JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 9 (2021), Issue 11 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Reviews

Original Papers

Corrigenda and Addendas

Review

Machine Learning Applications in Mental Health and Substance Use Research Among the LGBTQ2S+ Population: Scoping Review

Anasua Kundu[1], MSc; Michael Chaiton[1,2], PhD; Rebecca Billington[3], MSW; Daniel Grace[2], PhD; Rui Fu[4], PhD; Carmen Logie[3,5], PhD; Bruce Baskerville[6,7], PhD; Christina Yager[1], MSW; Nicholas Mitsakakis[2,8], P Stat, PhD; Robert Schwartz[1,2,4], PhD

[1]Centre for Addiction and Mental Health, Toronto, ON, Canada
[2]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[3]Factor-Inwentash Faculty of Social Work, University of Toronto, Toronto, ON, Canada
[4]Sunnybrook Research Institute, University of Toronto, Toronto, ON, Canada
[5]Women's College Research Institute, Toronto, ON, Canada
[6]Canadian Institutes of Health Research, Government of Canada, Ottawa, ON, Canada
[7]School of Pharmacy, Faculty of Science, University of Waterloo, Kitchener, ON, Canada
[8]Children's Hospital of Eastern Ontario Research Institute, Ottawa, ON, Canada

Corresponding Author:
Anasua Kundu, MSc
Centre for Addiction and Mental Health
1000 Queen Street West
Toronto, ON, M6J 1H4
Canada
Phone: 1 6476326493
Email: anasua.kundu@mail.utoronto.ca

Abstract

Background: A high risk of mental health or substance addiction issues among sexual and gender minority populations may have more nuanced characteristics that may not be easily discovered by traditional statistical methods.

Objective: This review aims to identify literature studies that used machine learning (ML) to investigate mental health or substance use concerns among the lesbian, gay, bisexual, transgender, queer or questioning, and two-spirit (LGBTQ2S+) population and direct future research in this field.

Methods: The MEDLINE, Embase, PubMed, CINAHL Plus, PsycINFO, IEEE Xplore, and Summon databases were searched from November to December 2020. We included original studies that used ML to explore mental health or substance use among the LGBTQ2S+ population and excluded studies of genomics and pharmacokinetics. Two independent reviewers reviewed all papers and extracted data on general study findings, model development, and discussion of the study findings.

Results: We included 11 studies in this review, of which 81% (9/11) were on mental health and 18% (2/11) were on substance use concerns. All studies were published within the last 2 years, and most were conducted in the United States. Among mutually nonexclusive population categories, sexual minority men were the most commonly studied subgroup (5/11, 45%), whereas sexual minority women were studied the least (2/11, 18%). Studies were categorized into 3 major domains: web content analysis (6/11, 54%), prediction modeling (4/11, 36%), and imaging studies (1/11, 9%).

Conclusions: ML is a promising tool for capturing and analyzing hidden data on mental health and substance use concerns among the LGBTQ2S+ population. In addition to conducting more research on sexual minority women, different mental health and substance use problems, as well as outcomes and future research should explore newer environments, data sources, and intersections with various social determinants of health.

(JMIR Med Inform 2021;9(11):e28962) doi:10.2196/28962

KEYWORDS

sexual and gender minorities; mental health; mental disorders; substance-related disorders; machine learning

Introduction

Background

Members of the lesbian, gay, bisexual, transgender, queer or questioning, and two-spirit (LGBTQ2S+) population experience significant mental health disparities and are at a higher risk of substance use problems compared with their heterosexual and cisgender peers [1-5]. A meta-analysis of 25 studies revealed that lesbian, gay, and bisexual individuals had 2.47 times increased lifetime risk of attempting suicide, 1.5 times increased risk of depression and anxiety disorders, and 1.5 times increased risk of alcohol and other substance dependence over a 12-month period [2]. Recent statistics from the 2015 National Survey on Drug Use and Health in the United States reported that the sexual minority population have an increased likelihood of past year use of illicit drugs, marijuana, and opioids; current use of cigarettes and alcohol; and past year diagnosis of any mental illness compared with sexual majority groups [6]. Members of the LGBTQ2S+ population also use mental health services and substance use treatment more frequently than cisgender and heterosexual individuals [6,7].

There is a robust evidence base documenting sexual orientation and gender identity as social determinants of health, whereby members of the LGBTQ2S+ population experience stressors from stigma, social, and economic exclusion that contribute to increased mental health challenges and resultant coping strategies, including problematic substance use [8-10]. In addition, intersecting experiences of marginalization such as race, ethnicity, disability, and homelessness; lack of familial and peer support; various acts of bullying, harassment, and hate crimes; and experience of self-stigmatization, such as internalized homophobia, biphobia, and transphobia, contribute to further deterioration of mental health and substance use concerns [8,11-16].

With advances in technology, novel statistical methods, such as machine learning (ML), have emerged as promising means of analyzing a vast range of complex data in public health informatics [17,18]. ML uses computational power to identify or *mine* hidden data patterns and has been increasingly used for content analysis and as a predictive modeling technique [17]. These characteristics are particularly important for investigating mental health and substance use issues among the LGBTQ2S+ population, where social stigma and institutional barriers make sexual and gender identity disclosure difficult, rendering the data invisible [19-21].

There are 3 major types of ML, including (1) supervised learning, (2) unsupervised learning, and (3) semisupervised learning. Supervised learning aims to learn from labeled data to predict the class of unlabeled input data or outcome variables [22]. Unsupervised learning does not require an outcome variable, thereby allowing the algorithm to freely detect and recognize hidden patterns with minimal human interference [22,23]. Semisupervised learning learns from both labeled and unlabeled data, where it can use readily available unlabeled data to improve supervised learning tasks when the labeled data are scarce or expensive [24]. A more advanced form of ML, deep learning, has gained popularity in health research in recent years and uses an artificial neural network model with multiple layers to hierarchically define and process data [25]. These ML methods provide the opportunity to understand data more thoroughly and effectively, as well as yield meaningful predictions beyond traditional statistical methods.

Several reviews, including 3 recent systematic reviews, have been conducted to summarize the application of ML in substance use and mental health issues [23,26-28]. These systematic reviews have reported ML applications in 54 articles on mental health, 87 articles on suicidal behavior, and 17 articles on addiction research and reported good performance in predicting human behavior [23,26,28]. However, most of these reviews and studies focused on broad categories and the general population or patient records.

Objectives

Although one scoping review has explored studies that predict population-specific health with ML [29], the study did not identify ML applications among the LGBTQ2S+ population. There is a substantial gap in the literature, with no existing review focused on ML studies examining mental health and substance use among the LGBTQ2S+ population. As a result, we conducted a scoping review to address these knowledge gaps with the aim of mapping the current status of ML studies, focusing on this field and identifying the research gap to facilitate future research. Regarding persistent mental health and problematic substance use concerns and disparities among the LGBTQ2S+ population, the findings from this review will provide useful insights to inform research and programs.

Methods

Objectives and Methodology Framework

This review aims to conduct a comprehensive search of studies using ML to investigate mental health or substance use among LGBTQ2S+ communities and to determine the scope of future research. We used the following 5-stage methodological framework developed by Arksey and O'Malley [30]: (1) identifying specific research questions; (2) identifying relevant studies through a comprehensive search of different sources; (3) study selection by applying inclusion and exclusion criteria; (4) data charting using custom-made data extraction forms; and (5) collating, summarizing, and reporting the results. We also used an extension of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for scoping reviews [31] to present our findings, and the Joana Briggs Institute proposed methodology of scoping reviews [32] to narrate the implications for future research. The review protocol was registered on the Open Science Framework [33] on December 17, 2020, to facilitate transparency and reproducibility of the study.

Identifying Research Questions

Initially, we identified a broad set of preliminary questions for this scoping review:

- What is the volume of the literature that used machine learning analysis in the field of mental health and substance use among the LGBTQ2S+ population?

- What are the fields of mental health and substance use among the LGBTQ2S+ population that have been studied by machine learning?
- Which subgroups of the LGBTQ2S+ population have been investigated? Are there any specific subgroups that have been studied using machine learning analysis?
- What types of machine learning methods (eg, supervised, unsupervised, semisupervised, and deep learning) and algorithms (eg, decision trees, random forest, logistic regression, and penalized regression) have been used to study LGBTQ2S+ mental health and substance use?
- What are the real-world implications of these studies? Are there any knowledge gaps or untouched domains that should be addressed in future research?

Identifying Relevant Studies

To gather a large quantity of relevant literature, we followed previous review studies with similar objectives [27,29] and searched the following databases: MEDLINE (Ovid), Embase (Ovid), CINAHL Plus, APA PsycINFO (Ovid), PubMed, and IEEE Xplore. We also searched the Summon (ProQuest) database used by the University of Toronto Libraries, which searches across many other databases, journal packages, e-book collections, and other resources. Information technology databases such as IEEE Xplore were selected as a potential source of ML-related literature. Literature searches involved a combination of keywords (eg, *mental health*, *mental disease, mental health service, substance abuse, ML, sexual and gender minorities, LGBT, lesbian, gay, men who have sex with men, bisexual, queer, two-spirit, intersex, and transgender*) and medical subject headings, if applicable. A librarian was consulted regarding the keywords and search terms.

Two reviewers (AK and RB) conducted the database search from November 25 to December 13, 2020, and imported all citations to the Covidence web platform, where duplicate papers were removed automatically. The databases were searched from the date of inception of the databases to the year 2020, with no filter in place for publication year. The bibliography lists of the included studies and review papers were reviewed on December 13, 2020, to identify any potential studies. The full Embase search strategy, representing an example of the search query applied to all other databases, is presented in Multimedia Appendix 1.

Study Selection

We included studies that used ML to investigate mental health or substance use behaviors of people within the LGBTQ2S+ population. Studies in which ML was used partially, but not for the main statistical analysis, were included in the review. We only included empirical investigations, thereby excluding editorials, opinion pieces, and reviews. We also excluded papers that used logistic regression analyses, not as a ML algorithm, but to determine LGBTQ2S+ identity status. In addition, studies in which full texts could not be retrieved with institutional license, and studies of genomics, pharmacokinetics, and those that were not directly relevant to humans were excluded.

Two reviewers (AK and RB) independently screened each title and abstract based on the eligibility criteria and completed full-text screening of the remaining studies. Disagreements were resolved through discussions among the 3 reviewers (AK, RB, and MC) to yield a list of final included studies.

Data Charting

To facilitate data charting and reporting, individual reviewers (AK and RB) first reviewed all studies and extracted key phrases and concepts from each study. We based our data extraction items on features identified in a recent biomedical guideline for reporting ML studies [34]. Custom-made data extraction forms were developed from this guideline, which included major extraction categories such as general study characteristics (ie, author, year, country, target population, source of data, sample size, field of study, ML domains, ML methods, algorithms, and outcomes), key components of model development (ie, whether the studies discussed methods of feature selection, resampling, model performance metrics, and method of validation), and discussion of study findings (ie, importance ranking of features, intersectionality, and other procedures or features applied).

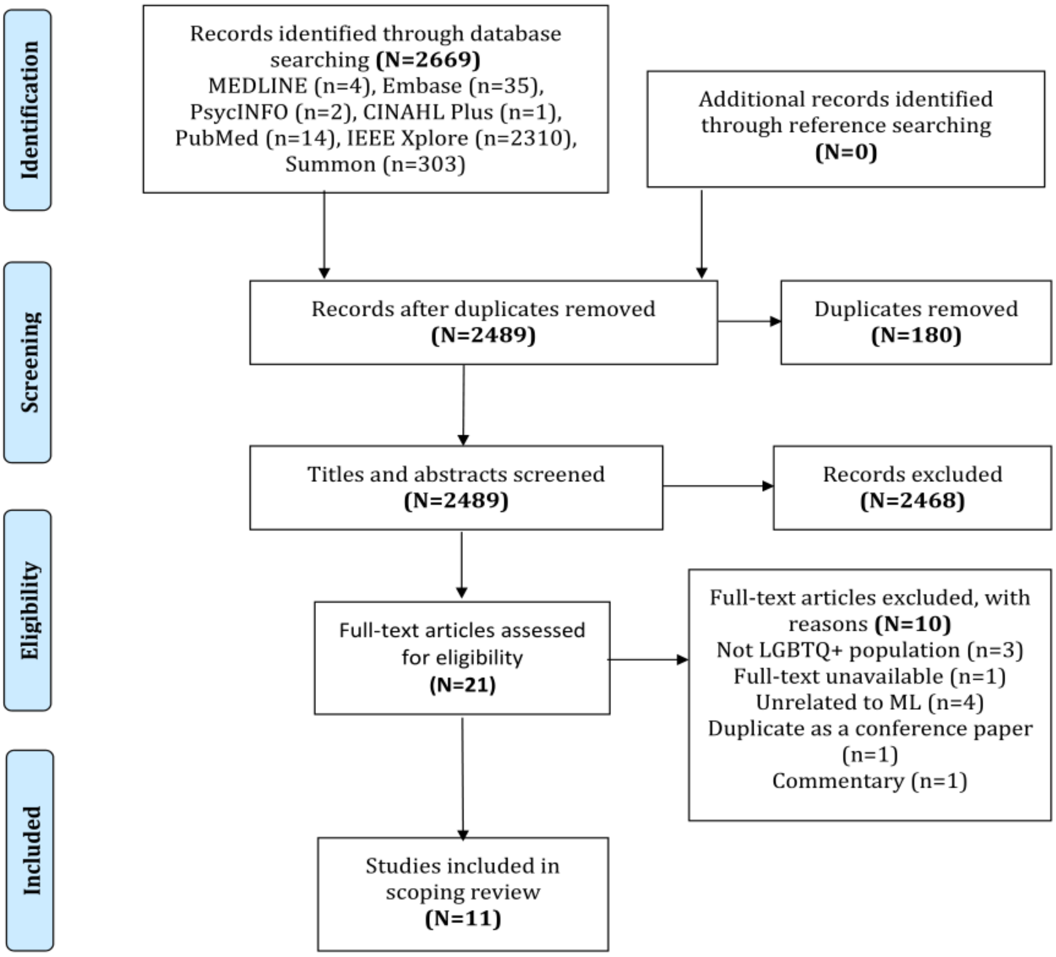
Collating, Summarizing and Reporting Results

We presented descriptive statistics for the extracted data sets by calculating the total number and percentage of all studies in each category. To provide a visual overview of the range of data, we presented a bar chart that showed the frequency analysis of studies according to the field of study and a pie chart that demonstrated the proportion of studies in the major domains of ML. We used a narrative synthesis approach [35] to describe the findings of the studies in the different ML domains and explored relationships in the data. Finally, we discussed research gaps to facilitate future research.

Results

The initial search of databases yielded 2669 articles, of which 2489 were retrieved after removing duplicates. We also searched the reference lists of potentially eligible articles and previous reviews but could not identify any studies that matched our inclusion criteria. After title and abstract screening, 21 articles were selected for full-text screening. Of these, we excluded articles that did not meet the target population criteria of the LGBTQ2S+ population (3/21, 14%), full-texts could not be retrieved (1/21, 4%), unrelated to ML (4/21, 19%), duplicate article published in a conference proceeding (1/21, 4%), and a commentary (1/21, 4%). This resulted in 11 studies being included in the final review [36-46]. The detailed selection process of the articles is presented in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram (Figure 1).

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram documenting study exclusion. LGBTQ+: lesbian, gay, bisexual, transgender, queer, or questioning; ML: machine learning.


Identification
Screening
Eligibility
Included
Records identified through database searching (N=2669)
MEDLINE (n=4), Embase (n=35), PsycINFO (n=2), CINAHL Plus (n=1), PubMed (n=14), IEEE Xplore (n=2310), Summon (n=303)
Additional records identified through reference searching (N=0)
Records after duplicates removed (N=2489)
Duplicates removed (N=180)
Titles and abstracts screened (N=2489)
Records excluded (N=2468)
Full-text articles assessed for eligibility (N=21)
Full-text articles excluded, with reasons (N=10)
Not LGBTQ+ population (n=3)
Full-text unavailable (n=1)
Unrelated to ML (n=4)
Duplicate as a conference paper (n=1)
Commentary (n=1)
Studies included in scoping review (N=11)

Study Characteristics

All 11 included studies [36-46] were published within the last 2 years (Table 1). Most of the studies were carried out in the United States (7/11, 63%) [36,38,39,41-43,45]. Among the target population categories that were not mutually exclusive, sexual minority men (gay, men who have sex with men, bisexual) were the most commonly studied (5/11, 45%) subgroups [37,40,42-44], followed by transgender (3/11, 27%) [39,45,46] and LGBTQ+ (3/11, 27%) [36,38,41] people at large, whereas sexual minority women (lesbian and bisexual) (2/11, 18%) [43,45] were the least commonly represented populations. None of the studies included two-spirit persons as their target population (Table 1).

Table 1. Summary statistics of included studies (N=11) [36-46].[a]

Characteristics	Number of studies, n (%)
Countries	
United States	7 (63)
China	2 (18)
Sweden	1 (9)
Australia	1 (9)
Years published	
2019	5 (45)
2020	6 (54)
Field of study	
Mental health (n=9)	
Suicide or self-injury	2 (18)
Depression	2 (18)
Mood or affect processes	3 (27)
Minority stress	1 (9)
Gender incongruence	1 (9)
Substance use (n=2)	
Tobacco	1 (9)
Poppers or alkyl nitrites	1 (9)
Target population[b]	
Sexual minorities: male (gay, MSM[c], bisexual)	5 (45)
Sexual minorities: female (lesbian, bisexual)	2 (18)
Transgender or gender minorities	3 (27)
LGBT/LGBTQ+[d]	3 (27)
Domains of ML[e]	
Web content analysis	6 (55)
Prediction modeling	4 (36)
Imaging study	1 (9)
Type of ML	
Supervised	9 (82)
Unsupervised	3 (27)
Deep	1 (9)
ML algorithms	
LDA[f]	3 (27)
RF[g]	2 (18)
SVM[h]	2 (18)
CNN[i]	1 (9)
MLP[j]	1 (9)
NB[k]	1 (9)
Penalized regression (LASSO[l], elastic net regularized regression, ridge regression)	2 (18)
Logistic regression	1 (9)

Characteristics	Number of studies, n (%)
Boosting (XGBoost[m], AdaBoost[n], GBM[o])	3 (27)
Classification tree	2 (18)
Feature selection	
Yes	7 (64)
No	4 (36)
Discussed model performance	
Used performance metrics	9 (82)
Didn't use performance metrics	1 (9)
Didn't discuss performance	1 (9)
Method of validation	
Hold-out	2 (18)
Cross-validation	7 (64)
External validation	2 (18)
Unspecified	4 (36)

[a] Multiple response options were possible for some study characteristics.

[b] Categories are not mutually exclusive.

[c] MSM: men who have sex with men.

[d] LGBT/LGBTQ+: lesbian, gay, bisexual, and transgender/lesbian, gay, bisexual, transgender, queer, or questioning.

[e] ML: machine learning.

[f] LDA: latent Dirichlet allocation.

[g] RF: random forest.

[h] SVM: support vector machine.

[i] CNN: convolutional neural network.

[j] MLP: multilayered perceptron.

[k] NB: Naive Bayes.

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

[m] XGBoost: eXtreme Gradient Boosting.

[n] AdaBoost: Adaptive Boosting.

[o] GBM: Generalized Boosted Model.

Most of the studies focused on mental health (9/11, 82%) [36-42,45,46], and only 18% (2/11) studies [43,44] focused on substance use concerns. Most studies examined several mental health issues, such as depression, suicide, mood or affect processes, minority stress, and gender incongruence [36-42,45,46], whereas other studies that focused on substance use only examined tobacco and poppers or alkyl nitrites use [43,44]. No study looked into mental health issues and substance use concerns among the LGBTQ2S+ population simultaneously (Table 1).

The studies were categorized into 3 major ML domains: web content analysis, prediction modeling, and imaging study. Over half of the studies (6/11, 55%) were identified as web content analysis [36-41], and 36% (4/11) were identified as prediction modeling [42-45]; 1 study (9%) was identified as an imaging study [46] (Table 1).

The most commonly used class of ML methods was supervised (9/11, 82%) [37-39,41-46], followed by unsupervised (3/11, 27%) [36,37,40] and deep learning (1/11, 9%; Table 1) [41]. The most frequently used ML algorithms were latent Dirichlet allocation (3/11, 27%) and boosting (3/11, 27%), followed by random forest, support vector machines, penalized regression (ie, least absolute shrinkage and selection operator, elastic net regularized regression, and ridge regression), classification tree, logistic regression, naive Bayes, multilayered perceptron, and convolutional neural network (Table 1).

Approximately two-thirds (7/11, 64%) of the studies [37,38,42-46] discussed their methods of feature selection, among which the median number of features used was 19. Most of the studies used cross-validation methods (7/11, 64%) [37-39,41,44-46], especially 10-fold cross-validation. Furthermore, 18% (2/11) of the articles used the hold out method [39,41], 18% (2/11) used external validation [37,41], and 36% (4/11) articles [36,40,42,43] did not report how they validated their method. Most studies (9/11, 82%) [36-39,41-43,45,46] used at least one performance metric (eg, area under ROC curve, precision-recall, or F1 score) to discuss model performance. However, the remaining studies either did not use any performance metric [44] or did not discuss any model performance [40] (Table 1).

Machine Learning Domains

Multimedia Appendix 2 summarizes the characteristics of the final 11 included studies [36-46] and Multimedia Appendix 3 [36-46] presents the ML methodology used in the studies.

The 54% (6/11) studies [36-41] in the web content analysis domain obtained their data from social media sources such as Twitter, Blued, Tumblr, Reddit, and LGBT Chat and Forums. The volume of data used ranged from 12,000 to 41 million web posts. Half of the studies used their data to analyze the mood or affect processes of the users related to their sexual and gender identities [39-41] (Multimedia Appendix 2).

Among the 4 studies in the prediction modeling domain, 50% (2/4) of the studies analyzed data on adult participants [42,44] and 50% (2/4) on adolescents [43,45]. Only 1 study used a public health data set of 28,811 participants [43]; other studies used either cross-sectional or cohort data from longitudinal studies [42,44,45]. Half of the studies focused on mental health (depression and suicide) [42,45] and half on substance use behavior (cigarette, e-cigarette, and poppers use) [43,44] (Multimedia Appendix 2). Of the 4 studies, only 25% (1/4) study [45] ranked their feature importance, and 50% (2/4) studies [42,45] examined intersectionalities (Multimedia Appendix 3). One study investigated the intersection of income and other social and environmental stressors with racial or ethnic disparities and its impact on depressive symptomology among men who have sex with men [42], whereas the other focused on the intersection between various social and behavioral determinants of health (self-image, race, education, socioeconomic status, family support, friends, stigma, discrimination, etc) as risk factors of self-injurious behaviors among sexual and gender minority women [45].

One imaging trial study used clinical and functional magnetic resonance imaging data of 25 transgender adults to identify the relationship between pretherapy functional brain connectivity and posthormone therapy body congruence [46]. All 4 studies [42-45] of the prediction modeling domain and 1 imaging study [46] used the supervised method of ML, whereas studies in the web content analysis domain [36-41] used supervised (4/11, 36%), unsupervised (3/11, 27%), and deep learning (1/11, 9%) methods (Multimedia Appendix 3).

Discussion

Principal Findings

Our results show that the application of ML to assess mental health and substance use behavior among the LGBTQ2S+ population is still new in health research, compared with the increasing use of ML techniques in other health research domains. Although there is continued criminalization and lack of LGBTQ2S+ rights protection in 67 United Nations member states at the end of 2020 [47], there appears to be an increasing acceptance of sexual and gender minority people in diverse contexts such as in North American countries and Western Europe [48]. However, very few of the included studies were conducted outside the United States (Table 1).

Only a few mental health problems were addressed across the few relevant ML studies conducted to date (Table 1). Although there is evidence of a higher prevalence of anxiety disorders, posttraumatic stress disorder, and various mood disorders (eg, mania and persistent depressive disorder) among the LGBTQ2S+ population compared with cisgender and heterosexual counterparts [4], no studies have been conducted on these issues. Compared with mental health issues, substance use problems among the LGBTQ2S+ population were almost untouched. Moreover, both of the included substance use related studies predicted the present use of substances [43,44], and no studies have examined future substance use, cessation, or substance use treatment-seeking behavior.

Underlying factors behind the low number of ML studies on mental health and substance use issues among the LGBTQ2S+ population may be sex and gender identity-related data invisibility and social and institutional bias [21,49]. Electronic health records have been used as a common and promising data source for ML techniques to predict population health in other research areas [27,29]. However, binary representation of sex and gender (ie, man or woman) in the electronic health records system makes some data unavailable for analysis by ML, which can underrepresent the actual problem [21,50,51]. Adopting inclusive gender, sex, and sexual orientation (GSSO) information practices, collecting sexual and gender diversity, has the potential to ensure data justice, alleviate unintentional bias, and reduce health inequity [49]. A good example of inclusive GSSO information practice could be the proposed equity stratifiers by the Canadian Institute of Health Information [52]. However, other potential data sources of ML applications, such as social media, cross-sectional survey data, longitudinal cohort, and administrative data sets were used in the included studies (Multimedia Appendix 2).

Most studies were in the web content analysis domain, indicating social media to be a potentially useful epidemiological resource for collecting data on LGBTQ2S+ people and analyzing the data using ML (Multimedia Appendix 2). We observed that unsupervised ML has also been applied in these studies with data drawn from social media [36,37,40], thus holding the potential to support qualitative research by handling large textual data sets with its computational power. This is particularly useful in LGBTQ2S+ health research, given the stigma-related and structural barriers toward identity disclosure that may inhibit data collection through other methodologies [50,51,53,54]. The use of ML in these studies has shown potential for automated identification of at-risk individuals for crisis suicide prevention and intervention [36], depressive emotions [37], minority stressors [38], negative emotions [40], and mental health signals [41] among the LGBTQ2S+ community. In addition, the sequence of transgender identity disclosure identified in a study by Haimson et al [39] may guide resource allocation and provide support through gender transition. However, self-reported mental health problems on social media might not reflect clinical diagnoses or symptomologies.

Although there is evidence of the influence of intersections of various social and behavioral determinants of health on the increased prevalence of mental health and substance use concerns among the LGBTQ2S+ population [11-16], only 2 studies examined the intersection of sexual and gender identity with ethno-racial identities, and several social, economic, and behavioral factors (ie, income, social stigma, discrimination, and family support), and their impact on depression and self-injurious behaviors [42,45]. No such studies in our review explored intersectionality in the field of substance use. Identifying these intersections by leveraging ML techniques would have practical implications by determining risk and protective factors as well as informing strategies for promoting mental well-being and substance use prevention and intervention with and for LGBTQ2S+ people. In the context of various techniques used in intersectional research, both qualitative and quantitative, and recent trends in mixed methods research [55], ML can be a very useful tool for processing vast quantities of data, data mining and clustering, and classifying attribute relationships [56,57]. Apart from the partial dependency-based measures, newer techniques and methods [58,59] in ML have emerged for analyzing interaction effects and are more suitable for assessing intersectionality.

Following the current guidelines for reporting ML studies in biomedical research [34], we documented a range of explanatory findings seen in the included studies and found that most studies mentioned their performance metrics, method of feature selection, and method of validation of their model (Table 1 and Multimedia Appendix 3). However, only 27% (3/11) studies [37,38,45] adopted the approach of approximating a relative importance score of individual features that reflected their overall contributions to the model (Multimedia Appendix 3). The implications of providing an importance score to features are particularly valuable for predictive modeling studies, where the most important predictors are targeted for future strategy adoption. Another notable finding was about half (n=2) [42,43] of the predictive modeling studies did not report any method of validation, and none of them conducted external validation of the resulting model on a different population (Multimedia Appendix 3). Validation is an important aspect of the predictive modeling process, which increases the reproducibility and generalizability of the model [60]. Hence, future studies in this domain should follow existing guidelines to validate their models [34]. Moreover, half of the predictive modeling studies had small sample sizes (<1000) (Multimedia Appendix 2). Small data sets can affect the model performance [61]. Using large population-based data sets for future research can overcome this problem and fully leverage the benefits of ML.

Compared with the other 2 domains, there was a significant gap in ML research using imaging data (ie, functional magnetic resonance imaging or electroencephalography) to examine mental health and substance use among the LGBTQ2S+ population (Table 1). Although a single identified imaging study [46] predicted cross-sex hormonal therapy responsiveness in the transgender population, which is useful for guiding and selecting candidates for therapy, the sample size was small, limiting the generalizability of the findings.

Future Research Directions

We detected significant research gaps in ML applications for mental health and substance use research among the LGBTQ2S+ population. First, future research should investigate other mental health issues (ie, anxiety disorders and mood disorders) and substance use behavior and problems (ie, alcohol, opioids, and illicit drugs) among the LGBTQ2S+ population. Second, the potential of ML applications in predicting substance use related outcomes (ie, cessation, overdose events, routes of administration, driving impairments, and other adverse reactions), mental health service access, and mental health-related outcomes (ie, disabilities, symptom management, suicide and suicide attempts, economic burden, and health care costs) should be explored.

Third, further research is needed on sexual minority women. The small number of studies included (Table 1) did not allow exploration of shared and different health needs and priorities between and within the LGBTQ2S+ population. Fourth, as the legal and societal context in which the LGBTQ2S+ population lives differ significantly between countries [48], more research should be conducted in countries outside the United States. Fifth, specific research initiatives targeted at investigating the intersection of sexual and gender minority identity with other social determinants of health (ie, race, ethnicity, citizenship, socioeconomic status, and housing condition) are necessary to better understand their potential for fostering risk and resilience regarding mental health and substance use. Finally, different data sources should be used in ML studies. Large-population-level administrative data sets should be used for prediction modeling studies for the accurate application of ML models. In addition, with the advancement of technology, the digitalization of health care, and where LGBTQ2S+ status is captured in electronic health records, these health records can be a potential data resource for ML studies with real-world clinical implications for LGBTQ2S+ people.

Strength and Limitations

To the best of our knowledge, our review is the first of its kind to explore the use of ML applications in examining mental health and substance use among LGBTQ2S+ populations. We adopted a comprehensive search strategy, including searching various multidisciplinary peer-reviewed databases to identify relevant articles as much as possible. The findings of our review need to be interpreted with consideration of one key limitation. Owing to the small number of studies, highly heterogeneous characteristics of the included studies, and inconsistent reporting of model development and validation, we could not perform a critical appraisal of the studies and therefore could not comment significantly on the overall performance of the ML techniques. However, we followed the approaches of previous scoping reviews with similar objectives [27,29] and were interested in understanding the general topics or areas being investigated by ML in the field of mental health and substance use among the LGBTQ2S+ population (ie, most commonly used data sources, study countries, and study populations) and identifying research gaps to inform future research.

