* 2020;66(3):261-267. [doi: [10.14188/j.1671-8836.2018.0217](https://doi.org/10.14188/j.1671-8836.2018.0217)]
31. Sun Q. Knowledge extraction and alignment for respiratory disease. Harbin Institute of Technology. 2019. URL: [https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD\\_202001&filename=1019646460.nh](https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD_202001&filename=1019646460.nh) [accessed 2021-03-30]
32. Liu X, Jin J, Ruan T, Gao D, Yin Y, Ge X. Construction of an open dataset for clinical event graph. *J Chinese Inf Process* 2020;11:37-48.
33. Balaur I, Mazein A, Saqi M, Lysenko A, Rawlings CJ, Auffray C. Recon2Neo4j: applying graph database technologies for managing comprehensive genome-scale networks. *Bioinformatics* 2017 Apr 01;33(7):1096-1098 [FREE Full text] [doi: [10.1093/bioinformatics/btw731](https://doi.org/10.1093/bioinformatics/btw731)] [Medline: [27993779](https://pubmed.ncbi.nlm.nih.gov/27993779/)]

## Abbreviations

- BERT:** bidirectional encoder representations from transformer  
**CCNLP:** Chinese Clinical Natural Language Processing System.  
**CEMR:** Chinese electronic medical record  
**EMR:** electronic medical record  
**HPI:** history of present illness  
**ICD:** International Classification of Diseases  
**KGPA:** knowledge graph for pituitary adenoma  
**PRL:** prolactin

*Edited by T Hao; submitted 25.02.21; peer-reviewed by S Zhu, H Lin; comments to author 12.03.21; revised version received 11.04.21; accepted 30.05.21; published 22.07.21.*

*Please cite as:*

Fang A, Lou P, Hu J, Zhao W, Feng M, Ren H, Chen X

Head and Tail Entity Fusion Model in Medical Knowledge Graph Construction: Case Study for Pituitary Adenoma

*JMIR Med Inform* 2021;9(7):e28218

URL: <https://medinform.jmir.org/2021/7/e28218>

doi: [10.2196/28218](https://doi.org/10.2196/28218)

PMID: [34057414](https://pubmed.ncbi.nlm.nih.gov/34057414/)

©An Fang, Pei Lou, Jiahui Hu, Wanqing Zhao, Ming Feng, Huiling Ren, Xianlai Chen. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org/>), 22.07.2021. 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/>