As more studies are published on this research topic in the future, a systematic review with critical appraisal of relevant literatures should be conducted as the next step in research. Researchers are attempting to expand established reporting guidelines to include items that accommodate ML studies, such as the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis statement specific for M [62], the Artificial Intelligence extension for Consolidated

Standards of Reporting Trials [63], and Artificial Intelligence extension for Standard Protocol Items: Recommendations for Interventional Trials [63] guidelines. Once developed, these guidelines can be used as critical appraisal tools for studies that adopt ML-based data analysis. There is also an opportunity to incorporate fairness and equity considerations in the development of appraisal tools for ML studies. Preliminary research has already developed mathematical metrics to measure the fairness of a ML algorithm, and if intersectionalities are met in the models [64].

Conclusions

Although there is an exponential growth of ML applications in other health research sectors, few studies have used these techniques in the field of mental health and substance use among the LGBTQ2S+ population. In addition to undertaking more research, future researchers should focus on applying ML algorithms with considerations for real-world implications through public health interventions and adopting policies that aim to improve health equity.

Acknowledgments

The authors would like to thank Elena Springall, a librarian at the Gerstein Science Information Centre, University of Toronto, for her support in reviewing the database search strategies. The study was funded by the Canadian Institutes of Health Research, grant number 1000993. The funder had no role in the study design, collection, analysis, or interpretation of the data, writing the manuscript, or the decision to submit the paper for publication.

Authors' Contributions

MC contributed to the study design and obtained funding and supervision. AK and RB conducted the database search, article screening, and data extraction. AK conducted the data analysis and primary drafting of the manuscript. All authors, AK, MC, RB, DG, RF, CHL, BB, CY, NM, and RS, contributed to the conceptualization, drafting, review, and approval of the manuscript for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Embase search query.
[DOCX File , 13 KB - medinform_v9i11e28962_app1.docx]

Multimedia Appendix 2
Summary of studies using machine learning analysis in mental health and substance use among lesbian, gay, bisexual, transgender, queer or questioning, and two-spirit population (N=11).
[DOCX File , 17 KB - medinform_v9i11e28962_app2.docx]

Multimedia Appendix 3
Summary of characteristics of machine learning methods used (N=11).
[DOCX File , 18 KB - medinform_v9i11e28962_app3.docx]

References

1. Marshal MP, Friedman MS, Stall R, King KM, Miles J, Gold MA, et al. Sexual orientation and adolescent substance use: a meta-analysis and methodological review. Addiction 2008 Apr;103(4):546-556 [FREE Full text] [doi: 10.1111/j.1360-0443.2008.02149.x] [Medline: 18339100]
2. King M, Semlyen J, Tai SS, Killaspy H, Osborn D, Popelyuk D, et al. A systematic review of mental disorder, suicide, and deliberate self harm in lesbian, gay and bisexual people. BMC Psychiatry 2008 Aug 18;8:70 [FREE Full text] [doi: 10.1186/1471-244X-8-70] [Medline: 18706118]
3. Marshal MP, Dietz LJ, Friedman MS, Stall R, Smith HA, McGinley J, et al. Suicidality and depression disparities between sexual minority and heterosexual youth: a meta-analytic review. J Adolesc Health 2011 Aug;49(2):115-123 [FREE Full text] [doi: 10.1016/j.jadohealth.2011.02.005] [Medline: 21783042]
4. Institute of Medicine. The Health of Lesbian, Gay, Bisexual, and Transgender People: Building a Foundation for Better Understanding. Washington (DC): National Academies Press (US); 2011.
5. National survey on LGBTQ youth mental health. The Trevor Project. 2019. URL: https://www.thetrevorproject.org/wp-content/uploads/2019/06/The-Trevor-Project-National-Survey-Results-2019.pdf [accessed 2021-10-19]
6. Medley G, Lipari R, Bose J, Cribb D, Kroutil L. Sexual orientation and estimates of adult substance use and mental health: results from the 2015 national survey on drug use and health. National Survey on Drug Use and Health. 2016. URL: https:/

/www.samhsa.gov/data/sites/default/files/NSDUH-SexualOrientation-2015/NSDUH-SexualOrientation-2015/NSDUH-SexualOrientation-2015.htm [accessed 2021-10-19]

7. Abramovich A, de Oliveira C, Kiran T, Iwajomo T, Ross LE, Kurdyak P. Assessment of health conditions and health service use among transgender patients in Canada. JAMA Netw Open 2020 Aug 03;3(8):e2015036 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.15036] [Medline: 32857149]
8. Wilson C, Cariola L. LGBTQI+ youth and mental health: a systematic review of qualitative research. Adolescent Res Rev 2019 May 21;5(2):187-211 [FREE Full text] [doi: 10.1007/s40894-019-00118-w]
9. Logie C. The case for the World Health Organization's commission on the social determinants of health to address sexual orientation. Am J Public Health 2012 Jul;102(7):1243-1246. [doi: 10.2105/AJPH.2011.300599] [Medline: 22594723]
10. Pega F, Veale JF. The case for the World Health Organization's Commission on Social Determinants of Health to address gender identity. Am J Public Health 2015 Mar;105(3):58-62. [doi: 10.2105/AJPH.2014.302373] [Medline: 25602894]
11. Meyer IH. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. Psychol Bull 2003 Sep;129(5):674-697 [FREE Full text] [doi: 10.1037/0033-2909.129.5.674] [Medline: 12956539]
12. Ryan C, Huebner D, Diaz RM, Sanchez J. Family rejection as a predictor of negative health outcomes in white and Latino lesbian, gay, and bisexual young adults. Pediatrics 2009 Jan;123(1):346-352. [doi: 10.1542/peds.2007-3524] [Medline: 19117902]
13. Burns MN, Ryan DT, Garofalo R, Newcomb ME, Mustanski B. Mental health disorders in young urban sexual minority men. J Adolesc Health 2015 Jan;56(1):52-58 [FREE Full text] [doi: 10.1016/j.jadohealth.2014.07.018] [Medline: 25294230]
14. Duncan DT, Hatzenbuehler ML. Lesbian, gay, bisexual, and transgender hate crimes and suicidality among a population-based sample of sexual-minority adolescents in Boston. Am J Public Health 2014 Feb;104(2):272-278. [doi: 10.2105/AJPH.2013.301424] [Medline: 24328619]
15. Kosciw J, Greytak E, Palmer N, Boesen M. The 2013 National School Climate Survey: the experiences of lesbian, gay, bisexual and transgender youth in our nation's schools. GLSEN. 2014. URL: https://www.glsen.org/research/2013-national-school-climate-survey [accessed 2021-10-19]
16. Choi S, Wilson B, Shelton J, Gates G. Serving our youth 2015: the needs and experiences of lesbian, gay, bisexual, transgender, and questioning youth experiencing homelessness. The Williams Institute with True Colors Fund. 2015. URL: https://escholarship.org/uc/item/1pd9886n [accessed 2021-10-19]
17. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015 Jul 17;349(6245):255-260. [doi: 10.1126/science.aaa8415] [Medline: 26185243]
18. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. Biomed Inform Insights 2016;8:1-10 [FREE Full text] [doi: 10.4137/BII.S31559] [Medline: 26843812]
19. Let's discuss stigma and discrimination around mental health and substance use problems. Canadian Mental Health Association, British Columbia Division. 2014. URL: https://www.heretohelp.bc.ca/sites/default/files/stigma-and-discrimination-around-mental-health-and-substance-use-problems.pdf [accessed 2021-10-19]
20. Committee on the Science of Changing Behavioral Health Social Norms, Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education, National Academies of Sciences, Engineering, and Medicine. Ending Discrimination Against People with Mental and Substance Use Disorders: The Evidence for Stigma Change. Washington (DC): National Academies Press (US); 2016:1-170.
21. Ruberg B, Ruelos S. Data for queer lives: how LGBTQ gender and sexuality identities challenge norms of demographics. Big Data Soc 2020 Jun 18;7(1):205395172093328. [doi: 10.1177/2053951720933286]
22. Naqa IE, Murphy MJ. What is machine learning? In: Naqa IE, Li R, Murphy MJ, editors. Machine Learning in Radiation Oncology. Cham: Springer; 2015:3-11.
23. Mak KK, Lee K, Park C. Applications of machine learning in addiction studies: a systematic review. Psychiatry Res 2019 May;275:53-60. [doi: 10.1016/j.psychres.2019.03.001] [Medline: 30878857]
24. Zhu X, Goldberg AB. Introduction to semi-supervised learning. Synth Lect Artif Intell Mach Learn 2009 Jan;3(1):1-130 [FREE Full text] [doi: 10.2200/s00196ed1v01y200906aim006]
25. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
26. Thieme A, Belgrave D, Doherty G. Machine learning in mental health: a systematic review of the HCI literature to support the development of effective and implementable ML systems. ACM Trans Comput-Hum Interact 2020 Oct 05;27(5):1-53 [FREE Full text] [doi: 10.1145/3398069]
27. Shatte AB, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med 2019 Jul;49(9):1426-1448. [doi: 10.1017/S0033291719000151] [Medline: 30744717]
28. Bernert R, Hilberg A, Melia R, Kim J, Shah N, Abnousi F. Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. Int J Environ Res Public Health 2020 Aug 15;17(16):5929 [FREE Full text] [doi: 10.3390/ijerph17165929] [Medline: 32824149]

29. Morgenstern JD, Buajitti E, O'Neill M, Piggott T, Goel V, Fridman D, et al. Predicting population health with machine learning: a scoping review. BMJ Open 2020 Oct 27;10(10):e037860 [FREE Full text] [doi: 10.1136/bmjopen-2020-037860] [Medline: 33109649]
30. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol Theory Pract 2005 Feb;8(1):19-32 [FREE Full text] [doi: 10.1080/1364557032000119616]
31. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
32. Peters M, Godfrey C, McInerney P, Soares C, Khalil H, Parker D. The Joanna Briggs Institute Reviewers’ Manual 2015: methodology for JBI scoping reviews. Joanna Briggs Institute. 2015. URL: https://nursing.lsuhsc.edu/JBI/docs/ReviewersManuals/Scoping-.pdf [accessed 2021-10-19]
33. Kundu A, Billington R, Chaiton M. Machine learning applications in mental health and substance use research among LGBTQ2S+ population: protocol for a scoping review. Open Sci Framework 2020:A. [doi: 10.17605/OSF.IO/TMPV3]
34. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
35. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews: a product from the ESRC Methods Programme - Version 1. Peninsula Medical School, Universities of Exeter and Plymouth. 2006. URL: https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/fhm/dhr/chir/NSsynthesisguidanceVersion1-April2006.pdf [accessed 2021-10-19]
36. Liang C, Abbott D, Hong Y, Madadi M, White A. Clustering help-seeking behaviors in LGBT online communities: a prospective trial. In: Meiselwitz G, editor. Social Computing and Social Media. Design, Human Behavior and Analytics. Cham: Springer; 2019:345-355.
37. Li Y, Cai M, Qin S, Lu X. Depressive emotion detection and behavior analysis of men who have sex with men social media. Front Psychiatry 2020;11:830 [FREE Full text] [doi: 10.3389/fpsyt.2020.00830] [Medline: 32922323]
38. Saha K, Kim SC, Reddy MD, Carter AJ, Sharma E, Haimson OL, et al. The language of LGBTQ+ minority stress experiences on social media. Proc ACM Hum Comput Interact 2019 Nov;3(CSCW):89 [FREE Full text] [doi: 10.1145/3361108] [Medline: 32935081]
39. Haimson OL, Veinot TC. Coming out to doctors, coming out to "Everyone": understanding the average sequence of transgender identity disclosures using social media data. Transgend Health 2020;5(3):158-165 [FREE Full text] [doi: 10.1089/trgh.2019.0045] [Medline: 32923666]
40. Huang G, Cai M, Lu X. Inferring opinions and behavioral characteristics of gay men with large scale multilingual text from blued. Int J Environ Res Public Health 2019 Sep 26;16(19):3597 [FREE Full text] [doi: 10.3390/ijerph16193597] [Medline: 31561423]
41. Zhao Y, Guo Y, He X, Wu Y, Yang X, Prosperi M, et al. Assessing mental health signals among sexual and gender minorities using Twitter data. Health Informatics J 2020 Jun;26(2):765-786 [FREE Full text] [doi: 10.1177/1460458219839621] [Medline: 30969146]
42. Barrett B, Abraham A, Dean L, Plankey M, Friedman M, Jacobson L, et al. Social inequalities contribute to racial/ethnic disparities in depressive symptomology among men who have sex with men. Soc Psychiatry Psychiatr Epidemiol 2021 Feb;56(2):259-272 [FREE Full text] [doi: 10.1007/s00127-020-01940-7] [Medline: 32780176]
43. Azagba S, Latham K, Shan L. Cigarette smoking, e-cigarette use, and sexual identity among high school students in the USA. Eur J Pediatr 2019 Sep;178(9):1343-1351. [doi: 10.1007/s00431-019-03420-w] [Medline: 31292730]
44. Demant D, Oviedo-Trespalacios O. Harmless? A hierarchical analysis of poppers use correlates among young gay and bisexual men. Drug Alcohol Rev 2019 Jul;38(5):465-472. [doi: 10.1111/dar.12958] [Medline: 31209963]
45. Smith DM, Wang SB, Carter ML, Fox KR, Hooley JM. Longitudinal predictors of self-injurious thoughts and behaviors in sexual and gender minority adolescents. J Abnorm Psychol 2020 Jan;129(1):114-121. [doi: 10.1037/abn0000483] [Medline: 31657599]
46. Moody T, Feusner J, Reggente N, Vanhoecke J, Holmberg M, Manzouri A, et al. Predicting outcomes of cross-sex hormone therapy in transgender individuals with gender incongruence based on pre-therapy resting-state brain connectivity. Neuroimage Clin 2021;29:102517 [FREE Full text] [doi: 10.1016/j.nicl.2020.102517] [Medline: 33340976]
47. Mendos L, Botha K, Lelis R, Tan D, de la Peña E, Savelev I, et al. State-sponsored homophobia : global legislation overview update. ILGA, Geneva. 2020. URL: https://ilga.org/downloads/ILGA_World_State_Sponsored_Homophobia_report_global_legislation_overview_update_December_2020.pdf [accessed 2021-10-19]
48. Poushter J, Kent N. The global divide on homosexuality persists, but increasing acceptance in many countries over past two decades. Pew Research Center. 2020. URL: https://www.pewresearch.org/global/wp-content/uploads/sites/2/2020/06/PG_2020.06.25_Global-Views-Homosexuality_FINAL.pdf [accessed 2021-10-19]

49. Davison K, Queen R, Lau F, Antonio M. Culturally competent gender, sex, and sexual orientation information practices and electronic health records: rapid review. JMIR Med Inform 2021 Feb 11;9(2):e25467 [FREE Full text] [doi: 10.2196/25467] [Medline: 33455901]
50. Sokkary N, Awad H, Paulo D. Frequency of sexual orientation and gender identity documentation after electronic medical record modification. J Pediatr Adolesc Gynecol 2021 Jun;34(3):324-327 [FREE Full text] [doi: 10.1016/j.jpag.2020.12.009] [Medline: 33333261]
51. Lau F, Antonio M, Davison K, Queen R, Bryski K. An environmental scan of sex and gender in electronic health records: analysis of public information sources. J Med Internet Res 2020 Nov 11;22(11):e20050 [FREE Full text] [doi: 10.2196/20050] [Medline: 33174858]
52. Canadian Institute for Health Information. In Pursuit of Health Equity: Defining Stratifiers for Measuring Health Inequality - A Focus on Age, Sex, Gender, Income, Education and Geographic Location. Ottawa, ON: CIHI; 2018.
53. Owen-Smith AA, Woodyatt C, Sineath RC, Hunkeler EM, Barnwell LT, Graham A, et al. Perceptions of barriers to and facilitators of participation in health research among transgender people. Transgend Health 2016;1(1):187-196 [FREE Full text] [doi: 10.1089/trgh.2016.0023] [Medline: 28861532]
54. Lucassen M, Fleming T, Merry S. Tips for research recruitment: the views of sexual minority youth. J LGBT Youth 2017 Jan 13;14(1):16-30 [FREE Full text] [doi: 10.1080/19361653.2016.1256246]
55. Hankivsky O, Grace D. Understanding and emphasizing difference and intersectionality in multimethod and mixed methods research. In: Hesse-Biber SN, Johnson RB, editors. The Oxford Handbook of Multimethod and Mixed Methods Research Inquiry. Oxford, United Kingdom: Oxford University Press; 2015.
56. Pastrana JL, Reigal RE, Morales-Sánchez V, Morillo-Baro JP, de Mier RJ, Alves J, et al. Data mining in the mixed methods: application to the study of the psychological profiles of athletes. Front Psychol 2019;10:2675 [FREE Full text] [doi: 10.3389/fpsyg.2019.02675] [Medline: 31866896]
57. Leavitt A. Human-centered data science: mixed methods and intersecting evidence, inference, and scalability. Annenberg School for Communication & Journalism, University of Southern California. 2016. URL: https://cscw2016hcds.files.wordpress.com/2015/10/29_alexleavitt.pdf [accessed 2021-10-19]
58. Schiltz F, Masci C, Agasisti T, Horn D. Using regression tree ensembles to model interaction effects: a graphical approach. Appl Econ 2018 Jul 05;50(58):6341-6354 [FREE Full text] [doi: 10.1080/00036846.2018.1489520]
59. Oh S. Feature interaction in terms of prediction performance. Appl Sci 2019 Nov 29;9(23):5191 [FREE Full text] [doi: 10.3390/app9235191]
60. Han K, Song K, Choi BW. How to develop, validate, and compare clinical prediction models involving radiological parameters: study design and statistical methods. Korean J Radiol 2016;17(3):339-350 [FREE Full text] [doi: 10.3348/kjr.2016.17.3.339] [Medline: 27134523]
61. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014 Dec 22;14:137 [FREE Full text] [doi: 10.1186/1471-2288-14-137] [Medline: 25532820]
62. Collins GS, Moons KG. Reporting of artificial intelligence prediction models. Lancet 2019 Apr;393(10181):1577-1579. [doi: 10.1016/s0140-6736(19)30037-6]
63. CONSORT-AISPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nat Med 2019 Oct 24;25(10):1467-1468. [doi: 10.1038/s41591-019-0603-3] [Medline: 31551578]
64. Foulds J, Islam R, Keya K, Pan S. An intersectional definition of fairness. arXiv. 2019 Sep 10. URL: http://jfoulds.informationsystems.umbc.edu/papers/2020/Foulds%20(2020)%20-%20An%20Intersectional%20Definition%20of%20Fairness%20(ICDE).pdf [accessed 2019-10-19]

Abbreviations

LGBTQ2S+: lesbian, gay, bisexual, transgender, queer or questioning, and two-spirit
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by C Lovis; submitted 20.03.21; peer-reviewed by M Cai, K Davison; comments to author 10.07.21; revised version received 02.09.21; accepted 03.10.21; published 11.11.21.

Please cite as:
Kundu A, Chaiton M, Billington R, Grace D, Fu R, Logie C, Baskerville B, Yager C, Mitsakakis N, Schwartz R
Machine Learning Applications in Mental Health and Substance Use Research Among the LGBTQ2S+ Population: Scoping Review
JMIR Med Inform 2021;9(11):e28962
URL: https://medinform.jmir.org/2021/11/e28962
doi:10.2196/28962
PMID:34762059

Original Paper

The Role of Physicians in Digitalizing Health Care Provision: Web-Based Survey Study

Anja Burmann[1,2], MSc; Max Tischler[3,4]; Mira Faßbach[4,5]; Sophie Schneitler[4,6,7], Dr med; Sven Meister[1,2], Prof Dr

[1]Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany
[2]Witten/Herdecke University, Witten, Germany
[3]Hautärzte am Markt, Dortmund, Germany
[4]Bündnis Junge Ärzte, Berlin, Germany
[5]Helios Klinikum Duisburg, Duisburg, Germany
[6]Saarland University Hospital, Homburg, Germany
[7]German Society for Tropical Medicine, Travel Medicine and Global Health, Hamburg, Germany

Corresponding Author:
Anja Burmann, MSc
Fraunhofer Institute for Software and Systems Engineering
Emil-Figge-Str 91
Dortmund, 44227
Germany
Phone: 49 2319 7677435
Email: anja.burmann@isst.fraunhofer.de

Abstract

Background: Digitalization affects all areas of society, including the health care sector. However, the digitalization of health care provision is progressing slowly compared to other sectors. In the professional and political literature, physicians are partially portrayed as digitalization sceptics. Thus, the role of physicians in this process requires further investigation. The theory of “digital natives” suggests a lower hurdle for younger generations to engage with digital technologies.

Objective: The objective of this study was to investigate the role of physicians in the process of digitalizing health care provision in Germany and to assess the age factor.

Methods: We conducted a large-scale study to assess the role of this professional group in the progress of the digital transformation of the German health care sector. Therefore, in an anonymous online survey, we inquired about the current digital penetration of the personal working environment, expectations, attitude toward, and concerns regarding digitalization. Based on these data, we studied associations with the nominal variable age and variations across 2 age groups.

Results: The 1274 participants included in the study generally showed a high affinity towards digitalization with a mean of 3.88 on a 5-point Likert scale; 723 respondents (56.75%) stated they personally use mobile apps in their everyday working life, with a weak tendency to be associated with the respondents’ age (η=0.26). Participants saw the most noticeable existing benefits through digitalization in data quality and readability (882/1274, 69.23%) and the least in patient engagement (213/1274, 16.72%). Medical practitioners preponderantly expect further improvements through increased digitalization across almost all queried areas but the most in access to medical knowledge (1136/1274, 89.17%), treatment of orphan diseases (1016/1274, 79.75%), and medical research (1023/1274, 80.30%).

Conclusions: Respondents defined their role in the digitalization of health care provision as ambivalent: “scrutinizing” on the one hand but “active” and “open” on the other. A gap between willingness to participate and digital sovereignty was indicated. Thus, education on digitalization as a means to support health care provision should not only be included in the course of study but also in the continuing process of further and advanced training.

(JMIR Med Inform 2021;9(11):e31527) doi:10.2196/31527

KEYWORDS

digitalization; digital transformation; health care; human factor; physicians; digital natives; web-based survey; digital health

Introduction

Background

The theoretical description of digitalization in health care promises the potential to improve quality of care, save time, streamline documentation, and support access through natural forms of interface design [1,2]. Digitalization might thus enable the health care domain to cope with globally occurring challenges like cost, efficiency, complexity, and reform pressure [3]. However, according to an annual cross-sectoral investigation in Germany, the health care domain is not exploiting the potential to the same extent as are other domains [4]. Furthermore, within the domain (eg, between different hospitals), the digitalization status varies considerably [5,6]. Health care institutions, admittedly, contain particularities that distinguish them from classic value-creating companies: they heavily rely on "highly specialized human capital" [7]. Human acceptance factors in the health care context have been investigated using hospital information systems as an example [8]. The heterogenous digitalization success of health care institutions calls for further investigation into the underlying causes and effects of this variability [9]. The German health care system is decisively governed in a self-administered manner [10]. Although all self-governing institutions are under the legal supervision of the state and are bound by the state's framework legislation, they are not under the professional supervision of the state. Representatives of health insurance companies, health care service providers, and patient representatives negotiate and determine medical services that are covered by the statutory health insurance. Advocacy groups thus have an important role in balancing the stakeholder's interests for the benefit of the common good. Additionally, several studies have pointed out the importance of humans as potentially the greatest obstacle to or the greatest promoter of digitalization in health care processes [11-13]. In the professional and political literature, physicians are partially portrayed as digitalization skeptics [14]. Individual physicians' organizations generally position themselves against efforts to increase the digitalization of health care processes [15,16], while other studies present a low digital penetration rate and a need for action [17,18]. Thus, the aim of this study was to further investigate the role of physicians in the process of digitalization as one of the key stakeholder groups in health care provision.

Prior Work

Digitalization is a disruptive change that affects all areas of society [19]. However, there is currently no consensus on a generally applicable definition of this term. With regard to an original technical understanding, digitization means the "conversion of analogue data (image, text, sound, etc) into digital data" [20]. Definitions of digitalization range from the "replacement of analogue service provision […] in whole or partly by service provision in a digital, computer-manageable" way [21], to the integration of all involved actors and data through digital technologies that influences the entire value chain [22]. Regarding health care organizations, Meister et al [19] describe digitalization as a "continuous change process," which combines the incorporation of digital technology and the ability to constantly adapt to changing conditions.

The importance of the human factor in health care processes is highlighted in the concept of health care–providing institutions as "expert organizations" [11,13,23]. Expert organizations are defined as "knowledge and competence-intensive service organizations whose value creation is primarily based on the recruitment, refinement and use of highly specialized human capital" [7]. Experts are thus individuals who are highly qualified, have a strong position in their institution, and strongly identify with their profession. Furthermore, they have a high degree of autonomy in decision-making and create complex services or products [23]. The integration of interprofessional knowledge and skills of experts participating in the clinical treatment process constitutes "the most important capital" in health care provision [23]. Digital process support requires a full integration across all contributors and change on different levels of hitherto established structures of health care institutions [24]. Child [25] states that experts are especially likely to be suspicious toward change of their established routines. As stated above, individual physicians' organizations, as stakeholders of self-administration, have raised concerns regarding digitalization [15,16]. This might partly be due to a general skepticism toward change in humans [26]. In the field of digitalization, however, the term "digital natives" is often used, which assumes a lower hurdle for younger generations to engage with digital technologies [27]. The role that the factor of age has indeed been investigated in the field of technology acceptance in general [28] but also with regard to the digitalization of hospitals. Hospital employees themselves suspect age to be a decisive factor in whether the digital transformation of their working environment is accepted or not [29].

Objective of This Study

The purpose of this study was to examine the role of physicians in the digital transformation of health care with a specific focus on the variable of age. Therefore, the following 3 research questions were investigated: (1) How do physicians perceive opportunities and risks of the digital transformation of their working environment? (2) How do physicians see their own role in digitalizing health care provision? (3) What role does age play in the perception of digitalization of health care provision and the personal role within this process?

In order to examine these issues, a nationwide survey among physicians in Germany was conducted.

Methods

Survey Design

The survey was designed in an iterative manner by scientists in the field of digital health and members of the Bündnis Junge Ärzte (BJÄ, Alliance of Young Physicians), a union of representatives of young physicians from 25 medical associations and medical societies in Germany. We followed the survey principles outlined by Dillman et al [30] and Schleyer and Forrest [31], while the results are reported in accordance with Eysenbach [32]. The survey design resulted in a structured format comprising a maximum of 42 questions, with adaptive questioning being used to reduce complexity and volume for the participants. Single- and multiple-choice questions were included with answer types assigned to nominal, ordinal, and ratio scales. Free-text fields were provided for further explanatory comments. The full translated questionnaire can be found in the Multimedia Appendix 1.

On the survey landing page, we describe the survey topics and length, goals and target group, and the inquiring organizations, and provide information on the data handling according to the European General Data Protection Regulation (GDPR). The survey was voluntary, nonincentivized, and fully anonymous. None of the participant information requested could be used to identify the participant, and no technical identifiers (eg, IP address) were stored. To start the survey, participants were required to express consent to the procedure. The first survey section included demographic questions related to the respondent's age, gender, professional position and type of employment, medical specialization, and general digital affinity derived from items provided by the technology affinity questionnaire from Karrer et al [33]. The following section "Status Quo," comprised questions regarding degree of digital process support in the respondent's current working environment, including internal and intersectoral data handling. The medical process steps queried in this section were derived from the best practice report provided by Kılıç [34], which describes digital health care processes, as well as the approach by Burmann et al [35], who describe different maturity states of digital health care provision. Moreover, the already noticeable benefits through digitalization and the areas of untapped potential were addressed. These areas, where advantages through digitalization are anticipated, were adapted from the industry and hospital 4.0 paradigm [36,37]. Following this, the role of medical professionals was examined. In response to the controversial description of medical practitioners as, by profession, not being capable of orchestrating digitally supported health care supply chains [38], the view on hindrances to digitalization of the mentioned group was queried. Additionally, the respondents were asked to assess their own familiarity with current technological, processual, and legal topics with regard to digitalization of the German health care sector. The following section, "Mobile Health Apps," was dedicated to general professional mobile app use and digital health apps. The latter was involved due to its facilitation of medical prescriptions for digital health apps (Digitale Gesundheitsanwendungen [DiGA]) by law through the digital health care act (Digitale-Versorgung-Gesetz [DGV]), which came into effect in Germany just when the survey was launched [39]. The last survey section, "Future," detailed the respondents' perspectives on the future of digital health supply, including expectations and the personal role the professionals within this current change. Each survey section was presented on a single page, resulting in a total number of 5 pages including the welcome message. A total of 42 questions were partitioned across 4 questionnaire pages. Comprehensibility, usability, and technical functionality were tested before the survey launch with a group of members of the BJÄ.

Recruitment

The target group of the survey was practicing and prospective physicians. In order to effectively use the distribution channels of the BJÄ for acquiring a convenience sample, the survey was held open. To prevent multiple participation, a cookie was set with submission of the questionnaire. The survey was administered from October 16, 2020, to December 18, 2020, via the online survey platform LimeSurvey. During this period, the survey was publicly available and repeatedly announced through various online and personal channels, including social media accounts (Twitter, LinkedIn, Facebook), press releases, and mailing lists from the BJÄ and its 25 member associations, as well as magazines and newsletters for the health care sector. Furthermore, the personal approaches of the professional networks of the actors involved were used. We provided a dedicated URL redirect which led to the survey via a link. The first contact points with participants were professional networks, or a personal or direct approach via medical associations, all mainly through online channels.

Data Exclusion

We included only those questionnaires that were complete and from respondents with a professional background as medical practitioners. The latter included physicians either in training or in practice in the health care sector. Responses from medical practitioners in retirement or employed in the industrial context, as well as actors with other professional backgrounds were thus excluded from further analysis. From 1940 initial questionnaires, 651 were excluded due to incompletion or missing values, resulting in a completion rate of 66.44%. A further 15 questionnaires were then excluded due to the aforementioned exclusion criteria, resulting in 1274 included data sets.

Data Analysis

The main outcome variables of the survey were the perceived digitalization hindrances, the anticipated role of digitalization in the future health care process, and the respondent's role within this change process. The first aspect was assessed through multiple-choice questions, while the latter 2 were assessed with both single-choice and multiple-choice questions. All mentioned outcome variables were assigned to nominal scales. Descriptive variables included nominal scales (gender, working environment, medical specialization), ordinal scales (professional level, Likert-type digital affinity), interval (Likert scale digital affinity), and ratio scales (age). In order to examine the relation between the outcome variables and age as the primary covariate, we carried out statistical parametric tests for metric scales and nonparametric tests for investigating the association of categorical and metric variables [40]. Depending on the respective scale of the covariate, Pearson correlation coefficient [41], *t* test [42], effect analysis with Cohen *d* [43], and η coefficient [44] was calculated. For accompanying questions, the percentage of respondents who chose each item was calculated. The descriptive data analysis was carried out using Microsoft Excel (Microsoft Corp). For the investigation of associations via Pearson correlation coefficient and *t* test, along with effect analysis with Cohen *d* and η, the open source software PSPP (GNU project) was used. For parametric testing via Pearson correlation coefficient, the assumption of linearity, related pairs, absence of outliers, and suitable measurement scales were investigated. For the application of *t* tests, the assumptions of suitable measurement scales, adequacy of sample size, and homogeneity of variance were examined [45].

Results

User Statistics

The total of 1274 complete and included responses comprised 567 (44.51%) female respondents. The age of the respondents ranged from 22 to 67 years (mean 45.09, SD 12.06). The professional level of the respondents included first level, medical students (24/1274, 1.88%); second level, physicians in specialist training (328/1274, 25.75%); third level, medical specialists with <5 years professional experience (180/1274, 14.13%); fourth level, medical specialists with >5 years professional experience (732/1274, 57.46%); and other (10/1274, 0.78%). The 2 major shares of respondent's working environment was split approximately evenly, with one-half working in clinical environments (593/1274, 46.55%) and one-half in physician's offices (594/1274, 46.62%), with 87 others (6.83%). The 5 most-represented medical specializations were general internal medicine (184/1274, 14.44%), dermatology and venerology (138/1274, 10.83%), ophthalmology (133/1274, 10.44%), urology (122/1274, 9.58%), and general medicine (103/1274, 8.08%). Further demographic data of the respondents are depicted in Table 1.

Table 1. Respondent's demographics (N=1274).

Characteristic	Value, n (%)
Gender	
Female	567 (44.51)
Male	706 (55.42)
Other	1 (0.08)
Age	
≤35 years	382 (29.98)
36-45 years	290 (22.76)
46-55 years	276 (21.66)
≥56 years	326 (25.59)
Professional level	
First: medical student	24 (1.88)
Second: specialist training	328 (25.75)
Third: specialist <5 years	180 (14.13)
Fourth: specialist >5 years	732 (57.46)
Other	10 (0.78)
Working environment	
Clinic	593 (46.62)
University hospital	192 (32.43)
Public hospital	165 (27.87)
Nonprofit hospital	134 (22.64)
Privat hospital	92 (15.54)
No answer	9 (1.52)
Physician's office	594 (46.55)
Self-employed	465 (78.28)
Employee	129 (21.72)
Other	87 (6.83)
Volume of employment	
Full-time	1008 (79.12)
Part-time	227 (17.82)
Marginal employment	13 (1.02)
No answer	26 (2.04)

The demographic factor "digitalization affinity" was also measured. For this, 4 suitable theses from a general technology affinity questionnaire [33] were taken and adapted to the focus of digitalization. These 4 theses included affinity toward exploring digital services, perceived ease of access, impact on everyday convenience, and impact on communication. All these were queried in a 5-point Likert-type scale (“strongly disagree”=1 to “totally agree”=5). After investigating internal consistency (Cronbach α=.68) [46], we combined these items into a single Likert scale by calculating the mean value per respondent [47].

For further investigation, we also split the respondents into 2 groups: based on the age limit of the BJÄ and the definition of the German medical associations, we placed participants who were 45 years of age and younger into group 1 (672/1274, 52.75%) and those who were older than 45 years into group 2 (602/1274, 47.25%).

Descriptive Outcomes

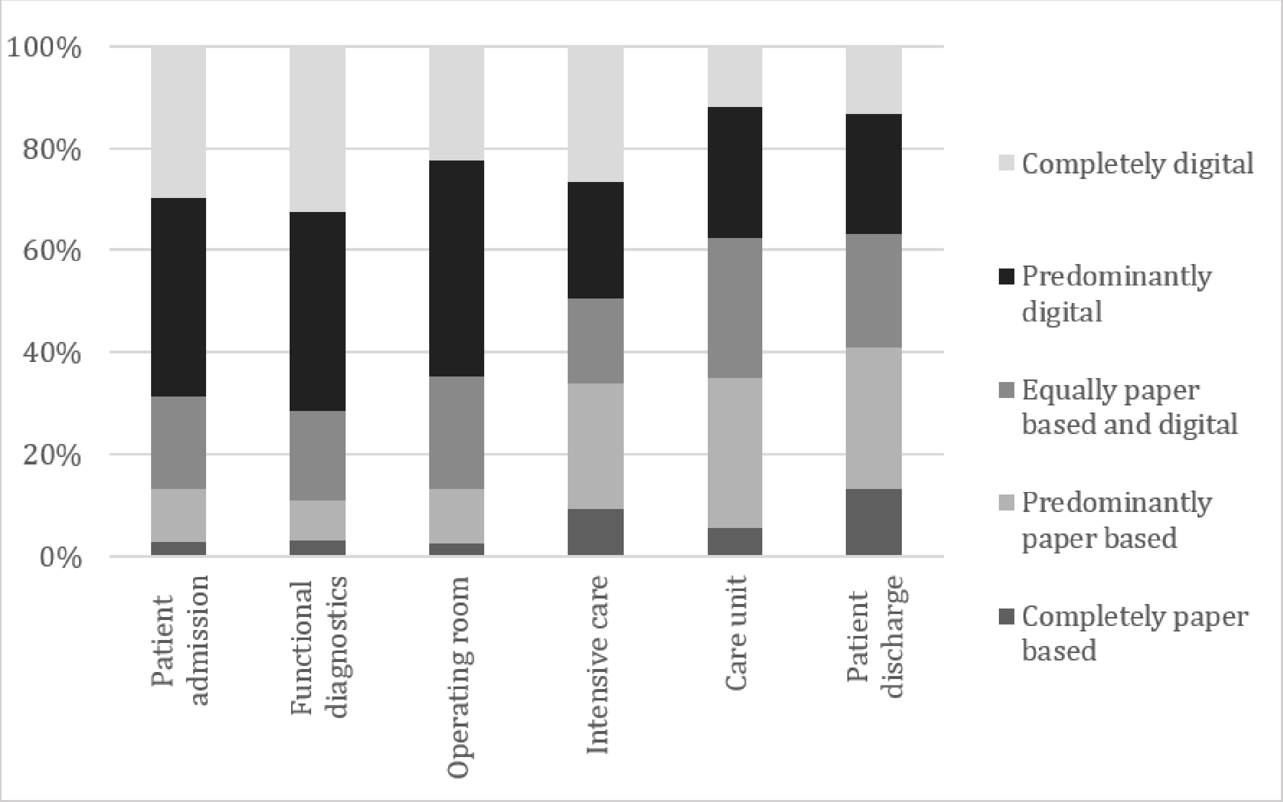
The digital affinity variable, comprising 4 five-point Likert-type items within a Likert scale resulted in a moderate tendency toward a positive perception of digitalization. The mean score across all respondents was 3.88 (SD 0.67). In the following subsections (*Status Quo*, *Mobile Health Apps*, and *Future*), we present a descriptive analysis of these 3 areas of the questionnaire.

Status Quo

First, we identified the status quo of use of digital systems in the respondents’ everyday working life. We queried the 4 segments of internal process support (including applications and data administration), interorganizational data exchange, professional communication, other digital services for internal organization (that do not directly concern patients, such as professional training or e-learning, duty planning, worktime recording), and other services addressed to patients (eg, appointment scheduling, virtual consultation hours, medication plan, access to patient data, mobile apps).

The digitalization of internal processes was led by functional diagnostics (radiography, laboratory), with 749/1187 (68.72%, adjusted by the share of respondents who stated that this was not relevant for them) respondents indicating that they organize completely or predominantly digitally. This was directly followed by the areas of patient admission (749/1090, 68.71%), operating room (398/616, 64.61%), and intensive care (207/418, 49.52%). Care unit (240/637, 37.68%) and patient discharge (264/716, 36.87%) were the least digitally organized areas. The adjusted percentages of responses are shown in Figure 1.

Figure 1. Status quo: respondents' assessment of digitalization of internal processes.



Interorganizational data exchange still is a primarily paper-based process, as only 132 of 1259 respondents (10.48%) stated they receive data completely or predominantly digitally from other service providers, while 161 of 1258 (12.80%) transfer data themselves mainly in a digital format to other service providers.

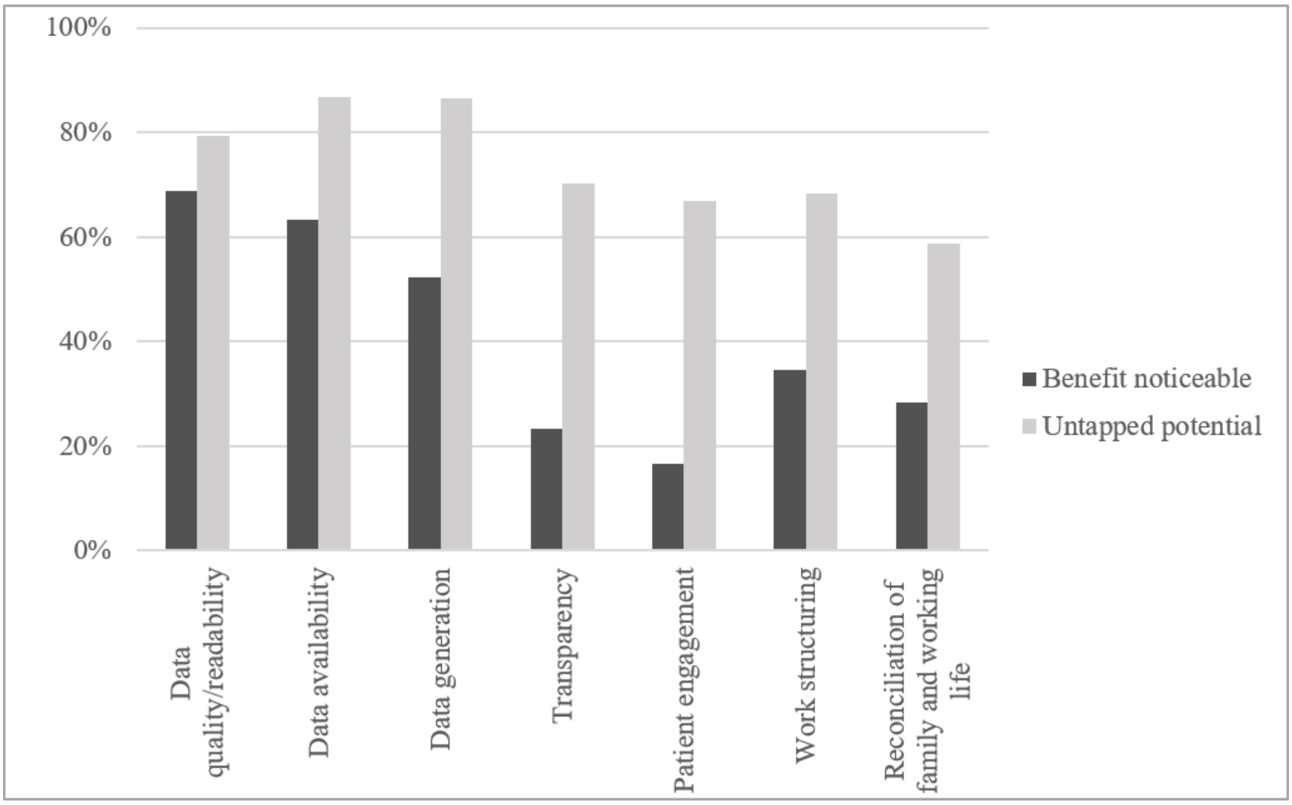
As expected, regarding professional communication, the phone call was still the predominant tool for interaction, as 1248 of 1274 respondents indicated using it for professional communication (97.96%). Fax (1082/1274, 84.93%), mail (967/1274, 75.90%), and email (976/1274, 76.61%) were also used by a substantial majority. Meanwhile, medical platforms (115/1274, 9.03%), messaging apps for specific medical purposes (142/1274, 11.15%), and generic messaging apps (332/1274, 26.06%) were ranked at the bottom of the list.

Patient distant digital services use was relatively widespread: 1006 participants (78.96%) stated that they used digital services for professional training or e-learning, 776 (60.91%) planned their duty in a digital system, and 579 (45.45%) recorded their worktime electronically.

Interestingly, services addressed to patients did not show a high degree of dissemination. Digitalized appointment scheduling ranked highest, with 24.49% (312/1274) of the respondents stating that they offered this service, followed by the provision of an electronic medication plan (235/1274, 18.45%), access to personal data (228/1274, 17.90%), virtual consultation (203/1274, 15.93%), and mobile apps (47/1274, 3.67%). The provision of none of these services without mentioning alternatives in use was the only option selected more frequently (580/1274, 45.53%). One question was then aimed at the proactive offering of health-related data for assessment through patients themselves, acquired by, for instance, wearables or apps. Of the 1274 respondents, 51 (4%) indicated experiencing this regularly, 197 (15.46%) occasionally, 448 (35.16%) sporadically, and 553 (43.41%) had never encountered this situation. Of the 696 respondents who had encountered self-acquired patient data to a varying extent, 41 (5.90%) generally refused to incorporate this kind of information into their medical investigation, 314 (45.11%) stated they verify only acutely relevant data, and 341 (48.99%) indicated being generally open to data from consumer products provided by the patient.

We queried perceptions regarding the already existing benefits of digitalization and untapped potential in the 7 categories of data quality and readability, data availability, data generation, transparency, patient engagement, work structuring, and reconciliation of family and working life, and responses varied considerably. Affirming the comparatively low usage of digital services for patients, only 213 of 1274 respondents (16.72%) already noticed benefits through digitalization in the category of patient engagement. An only slightly higher perception of utility was indicated for transparency, while data quality and readability and data availability were indicated to have received the most benefit thus far. However, the believed untapped potential exceeded the already noticeable benefit in all queried categories. Data availability, generation, and quality or readability ranked highest while optimism for digitalization improving everyday working life (working structure and reconciliation of family and working life) was also present, but not quite on the same level. The detailed data concerning the perceived benefits and potential of digitalization are displayed in Figure 2.

Figure 2. Shares of respondents who see noticeable benefits and untapped potential through digitalization across different categories.

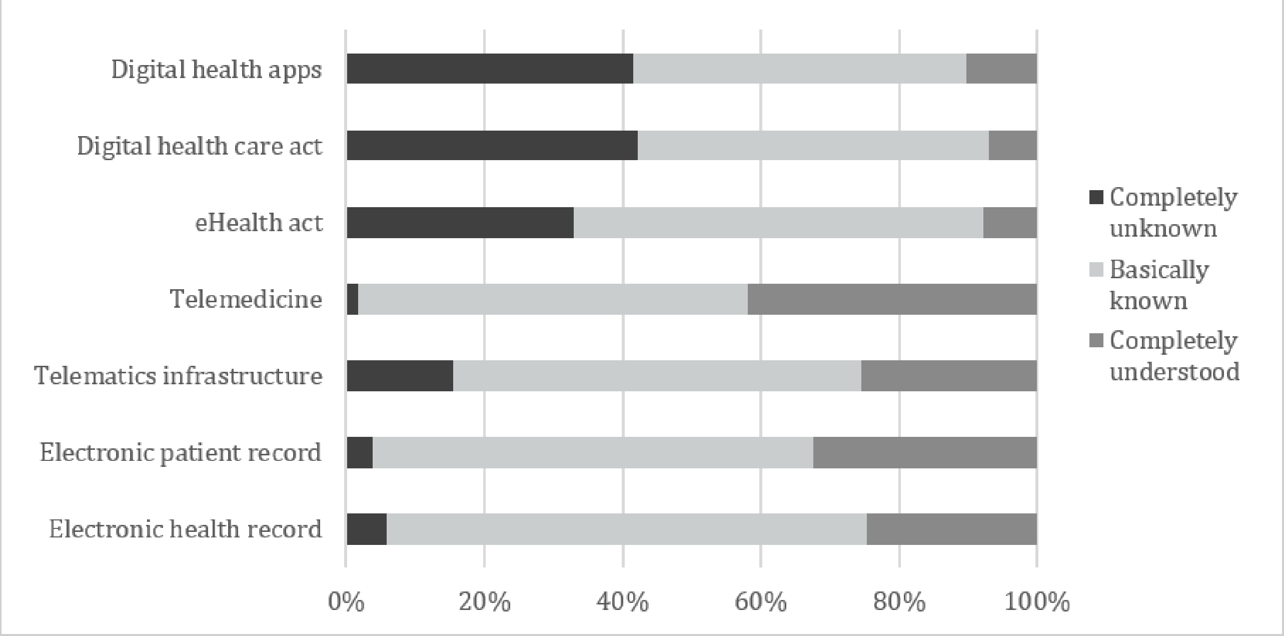


The next 2 multiple choice questions queried obstructions and constraints in digitalization regarding the user and technology side. For the user side, we asked the respondents to indicate whether they perceived 5 different items to be a “major hindrance” for digitalization. Almost half of the respondents (626/1274, 49.14%) stated a lack of noticeable saving of time to be a major impeding factor. Slightly fewer respondents considered insufficient digital literacy or sovereignty (530/1274 41.60%), fear of surveillance (508/1274, 39.87%), and an unwillingness to change (461/1274, 36.19%) to also be hindrances. Fear of loss of importance was indicated to be a major limiting factor by 99 respondents (7.77%), and 210 respondents stated they did not perceive these hindrances to be present in themselves or their age group. When “other” was indicated, these responses referred to data security concerns, loss of trust between patients and physicians, and insufficient user integration.

The most-frequently chosen major technical hindrance was "insufficient system integration" (798/1274, 62.64%). Almost half of the respondents perceived insufficient software functionality (575/1274, 45.13%) to be a major issue, followed by insufficient hardware (503/1274, 39.48%), insufficient budget (453/1274, 35.56%), legal concerns regarding the exchange of medically sensitive data (341/1274, 26.77%), and insufficient cooperation by system providers (247/1274, 19.39%). When "other" was indicated, these response referred to data security concerns, system availability or performance issues, and a lack of user-centered system design.

Subsequently, the respondents provided an assessment of their familiarity with the current trending topics regarding digitalization of the health care sector in Germany. This included electronic health and electronic patient records, telematics infrastructure, telemedicine, the eHealth act, the digital health care act (DGV), and digital health apps (DiGA). For each topic, participants were required to indicate if it was completely unknown, basically known, or completely understood by them. The distributions of responses across these 7 topics are summarized in Figure 3. It is important to note that at the time of the survey (November 2020 to December 2020) some currently trending topics (eg, digital health apps, the digital health care act, and eHealth act) were less well known than were others.

Figure 3. Respondents' assessment of familiarity with current trending digitalization topics.



Mobile Health Apps

Of the 1274 respondents, 723 (56.75%) stated that they personally used mobile apps in their everyday work life. The most mentioned field of use was pharmaceutical information (516/723, 71.37%) and diagnosis (386/723, 53.39%), followed by training (317/723, 43.85%) and communication (300/723, 41.49%). When asked whether they trust in digital health apps, 425 stated yes (33.36%), 196 stated no (15.38%), and 653 stated "that depends" (51.26%).

Only a small portion of respondents (223/1274, 17.50%) stated that they had recommended specific mobile health apps to their patients. Of the 1051 respondents who had not yet recommended an app, 420 (39.96%) stated that their reasons for not having done so included "insufficient validity", 286 (27.21%) stated it was "not relevant in my area," and 266 (25.31%) indicated "insufficient data protection." However, 80.46% (1025/1274) of the participants expect mobile or digital health apps to play a role in health care provision in the future.

Regarding the main sources of information for mobile or digital health apps, 812 (63.74%) indicated medical societies as the main source, 671 (52.67%) the internet, and 622 (48.82%) colleagues. A much smaller proportion of respondents indicated public bodies (182/1274, 14.29%) and developers (104/1274, 8.16%) as playing important roles in information acquisition.

Future

In the section, "Future," the survey participants were first asked to rate their expectation for the impact of digitalization on 10 health care provision–related areas from "worsening" over "no change" to "improving.". The highest positive expectations were shown in the areas "access to knowledge" (1136/1274, 89.17%), "medical research" (1023/1274, 80.30%), and "treatment of rare diseases" (1016/1274, 79.75%). The greatest doubts were expressed in the areas of "physician-patient relationship" (397/1274, 31.16%), "administration" (265/1274, 20.80%), and "attractiveness of the profession" (237/1274, 18.60%; see Table 2).

Table 2. Expected impact of digitalization on health care provision (N=1274).

Area	Worsening, n (%)	No change, n (%)	Improving, n (%)
Early detection of diseases	51 (4)	549 (43.09)	674 (52.91)
Medical quality	156 (12.24)	422 (33.12)	696 (54.63)
Access to knowledge	10 (0.78)	128 (10.05)	1136 (89.17)
Treatment of rare diseases	8 (0.63)	250 (19.62)	1016 (79.75)
Administration	265 (20.80)	207 (16.25)	802 (62.95)
Patient adherence	91 (7.14)	691 (54.24)	492 (38.62)
Physician-patient relationship	397 (31.16)	672 (52.75)	205 (16.09)
Interdisciplinary collaboration	78 (6.12)	334 (26.22)	862 (67.66)
Attractiveness of the profession	237 (18.60)	644 (50.55)	393 (30.85)
Medical research	20 (1.57)	231 (18.13)	1023 (80.30)

Subsequently, the respondents rated their attitude towards upcoming changes through digitalization from “mainly positive” (567/1274, 44.51%) to “with mixed feelings” (557/1274, 43.72%) to “mainly negative” (130/1274, 10.20%).

When asked for multiple adjectives to describe their personal role in digitalizing health care provision, 36.50% (465/1274) of respondents assessed themselves as “scrutinizing,” 30.06% (383/1274) as “active,” 29.51% (376/1274) as “open,” and 25.20% (321/1274) as “critical.” Only 1.73% (22/1274) stated that they were “indifferent.”

Investigation of Age Associations

A Pearson correlation coefficient (r=–0.30; P<.001) revealed a significant negative linear relationship between age and the digital affinity variable. The Likert scale resulted in a mean of 4.06 for group 1 (SD 0.55) and 3.68 for group 2 (SD 0.72). A significant difference between the 2 groups was found in a 2-tailed t test (t_{1122}=10.64; P<.001) with a medium effect size (Cohen d=0.61). Inhomogeneity of variances was presumed based on a Levene test (P<.001).

Status Quo

We assumed that the already existing penetration of digital systems, mainly queried in the section, “Status Quo,” was substantially dependent on other factors (eg, working environment, financial situation of the employing organization, career stage). Thus, we focused this investigation on areas presumably in the sphere of influence of the respondents.

Regarding the communication medium of choice, no noticeable differences were found between the 2 age groups. Fax was used by 86.61% (582/672) of group 1 and 83.01% (500/602) of group 2, specific medical messaging apps were used by 10.42% (70/672) of group 1 and 11.96% (72/602) of group 2, and generic messengers were used by 26.34% (177/672) of group 1 and 25.75% (155/602) of group 2.

The perception of already noticeable benefit through digitalization was almost equally distributed in the 2 age groups. In the 7 queried categories (data quality and readability, data availability, data generation, transparency, patient participation, work structuring, and reconciliation of family and working life) an average of 41.48% (279/672, SD 22.71%) of age group 1 stated that they already noticed benefits of digitalization while an average of 41.03% (247/602, SD 18.53%) in age group 2 stated the same. Moreover, the η coefficient showed no or negligible association between age and the assessments of noticeable benefits within these 7 categories. However, when asked about the untapped potential of digitalization, the 2 age groups showed differences. In age group 1, an average of 83.25% (559/672, SD 8.71%) saw untapped potential across these categories, while the average in age group 2 was 64.29% (387/602, SD 13.06%). A 2-tailed t test (t_{12}=3.20; P=.003) underlined the significance of this difference between the 2 groups, and Cohen d showed a strong effect size (d=1.71). Homogeneity of variances was asserted using a Levene test, which showed that equal variances could be assumed (P=.17). Additionally, the singular assessment of each category, except for “data generation,” showed an association with age. The η associations of noticed benefits, perceived untapped potential, and the age of the respondents are depicted in Table 3, along with the number of affirmations per group.

Table 3. Association between noticed benefits and potentials across categories and affirmation numbers per group.

Category	Association with age (η)	Group 1, n (%) (N=672)	Group 2, n (%) (N=602)
Noticeable benefits			
Data quality/readability	0.08	486 (72.32)	396 (65.78)
Data availability	0.05	449 (66.82)	362 (60.13)
Data generation	0.08	372 (55.36)	298 (49.50)
Transparency	0.00	160 (23.81)	139 (23.09)
Patient participation	0.03	115 (17.11)	98 (16.28)
Work structuring	0.11	201 (29.91)	241 (40.03)
Reconciliation of family and working life	0.09	168 (25)	195 (32.39)
Untapped potential			
Data quality/readability	*0.26* [a]	601 (89.43)	417 (69.27)
Data availability	*0.27* [a]	638 (94.94)	474 (78.74)
Data generation	0.16	614 (91.37)	494 (82.06)
Transparency	*0.23* [a]	533 (79.32)	367 (60.96)
Patient participation	*0.25* [a]	518 (77.08)	339 (56.31)
Work structuring	*0.25* [a]	534 (79.46)	343 (56.98)
Reconciliation of family and working life	*0.27* [a]	478 (71.13)	275 (45.68)

[a] Italics indicate a significant difference between the 2 age groups.

The attitude toward patient-provided consumer data was generally positive in both groups: of the respondents who had encountered this, 94.38% (336/356) in group 1 were willing to incorporate these data into their medical investigation while 93.82% of group 2 (319/340) showed the same willingness. A negligible association between age and willingness was found to be related to the willingness to incorporate this type of data (η=0.13).

Regarding the perception of major hindrances for digitalization on the user side, the 2 groups showed differences in 3 categories. The ratings in lacking noticeable saving of time, insufficient digital literacy or sovereignty, and no perception of such hindrances in themselves and their age group showed a weak association in the η coefficient with the nominal variable of age (see Table 4).

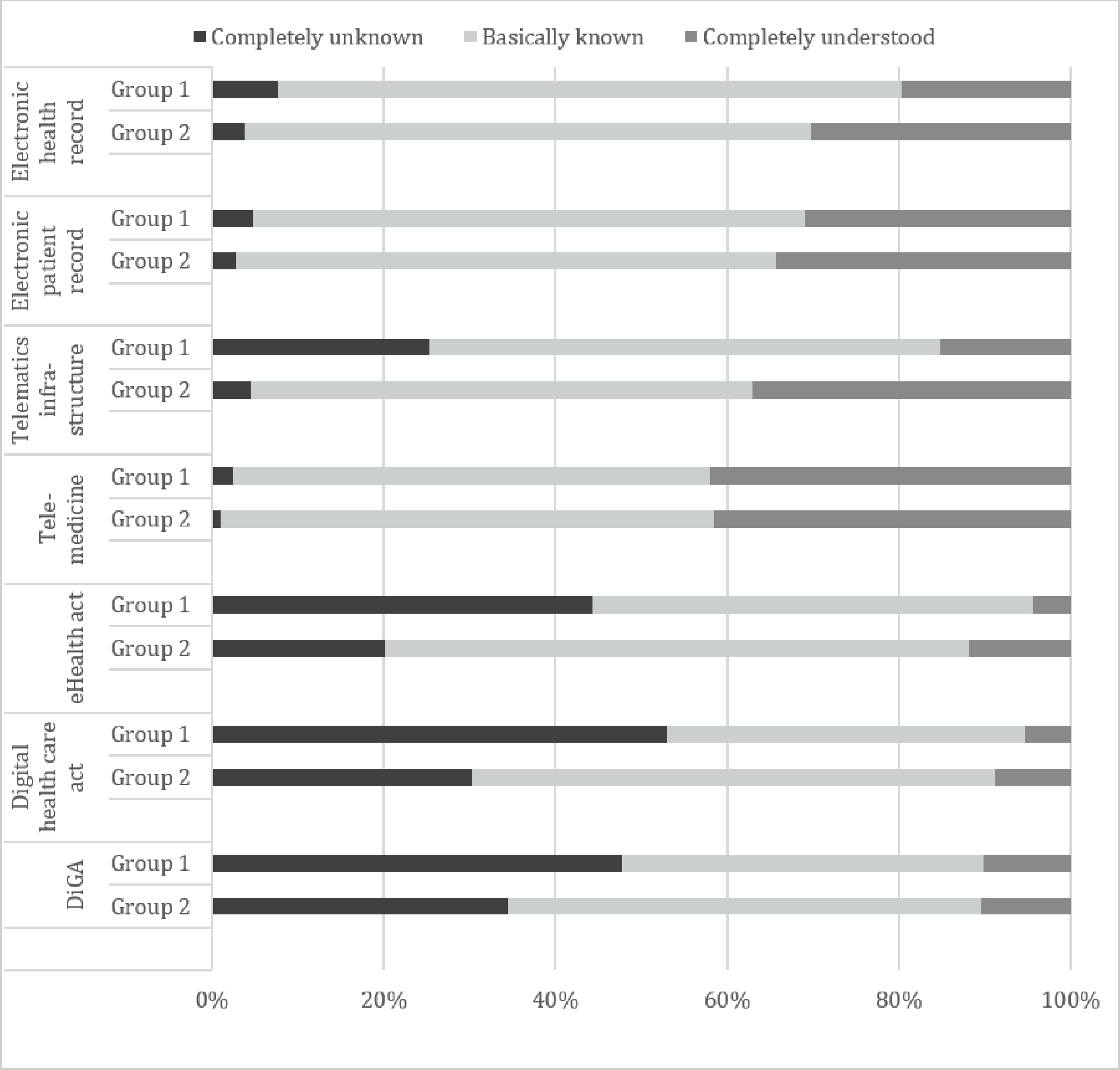
Table 4. Association of respondents' perception of major hindrances for digitalization with age and the agreement numbers per age group.

Hindrance	Association with age (η)	Group 1, n (%) (N=672)	Group 2, n (%) (N=602)
User side			
Insufficient digital literacy/sovereignty	*0.21* [a]	223 (33.18)	307 (51)
Lack of willingness to change	0.01	247 (36.76)	214 (35.55)
Lack of noticeable saving of time	*0.24* [a]	260 (38.69)	366 (60.80)
Fear of loss of importance	0.02	57 (8.48)	42 (6.98)
Fear of surveillance	0.15	222 (33.04)	286 (47.51)
No such hindrances	*0.27* [a]	172 (25.60)	38 (6.31)
Technology side			
Insufficient hardware	*0.26* [a]	343 (51.04)	160 (26.58)
Insufficient software functionality	0.08	323 (48.07)	252 (41.86)
Insufficient system integration	0.10	444 (66.01)	354 (58.80)
Insufficient budget	0.03	186 (27.68)	144 (23.92)
Insecurity with legal framework regarding data exchange	0.05	164 (24.40)	177 (29.40)
Insufficient cooperation by system providers	0.12	99 (14.73)	148 (24.58)
No such hindrances	0.08	8 (1.19)	16 (2.66)

[a] Italics indicate a significant difference between the 2 age groups.

The distribution of familiarity with current trending topics regarding digitalization of the health care sector in Germany across age groups 1 and 2 is displayed in Figure 4. Age group 1 consider themselves to be less well informed across all topics, except telemedicine. However, the association of "informedness" with age was negligible in all categories except telematics infrastructure (η=0.36), the eHealth act (η=0.31), and the digital health care act (η=0.25), where a weak association between age and familiarity was found.

Figure 4. Respondents' assessment of familiarity with current trending digitalization topics by age group. DiGA: Digitale Gesundheitsanwendungen.



Mobile Health Apps

Respondents’ personal use of mobile apps in their everyday working life was more common in age group 1 (456/672, 67.86%) than in group 2 (267/602, 44.35%), which showed a weak tendency (η=0.26) in the association with the respondents’ age.

The fields of use also showed differences between the 2 groups. The use of mobile apps for information on pharmaceuticals and as a diagnosis aid showed a weak association with the age (η=0.29 and η =0.23, respectively; Table 5.)

Table 5. Association of professional usage fields of mobile apps and occurrence per age group.

Professional usage fields of mobile apps	Association with age (η)	Group 1, n (%) (N=672)	Group 2, n (%) (N=602)
Communication	0.03	148 (41.49)	152 (56.93)
Training	0.13	195 (42.76)	122 (45.69)
Information on pharmaceuticals	*0.29* [a]	354 (77.63)	162 (60.67)
Diagnosis	*0.23* [a]	264 (57.89)	122 (45.69)

[a] Italics indicate a significant difference between the 2 age groups.

Regarding trust in digital health apps, we also saw a tendency of group 1 to have more confidence in these (yes: 305/672, 45.39%; no: 42, 6.25%; that depends: 325, 48,36%) compared to group 2 (yes: 120/602, 19.93%; no: 154, 25.58%; that depends: 328, 54.49%). The η coefficient (η=0.35) showed a weak association with the age.

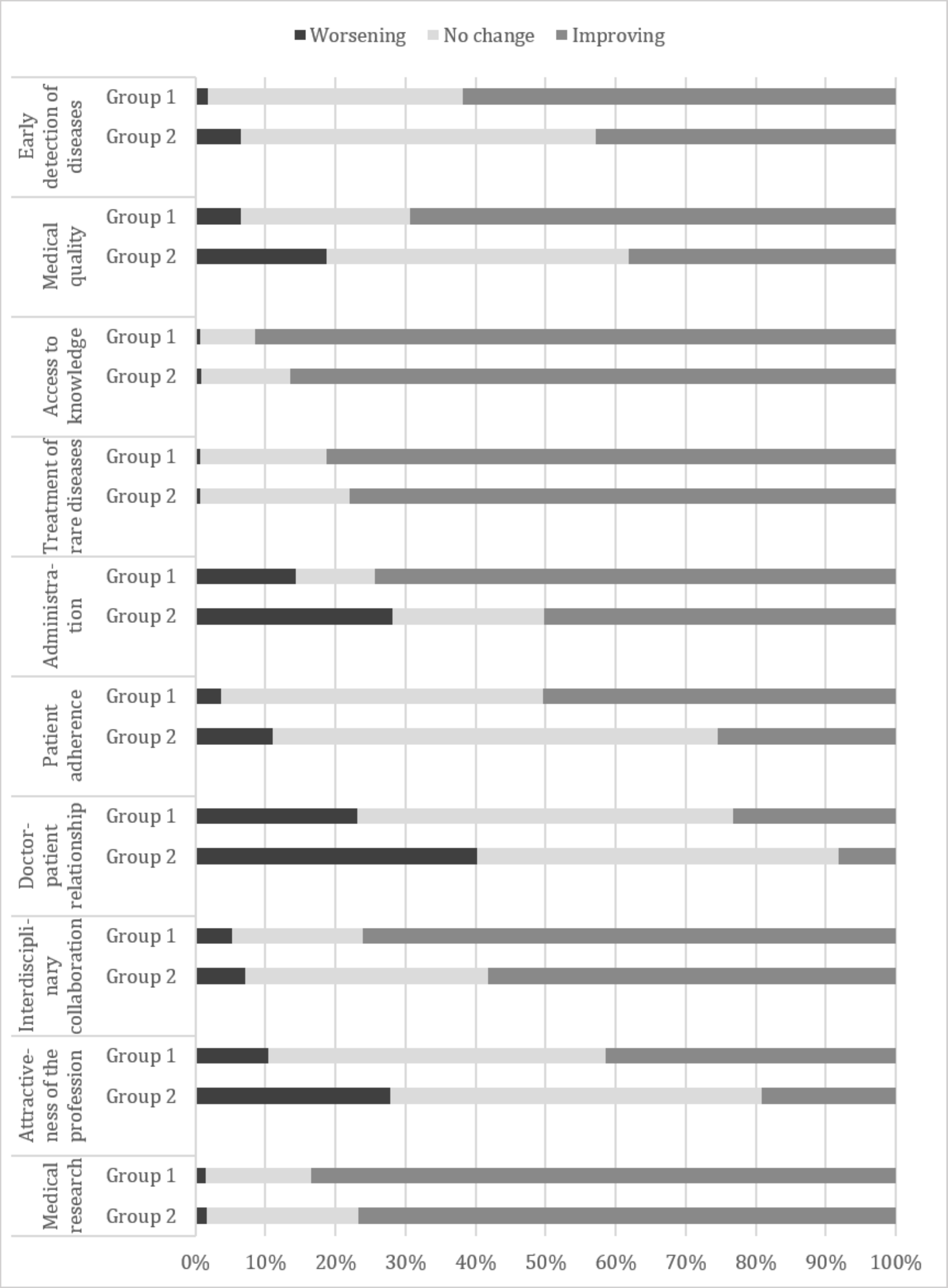
The recommendation rate of mobile health apps to patients did not relate noticeably with respondent age (η=0.02). As a reason for not having recommended an app, group 1 mentioned "insufficient data protection" (84/672, 15.19%) at a lower proportion than did group 2 (182/602, 36.55%) with η=0.21 for the nominal association with age, while other reasons showed no or negligible association with the age. Moreover, the belief that mobile apps will be relevant for future health care provision was somewhat stronger in group 1 (603/672, 89.73%) than in group 2 (422/602, 70.10%), with a weak relation with the nominal variable (η=0.21).

As an important source of information, the internet (η=0.14), public bodies (η=0.09), medical societies (η=0.15), and developers (η=0.04) were valued without considerable associations with the age. Only the selection of colleagues as an important information source had a slightly increased importance in group 1 (398/672, 59.23%) compared to group 2 (224/602, 37.21%; η=0.26).

Future

The assessment of the expected impact of digitalization on 10 health care provision–related areas showed a generally positive attitude. Mixed emotions became apparent regarding the physician-patient relationship, administration, and the attractiveness of the profession. In Figure 5, the assessments are displayed by age group. The areas "access to knowledge," "treatment of rare diseases," and "medical research" were assessed equally by both groups. However, weak associations between assessment and age were found in "medical quality" (η=0.30), "attractiveness of the profession" (η=0.28), "administration" (η=0.27), "patient adherence" (η=0.27), "physician-patient relationship" (η=0.25), "early detection of diseases" (η=0.21), and "interdisciplinary collaboration" (η=0.21).

Figure 5. Expected impact of digitalization on health care provision by age group.



The subsequent rating of the personal attitude towards upcoming changes through digitalization also showed a weak association with the age (η=0.36), with a rather positive trend in age group 1 (mainly positive: 399/672, 59.38%; with mixed feelings: 241/672, 35.86%; mainly negative: 25/672, 3.72%) compared to age group 2 (168/602, 27.92%; 316/602, 52.49%; and 105/602, 17.44%, respectively). For adjectives used to describe self-perceived roles, only the description of "critical" (group 1: 91/672, 13.54%; group 2: 230/602, 38.21%) and "open" (group 1: 260/672, 28.69%; group 2: 116/602, 19.27%) showed an association, albeit a weak one, with age (η=0.27 and η=0.21, respectively).

Discussion

Principal Results

In this study, one of the main stakeholder groups when it comes to digitalizing health care provision, physicians, showed a general affinity toward digitalization, with a negative linear tendency with decreasing age of the participants. Considering length and complexity of the questionnaire, the completion rate of 66.44% confirms a high interest of the sample in the enquired topic [48]. Survey dropout mainly occurred on survey pages 0 (welcome and consent, 223/666, 33.48%) and 1 (352/666, 52.85%).

The penetration of already existing digital process support was found to be heterogeneous for intraorganizational process areas, while interorganizational processes in general are still primarily paper based.

Also, digital services for professional communication have not yet reached a high adoption rate, with no association to users' age. Other services for organizing working life, such as duty planning or e-learning, show relatively widespread use. This contrasts with digital offers for patients, which reach a maximum usage rate of a quarter of the respondents for appointment scheduling, while other services show much lower proportions. Meanwhile, more than half of the respondents indicated that they had experienced proactive offering of self-acquired data by patients, with the majority being willing to incorporate these data into medical decision-making. The age groups did not show differences in this regard.

The greatest perceived benefits of digitalization were data quality and readability. Perceived benefits were not associated with respondents' age. However, participants saw untapped potential in all queried areas, with a relation with age to all categories except for data generation.

As major hindrances for digitalization, participants indicated a lack of a noticeable saving of time, followed by insufficient digital literacy or sovereignty as the dominant human factors. The association with the nominal variable of age in the category of insufficient digital literacy or sovereignty and no perception of such hindrances in respondents and their age group was noteworthy. Regarding technology, age groups agreed in most areas on its relevance, and rated insufficient system integration as the major obstacle. Only insufficient hardware was identified more frequently in group 1 compared to group 2, with a weak association with the nominal age.

Familiarity of the respondents with current trending topics regarding digitalization of the health care sector varied widely and seemed to decrease with recency of the discussion or initiative. Interestingly, age group 1 considered themselves to be less well informed across almost all topics. The association with the variable age was only relevant in the 3 topic areas of telematics infrastructure, the eHealth act, and the digital health care act.

More than half of the participants stated that they use mobile apps within their profession, with a weak tendency of an increasing adoption rate with decreasing age. Across all participants, most stated that the use of mobile apps was for information on pharmaceuticals, which was also weakly associated with age. Interestingly, the recommendation rate of mobile health apps to patients did not relate noticeably with the age (η=0.02), but was equally not very common. Only a share of 17.50% stated that they had recommended mobile health apps to patients. Insufficient services available was the main reason for not having done so yet for all participants, while insufficient data protection was a little more relevant for group 2 compared to group 1. The general belief of relevance of mobile apps for future health care provision was weakly associated with decreasing age.

The peer group "colleagues" as a source of information on mobile or digital health applications was slightly more important for younger respondents, while medical societies were the most relevant for all participants.

Respondents exhibited mainly positive expectations concerning the impact of digitalization on specific areas of health care. In particular, access to knowledge, medical research, and the treatment of rare diseases were associated with respondent optimism. Mixed feelings were expressed regarding the physician-patient relationship, administration, and attractiveness of the profession. The latter 3 categories, as well as medical quality, patient adherence, interdisciplinary collaboration, and early detection of diseases, showed a weakly increased optimism with the decreasing age of the respondents. The general attitudes toward upcoming changes through digitalization were split fairly evenly, with one-half having mainly positive feelings and the other having rather mixed feelings. The positive trend was once again weakly associated with age.

Regarding describing adjectives, being "critical" of digitalization was more associated with increased age, while being "open" was associated with decreased age. Indifference towards digitalization was hardly existent.

Limitations

This study is subject to limitations due to participant selection and thus representativeness. With the first contact point being the digital channels of professional networks, a selection bias can be assumed [49]. Inherently, a digital contact point with a survey on digitalization itself might have led to a sample with a greater affinity for digitalization. On the other hand, the German Society for Tropical Medicine, Travel Medicine and Global Health e.V., for example, reaches 1060 of its 1085 members via email. We thus presumed that the undercoverage of respondents with no internet access could be ignored, since the self-organization of professional societies via digital channels can be assumed for the majority to be a prerequisite for professional participation. Another limitation might involve the initially mentioned self-administration of the health care sector in Germany. A presumed participant awareness of a potential interest of political stakeholders on such an investigation might lead to a tendency toward more extreme expression of opinion. However, we assumed this occurs in both directions.

Additionally, the partitioning of responses for the investigation of association was based on the age limit of the BJÄ and the definition of the German medical associations. To complement this presentation of results, a calculation of the eta coefficient incorporated the nominal value of age as an independent variable, where applicable.

Conclusions

Physicians are emotionally involved in digitalizing health care provision, and they predominantly see opportunities as positive but also differentiated. The lower the involvement of second or third parties, such as patients or intersectoral service providers, was apparent, and the lower the GDPR sensitivity was assumed, and the higher was the apparent adoption rate of digital services. However, despite existing data security concerns, generic messaging apps were also found to be acceptable for professional communication from a quarter of the respondents, which supports the need for convenient and seamless solutions. Additionally, the need to personally perceive benefits through digitalization, like the saving of time, was expressed. Interestingly, this was more present with increasing age, which indicates an expectation of an automated and effortless transition. For younger generations, the handling of digital technology may be already inherent, and the conversion burden may thus not seem as onerous as that perceived by older colleagues. This theory might be supported by participants' critical assessment of digital literacy or sovereignty as a field of development, which was increasingly perceived as a major hindrance for digitalization with increasing age. Query related to the current trending topics regarding digitalization of the sector confirmed the presence of an education gap. However, this was slightly more prevalent with decreasing age. Information and education on digitalization as a mean to support health care provision should thus not only be included in the course of study, but also in the continuing process of further and advanced training. Medical societies, statutory health insurance companies, and professional associations were mentioned as desired and trustworthy information providers. This also raises the question of determination and empowerment: when legislative initiatives are unknown, how does the profession want to participate in shaping digital health?

The role physicians see for themselves in the digitalization of health care provision was mainly described as “scrutinizing,” “active,” and “open.” This represents the ambivalence and inner conflict between observant expectation and active participation. The role of individual physicians as multipliers and stakeholders of digitalization within their scope of operation should be acknowledged, as well as the general willingness to participate in this process. On the other hand, the need for guidance and orientation through trustworthy organizations has a right to be instituted in the self-administered health care sector. The incorporation of physicians into the digitalization of their working environment is essential for a functional cocreation of future processes. However, digitalization is a multidisciplinary process [50], and despite the fact that digital affinity seems to increase in each successive generation, in a self-administered system, responsibility for this upcoming change cannot be attributed solely to physicians. A transformation of the system must be collaboratively implemented by all professional stakeholder groups, service providers and organizations, and political groups and sponsors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey translation.
[PDF File (Adobe PDF File), 298 KB - medinform_v9i11e31527_app1.pdf]

References

1. Daum M. Digitalisierung und Technisierung der Pflege in Deutschland. In: Aktuelle Trends und ihre Folgewirkungen auf Arbeitsorganisation, Beschäftigung und Qualifizierung. Hamburg: DAA Stiftung Bildung und Beruf; 2017.
2. Bauernhansl T. Die Vierte Industrielle Revolution? Der Weg in ein wertschaffendes Produktionsparadigma. In: Bauenrhansl T, ten Hompel M, Vogel-Heuser B, editors. Industrie 4.0 in Produktion, Automatisierung und Logistik. Wiesbaden: Springer Vieweg; 2014:5-35.
3. Paré G, Sicotte C. Information technology sophistication in health care: an instrument validation study among Canadian hospitals. International Journal of Medical Informatics 2001 Oct;63(3):205-223. [doi: 10.1016/s1386-5056(01)00178-2]
4. Digitalisierungsindex Mittelstand 2020/2021. In: Der digitale Status quo des deutschen Mittelstands. Bonn: Deutsche Telekom AG; 2020.
5. Hoyt JP. European Hospitals EMRAM Maturity Overview. 2015 Presented at: HIMSS Europe CIO Summit; Oct 7-8, 2015; Valencia.
6. Hübner U, Esdar M, Hüsers J, Liebe JD, Rauch J, Thye J, et al. IT-Report Gesundheitswesen: Wie reif ist die IT in deutschen Krankenhäusern?. Osnabrück: Hochschule Osnabrück - IGW; 2018.
7. Rasche C, Braun von Reinersdorff A. Krankenhäuser als Expertenorganisationen. In: Pfannstiel MA, Rasche C, Mehlich H, editors. Dienstleistungsmanagement im Krankenhaus. Wiesbaden: Springer Fachmedien Wiesbaden; 2016:1-23.
8. Handayani PW, Hidayanto AN, Budi I. User acceptance factors of hospital information systems and related technologies: Systematic review. Inform Health Soc Care 2018 Dec;43(4):401-426. [doi: 10.1080/17538157.2017.1353999] [Medline: 28829650]

9. Burmann A, Meister S. Practical application of maturity models in healthcare: findings from multiple digitalization case studies. 2021 Presented at: 14th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF; Feb 11-13, 2021; Virtual. [doi: 10.5220/0010228601000110]
10. Das Prinzip der Selbstverwaltung.: Bundesministerium für Gesundheit; 2021. URL: https://www.bundesgesundheitsministerium.de/gesundheitswesen-selbstverwaltung.html [accessed 2021-10-19]
11. Green P, Pashayeva A. The concept of the expert organisation. In: Human Resource Management in Higher Education. Krems: Danube University Krems; 2014:8-10.
12. Meisterjahn C, Krins C, Koch J. Befähigung und Begleitung unternehmensinterner Change Enabler als Wegbereiter und Triebfedern der Digitalisierung. In: Bosse CK, Zink KJ, editors. Arbeit 4.0 im Mittelstand. Berlin, Heidelberg: Springer Berlin Heidelberg; 2019:105-120.
13. Schmerfeld K, Schmerfeld J. Interprofessionelle Kooperation im Krankenhaus 1 - Problembeschreibungen. Jahrbuch für Kritische Medizin 2000:1-21.
14. Keckley P, Coughlin S, Stanley E. Survey of U.S. Physicians: Physician Perspectives About Health Care Reform and the Future of the Medical Profession. Washington: Deloitte Center for Health Solutions; 2013.
15. Lüder S. eHealth und Telematikinfrastruktur: Was kommt 2021 auf uns zu?. 2020. URL: https://freie-aerzteschaft.de/ehealth-und-telematikinfrastruktur/ [accessed 2021-10-19]
16. Ärzteblatt. Klage gegen Kostenerstattung beim Betrieb des TI-Konnektors abgelehnt. 2020. URL: https://tinyurl.com/a3v744tm [accessed 2021-10-19]
17. Albrecht M, Temizdemir E, Nissing M, Bock H. Stand und Perspektiven der Digitalisierung der vertragsärztlichen und -psychotherapeutischen Versorgung. In: Praxisbarometer Digitalisierung 2019. Berlin: Kassenärztliche Bundesvereinigung; Oct 25, 2019.
18. Hartmannbund. Umfrage des Hartmannbundes unter Assistenzärzten. In: Ärztliche Arbeitswelten. Heute. Und Morgen. Berlin: Hartmannbund; 2017.
19. Meister S, Burmann A, Deiters W. Digital health innovation engineering enabling digital transformation in health care: introduction of an overall tracking tracing at the super hospital Aarhus Denmark. In: Urbach N, Röglinger M, editors. Digitalization Cases. Basel: Springer International Publishing; 2019:329-341.
20. Schallmo D, Williams C. Digital Transformation Now!. Basel: Springer International Publishing; 2018.
21. Wolf T, Strohschen J. Digitalisierung: Definition und Reife. Informatik Spektrum 2018 Jan 22;41(1):56-64. [doi: 10.1007/s00287-017-1084-8]
22. Bordeleau F, Felden C. Digitally transforming organisations: a review of change models of industry 4.0. 2019 Presented at: 27th European Conference on Information Systems (ECIS); June 8-14, 2019; Stockholm & Uppsala.
23. Conrad C. Organisation Krankenhaus - Balanceakt zwischen Spezialisierung und Koordination. In: Goepfert A, Conrad CB, editors. Unternehmen Krankenhaus. Stuttgart: Georg Thieme Verlag KG; 2013:107-122.
24. Fuchs-Frohnhofen P, Esser N, Kurt-Georg C, Warner N, Müller P. Chancen und Risiken des Einsatzes digitaler Technologien in der Altenpflege. Würselen: MA&T Sell & Partner GmbH; 2020.
25. Child J. Organization: Contemporary principles and practices. Southern Gate, Chinchester, West Sussex, UK: Wiley; 2015.
26. Wanous JP, Reichers AE, Austin JT. Cynicism about organizational change. Group & Organization Management 2016 Jul 26;25(2):132-153. [doi: 10.1177/1059601100252003]
27. Helsper EJ, Eynon R. Digital natives: where is the evidence? British Educational Research Journal 2010 Jun;36(3):503-520. [doi: 10.1080/01411920902989227]
28. Hülür G, Macdonald B. Rethinking social relationships in old age: digitalization and the social lives of older adults. Am Psychol 2020;75(4):554-566. [doi: 10.1037/amp0000604] [Medline: 32378949]
29. Bräutigam C, Enste P, Evans M, Hilbert J, Merkel S, Öz F. Digitalisierung im Krankenhaus: Mehr Technik - bessere Arbeit?. Düsseldorf: Hans-Böckler-Stiftung; 2017.
30. Dillman DA, Tortora RD, Bowker D. Principles for constructing web surveys. 1998 Presented at: Joint Meetings of the American Statistical Association; Aug 9, 1998; Dallas, TX URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.502.8329&rep=rep1&type=pdf
31. Schleyer TK, Forrest JL. Methods for the design and administration of web-based surveys. J Am Med Inform Assoc 2000;7(4):416-425 [FREE Full text] [doi: 10.1136/jamia.2000.0070416] [Medline: 10887169]
32. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
33. Karrer K, Glaser C, Clemens C, Bruder C. Technikaffinität erfassen? der Fragebogen TA-EG. In: Lichtenstein A, Stößel C, Clemens C, editors. Der Mensch im Mittelpunkt technischer Systeme. Düsseldorf: VDI Verlag GmbH; 2009:196-201.
34. Kiliç T. Digital hospital: an example of best practice. IJHRSP 2016 Jun 29;1(2):52-58. [doi: 10.23884/ijhsrp]
35. Burmann A, Deiters W, Meister S. Digital maturity of hospitals in practice: a qualitative design-approach. In: The 29th European Conference on Information Systems (ECIS). Marrakech, Morocco; 2021 Presented at: European Conference on Information Systems; June 14-16, 2021; Virtual.
36. Digitale Wirtschaft und Gesellschaft: Industrie 4.0. Bundesministerium für Bildung und Forschung - BMBF. 2018. URL: https://www.bmbf.de/de/zukunftsprojekt-industrie-4-0-848.html [accessed 2021-10-19]

37. Burmann A. Bedeutung, Chancen und Risiken des digitalen Krankenhauses. In: Tagungsband . Hamek - Kongress für Medizin- und Krankenhaustechnik Hamburg. 2017 Presented at: Kongress für Medizin- und Krankenhaustechnik; Sep 20, 2017; Hamburg, Germany p. 7-1.
38. Kuhn S, Bartmann F, Klapper B, Schwenk U. Neue Gesundheitsberufe für das digitale Zeitalter. In: Projektbericht. Berlin: Stiftung Münch; 2020.
39. Digitale-Versorgung-Gesetz: DGV. Berlin: Bundegesetzblatt Teil I; 2019:2562.
40. Sheskin DJ. Handbook of Parametric and Nonparametric Statistical Procedures. USA: Taylor & Francis Inc; 2011.
41. Pearson K. VII. Note on regression and inheritance in the case of two parents. Proc. R. Soc. Lond 1997 Jan 31;58(347-352):240-242. [doi: 10.1098/rspl.1895.0041]
42. Winter JD, Dodou D. Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. Practical Assessment, Research, and Evaluation 2010;15:1.
43. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hoboken: Taylor and Francis; 2013.
44. Siegel S, Castellan N. Nonparametric statistics for the behavioral sciences, 2nd ed. Boston, Mass: McGraw-Hill; 2003:0070573573.
45. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23:151-169. [doi: 10.1146/annurev.publhealth.23.100901.140546] [Medline: 11910059]
46. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess 2003 Feb;80(1):99-103. [doi: 10.1207/S15327752JPA8001_18] [Medline: 12584072]
47. Boone H, Boone D. Analyzing Likert data. Journal of Extension 2012;50:6 [FREE Full text]
48. Liu M, Wronski L. Examining completion rates in web surveys via over 25,000 real-world surveys. Social Science Computer Review 2017 Feb 23;36(1):116-124. [doi: 10.1177/0894439317695581]
49. Bethlehem J. Selection bias in web surveys. International Statistical Review 2010;78(2):161-188. [doi: 10.1111/j.1751-5823.2010.00112.x]
50. Crowder J, Carbone J, Demijohn R. Multidisciplinary Systems Engineering: Architecting the Design Process. Basel: Cham, Springer International Publishing; 2016.

Abbreviations

BJÄ: Bündnis Junge Ärzte (Alliance of Young Physicians)
DGV: Digitale-Versorgung-Gesetz (digital health care act)
DiGA: Digitale Gesundheitsanwendungen (digital health applications)
GDPR: General Data Protection Regulation

Edited by G Eysenbach; submitted 24.06.21; peer-reviewed by V Castillo; comments to author 16.07.21; revised version received 20.07.21; accepted 19.09.21; published 11.11.21.

Please cite as:
Burmann A, Tischler M, Faßbach M, Schneitler S, Meister S
The Role of Physicians in Digitalizing Health Care Provision: Web-Based Survey Study
JMIR Med Inform 2021;9(11):e31527
URL: https://medinform.jmir.org/2021/11/e31527
doi:10.2196/31527
PMID:34545813

Original Paper

Assimilation of Medical Appointment Scheduling Systems and Their Impact on the Accessibility of Primary Care: Mixed Methods Study

Guy Paré[1], PhD; Louis Raymond[2], PhD; Alexandre Castonguay[1], PhD; Antoine Grenier Ouimet[3], MSc; Marie-Claude Trudel[1], PhD

[1]Department of Information Technologies, HEC Montréal, Montréal, QC, Canada

[2]École de gestion, Université du Québec à Trois-Rivières, Trois-Rivières, QC, Canada

[3]Smith School of Business, Queen's University, Kingston, QC, Canada

Corresponding Author:
Alexandre Castonguay, PhD
Department of Information Technologies
HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal, QC, H3T 2A7
Canada
Phone: 1 514 340 6812
Email: alexandre.castonguay@hec.ca

Abstract

Background: The COVID-19 pandemic has prompted the adoption of digital health technologies to maximize the accessibility of medical care in primary care settings. Medical appointment scheduling (MAS) systems are among the most essential technologies. Prior studies on MAS systems have taken either a user-oriented perspective, focusing on perceived outcomes such as patient satisfaction, or a technical perspective, focusing on optimizing medical scheduling algorithms. Less attention has been given to the extent to which family medicine practices have assimilated these systems into their daily operations and achieved impacts.

Objective: This study aimed to fill this gap and provide answers to the following questions: (1) to what extent have primary care practices assimilated MAS systems into their daily operations? (2) what are the impacts of assimilating MAS systems on the accessibility and availability of primary care? and (3) what are the organizational and managerial factors associated with greater assimilation of MAS systems in family medicine clinics?

Methods: A survey study targeting all family medicine clinics in Quebec, Canada, was conducted. The questionnaire was addressed to the individual responsible for managing medical schedules and appointments at these clinics. Following basic descriptive statistics, component-based structural equation modeling was used to empirically explore the causal paths implied in the conceptual framework. A cluster analysis was also performed to complement the causal analysis. As a final step, 6 experts in MAS systems were interviewed. Qualitative data were then coded and extracted using standard content analysis methods.

Results: A total of 70 valid questionnaires were collected and analyzed. A large majority of the surveyed clinics had implemented MAS systems, with an average use of 1 or 2 functionalities, mainly “automated appointment confirmation and reminders” and “online appointment confirmation, modification, or cancellation by the patient.” More extensive use of MAS systems appears to contribute to improved availability of medical care in these clinics, notwithstanding the effect of their application of advanced access principles. Also, greater integration of MAS systems into the clinic’s electronic medical record system led to more extensive use. Our study further indicated that smaller clinics were less likely to undertake such integration and therefore showed less availability of medical care for their patients. Finally, our findings indicated that those clinics that showed a greater adoption rate and that used the provincial MAS system tended to be the highest-performing ones in terms of accessibility and availability of care.

Conclusions: The main contribution of this study lies in the empirical demonstration that greater integration and assimilation of MAS systems in family medicine clinics lead to greater accessibility and availability of care for their patients and the general population. Valuable insight has also been provided on how to identify the clinics that would benefit most from such digital health solutions.

(JMIR Med Inform 2021;9(11):e30485) doi:10.2196/30485

KEYWORDS

medical appointment scheduling system; electronic booking; e-booking; primary care; accessibility of care; availability of care; advance access; electronic medical record

Introduction

The accessibility of primary care services remains a global concern, and to address the underlying issues, considerable improvements must be made to health systems around the world [1,2]. Even well-developed universal health care systems that have been adopted by the most economically advanced and politically stable countries are not exempt from the need for major improvements [3].

Primary care clinics have adopted several organizational and technological solutions to maximize accessibility to care services. One such organizational solution used in primary care settings is called advanced access scheduling, a popular patient-centered scheduling model that enables patients to seek and receive medical care at a time that suits their needs and from the health care provider of their choice [4,5]. The main idea is to manage supply and demand efficiently by applying 5 key principles: balance supply and demand, reduce the backlog, review the appointment system, create contingency plans, and integrate interprofessional practices [6-9].

For their part, digital health technologies supporting medical appointment scheduling (MAS) represent innovations whose integration is critical for primary care clinics looking to improve accessibility [10-12]. Compelled to manage in-person and virtual consultations, family medicine practices face considerable challenges. MAS systems, also called electronic booking systems, are digital solutions that enable practices to streamline their management of medical scheduling through functionalities that optimize physician schedules according to preset parameters [13]. These systems also propose a convenient and accessible means for patients to manage their appointments with their care provider while facilitating online communication [14].

Health care authorities in some countries such as the United Kingdom and Canada have also sought to deploy interoperable medical appointment scheduling (iMAS) systems. Although MAS systems are proprietary, off-the-shelf software solutions that are customized to the needs of medical practices, iMAS systems are province-wide or country-wide platforms. Put simply, an iMAS can be defined as a one-stop, free, and centralized interface that allows individuals to book appointments at medical clinics of their choice. iMAS solutions extract data from proprietary MAS systems used in primary care clinics to display availabilities across clinics and regions. In Quebec, where this study was conducted, medical practices were encouraged to use the provincial iMAS system (Rendez-vous Santé Québec) instead of a proprietary MAS software solution. The deployment of centralized iMAS platforms aimed to increase the accessibility of medical care to the general population, including “orphan” patients who are not registered in any clinic, and to monitor the performance of primary care clinics regarding medical scheduling and appointments [15].

Considering the sizable impact of the COVID-19 pandemic on primary care and the well-known benefits of digital health tools to provide quality health care despite physical distancing, current data on the use of MAS systems and their impacts on access to family medicine care are essential. Prior empirical studies on this topic have mostly taken a user-oriented perspective, focusing on perceived patient outcomes such as satisfaction and individual impacts (eg, reduced waiting times) associated with various system features and user characteristics [16-19]. Further, MAS systems have mainly been studied from a technical perspective, focusing on optimizing medical scheduling algorithms [20,21]. However, less attention has been given to the extent to which family medicine practices have assimilated these systems and achieved impacts. Our study attempts to fill this gap. In particular, the assimilation of MAS systems is conceptualized as their integration into medical practices and extended use in their daily operations [22-24]. Moreover, the impacts of MAS assimilation are conceptualized as the accessibility and availability of medical care, a major dimension of organizational performance in family medicine settings [25]. In summary, this exploratory study aimed to provide answers to the following questions: (1) to what extent have primary care practices assimilated MAS systems into their daily operations? (2) what are the impacts of assimilating MAS systems on the accessibility and availability of primary care? and (3) what are the organizational and managerial factors associated with greater assimilation of MAS systems in family medicine clinics?

Methods

This research was designed as a mixed methods study that included both a quantitative and a qualitative component. It received final approval from HEC Montréal’s ethics committee on March 20, 2020 (#2020-3784).

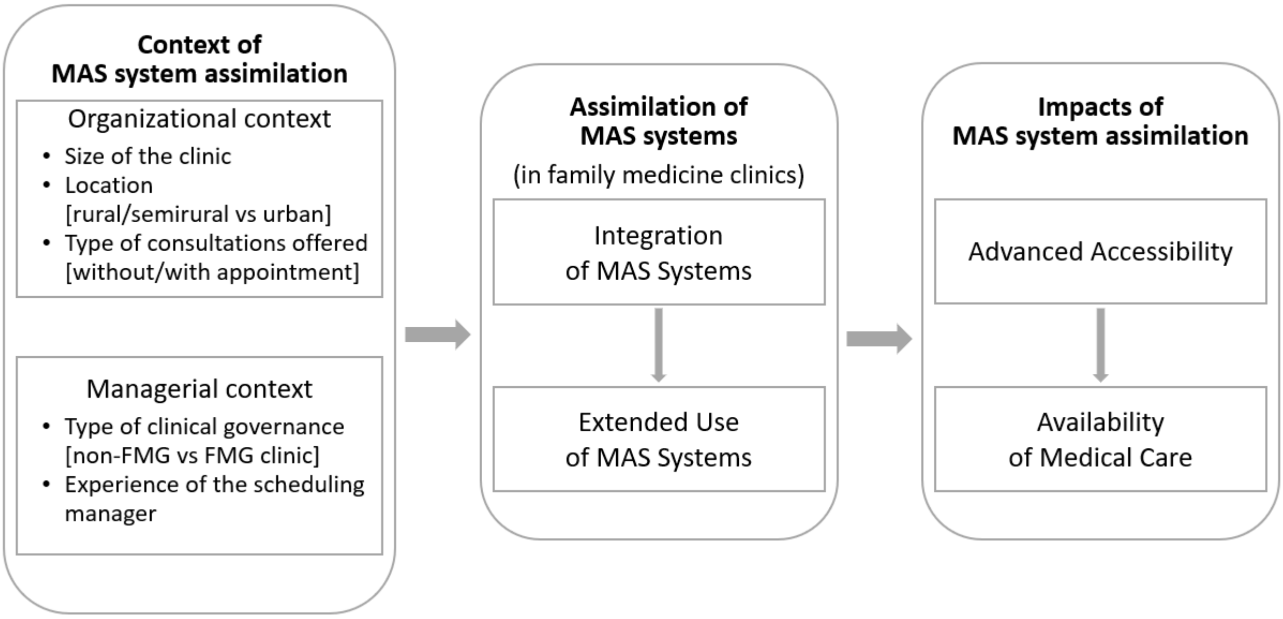
Theoretical Background

In line with the abovementioned research questions, a conceptual framework was developed to describe and explain the assimilation of MAS systems in primary care clinics, as well as the potential influencing factors and performance outcomes of such an assimilation, as shown in Figure 1. The framework is based on previous works on the information technology (IT) assimilation theory. The basic tenet of the IT assimilation theory is that using IT-based systems per se does not necessarily lead to improved organizational performance [22]. Instead, an integrated, extended, and innovative use of IT systems are expected to positively impact the performance of health care organizations [23]. Thus, our initial research proposition is that the more integrated and extended the use of MAS systems, the greater their positive impact on family medicine clinics in terms of the accessibility and availability of medical care [24].

Moreover, following the IT assimilation theory, we propose that a more integrated use of MAS systems will lead to more extended usage in clinics [24]. Furthermore, the theory assumes organizational munificence and managerial readiness, in terms of IT and non-IT resources and competencies, to be critical influencing factors of IT assimilation, with organizational size and managerial experience being often employed in empirical studies as proxies [13,14]. We thus propose that the clinics' organizational and managerial contexts will influence the integration and extended use of MAS systems. Importantly, based on previous work on organizational performance and advanced access scheduling in primary care settings [5], the last research proposition of our conceptual framework is that greater accessibility will lead to greater availability of medical care to the clinics' patients and to the population at large.

Following the preceding theoretical propositions, the conceptual framework guided the design of the survey instrument administered to find answers to the research questions presented in Figure 1. Thus, the operationalization of the following 6 research constructs originates from extant literature on health IT assimilation in general and MAS systems in particular: organizational context [13,22,23], managerial context [13,24,26], integration of MAS systems [22,23], extended use of MAS systems [13,20,24], availability of medical care [13,17,21,25,26], and advanced accessibility [6-9,27].

Figure 1. Conceptual framework. MAS: medical appointment scheduling, FMG: family medicine group.


Context of MAS system assimilation
Organizational context
• Size of the clinic
• Location [rural/semirural vs urban]
• Type of consultations offered [without/with appointment]
Managerial context
• Type of clinical governance [non-FMG vs FMG clinic]
• Experience of the scheduling manager
Assimilation of MAS systems
(in family medicine clinics)
Integration of MAS Systems
Extended Use of MAS Systems
Impacts of MAS system assimilation
Advanced Accessibility
Availability of Medical Care

Mixed Methods Research Design

This study was primarily designed as a web-based survey. As described below, we followed best practices in web-based survey methodologies [28,29]. A pretest of the questionnaire was conducted at 5 primary care clinics, resulting in minor adjustments being made to the survey instrument. As a complementary qualitative approach, we interviewed 6 experts in MAS systems between February 12 and March 16, 2021. These experts include 2 physicians working full time in medical clinics, 1 family medicine group (FMG) administrator, 2 managers of the Quebec government health insurance system (Régie de l'Assurance Maladie du Québec), and 1 private MAS solution provider. Based on their extensive experience in Quebec's medical sector, these respondents were asked to provide their interpretation of this study's findings and evaluate whether these results were representative of the situation in medical clinics throughout the province. The web-based interviews were recorded and then transcribed. Qualitative data were coded and extracted using standard content analysis methods [30] and the NVivo software (version 1.4, QSR International). The experts' insights were used mainly to enrich our discussion and interpretation of the survey findings.

Sampling Procedure

The quantitative, survey-based component of our study targeted the 358 FMGs operating in Quebec, Canada. FMGs are primary care clinics structured and organized to provide Quebec's population with greater accessibility to health care services. The basic tenets of FMG governance include teamwork and interdisciplinarity in the rendering of health care services to patients, registration of each patient to 1 family practitioner within the group for care and follow-up, and providing an option that allows registered patients to benefit from services that are integrated and offered nearby with extended office hours. Deployed on the Qualtrics web-based survey platform [31], the survey questionnaire was addressed to the individual responsible for managing medical schedules and appointments at these clinics. An invitation to participate in the study was sent on November 10, 2020, through a health care services newsletter whose distribution list comprises the FMGs in Quebec. The invitation contained a hyperlink directing the participants to the questionnaire through a secure website. This invitation was renewed 14 days later via the same newsletter.

Statistical Analysis

Following basic descriptive statistics, component-based structural equation modeling (SEM) was used to empirically explore the causal paths implied in our conceptual framework.

As implemented in SmartPLS software (version 2.0, SmartPLS GmBH), the partial least squares (PLS) SEM technique was chosen for its robustness regarding the distribution of residuals and its greater affinity for exploratory rather than confirmatory research purposes when compared to covariance-based SEM techniques [32]. A cluster analysis was also performed to complement the causal analysis.

Results

Sample Characteristics

A total of 70 valid questionnaires were collected from November 11 to December 20, 2020. Specifically, 60 were received from FMGs (17% response rate) and 10 from non-FMG clinics. Low response rates raise the possibility of nonresponse bias, so the potential for such bias was assessed by comparing the 47 "late" respondents (ie, those that responded after more than 14 days following the initial invitation) with the 23 "early" respondents [33]. An analysis of variance found no significant differences between these two sets of responding clinics in terms of the context and impacts of their assimilation of MAS systems, thus indicating a low probability of the presence of nonresponse bias in the study.

Of the 70 family medicine clinics sampled, 18 (26%) used a proprietary MAS software solution and Quebec's iMAS provincial solution, 24 (36%) used only the iMAS system, 16 (23%) used only a proprietary MAS system, and 12 (17%) used no such system. At present, primary care clinics in Quebec are encouraged to use the iMAS system but are not obligated to do so, but this has not always been the case. Some FMGs have been obliged to use Quebec's iMAS solution since its initial deployment. In the iMAS and MAS systems, the most used functionalities are "automated appointment confirmation and reminders" and "online appointment confirmation, modification, or cancellation by the patient." The least used functionality was "invoicing compensatory fees for missed appointments." Moreover, in the 42 clinical settings in which the iMAS system has been implemented, it was integrated into the clinics' electronic medical record (EMR) system. Data transfers between the 2 systems were automated in most cases (n=38, 90%). Among the 34 clinics using the MAS system, 23 (67%) clinics had integrated it into their EMR system. The descriptive statistics of the research variables are presented in Tables 1 and 2.

We then characterized the 70 participating clinics based on the contexts and impacts of their assimilation of MAS systems. First, in terms of the organizational context, most (n=40, 57%) of the clinics were large (25,000 or more consultations per year), whereas small clinics (fewer than 5000 consultations per year) were the least represented in the sample (n=6, 8%). Many of the clinics (n=27, 39%) were located in rural, semirural, or remote regions. In terms of their openness to receiving "walk-in" patients, a significant proportion (n=16, 23%) of the clinics reported that half or more of their medical consultations were made without a prior appointment. Second, in terms of the managerial context, 86% (n=60) of the sampled clinics were found to be FMGs. Moreover, most (n=48, 69%) of the managers responsible for medical scheduling in the sampled clinics had 4 or more years of experience in their position.

More than half (n=36, 51%) of the 70 sampled clinics had either not implemented advanced access scheduling (14/70, 20%) or had done so in a tentative manner (22/70, 31%) by applying only 1 or 2 of its core principles (Table 2). In this regard, "rethinking the appointment system" by opening physicians' working hours over a period of approximately 2 to 4 weeks was the only principle to have been applied by more than half of the sampled clinics (n=45, 64%). Approximately half of the sampled clinics reported making efforts to balance supply and demand as well as incorporate interprofessional practices (directing the patient to the right professionals according to their needs). Only a minority (n=11, 16%) of the clinics reported having prepared contingency plans for overflow periods.

The scheduling performance of the 70 clinics varied significantly, as 31% (n=22) scored themselves at ≤3 and 23% (n=16) at ≥4 on a 5-point scale, where 5 represents the highest performance level. On average, the sampled clinics considered their performance to be at its worst during enrollment of "orphan" patients, and their performance to be at its best when managing physician schedules. Overall, the respondents tended to think that their clinics performed at a rather high level. Moreover, the patient attendance rate was perceived to range from 80% to 89% for most of the sampled clinics (n=40, 57%), with a significant proportion (n=16, 23%) reaching an attendance rate of 90% or more.

Table 1. Descriptive statistics of the research variables (N=70).

Variable	Value
Organizational context	
Size of the clinic in terms of number of consultations, n (%)	
Less than 5000	6 (8)
5000 to 9999	11(16)
10,000 to 14,999	13 (19)
15,000 to 19,999	23 (33)
20,000 to 24,999	6 (8)
More than 25,000	11 (16)
Location of the clinic, n (%)	
Nonurban (rural, semirural, or remote)	27 (39)
Urban	43 (61)
Type of consultations offered (without appointment), n (%)	20 (28)
Managerial context	
Type of clinical governance, n (%)	
FMG[a]	49 (70)
Non-FMG	21 (30)
Experience of the scheduling manager, n (%)	
Less than 1 year	5 (7)
1 to 3 years	17 (25)
4 to 6 years	22 (31)
7 to 9 years	7 (10)
More than 10 years	19 (27)
Integration of MAS[b] systems, n (%)	
Clinic with implemented systems	
EMR[c]	69 (98)
iMAS[d]	42 (60)
iMAS integrated with the EMR	38 (54)
MAS	34 (49)
MAS integrated with the EMR	23 (33)
Integration of iMAS and MAS systems with the EMR	18 (26)
Extended use of MAS systems, mean (SD)	
iMAS system (RVSQ[e]) functionalities used[f]	0.8 (1.0)
MAS system functionalities used[f]	1.6 (2.2)
Advanced accessibility, mean (SD)	
Advanced access scheduling principles applied[f]	2.4 (1.8)
Availability of medical care, mean (SD)	
Scheduling performance[g,h]	3.4 (0.7)
Patient attendance[i]	1.6 (1.0)

[a] FMG: family medical group.

[b] MAS: medical appointment scheduling.

[c] EMR: electronic medical record.

[d] iMAS: interoperable medical appointment scheduling.

[e] RVSQ: Rendez-vous Santé Québec.

[f] See Table 2 for the distribution of this variable.

[g] Cronbach alpha coefficient of reliability (α=.76).

[h] 1=totally disagree, 2=rather disagree, 3=neither disagree nor agree, 4=rather agree, and 5=totally agree.

[i] 0=less than 80%, 1=80% to 84%, 2=85% to 89%, 3=90% to 94%, and 4=95% or more.

Table 2. Operationalization and distribution of the research variables (N=70).

Variable	Value
iMAS[a] system functionalities used, n (%)	
Automated appointment confirmation and reminder by email, SMS[b] text messaging, or telephone	28 (67)
Confirmation, modification, or cancellation of the appointment via the internet by the patient	21 (50)
Invoicing compensatory fees for missed appointments	1 (2)
Optimization of web-based appointment scheduling according to predetermined parameters (eg, adapted access, patient attendance, and avoidance of gaps in the schedule)	9 (21)
MAS[c] system functionalities used, n (%)	
Offer of appointments by automated telephone messages	20 (59)
Automated appointment confirmation and reminder(s) by email, SMS text messaging, and telephone	22 (65)
Confirmation, modification, or cancellation of the appointment via the internet by the patient	21 (62)
Internet-based preconsultation questionnaire completed by the patient (reason for the consultation)	12 (35)
Invoicing compensatory fees for missed appointments	5 (15)
Optimization of web-based appointment scheduling according to predetermined parameters (eg, advanced access, attendance, avoidance of gaps in the schedule, and automated appointment scheduling for patients on the waiting list)	10 (29)
Restriction of the appointment offer for certain patients (registered vs unregistered)	20 (59)
Advanced access scheduling principles applied, n (%)	
Balancing supply and demand	34 (49)
Reducing accumulated backlog[d]	27 (39)
Rethinking the appointment system[e]	45 (64)
Developing contingency plans	
Schedule planning based on absences	18 (26)
Planning for overflow periods	11 (16)
Incorporating interprofessional practice	32 (46)
Scheduling performance[f], mean (SD)	
The number of missed appointments (“no-shows”) at my clinic is not a problem	3.4 (1.1)
My clinic is still enrolling a large number of “orphan” patients	3.0 (1.3)
The management of schedules by the administrative staff is very efficient	3.7 (1.0)
Web-based appointment booking by the administrative staff is very efficient	3.3 (1.2)
The satisfaction of the administrative staff in my clinic with regard to scheduling and making appointments is very high	3.3 (1.1)
The satisfaction of the doctors in my clinic with regard to scheduling and making appointments is very high	3.6 (1.0)
Patient satisfaction in my clinic with the way appointment scheduling works is very high	3.3 (1.1)
Registered patients may obtain a consultation at my clinic within a very short time	3.6 (1.2)
Unregistered patients may obtain a consultation at my clinic within a very short time	3.3 (1.3)

[a] iMAS: interoperable medical appointment scheduling.

[b] SMS: short message service.

[c] MAS: medical appointment scheduling.

[d] Using patient empowerment (eg, patient confirmation of appointment, missed appointment fee).

[e] Opening hours over a period of approximately 2 to 4 weeks.

[f] 1=totally disagree, 2=rather disagree, 3=neither disagree nor agree, 4=rather agree, and 5=totally agree.

Causal Analysis

In this study, all 6 research constructs were modeled as being “formative” given their composite and multidimensional nature (Figure 1). The first step in the data analysis consists of simultaneously estimating the measurement and theoretical models using the PLS SEM technique [32].

Assessment of the Measurement Model

The metric properties of the research constructs were assessed within the context of the theoretical model. As the standard reliability and validity criteria applicable to reflective constructs do not apply to formative constructs, one must first verify that there is no collinearity among the formative construct's indicators. To do so, the variance inflation factor (VIF) statistic is used, the rule being that the VIF must not be greater than 3.3 [34]. As presented in Table 3, the VIF values for all 11 indicators (research variables) were below this threshold (ranging between 1.00 and 1.12), confirming the absence of any multicollinearity. Once the validity of the measures has been assessed, the last property to be verified is discriminant validity, which shows the extent to which each research construct, as measured, is unique and different from the other five. This validity was confirmed by the fact that each construct shared less than 50% of its variance with any other construct (an interconstruct correlation of >0.7), as shown in Table 3.

Table 3. Validity of the research constructs[a].

Research construct	Construct indicators[b]			Interconstruct correlation matrix				
	VIF_1	VIF_2	VIF_3	1	2	3	4	5
1. Organizational context	1.11	1.12	1.02	—[c]	—	—	—	—
2. Integration of MAS[d] systems	1.00	—	—	0.46	—	—	—	—
3. Managerial context	1.00	1.00	—	–0.63	–0.44	—	—	—
4. Extended use of MAS systems	1.05	1.05	—	0.51	0.64	–0.38	—	—
5. Advanced accessibility	1.00	—	—	0.19	0.29	–0.18	0.14	—
6. Availability of medical care	1.00	1.00	—	0.54	0.21	–0.35	0.26	0.39

[a] Assessing composite reliability and average variance extracted is inappropriate for formative constructs.

[b] VIF_i: variance inflation factor of the construct's *i*th indicator.

[c] Not applicable.

[d] MAS: medical appointment scheduling.

Assessment of the Theoretical Model

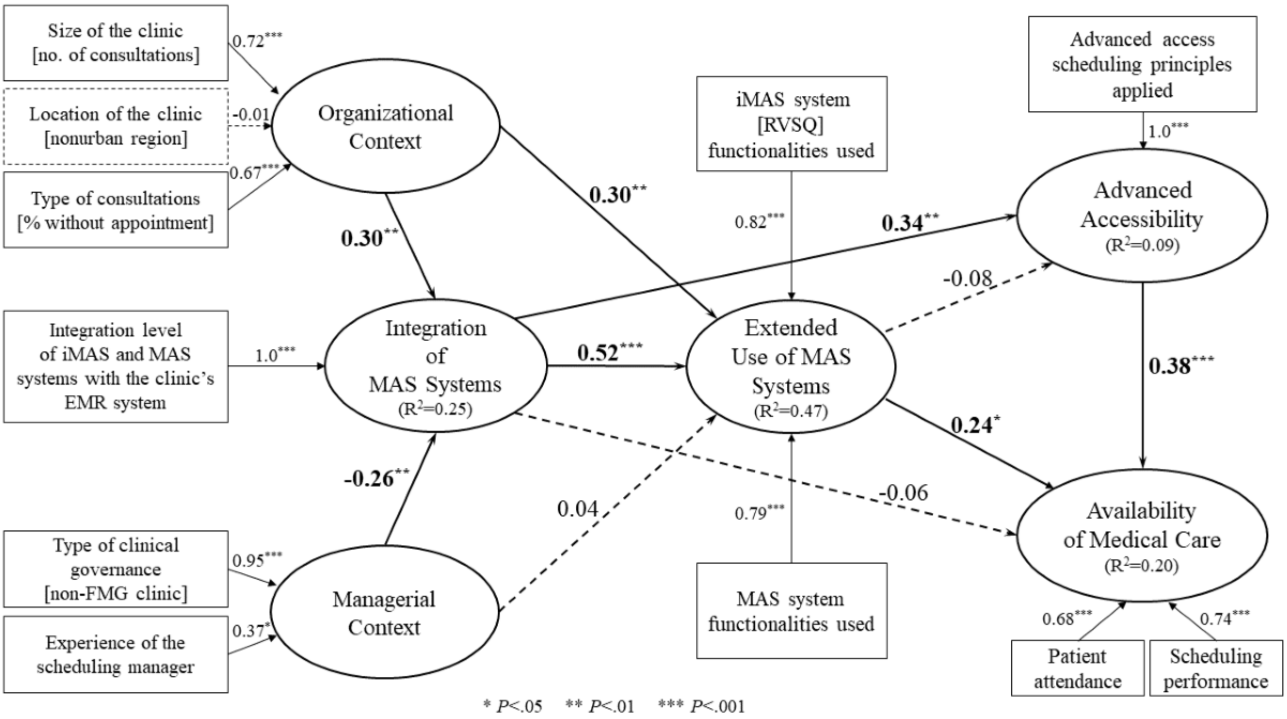
As shown in Figure 2, the relationships inferred from the conceptual framework were tested by assessing the path coefficients (β) estimated by the SEM procedure, as executed by SmartPLS. The performance of the theoretical model highlighting interrelationships among the 6 research constructs was assessed by the strength and significance of the path coefficients (β) and the proportion of explained variance (R^2), befitting the focus of the PLS method on the prediction and generalization [34].

Given the results of the causal analysis provided by the SEM procedure, the initial finding concerned the positive and significant path coefficients linking the sampled clinics' organizational context to their integration (β=.3, *P*=.009) and extended use (β=.3, *P*=.009) of MAS systems. Hence, family medicine clinics that were larger in size and more open to walk-in patients showed greater assimilation of MAS technology in their daily operations. The second finding was that differences in the managerial context in terms of the clinics' governance and the level of experience of their scheduling manager directly influenced their integration of MAS systems with their EMR systems (β=.26, *P*=.006). Here, the data showed that clinics whose governance was not of the FMG type and whose scheduling manager was more experienced had a lower level of system integration. Although the managerial context was found to have no direct effect on the extended use of MAS systems (β=.04, *P*=.38), it was nevertheless shown to have an indirect effect [35] through the mediating effect of MAS system integration.

Our results also suggested that the main precursor to the extended use of MAS systems is the sampled clinics' integration of these systems with their EMR systems (β=.52, *P*<.001). Moreover, their integration of MAS systems with their EMR systems was also found to have a positive impact on their use of advanced access scheduling principles (β=.34, *P*=.005). Greater system integration appears to enable increased application of advanced access scheduling principles by these clinics, such as interprofessional work and joint monitoring. Although integration had no direct effect on the availability of medical care (*P*=.39), it did have significant indirect effects through the mediating impacts of advanced accessibility and extended use of MAS systems.

A positive and significant path coefficient was found between the use of MAS systems by family medicine clinics and the availability of medical care in these clinics (β=.24, *P*=.047). In contrast with the effect associated with the integration of MAS systems, the clinics' use of MAS system functionalities appeared to have no significant impact on the extent to which they implemented advanced access scheduling principles (*P*=.28). One additional finding on the impacts of MAS system assimilation in family medicine clinics highlighted the role of advanced access scheduling as a precondition to increased availability of medical care in these clinics (β=.38, *P*<.001).

Figure 2. Results of the causal analysis. iMAS: interoperable medical appointment scheduling, MAS: medical appointment scheduling, EMR: electronic medical record, FMG: family medicine group, RVSQ: Rendez-vous Santé Québec.


Size of the clinic [no. of consultations]
0.72***
Location of the clinic [nonurban region]
-0.01
Type of consultations [% without appointment]
0.67***
Organizational Context
0.30**
0.30**
Integration level of iMAS and MAS systems with the clinic's EMR system
1.0***
Integration of MAS Systems (R^2=0.25)
0.52***
iMAS system [RVSQ] functionalities used
0.82***
Extended Use of MAS Systems (R^2=0.47)
0.34**
-0.08
Advanced access scheduling principles applied
1.0***
Advanced Accessibility (R^2=0.09)
0.38***
0.24*
-0.06
-0.26**
0.04
0.79***
Type of clinical governance [non-FMG clinic]
0.95***
Experience of the scheduling manager
0.37*
Managerial Context
MAS system functionalities used
Availability of Medical Care (R^2=0.20)
0.68***
0.74***
Patient attendance
Scheduling performance
* P<.05 ** P<.01 *** P<.001

During causal analysis, we found significant and important indirect effects of the organizational context on the extended use of MAS systems (*P*=.03, 34% of total effects), namely through the integration of MAS systems, as shown in Table 4. There were also important indirect effects of integration on the availability of medical care (*P*=.10) through extended use. These findings highlight the "mediating" role played by the clinics' assimilation of MAS systems in improving organizational performance in terms of medical care accessibility and availability.

Table 4. Breakdown of the total effects of the research constructs.

Relationship between the research constructs	Direct effects	Indirect effects	Total effects
Organizational context → Integration of MAS[a] systems	0.295	0.000	0.295
Managerial context → Integration of MAS systems	–0.257	0.000	–0.257
Organizational context → Extended use of MAS systems	0.303	0.153	0.456
Managerial context → Extended use of MAS systems	0.041	–0.134	–0.093
Integration of MAS systems → Extended use of MAS systems	0.520	0.000	0.520
Organizational context → Advanced accessibility	0.000	0.064	0.064
Managerial context → Advanced accessibility	0.000	–0.081	–0.081
Integration of MAS systems → Advanced accessibility	0.258	0.042	0.300
Extended use of MAS systems → Advanced accessibility	–0.080	0.000	–0.080
Organizational context → Availability of medical care	0.000	0.114	0.114
Managerial context → Availability of medical care	0.000	–0.037	–0.037
Integration of MAS systems → Availability of medical care	–0.057	0.235	0.178
Extended use of MAS systems → Availability of medical care	0.235	–0.030	0.205
Advanced accessibility → Availability of medical care	0.377	0.000	0.377

[a] MAS: medical appointment scheduling.

Cluster Analysis

To further explain the assimilation of MAS systems into primary settings, we applied an alternative approach to analyze our survey data. We followed a "case-oriented" approach as a complement to the preceding "variable-oriented" approach [36]. The case-oriented or configurational approach makes no assumptions about the statistical distribution of the research variables and the linearity of the relationships between these variables [37]. As it is operationalized with methods such as cluster analysis, this approach is meant to provide all-encompassing and holistic views regarding the use of MAS systems by family medicine clinics. It allows us to verify whether these clinics are better represented as a whole or as members of distinct subgroups, and to explore why and in what regard. A cluster analysis was thus conducted to group the surveyed clinics based on profiles that characterize the impacts of their assimilation of MAS systems on the accessibility and availability of medical care. A 3-cluster solution was found to be the most interpretable and meaningful one in terms of identifying profiles that could be clearly distinguished from one another. Given the exploratory research nature of this study and the goal of finding unknown groups (ie, groups not explicitly labeled in the data), the k-means clustering algorithm was used (SPSS Quick Cluster procedure) [38].

As shown in Table 5, among the 70 clinics, 15 (21%) clinics in the first profile were named “low performance” or “low” clinics (low levels of accessibility and availability). A second group of 25 (36%) clinics was named “mixed performance” or “mixed” (a low level of accessibility but a high level of availability). The third profile consisted of 30 (43%) clinics that were named “high performance” or “high” (high levels of accessibility and availability).

Table 5. Descriptive statistics and analysis of variance results for clinic profiles (N=70).

Research construct	Clinic profiles[a]			*F* value	*P* value
	Low (n=15)	Mixed (n=25)	High (n=30)		
Organizational context, mean					
Size of the clinic[b] (number of consultations per year)	2.5#	5.6^	4.7*	21.8	<.001
Location of the clinic (1=rural/semirural, 0=urban)	0.3	0.3	0.5	0.7	.49
Type of consultations offered (% without appointment)	21.3	28.3	30.2	0.9	.92
Managerial context, mean					
Type of clinical governance (1=non-FMG[c], 0=FMG)	0.5^	0.0*	0.1*	11.1	<.001
Experience of the scheduling manager[d]	3.1	3.6	3.0	1.4	.25
Implementation of MAS[e] systems, mean					
iMAS[f] system implemented (1=yes, 0=no)	0.27*	0.64^	0.73^	5.2	.008
MAS system implemented (1=yes, 0=no)	0.47	0.52	0.47	0.1	.43
Extended use of MAS systems, mean					
iMAS system functionalities used	0.2#	1.1^	0.9^	4.2	.02
MAS system functionalities used	1.3	1.5	1.8	0.3	.76
Integration of MAS systems, mean					
Integration of iMAS and MAS systems with the EMR[g,h]	0.9*	2.0^	2.2^	7.4	<.001
Advanced accessibility, mean					
Advanced access scheduling principles applied	0.2#	1.5*	4.2^	118.3	<.001
Availability of medical care, mean					
Scheduling performance[i]	3.0*	3.4^	3.6^	4.9	.01
Patient attendance[j]	0.5#	2.3^	1.7*	23.8	<.001

[a] Within rows, different symbols (#, *, and ^) indicate significant (*P*<.05) pairwise differences between means (Tamhane T2 test).

[b] 1: less than 5000, 2: 5000 to 9999, 3: 10,000 to 14,999, 4: 15,000 to 19,999, 5: 20,000 to 24,999, and 6: more than 25,000.

[c] FMG: family medicine group.

[d] 1=less than 1 year, 2=1 to 3 years, 3=4 to 6 years, 4=7 to 9 years, and 5=more than 10 years.

[e] MAS: medical appointment scheduling.

[f] iMAS: interoperable medical appointment scheduling.

[g] iMAS system integration (no=0, manual=1, and automated=2) + MAS system integration (no=0, manual=1, and automated=2).

[h] EMR: electronic medical record.

[i] 1=totally disagree, 2=rather disagree, 3=neither disagree nor agree, 4=rather agree, and 5=totally agree.

[j] 0=less than 80%, 1=80% to 84%, 2=85% to 89%, 3=90% to 94%, and 4=95% or more.

To identify the technological, organizational, and managerial correlations of the sampled clinics' medical care accessibility and availability, we contextualized the 3 profiles that emerged from the cluster analysis, as shown in Table 3. First, it can be noted that all clinics showed a similar organizational and managerial context except for low clinics, which tended to present a significantly lower number of consultations per year (between 5000 and 9999 compared to between 15,000 and 24,999). This implies smaller-sized clinics and a higher proportion of non-FMG clinics (approximately 50% compared to approximately 10% or less) in comparison to the clinics in the mixed and high groups. Second, the low clinics presented the lowest levels of iMAS system implementation and functionality usage, as well as the lowest levels of integration of iMAS and MAS systems with their EMR systems. In this regard, the results obtained for the mixed and high groups were comparable. There were no differences between them in terms of MAS system implementation. Third, concerning advanced access principles, the clinics in the high group showed a significantly higher level of application (4 principles applied on average out of a maximum of 6), followed by the mixed group (1.5 principles applied on average), which is still significantly higher than the low group level (0.2 principles applied on average). Finally, in terms of medical care availability, the low group presented the lowest results for scheduling performance and patient attendance. Although the high and mixed groups presented similar levels of scheduling performance, the mixed group had the highest patient attendance rate.

Discussion

Main Results

Using a mixed methods approach, this study surveyed a sample of medical clinics with the objective of gaining further knowledge on their assimilation of MAS systems. We also sought to gain insight into the performance benefits obtained by the clinics from their use of such systems. Notwithstanding the small sample size, the 6 experts we consulted stated that the results of this study accurately portrayed the situation in the family medicine clinics in Quebec with regard to medical appointment scheduling.

Clinics' Assimilation of MAS Systems and Their Impacts on Accessibility of Care

The factors that contribute to the adoption of digital health technologies in general and MAS systems in particular are still unclear. An initial observation made in this study was that 17% of all clinics do not use any MAS system. A plausible explanation might be that some practices are more affected by the demographics of their patient population than others. In this line of thought, prior research found several patient-level barriers to the uptake of digital health technologies, including computer literacy, no or poor internet connection, and fear of using or lack of interest in technology [39,40]. Recent research also observed disparities in digital health technology adoption across sociodemographic subgroups, highlighting a persistent digital divide [41]. The rapid shift to digital care prompted by the COVID-19 pandemic demands research and action to ensure that underserved populations are not left behind [42].

Although MAS systems have been widely implemented in the surveyed clinics, these systems are only being used at a fraction of their full potential. Indeed, very few system functionalities, ranging from 1 to 2 on average, are being used. Clinics are mainly using those solutions that allow patients to manage their appointments via the internet and to automate appointment confirmations and reminders. Interestingly, the clinics that use more functionalities also tend to show improved performance in terms of accessibility of medical care. These findings are consistent with a previous study [13] showing that extended use of MAS systems in medical clinics is associated with improved care, as represented by an overall higher level of patient satisfaction and a lower level of missed appointments.

Another key observation is that despite considerable efforts to promote the use of advanced access principles throughout Quebec's primary care clinics, our results showed that the 5 advanced access principles are being rather weakly applied. In fact, the 2 main principles applied in the sampled clinics included opening the physicians' schedules over a "short" period of approximately 2 to 4 weeks, followed by balancing supply and demand, and incorporating interprofessional practice. A recently published study protocol [43] underscored the need for more detailed and representative data in this regard. Nevertheless, our study supports the idea that the greater the application of the advanced access principles, the higher the accessibility of medical services, and the better the clinics' patient attendance rate and scheduling performance.

As observed in Figure 2, the results of our causal analysis also confirmed the proposition that integrating the MAS and EMR systems is a precondition to the successful assimilation of MAS systems in family medicine clinics. This integration provides a facilitating technological context for clinics seeking more integrated and extended usage of these systems [44]. Moreover, to explore the organizational and managerial factors that drive family medicine clinics to assimilate MAS systems, a cluster analysis was performed as a complement to the preceding causal analysis.

Characterization of Clinics Urgently Requiring Improvements Through MAS Systems

The cluster analysis allowed us to holistically characterize the family medicine clinics whose organizational performance urgently requires improvement in terms of the accessibility and availability of medical care they provide to their patients and to the general population. In this regard, clinics were categorized into 1 of 3 performance profiles: low, mixed, or high (accessibility and availability). When comparing how each group differs from the other two, our analysis showed that clinics in the low group had an organizational, managerial, and technological context that differed substantially from the mixed and high groups. These clinics may be characterized as "late implementers" with regard to their implementation of the iMAS system and as "beginners" (as opposed to "advanced users") with regard to their assimilation of MAS systems. Low clinics also lagged behind in terms of MAS-EMR integration, irrespective of their interoperability, and their application of advanced access principles. In addition, they tended to be smaller in size and were not FMG clinics. Therefore, one may conclude that smaller clinics that are not FMGs are less likely to adopt and use an iMAS system or integrate their EMR into their MAS solution.

Although Quebec's iMAS system has always been free for both clinics and patients, MAS systems are proprietary solutions available at a cost for clinics and sometimes for patients. It could be that smaller clinics, having less financial resources at their disposal, are less likely to adopt new technological solutions because of their greater relative cost. However, we found no statistical difference between the 3 groups of clinics in terms of MAS adoption and use. In addition to the iMAS solution being free, considerable efforts have been made by Quebec's health care authorities to encourage clinics to adopt its centralized solution, including placing some "political" pressure on FMGs. In fact, we did find that a high proportion of FMGs had adopted Quebec's iMAS system.

The experts we interviewed proposed additional explanations as to why the low group of clinics had a lower iMAS implementation rate. Importantly, the overall rate of approximately 60% was in line with the experts' perceptions of the overall situation in family medicine clinics in Quebec. This could be explained in different ways. One is that the iMAS solution, although free, might be perceived by physicians as a means for the government to control how they manage their patient schedules. Irrespective of the validity of this perception, it appears to be real and widespread among general practitioners in Quebec. Physicians might also be reluctant to use an "imposed" MAS solution in a medical software market where other alternative solutions exist. It is also possible that such alternative solutions, albeit offered by the private sector, offer services and provide functionalities that are better suited to the needs of physicians and their clinics. Moreover, in Quebec, FMG clinics receive financial aid for supporting their digitalization that is based on their size. This might partially explain why larger FMG clinics tend to show a higher MAS implementation rate. According the experts' opinions, smaller clinics prefer having experienced administrative personnel manage their physicians' schedules, believing that they best know how to optimize each physician's schedule based on their patients' needs for care. Another possible explanation is that MAS systems provide true added value once the number of medical appointments reaches a certain threshold. This threshold may be the maximum number of daily appointments that a scheduling administrator can manage efficiently.

Since the clinics in the mixed group tended to be the largest and applied significantly fewer advanced access principles when compared to those in the high group, one could surmise from a strategic perspective that the larger clinics, operating at full capacity and being unable to cope with a growing demand for medical consultations, would want to limit rather than increase their accessibility, regardless of their assimilation of MAS systems. Although this last conjecture is in line with the surveyed experts' opinions, it needs to be supported by further research on the strategic management of health IT in primary care settings [45].

Contributions and Implications

The main contribution of this study is that it supports the idea that the adoption of MAS systems in family medicine clinics is by itself not sufficient to promote availability of care to patients and to the general population. Rather, it is the greater assimilation of the multiple functionalities of these systems that makes the difference in fostering patient attendance and enhancing scheduling performance. Whether the goal is to promote the application of advanced access principles in these clinics or their assimilation and the optimal use of MAS systems, further knowledge of the organizational, managerial, and, above all, technological factors that differentiate high-performing clinics from low-performing ones is required. This would allow clinic managers and physicians as well as consultants and governments to make better-informed plans and better-targeted recommendations on the development, promotion, adoption, and assimilation of MAS systems. Furthermore, this implies that it is mainly the family medicine clinics in the low group, and to a lesser extent in the mixed group, that should be targeted for improvement by national health authorities and health IT researchers and practitioners. Such efforts should support the following: (1) increased implementation of advanced access scheduling and (2) adoption and greater assimilation of MAS systems, particularly iMAS systems. This also implies that resources should be allocated to the low clinics for organizational learning purposes, and customized counseling and support should be offered throughout the MAS implementation process. As for the mixed clinics, counseling and support should lead them to optimize their use of MAS systems or iMAS solution. Another implication of our findings for future research lies in the renewed affirmation that in this digital health care era, the "assimilation" of medical systems [46]—rather than their mere adoption—and their "extended use" [47]—rather than mere use—are the key factors to achieving a better understanding of the impacts of such systems. Finally, this study makes a significant contribution toward research methodologies by analyzing the survey data through a combination of variable-oriented (SEM) and case-oriented (cluster analysis) approaches, thereby benefiting from the complementarity of these approaches to generate richer insights for researchers and practitioners [48].

Limitations

The results of this study must be interpreted with care due to some inherent limitations. First, although 17% represents a good response rate for an online survey and all the surveyed experts confirmed that the study results matched their knowledge of family medicine clinics in Quebec, the limited sample size calls for some caution with respect to the generalizability of our findings. The second limitation pertains to how the performance of the sampled clinics was assessed. Ideally, it would have been preferable to survey more than 1 respondent per clinic to minimize common-method bias. Third, asking scheduling managers to evaluate their own performance may have somewhat biased the results; hence, a more objective assessment could have been made if iMAS, MAS, and EMR system data were available. Fourth, an intrinsic limitation of survey research is related to its cross-sectional nature, as true causality cannot be inferred. Future studies on the assimilation and impacts of MAS systems would thus warrant longitudinal designs.

Conclusions

Notwithstanding the recent upheavals in health care brought about by the COVID-19 pandemic, access to and availability of primary care for populations remain a global issue. MAS systems are among the digital health solutions addressing these concerns by facilitating and optimizing the scheduling of medical appointments. As our data were collected in the very midst of the ongoing pandemic, our findings should be interpreted in light of the profound changes that have emerged in primary care settings in response to this crisis.

The main contribution of this study lies in its empirical demonstration that greater integration and assimilation of MAS systems in family medicine clinics lead to greater accessibility and availability of care for their patients and for the general population. Valuable insight has also been provided on how to identify clinics that would benefit the most from public- and private-sector initiatives to improve their efficiency and effectiveness as primary care providers for better primary care accessibility. Advancement of knowledge on this topic could benefit from similar studies carried out in contexts other than family medicine clinics, such as specialized or ambulatory care clinics.

Acknowledgments

We sincerely thank Isabelle Girard (Plakett Services Cliniques), Myriam Brel (Canada Health Infoway), and Dr Julien Quinty for their assistance with participant recruitment. We are also grateful to Canada Health Infoway for providing financial support for this study.

Conflicts of Interest

None declared.

References

1. Transforming Our World, the 2030 Agenda for Sustainable Development. United Nations. 2015. URL: https://sdgs.un.org/2030agenda [accessed 2021-11-08]
2. Healthy Systems for Universal Health Coverage: A Joint Vision for Healthy Lives. Geneva, Switzerland: World Health Organization and the World Bank; 2018.
3. Barua B, Moir M. Comparing Performance of Universal Health Care Countries, 2020. Fraser Institute. 2020 Nov 10. URL: https://www.fraserinstitute.org/studies/comparing-performance-of-universal-health-care-countries-2020 [accessed 2021-11-10]
4. Ansell D, Crispo JAG, Simard B, Bjerre LM. Interventions to reduce wait times for primary care appointments: a systematic review. BMC Health Serv Res 2017 Apr 20;17(1):295 [FREE Full text] [doi: 10.1186/s12913-017-2219-y] [Medline: 28427444]
5. Rivas J. Advanced Access Scheduling in Primary Care: A Synthesis of Evidence. J Healthc Manag 2020;65(3):171-184. [doi: 10.1097/JHM-D-19-00047] [Medline: 32398527]
6. Murray M, Tantau C. Same-day appointments: exploding the access paradigm. Fam Pract Manag 2000 Sep;7(8):45-50 [FREE Full text] [Medline: 11183460]
7. Murray M, Berwick DM. Advanced access: reducing waiting and delays in primary care. JAMA 2003 Feb 26;289(8):1035-1040. [doi: 10.1001/jama.289.8.1035] [Medline: 12597760]
8. Breton M, Maillet L, Paré I, Abou Malham S, Touati N. Perceptions of the first family physicians to adopt advanced access in the province of Quebec, Canada. Int J Health Plann Manage 2017;32(4):e316-e332. [doi: 10.1002/hpm.2380] [Medline: 27605412]
9. Murray M, Bodenheimer T, Rittenhouse D, Grumbach K. Improving timely access to primary care: case studies of the advanced access model. JAMA 2003;289(8):1042-1046. [doi: 10.1001/jama.289.8.1042] [Medline: 12597761]
10. Matulis JC, McCoy R. Patient-Centered Appointment Scheduling: a Call for Autonomy, Continuity, and Creativity. J Gen Intern Med 2021;36(2):511-514 [FREE Full text] [doi: 10.1007/s11606-020-06058-9] [Medline: 32885369]
11. Digital Technologies: Shaping the Future of Primary Health Care. World Health Organization. 2018. URL: https://apps.who.int/iris/handle/10665/326573 [accessed 2021-11-10]
12. Brandenburg L, Gabow P, Steele G, Toussaint J, Tyson BJ. Innovation and Best Practices in Health Care Scheduling. NAM Perspectives 2015 Feb 11;5(2):1-22 [FREE Full text] [doi: 10.31478/201502g]
13. Paré G, Trudel MC, Forget P. Adoption, use, and impact of e-booking in private medical practices: mixed-methods evaluation of a two-year showcase project in Canada. JMIR Med Inform 2014;2(2):e24 [FREE Full text] [doi: 10.2196/medinform.3669] [Medline: 25600414]
14. Zhao P, Yoo I, Lavoie J, Lavoie BJ, Simoes E. Web-Based Medical Appointment Systems: A Systematic Review. J Med Internet Res 2017 Apr 26;19(4):e134 [FREE Full text] [doi: 10.2196/jmir.6747] [Medline: 28446422]
15. Green J, McDowall Z, Potts HWW. BMC Med Inform Decis Mak 2008 Aug 01;8:36 [FREE Full text] [doi: 10.1186/1472-6947-8-36] [Medline: 18673533]

16. Cao W, Wan Y, Tu H, Shang F, Liu D, Tan Z, et al. A web-based appointment system to reduce waiting for outpatients: a retrospective study. BMC Health Serv Res 2011;11:318 [FREE Full text] [doi: 10.1186/1472-6963-11-318] [Medline: 22108389]
17. Sliwa M, Okane J. Service quality measurement: appointment systems in U.K. GP practices. Int J Health Care Qual Assur 2011;24(6):441-452. [doi: 10.1108/09526861111150707] [Medline: 21916146]
18. Adedokun A, Idris O, Odujoko T. Patients' willingness to utilize a SMS-based appointment scheduling system at a family practice unit in a developing country. Prim Health Care Res Dev 2016;17(2):149-156. [doi: 10.1017/S1463423615000213] [Medline: 25851031]
19. Kurtzman GW, Keshav MA, Satish NP, Patel MS. Scheduling primary care appointments online: Differences in availability based on health insurance. Healthc (Amst) 2018 Sep;6(3):186-190. [doi: 10.1016/j.hjdsi.2017.07.002] [Medline: 28757308]
20. Ahmadi-Javid A, Jalali Z, Klassen KJ. Outpatient appointment systems in healthcare: A review of optimization studies. European Journal of Operational Research 2017 Apr;258(1):3-34. [doi: 10.1016/j.ejor.2016.06.064]
21. Klassen KJ, Yoogalingam R. Appointment scheduling in multi-stage outpatient clinics. Health Care Manag Sci 2019;22(2):229-244. [doi: 10.1007/s10729-018-9434-x] [Medline: 29404881]
22. Goh JM, Gao G, Agarwal R. Evolving Work Routines: Adaptive Routinization of Information Technology in Healthcare. Information Systems Research 2011 Sep;22(3):565-585. [doi: 10.1287/isre.1110.0365]
23. O'Connor Y, O'Rahailligh P, O'Donoghue J. Individual infusion of m-health technologies: Determinants and outcomes. 2012 Presented at: ECIS 2012 - the 20th European Conference on Information Systems; June 11, 2012; Barcelona, Spain p. 164 URL: https://aisel.aisnet.org/ecis2012/164
24. Raymond L, Paré G, Ortiz de Guinea A, Poba-Nzaou P, Trudel MC, Marsan J, et al. Improving performance in medical practices through the extended use of electronic medical record systems: a survey of Canadian family physicians. BMC Med Inform Decis Mak 2015 Apr 14;15:27 [FREE Full text] [doi: 10.1186/s12911-015-0152-8] [Medline: 25888991]
25. Haj-Ali W, Hutchison B, Primary Care Performance Measurement Steering Committee. Establishing a Primary Care Performance Measurement Framework for Ontario. Healthc Policy 2017 Feb;12(3):66-79 [FREE Full text] [Medline: 28277205]
26. McLean SM, Booth A, Gee M, Salway S, Cobb M, Bhanbhro S, et al. Appointment reminder systems are effective but not optimal: results of a systematic review and evidence synthesis employing realist principles. Patient Prefer Adherence 2016;10:479-499 [FREE Full text] [doi: 10.2147/PPA.S93046] [Medline: 27110102]
27. Rose KD, Ross JS, Horwitz LI. Advanced access scheduling outcomes: a systematic review. Arch Intern Med 2011;171(13):1150-1159 [FREE Full text] [doi: 10.1001/archinternmed.2011.168] [Medline: 21518935]
28. Kelley K, Clark B, Brown V, Sitzia J. Good practice in the conduct and reporting of survey research. Int J Qual Health Care 2003 Jun;15(3):261-266. [doi: 10.1093/intqhc/mzg031] [Medline: 12803354]
29. Parsons C. Web-Based Surveys: Best Practices Based on the Research Literature. Visitor Studies 2007 May 08;10(1):13-33. [doi: 10.1080/10645570701263404]
30. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
31. Snow J. Qualtrics Survey Software: Handbook for Research Professionals. Provo, UT: Qualtrics Labs Inc; 2012.
32. Gefen D, Rigdon EE, Straub D. Editor's Comments: An Update and Extension to SEM Guidelines for Administrative and Social Science Research. MIS Quarterly 2011;35(2):iii-xiv. [doi: 10.2307/23044042]
33. Hikmet N, Chen SK. An investigation into low mail survey response rates of information technology users in health care organizations. Int J Med Inform 2003 Dec;72(1-3):29-34. [doi: 10.1016/j.ijmedinf.2003.09.002] [Medline: 14644304]
34. Ringle CM, Sarstedt M, Straub DW. Editor's Comments: A Critical Look at the Use of PLS-SEM in "MIS Quarterly". MIS Quarterly 2012;36(1):iii-xiv. [doi: 10.2307/41410402]
35. Leth-Steensen C, Gallitto E. Testing Mediation in Structural Equation Modeling: The Effectiveness of the Test of Joint Significance. Educ Psychol Meas 2016;76(2):339-351 [FREE Full text] [doi: 10.1177/0013164415593777] [Medline: 29795869]
36. Ragin CC. Turning the tables: How case-oriented research challenges variable-oriented research. Comparative Social Research 1997;16(1):27-42. [doi: 10.4135/9781473915480.n15]
37. Fiss PC, Marx A, Cambré B. Chapter 1 Configurational Theory and Methods in Organizational Research: Introduction. In: Configurational theory and methods in organizational research. Bingley, UK: Emerald Publishing Limited; 2013:1-22.
38. Gan G, Ma C, Wu J. Data Clustering: Theory, Algorithms, and Applications. Philadelphia, PA: SIAM; 2007.
39. Whitelaw S, Pellegrini DM, Mamas M, Cowie M, Van Spall HGC. Barriers and facilitators of the uptake of digital health technology in cardiovascular care: a systematic scoping review. Eur Heart J Digit Health 2021;2(1):62-74 [FREE Full text] [doi: 10.1093/ehjdh/ztab005] [Medline: 34048508]
40. Dunn P, Hazzard E. Technology approaches to digital health literacy. Int J Cardiol 2019;293:294-296. [doi: 10.1016/j.ijcard.2019.06.039] [Medline: 31350037]
41. Mahajan S, Lu Y, Spatz ES, Nasir K, Krumholz HM. Trends and Predictors of Use of Digital Health Technology in the United States. Am J Med 2021;134(1):129-134. [doi: 10.1016/j.amjmed.2020.06.033] [Medline: 32717188]

42. Rodriguez JA, Clark CR, Bates DW. Digital Health Equity as a Necessity in the 21st Century Cures Act Era. JAMA 2020;323(23):2381-2382. [doi: 10.1001/jama.2020.7858] [Medline: 32463421]
43. Breton M, Maillet L, Duhoux A, Malham SA, Gaboury I, Manceau LM, et al. Evaluation of the implementation and associated effects of advanced access in university family medicine groups: a study protocol. BMC Fam Pract 2020;21(1):41 [FREE Full text] [doi: 10.1186/s12875-020-01109-w] [Medline: 32085728]
44. Peng C, Goswami P, Bai G. A literature review of current technologies on health data integration for patient-centered health management. Health Informatics J 2020;26(3):1926-1951 [FREE Full text] [doi: 10.1177/1460458219892387] [Medline: 31884843]
45. Berry LL, Beckham D, Dettman A, Mead R. Toward a strategy of patient-centered access to primary care. Mayo Clin Proc 2014;89(10):1406-1415. [doi: 10.1016/j.mayocp.2014.06.011] [Medline: 25199953]
46. Trudel MC, Marsan J, Pare G, Raymond L, Ortiz de Guinea A, Maillet E, et al. Ceiling effect in EMR system assimilation: a multiple case study in primary care family practices. BMC Med Inform Decis Mak 2017;17(1):46 [FREE Full text] [doi: 10.1186/s12911-017-0445-1] [Medline: 28427405]
47. Raymond L, Pare G, Marchand M. Extended use of electronic health records by primary care physicians: Does the electronic health record artefact matter? Health Informatics J 2019;25(1):71-82 [FREE Full text] [doi: 10.1177/1460458217704244] [Medline: 28434279]
48. Levallet N, Denford JS, Chan YE. Following the MAP (Methods, Approaches, Perspectives) in Information Systems Research. Information Systems Research 2021 Mar 01;32(1):130-146. [doi: 10.1287/isre.2020.0964]

Abbreviations

EMR: electronic medical record
FMG: family medicine group
iMAS: interoperable medical appointment scheduling
IT: information technology
MAS: medical appointment scheduling
PLS: partial least squares
SEM: structural equation modeling
VIF: variance inflation factor

Edited by G Eysenbach; submitted 02.06.21; peer-reviewed by M Jones, T Song; comments to author 27.06.21; revised version received 14.07.21; accepted 09.10.21; published 16.11.21.

Please cite as:
Paré G, Raymond L, Castonguay A, Grenier Ouimet A, Trudel MC
Assimilation of Medical Appointment Scheduling Systems and Their Impact on the Accessibility of Primary Care: Mixed Methods Study
JMIR Med Inform 2021;9(11):e30485
URL: https://medinform.jmir.org/2021/11/e30485
doi:10.2196/30485
PMID:34783670

Original Paper

Global Research on Coronaviruses: Metadata-Based Analysis for Public Health Policies

Thierry Warin[1], DPhil

[1] HEC Montréal, Montréal, QC, Canada

Corresponding Author:
Thierry Warin, DPhil
HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal, QC, H3T 2A7
Canada
Phone: 1 5146082106
Email: thierry.warin@hec.ca

Abstract

Background: Within the context of the COVID-19 pandemic, this paper suggests a data science strategy for analyzing global research on coronaviruses. The application of reproducible research principles founded on text-as-data information, open science, the dissemination of scientific data, and easy access to scientific production may aid public health in the fight against the virus.

Objective: The primary goal of this paper was to use global research on coronaviruses to identify critical elements that can help inform public health policy decisions. We present a data science framework to assist policy makers in implementing cutting-edge data science techniques for the purpose of developing evidence-based public health policies.

Methods: We used the EpiBibR (epidemiology-based bibliography for R) package to gain access to coronavirus research documents worldwide (N=121,231) and their associated metadata. To analyze these data, we first employed a theoretical framework to group the findings into three categories: conceptual, intellectual, and social. Second, we mapped the results of our analysis in these three dimensions using machine learning techniques (ie, natural language processing) and social network analysis.

Results: Our findings, firstly, were methodological in nature. They demonstrated the potential for the proposed data science framework to be applied to public health policies. Additionally, our findings indicated that the United States and China were the primary contributors to global coronavirus research during the study period. They also demonstrated that India and Europe were significant contributors, albeit in a secondary position. University collaborations in this domain were strong between the United States, Canada, and the United Kingdom, confirming the country-level findings.

Conclusions: Our findings argue for a data-driven approach to public health policy, particularly when efficient and relevant research is required. Text mining techniques can assist policy makers in calculating evidence-based indices and informing their decision-making process regarding specific actions necessary for effective health responses.

(JMIR Med Inform 2021;9(11):e31510) doi:10.2196/31510

KEYWORDS

COVID-19; SARS-CoV-2; natural language processing; coronavirus; unstructured data; data science; health 4.0

Introduction

Vaccines against the original SARS-CoV-2 strain have been developed. Public health policies are currently engaged in a battle against new waves of contamination and variants. The political logic is straightforward: the larger the population that has been immunized, the lower the probability of variants. Among their tools, they now have access to new data science tools (eg, machine learning–based analyses and big data, some of which are unstructured) and technological resources, such as high-performance computing platforms. Data science approaches are advantageous, not only for vaccine discovery but also for public health policies.

In this action research–type paper, we use data science techniques to collect and analyze real-time global scientific data. The objective is to examine how data science can be used to improve public health policies. Indeed, with these new tools and data sources, policy makers can (1) conduct the most accurate diagnosis of the current state of knowledge regarding SARS-CoV-2 and (2) act by assisting leading collaborative teams. As a result, decision-making processes at the national and international levels must be optimized. We propose a data science protocol in this paper that could be quickly implemented, for example, with the support of the World Health Organization (WHO), in order to optimize research collaboration across countries, universities, and researchers.

To our knowledge, this is the first paper describing a data science approach for better informing health policy decisions about coronaviruses based on global research.

One of the lessons learned from the SARS-CoV-2 outbreak is the critical nature of public policy responses. Health policy makers must be aware of global research activity. They can, for example, use this information to support some research groups that are closer to developing a vaccine. Another critical feature is that they have real-time access to information, which improves response efficiency. The COVID-19 outbreak exemplifies the critical need for more accurate and timely information. COVID-19 was first identified in late 2019 in Wuhan, China, and some studies were already using data science as a methodology [1]. On January 7, 2020, a novel coronavirus (2019-nCoV) was isolated. Since 2000, two coronavirus outbreaks have occurred: one caused by SARS-CoV and another by the Middle East respiratory syndrome coronavirus (MERS-CoV) [2]. Thus, time is critical.

Another critical factor is having access to the appropriate information. Governments have information about their research groups and their performance based on traditional data collection methods, such as annual reports. However, very few of the world's close to 200 countries possess this information. Primary sources, on the other hand, are available in the form of research publications. It would first require leveraging all of the metadata contained in these publications. Nowadays, this is possible through the use of natural language processing (NLP) techniques. Second, it would necessitate the development of algorithms to visualize the researchers, countries, and concept networks extracted from these publications. This paper illustrates the use of NLP and social network analysis (SNA) to map the aforementioned networks.

Therefore, our primary contribution is about the utility of a data science–based analysis of global coronavirus research for public health policies. We believe that a detailed map of global research on all coronaviruses is critical. Health care organizations may benefit from such a map. With today's technologies, this comprehensive mapping can be performed in real time, thanks to a code-based pipeline as illustrated in this paper, allowing for the detection of potential outbreaks of new variants and providing the information necessary to develop subsequent vaccines.

Secondly, a methodological contribution is made. Indeed, we employ metadata in order to conduct an algorithmic review of pertinent literature. In the Methods section, we go into detail about the methodology. It is, in our opinion, a necessary methodological complement to qualitative reviews and meta-analyses.

In short, the primary objective of this paper is to use global research on coronaviruses to identify critical elements that can help inform public health policy decisions. By its very nature, our research question is inscribed in action research. It is methodological and exploratory: in the context of COVID-19 and our technological development stage, how can public health policy makers benefit from machine learning techniques (ie, NLP and SNA) to assist them in their decision making?

Methods

Overview

A metadata analysis entails accumulating more articles than a traditional systematic literature review (SLR) and using algorithms to filter and sort the initial data set. We approach this problem in two ways: first, by extracting text-as-data information via NLP techniques, and second, by visualizing potential collaboration networks via SNA.

Combining these two methodological approaches is consistent with Cochrane Reviews' principle of generating new knowledge through primary research. The primary objective of Cochrane Reviews is to provide information to individuals making health or health care decisions. New research should be designed or commissioned only if it does not duplicate previously conducted research in an unnecessary manner [3]. As a result, an SLR is advantageous prior to initiating any new research, for example, by highlighting specific knowledge gaps or biases [4].

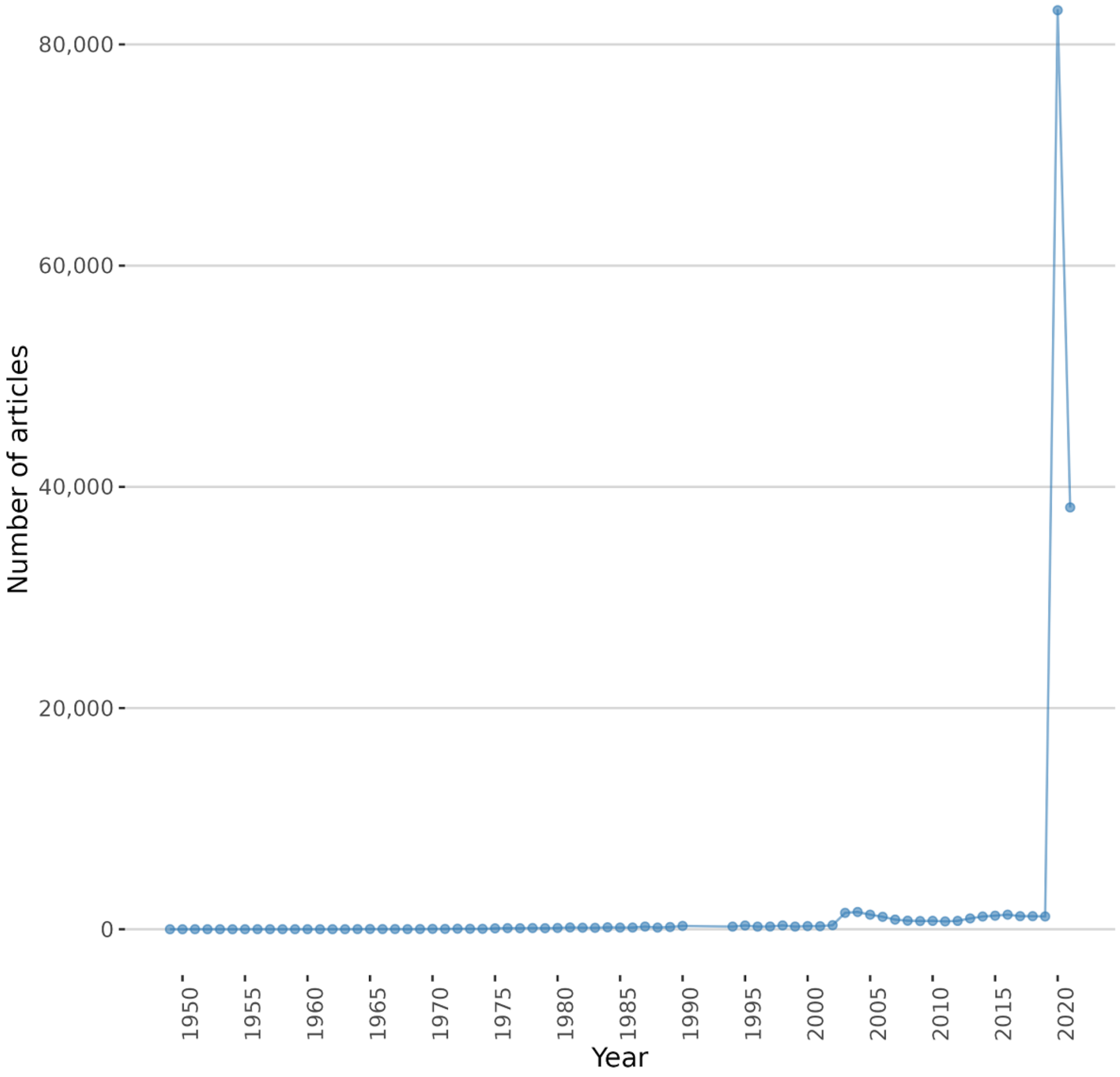
We were inspired by the guidelines for systematic reviews because we used a large data set of research documents. However, our distinction is that our objective was not to contribute to the development of a theoretical framework by identifying distinct research streams (ie, an academic objective) but to propose an example of applied research, more precisely action research.

All of these considerations were particularly pertinent during the COVID-19 period. Thus, the methodology presented in this paper was focused on using the largest data set possible and highlighting some of the mappings that were technologically possible via NLP and SNA.

We formulated two hypotheses about public health policies. First, policies require information about coronavirus research findings. This can assist governments and their various industrial partners in developing pandemic-related solutions. Second, they must be capable of supporting the ecosystems that generate these groundbreaking research findings. During a pandemic—but not exclusively—decision-making processes must be optimized to expedite the production of solutions based on research findings. This means that policy makers must be aware of the characteristics that contribute to the production of these research findings. Individuals (ie, single authors), groups of researchers (ie, multiauthored documents), interuniversity collaborations, or global collaborations are all examples of these characteristics.

The years 2020 and 2021 logically demonstrate exponential growth in research output (Figure 1).

Figure 1. Document count over time. The 2021 document count ended on May 4.



Protocol Development

As previously stated, our research question is methodological in nature and exploratory in scope. It is about whether and how public health policy makers can benefit from machine learning techniques to inform their decision-making process in the COVID-19 context and at our technological development stage.

We proposed a four-stage protocol: (1) the first stage required access to global research on coronaviruses, (2) the second stage used NLP techniques to convert the text from published research documents into data, (3) the third stage employed conventional statistical techniques, and (4) the fourth stage used SNA to identify key concepts and collaborators or universities. Interest in SNA has grown in recent years, despite the fact that it is a mathematical field that dates all the way back to the mid-1930s. SNA is predicated on the premise that the social contexts of actions matter [5]. When applied to epidemiology, this means that social contexts matter in coronavirus research, which policy makers should consider.

Each of these four stages would be computer intensive for a researcher but not for a national or international organization. We compiled the algorithms on a dedicated server built with an AMD Ryzen Threadripper processor (Advanced Micro Devices) with 32 cores (64 threads) at 3.2 GHz clock speed, with 128 GB memory.

The first stage involved the collection of data on coronavirus research conducted globally. In the fall of 2019, precisely zero scientists were investigating COVID-19, which was unknown at the time. SARS-CoV-2, the coronavirus that causes the disease, had not yet been identified or named. By the end of March 2020, the disease had spread to over 170 countries and sickened over 750,000 people, and thousands of researchers had shifted their focus away from whatever intellectual challenges had previously piqued their interest and toward the pandemic [6].

In this context, our data collection relied on the EpiBibR (epidemiology-based bibliography for R) package available on GitHub [7]. EpiBibR is a free resource based on open science principles (ie, reproducible research, open data, and open code). The package proposes 22 embedded metadata features and provides access to more than 120,000 references (N=121,231) from July 1, 1949, to May 4, 2021. Being a data package, it provides easy access to the data in order to be integrated efficiently in almost any researcher’s pipeline through the R language [8]. The references were collected via PubMed, a free resource that is developed and maintained by the National Center for Biotechnology Information at the US National Library of Medicine, located at the National Institutes of Health. PubMed includes over 30 million citations from biomedical literature. More specifically, the EpiBibR package adopted the procedure used by the Allen Institute for AI (artificial intelligence) for their COVID-19 Open Research Dataset (CORD-19) project. EpiBibR applies a similar query on PubMed with the following keywords: “COVID-19” OR “coronavirus” OR “corona virus” OR “2019-nCoV” OR “SARS-CoV” OR “MERS-CoV” OR “severe acute respiratory syndrome” OR “Middle East respiratory syndrome” [9]. To the best of our knowledge, the EpiBibR package is the only data package in R providing access to the global research on coronaviruses. This package is updated daily allowing us to build a real-time analysis. It is also the only one of this size. We were able to generate a data set of research documents as of May 4, 2021 (N=121,231). All of these references are accessible through the package [7]. We used the already-available metadata from the package and then, through NLP techniques, we also generated new metadata as explained further below.

For the second and fourth stages, we used the Bibliometrix package in R (version 3.1.4; The R Foundation) on top of our own algorithms, notably to perform disambiguation of authors’ names or to build the SNA [10]. We also created new metadata from the title, the abstract, the keywords, and the references. The latter was particularly computing intensive. Indeed, the algorithm scanned all the references in the references section of each paper. Metadata were generated using NLP techniques. To begin, we prepared the data set by choosing tokens and n-grams [10].

These attributes were required for conducting quantitative analysis on the sample. We were able to create a synthesis of research by using these machine learning tools in conjunction with other techniques, such as SNA. Additionally, the dynamics of research contributions, collaborations, idea generation, and dissemination were examined.

Study Design

The publishing landscape has shifted due to the introduction of new vehicles and practices, such as preprint servers and open data [11]. Technological advances have also provided access to new methods, such as NLP and machine learning, to complement more conventional SLRs or to present findings when a meta-analysis is not possible [12].

The SLR process is one that enables the collection of pertinent evidence on a given topic that meets predefined eligibility criteria and provides an answer to the formulated research questions. Meta-analyses employ descriptive and/or inferential statistical methods to pool data from multiple studies on a single subject. Thus, the techniques enable knowledge to be generated from a variety of qualitative and quantitative studies. The conventional method entails four basic steps: (1) search (define the search string and database types), (2) appraisal (use predefined criteria for literature inclusion and exclusion, as well as quality-assessment criteria), (3) synthesis (extract and categorize the data), and (4) analysis (narrate the results and, finally, reach a conclusion) [13].

The SLR process is defined as a “systematic, explicit, and reproducible method for identifying, evaluating, and synthesizing the existing body of completed and recorded work” [14]. According to Lasserson et al (page 1) [15], “A systematic review attempts to collate all the empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question.”

SLRs are not intended to be exhaustive or to be performed in real time. As a result, to complement SLRs, we proposed mapping the entire global research on coronaviruses, given the field’s rapid advancement. The large data set allowed us to analyze the metadata associated with the documents, such as the authors’ affiliations, universities, and references.

Another significant contribution of this new methodology is the computational treatment based on NLP techniques to convert the text to data. As such, NLP in systematic reviews is not new, and some articles have reflected on the interests of NLP techniques [16-18]. In particular, a first set of papers were about information extraction using NLP toolkits like scispaCy [19] or language-based models like BioBERT (bidirectional encoder representations from transformers for biomedical text mining) [20,21]. Another set of papers was about text classification and sentence extraction using BERT [22,23]. Using the CORD-19 data set from the Allen Institute for AI, some other papers have used paper titles and abstracts to build word pairs and co-occurrences to build knowledge graphs highlighting the existence of networks [24,25].

In this paper, we extended these NLP techniques by constructing a series of SNAs using the metadata. We were able to uncover research patterns, research history, and the actual research vehicles, as well as connect discoveries to institutions, to name a few examples. Co-occurrences in the titles and abstracts of each paper were used to highlight the findings from our SNAs.

Finally, another critical dimension was more specific and pertains to the use of each document’s references section. By concentrating on the metrics, researchers can decipher patterns of knowledge transmission. Due to the sheer volume of data being analyzed, this information can only be accessed via an algorithmic approach.

Additionally, we were cognizant of the exploratory nature of our research, employing tools and techniques whose validity had yet to be established. O’Mara-Eves et al [16] documented the biases introduced by machine learning techniques used in systematic reviews. Hopefully, this paper contributes like many others to this healthy and necessary trial and error exercise in terms of scientific validity [17]. Indeed, these new techniques may be used to save time by automating certain tasks, to act as a secondary screener, and to provide new analytical options, such as SNA. This latter point is precisely why this paper exists, particularly in the context of public health policies.

We organized the presentation of the results of these computations using the following theoretical framework. Aria and Cuccurullo [10] suggested examining three distinct structures in their study design—conceptual, intellectual, and social structures—which we did as follows:

1. The conceptual structures were concerned with leveraging the metadata to understand better which concepts and topics are used and how they have evolved in academic discourse.
2. The intellectual structures helped us in determining who originated these concepts, which journals aided in the establishment of this nascent literature, and which articles were most frequently cited in the establishment of this literature.
3. Finally, the social structures enabled us to investigate authors' collaborations and the knowledge support provided by universities and countries due to those collaborations.

Data Extraction and Quality Assessment

The relevant "universe" of the literature consists of references from EpiBibR (Table 1)**,** totaling 121,231 documents, most of which have been published in refereed journals (Table 2). The literature review covered the period between January 1, 2020, and May 2021.

The year 2020 has seen an exponential growth of papers on coronaviruses, and 2021 seems to be a replication of 2020. The average citations per document were 0.04 with the information we had. It is a low number, probably explained by the fact that these publications were published in the last few months. As a reference point, the total citations per paper in clinical medicine for the highly cited papers were 5.78 for the 2017-2021 period (Clarivate Analytics, 2021). As seen in Table 1, the documents were published within 7160 different sources, a diverse set of publication vehicles.

Table 2 summarizes the documents' classifications. The results may be conservative, as some references in the original data set may not contain all of the necessary information. Taking this limitation into account, articles dominated the sample for the entire period (Table 2), accounting for 88,374 occurrences, followed by 16,405 preprints and letters. There have been 120 SLRs published. To summarize, brief contributions (ie, articles and preprints) served as a proxy for the final product.

Consider the metadata generated from the authors' names and the keywords chosen by the authors of the documents. Coronavirus research on a global scale encompassed 5118 keywords during the overall period (Table 3). It is also worth noting for policy makers that this is a research agenda that interests 377,405 authors. There are a plethora of potential questions raised by these data in the context of public health policy. Additionally, the majority of publications were multiauthored, indicating the increasingly collaborative nature of domain research.

Additionally, the descriptive statistics analysis revealed an average of 3.11 authors and 7.15 coauthors for each publication (Table 4). The vast majority of documents were collaboratively written. Only 13,794 documents were written by a single individual (Table 4 [26]).

Now consider the three distinct structural components: conceptual, intellectual, and social. The first two are required to complete the descriptive statistics aspect.

Table 1. Preliminary information about data during the overall period and per year.

Information	Overall time period: 2020-2021	2020	2021
Sources (journals, books, etc), n	7160	6142	4982
Documents, n	121,231	83,090	38,141
Average years from publication	0.685	1	0
Average citations per document	0.04664	0.06746	0.001285
Average citations per year per document	0.02352	0.03373	0.001285

Table 2. Document type during the overall period and per year.

Type of document	Overall time period: 2020-2021, n	2020, n	2021, n
Case report	3294	2211	1083
Classical article	2	0	2
Clinical conference	7	5	2
Clinical study	2	2	0
Clinical trial	13	7	6
Clinical trial protocol	41	39	2
Clinical trial, phase II	1	1	0
Comparative study	69	58	11
Congress	8	5	3
Consensus development conference	5	4	1
Editorial	5766	4622	1144
English abstract	1664	1174	490
Equivalence trial	1	0	1
Evaluation study	14	11	3
Guideline	15	15	0
Historical article	22	21	1
Interview	32	27	5
Introductory journal article	6	6	0
Journal article	88,374	58,601	29,773
Lecture	2	2	0
Preprint or letter	16,405	13,068	3337
Meta-analysis	9	5	4
Published erratum	492	270	222
Retraction of publication	15	7	8
Review	1	1	0
Systematic review	120	65	55

Table 3. Document content and authors during the overall period and per year.

Document content	Overall time period: 2020-2021, n	2020, n	2021, n
Authors’ keywords	5118	4699	2044
Authors	377,405	266,579	188,900
Author appearances	866,589	569,924	296,665
Authors of single-authored documents	8819	6835	2580
Authors of multiauthored documents	368,586	259,744	186,320

Table 4. Details about authors' collaborations.

Collaboration measure	Overall time period: 2020-2021	2020	2021
Single-authored documents, n	13,794	10,324	3470
Documents per author, n	0.321	0.312	0.202
Authors-per-document index[a]	3.11	3.21	4.95
Coauthors per document, n	7.15	6.86	7.78
Collaboration index[b]	3.43	3.57	5.37

[a] The authors-per-document index was calculated by dividing the total number of authors by the total number of articles.

[b] The collaboration index was calculated by multiplying the total number of authors on multiauthored documents by the total number of multiauthored documents [26].

Results

Overview

As mentioned in the Methods section, we used Aria and Cuccurullo's [10] theoretical framework to present our findings. We present, respectively, the conceptual, intellectual, and social structures. For each structure, we present the relevant metrics that are available.

Additionally, as a proof of concept, we generated the necessary metadata and metrics based on the 121,231 total documents. We would encourage future researchers to filter the data set to address their own research questions, for example, by limiting their search to randomized controlled trial documents or even by content, such as proteins. Due to the fact that text is data, a new set of options becomes available.

Conceptual Structures of the Global Research on Coronaviruses

Overview

In the following subsections, we examined the conceptual structures of our sample by analyzing the keywords, their co-occurrences, and the evolution of the topics using a topic modeling technique. To create this conceptual framework, we created a matrix of the keywords and titles of the 121,231 documents.

Keyword-Based Metrics

The keyword section of Figure 2 highlights the most frequently used keywords by authors in their documents. Between 2020 and 2021, it was largely stable. Table 5 displays the top keywords in the overall sample and per year.

Figure 2. Evolution of the usage of authors' keywords.

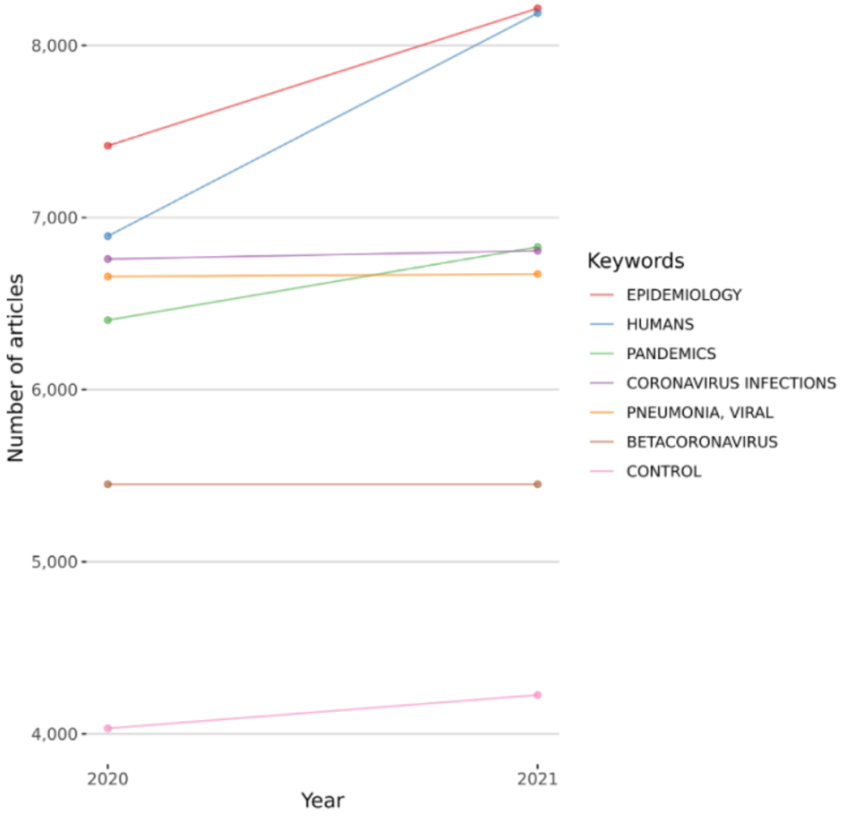


Table 5. Most relevant keywords during the overall period and per year.

Author keywords	Articles where keywords appear (N=121,231), n (%)
Overall time period: 2020-2021	
Epidemiology	8216 (6.8)
Humans	8188 (6.8)
Pandemics	6829 (5.6)
Coronavirus infections	6807 (5.6)
Pneumonia viral	6672 (5.5)
2021	
Humans	1296 (1.1)
COVID-19	1246 (1.1)
SARS-CoV-2	857 (0.1)
Epidemiology	799 (0.1)
Pandemics	425 (0.1)
2020	
Epidemiology	7417 (6.1)
Humans	6892 (5.7)
Coronavirus infections	6759 (5.6)
Pneumonia viral	6658 (5.5)
Pandemics	6404 (5.3)

Topic Modeling–Based Analyses Using Keywords

We added a new dimension to the analysis in the following section using structural topic modeling. The purpose of this section is to supplement the information gleaned from keyword co-occurrences. We illustrate this analysis in Figure 3 (overall period), Figure 4 (2020), and Figure 5 (2021). We discovered that the topics were classified into four categories: fundamental themes, emerging or declining themes, niche themes, and motor themes. The results in this case were based on the keywords solely to demonstrate the framework.

The analysis can be carried out using techniques for dimensionality reduction. The following sections make use of multiple correspondence analysis.

We augmented our field's conceptual structure with k-means clustering in order to identify clusters of documents expressing common concepts solely based on keywords. We used NLP to extract terms from the keywords section. In addition, the algorithm implemented the Porter stemming algorithm to reduce inflected, or sometimes derived, words to their word stem, base, or root form. Finally, we tokenized all the words, and we computed the latent variables to identify potential topics. Because of the necessary high computing power, we performed this analysis on the 2021 data set.

Figures 6 and 7 illustrate ample room for policy implications regarding social distancing and vaccination, respectively (red). The significant topic is population (ie, health status, age, and so on), which is depicted in blue in Figure 6 and red in Figure 7. The same analysis can be performed on additional terms, such as those found in titles, abstracts, or references. As a result, a plethora of potential classifications becomes available.

Following our examination of possible measures of conceptual structures, let us turn our attention to the analysis of intellectual structures.

Figure 3. Topic modeling for the overall period.

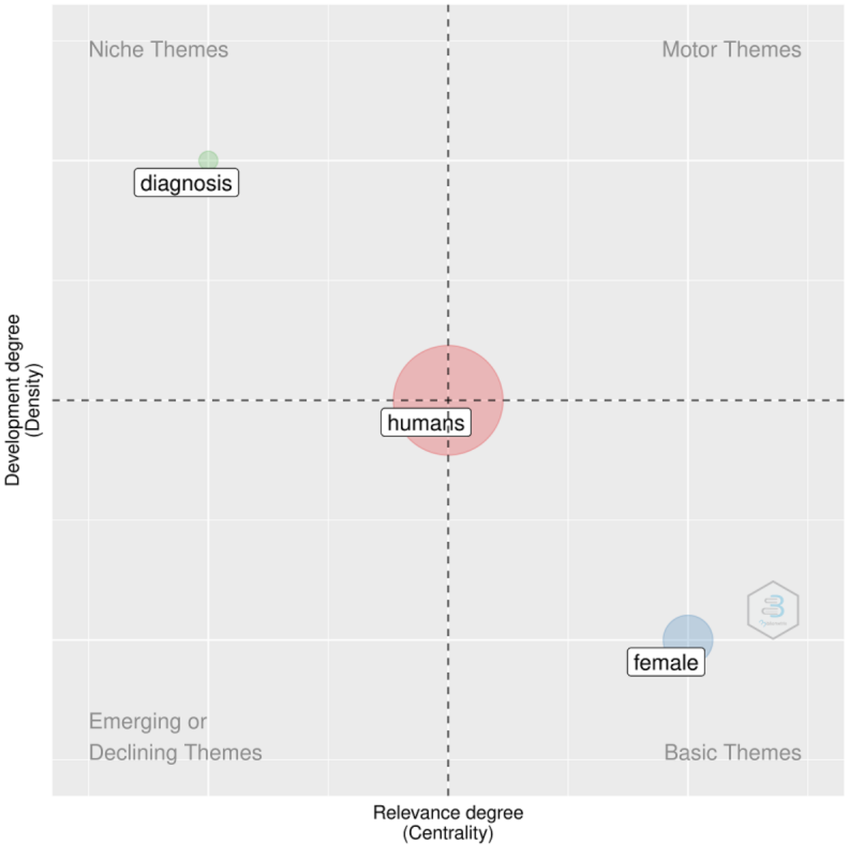


Figure 4. Topic modeling for 2020.

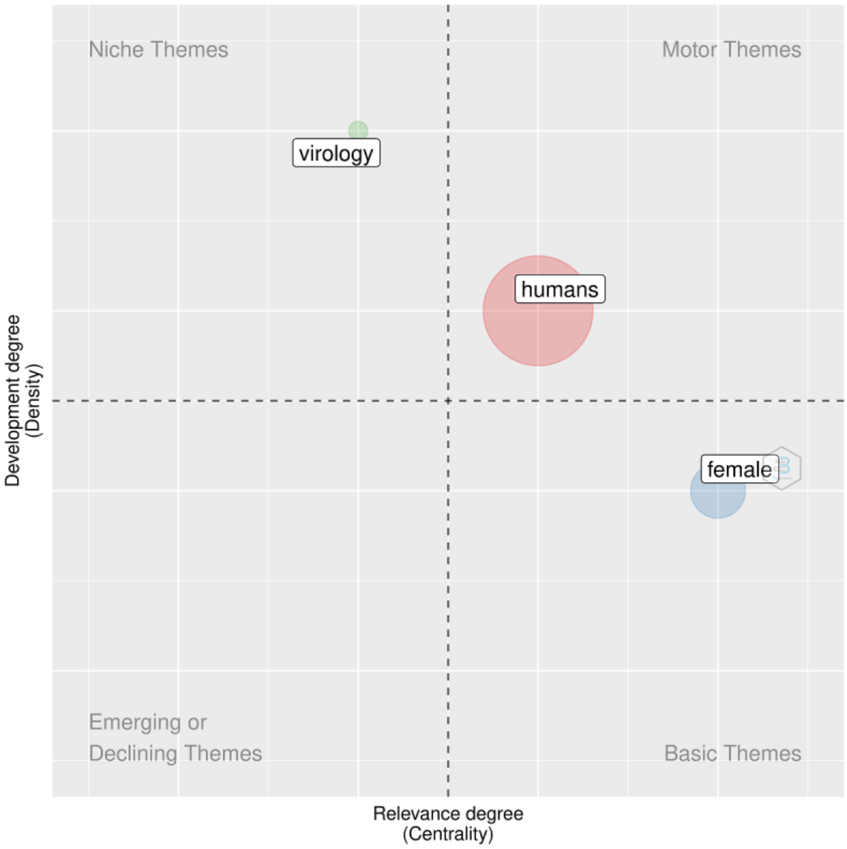


Figure 5. Topic modeling for 2021.

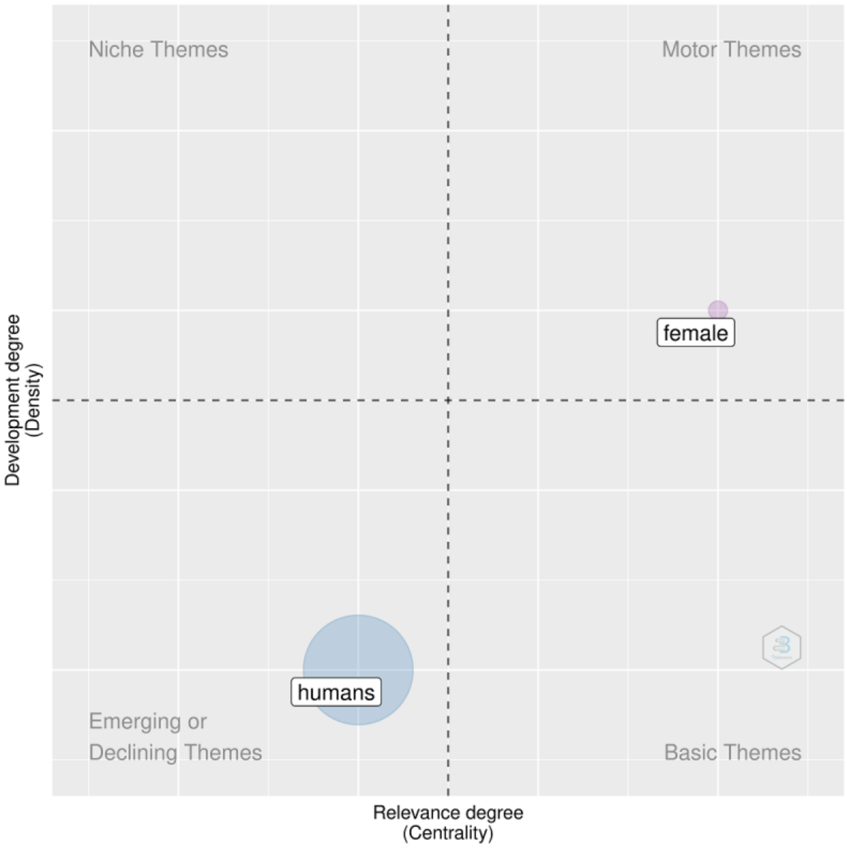


Figure 6. Conceptual structure map based on multiple correspondence analysis. Dim: dimension.

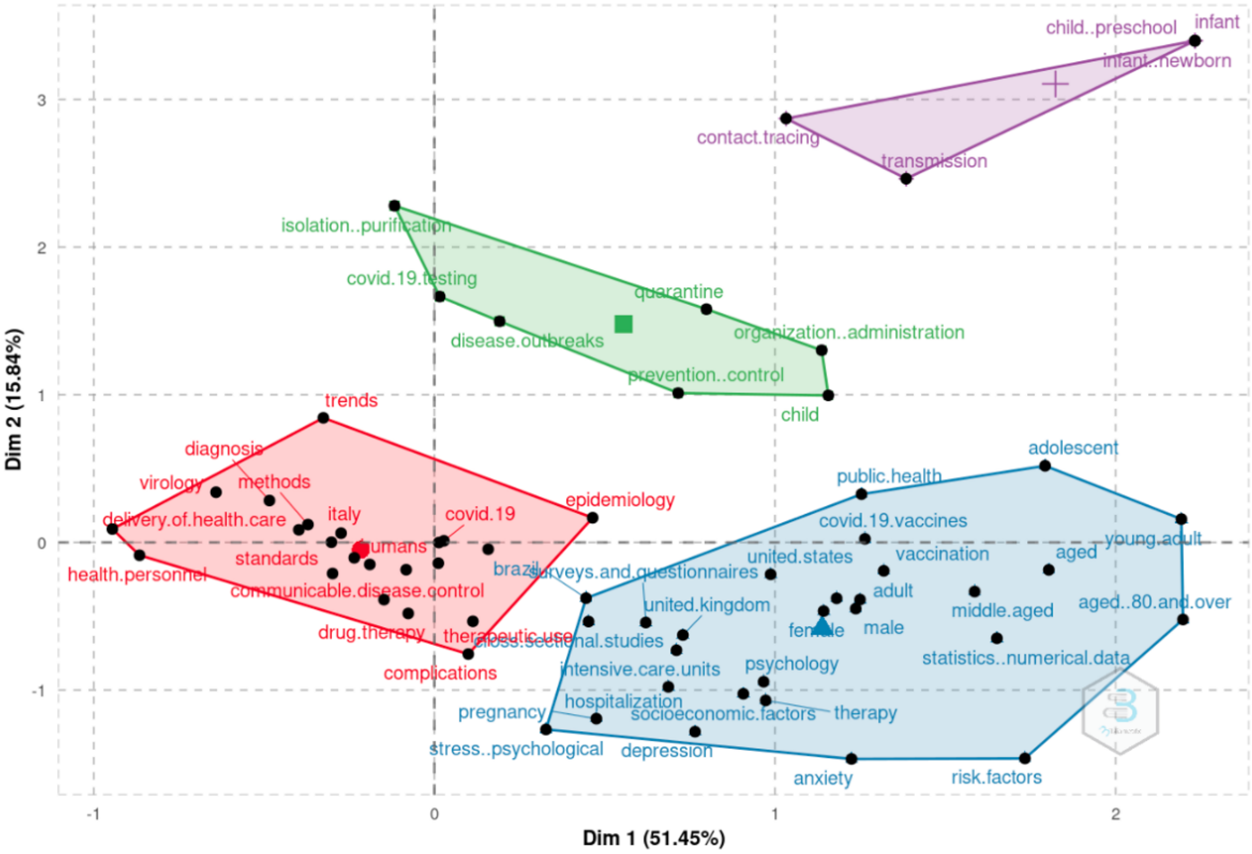
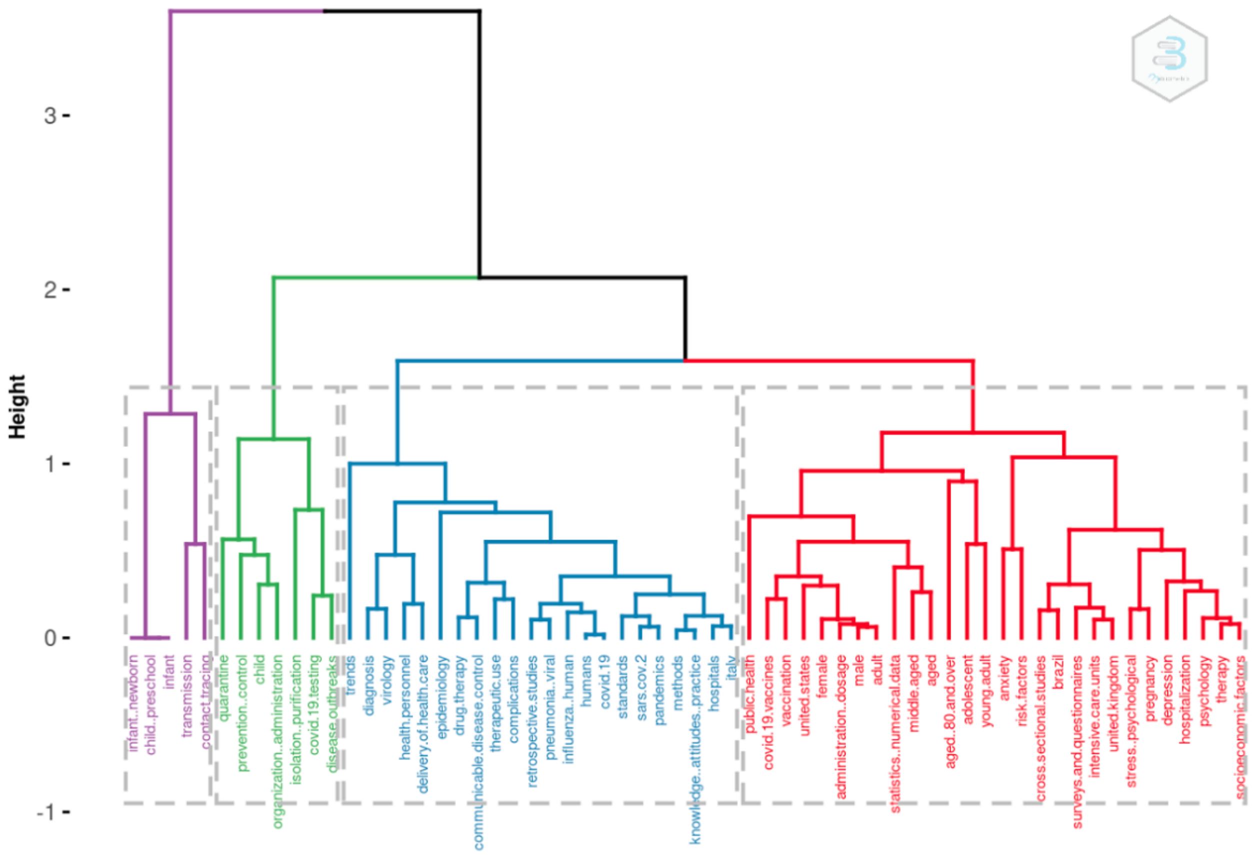


Figure 7. Topic dendogram.



Intellectual Structures of the Global Research on Coronaviruses

Another dimension leading to another interesting analysis is to know who, what journals, and which organizations are leaders in these topic dynamics.

Author-Based Metrics

In the intellectual structure, authors are interesting to consider for public policies. These metrics come with many biases, as some family names can be prevalent. An important dimension is equity, diversity, and inclusion (EDI). It is not the focus of this paper on public health policy. However, it is possible for future research to delve deeper into this author component of the intellectual structure. With this algorithmic approach and the available metadata, scholars can design EDI metrics to assess, for instance, gender-related questions, such as first and last authors; leadership positions in academia; among others [27-32]. An EDI-based analysis could also correct for the fact that fewer articles have females as the last author and these articles accrue fewer citations per publication [33]. With this metadata-based approach, scholars have access to these metrics. This is a subject that would require a more comprehensive examination of the field as a whole, which is beyond the scope of this work.

In Figures 8 and 9, respectively, we present the total count per name for the overall period and per year. It is important to note that homonymy is always an issue to correct. To correct for homonymy, several strategies exist. We could use the ORCID (Open Researcher and Contributor ID) numbers or any other unique identifier. Unfortunately, this information was not available in the original data set. Thus, we designed an algorithm that would associate an author's name with a university's name. We sorted the whole data set making sure there were unique pairs of authors and affiliations. Sometimes, university affiliations were written in different forms. We corrected them by creating a dictionary of affiliations to standardize the format.

Figure 8. Top authors in terms of production during the overall period.

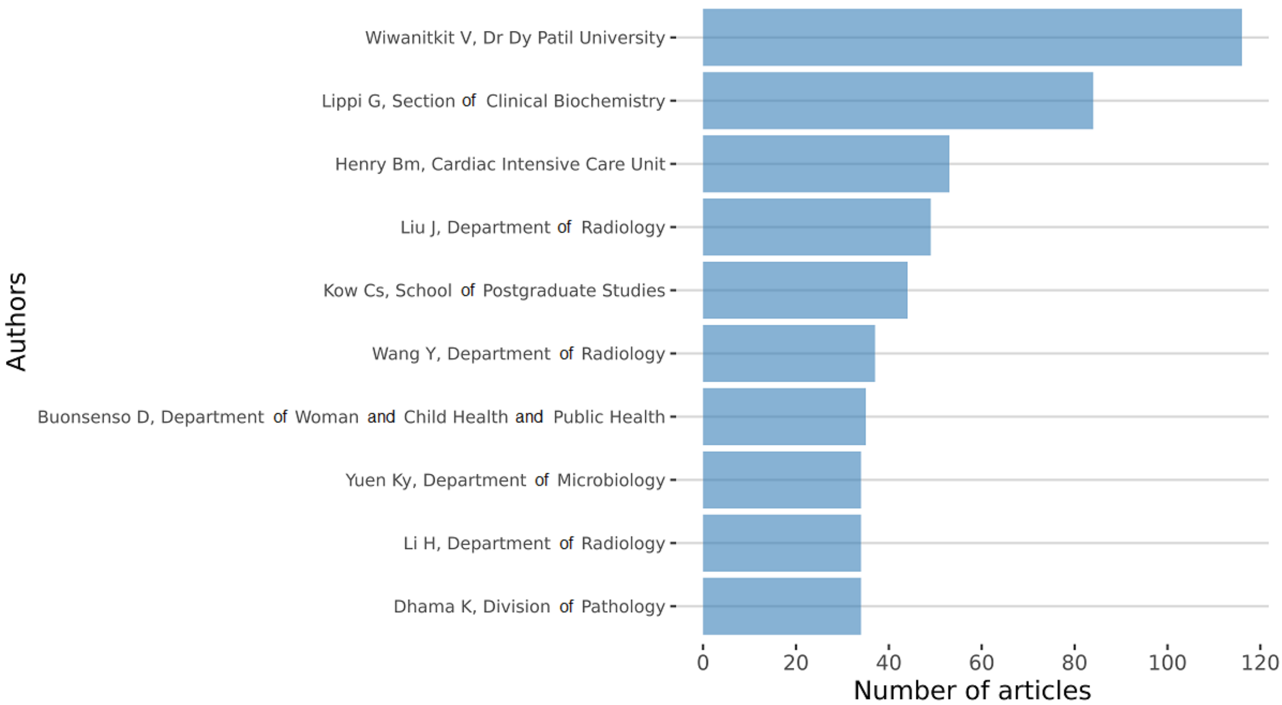
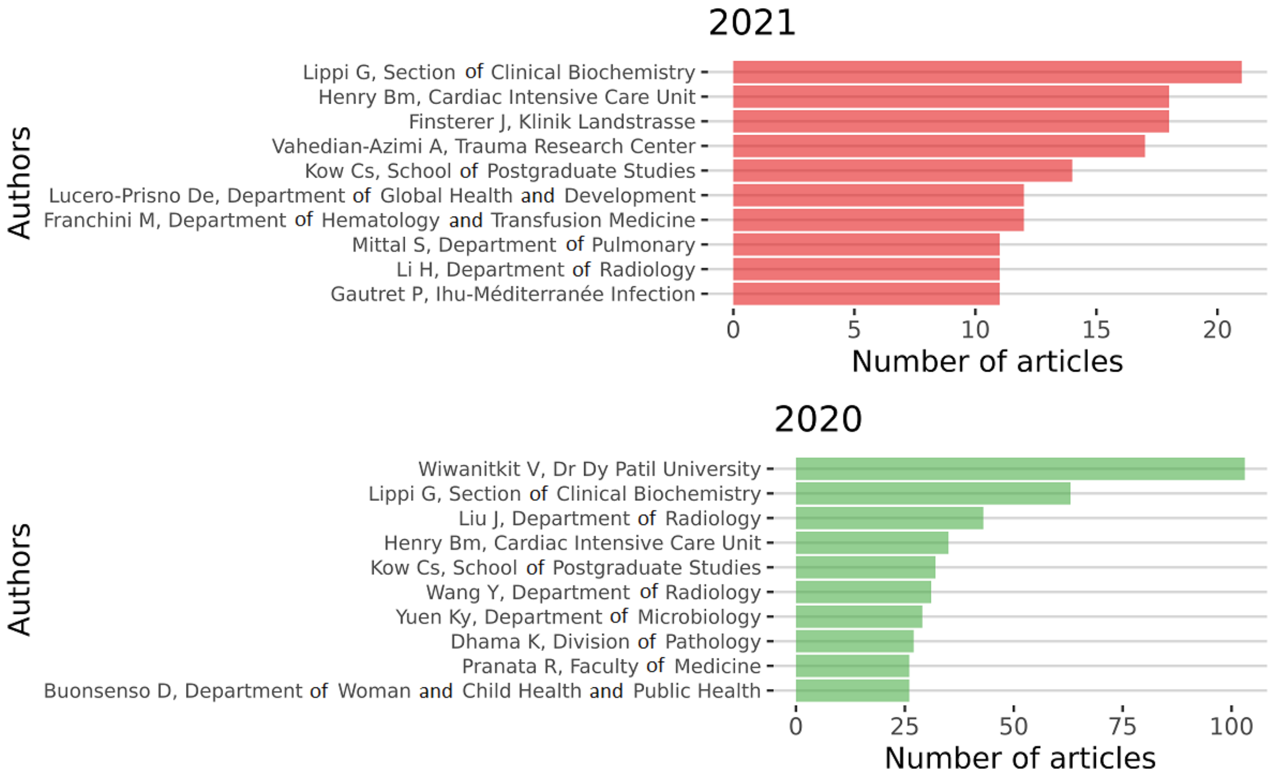


Figure 9. Top authors in terms of production per year.



We can go a little deeper and look at the average productivity of all the authors. One way to design better metrics would be to consider how many articles an author produces per year in our 2-year sample. In Figures 10-12, we computed the Lotka coefficient for the overall period, 2020, and 2021, respectively, to compare the scientific productivity of researchers to the Lotka theoretical coefficient [34]. The Lotka law describes the frequency of publication by authors as an inverse square law, where the number of authors publishing a certain number of articles is a fixed ratio to the number of authors publishing a single article. This assumption implies that the theoretical β coefficient of the Lotka law is equal to 2.

Figures 10-12 describe the share of authors having published a certain number of articles. Here, there was a statistically significant difference between the observed and the theoretical Lotka distributions, meaning that authors were more prolific in this research topic. This does not come as a surprise, considering the urgency of the topic.

Figure 10. Scientific productivity during the overall period.

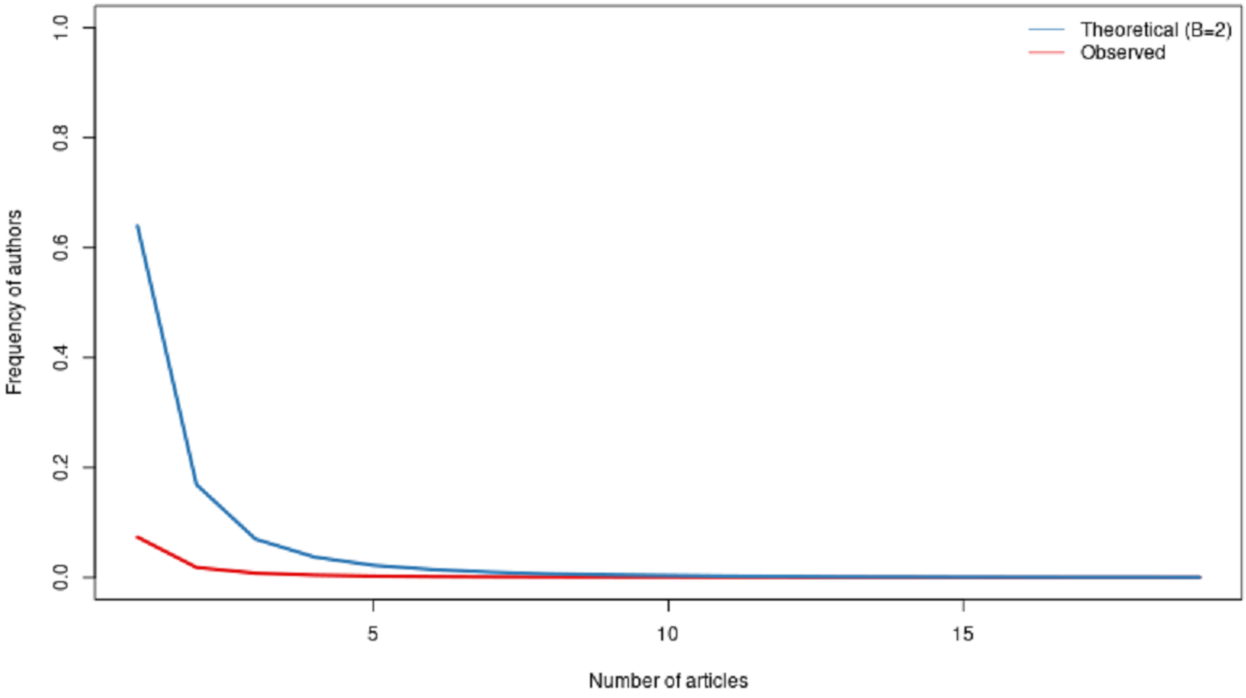


Figure 11. Scientific productivity during 2020.

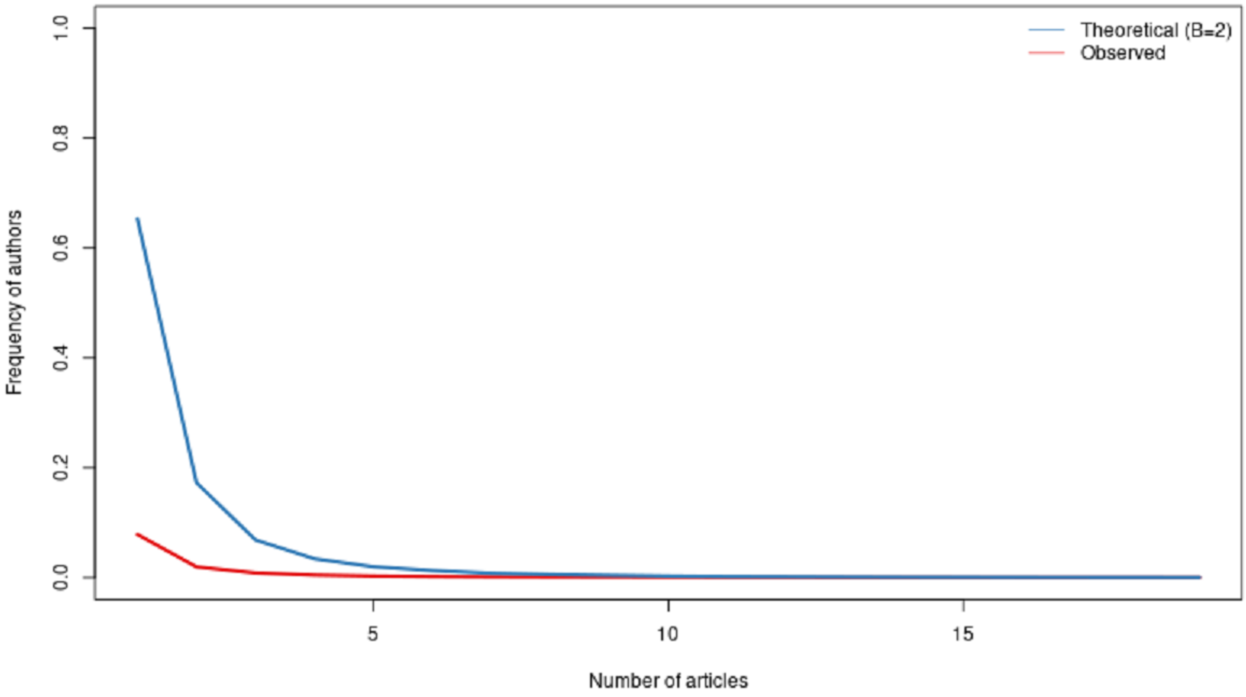
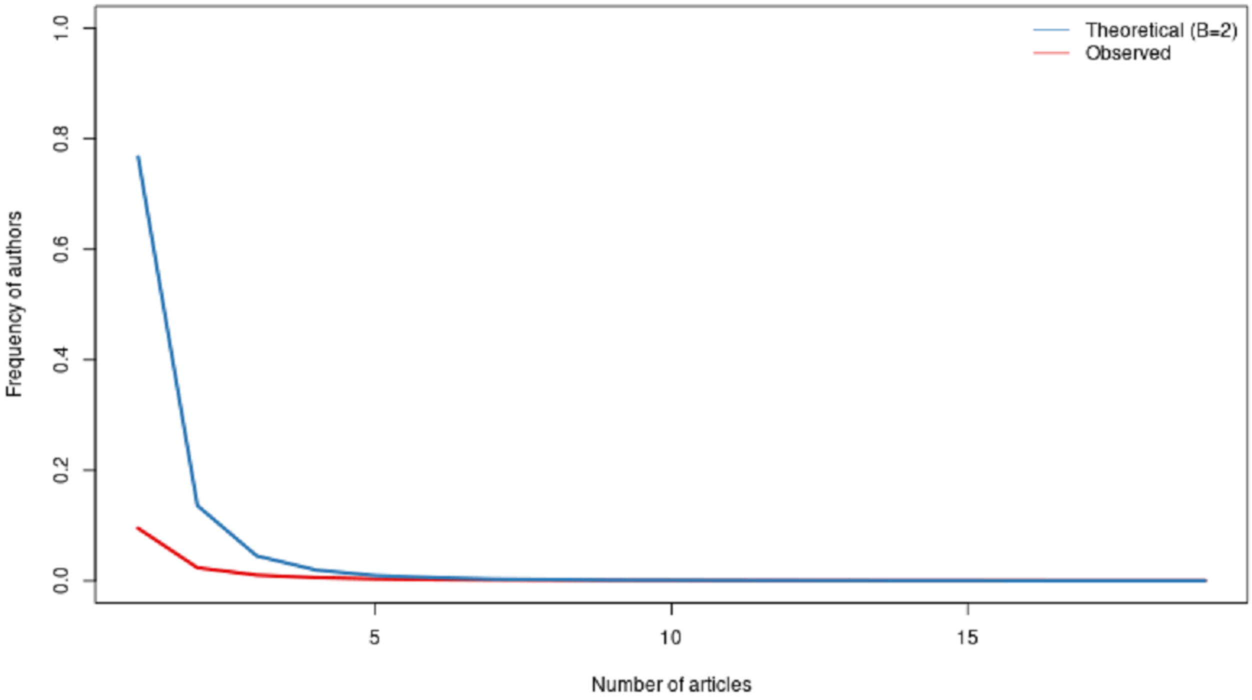


Figure 12. Scientific productivity during 2021.



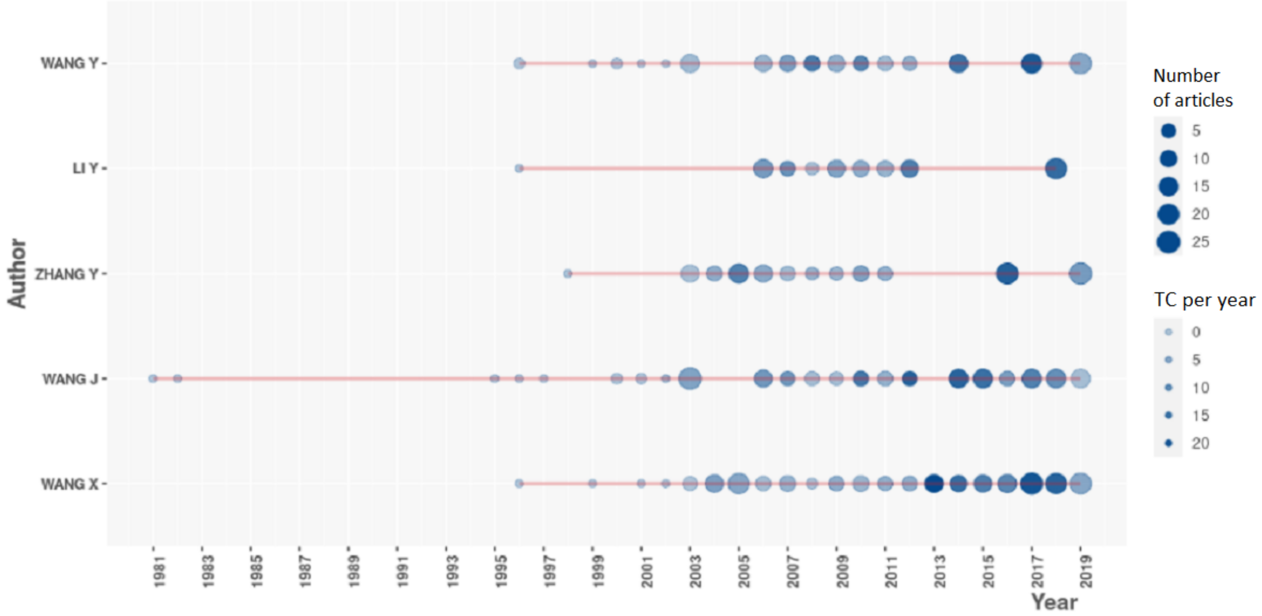
Due to the large size of the data set, our dedicated server was not powerful enough to compute the results. Our strategy was, thus, to extract a random sample for 2020 and 2021 of 25,000 documents each year. The 2021 sample corresponded to 65.5% of the total 2021 data set. The 2020 sample corresponded to 30.0% of the total 2020 data set.

To go further, we narrowed it down to specific groups of authors, institutions, or research teams and computed the scientific productivity. It may be relevant, indeed, to allocate resources, as a policy maker, to some of these dimensions.

To conclude, in Figure 13, we first filtered the original authors' list to authors having published fewer than 25 articles and to those who had fewer than 20 total citations per year. It was an arbitrary choice, and we could easily filter it differently, which is precisely in line with our main point: data science allows this agile adaptation.

Let us now move to the article element as another interesting dimension to measure intellectual structures.

Figure 13. Productivity of the top authors over time. TC: total citations.



Article-Based Metrics

We had a look at the citations from the data set (N=121,231). Authors represented interesting information regarding public health policies, including their productivity metrics, but we also found it interesting that the most cited manuscripts may help refine the metrics (Table 6).

Let us now go deeper and consider the social structures of the global research on coronaviruses.

Table 6. Most cited manuscripts.

Articles (author, year, journal)	Total citations, n	Total citations per year, n
Huang C, 2020, The Lancet	146	73.0
Zhu N, 2020, New England Journal of Medicine	102	51.0
Chen N, 2020, The Lancet	100	50.0
Li Q, 2020, New England Journal of Medicine	89	44.5
Chan JF, 2020, The Lancet	75	37.5
Veljkovic V, 2021, F1000Research	7	7.0
Endo A, 2021, Wellcome Open Research	6	6.0
Wang L, 2021, medRxiv	2	2.0
Fu L, 2021, Clinical Cardiology	1	1.0
Ackermann M, 2021, New England Journal of Medicine	1	1.0

Social Structures of the Global Research on Coronaviruses

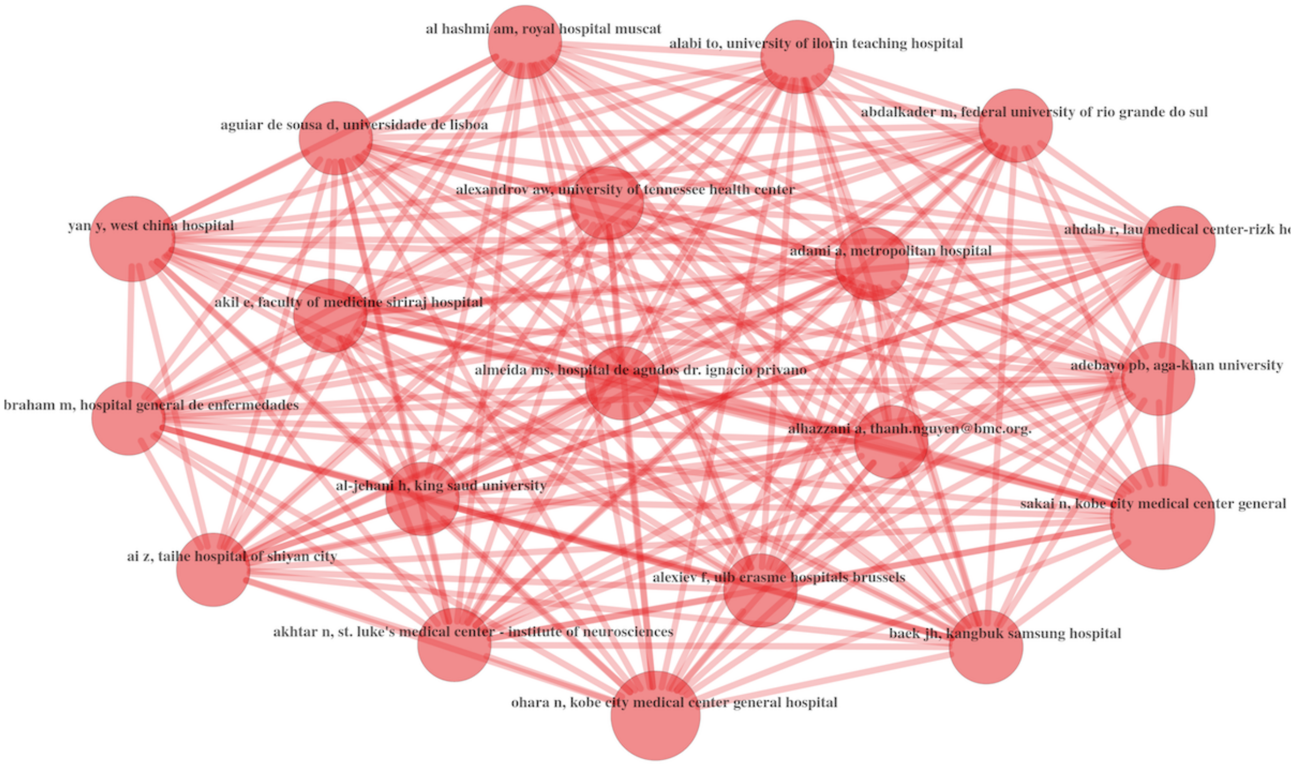
In this section, we focus on different measures to capture the social connections: the co-citations of authors, the co-citations of articles, the co-citations of journals, and the collaborations across institutions.

Authors' Collaboration Metrics

Figure 14 highlights the authors' collaborations. This figure shows the network of the top authors. Again, we can see a high level of collaboration and knowledge transfer. In further research, scholars could also perform the analyses with EDI in mind and use the metadata to have a metric of potential EDI metric imbalances [35]. This can be particularly useful in order to correct these imbalances.

Let us now move our discussion to the country level.

Figure 14. Authors’ collaboration networks in 2021.


al hashmi am, royal hospital muscat
alabi to, university of ilorin teaching hospital
abdalkader m, federal university of rio grande do sul
aguiar de sousa d, universidade de lisboa
alexandrov aw, university of tennessee health center
yan y, west china hospital
ahdab r, lau medical center-rizk h
adami a, metropolitan hospital
akil e, faculty of medicine siriraj hospital
almeida ms, hospital de agudos dr. ignacio privano
adebayo pb, aga-khan university
braham m, hospital general de enfermedades
alhazzani a, thanh.nguyen@bmc.org.
al-jehani h, king saud university
sakai n, kobe city medical center general
ai z, taihe hospital of shiyan city
alexiev f, ulb erasme hospitals brussels
akhtar n, st. luke's medical center - institute of neurosciences
baek jh, kangbuk samsung hospital
ohara n, kobe city medical center general hospital

Country-Based Metrics

It is also possible to extract country information from the documents. We mapped the top five countries per period. Most of the authors were residents of the United States, the People’s Republic of China, India, and Europe (Table 7).

Table 8 provides supplementary information on the total citations per country. Again, the United States and China dominated the ranking.

Figures 15 and 16 show an apparent increase in the contributions coming from Asia: China and India were at the forefront of academic production. Starting from a bibliographic matrix, two groups of descriptive measures were computed: (1) the summary statistics of the network and (2) the leading indices of centrality and prestige of vertices.

Table 7. Corresponding authors' countries during the overall period and per year.

Country	Articles (N=121,231), n (%)	Frequency	Single-country publications	Multiple-country publications	Multiple-country publications ratio
Overall time period: 2020-2021					
United States	15,904 (13.1)	0.1923	15,840	64	0.004024
China	11,471 (9.5)	0.1387	11,451	20	0.001744
Italy	7565 (6.2)	0.0915	7533	32	0.004230
India	5314 (4.4)	0.0643	5295	19	0.003575
France	3156 (2.6)	0.0382	3139	17	0.005387
2021					
United States	5483 (4.5)	0.2025	5433	50	0.00912
China	2859 (2.4)	0.1056	2843	16	0.00560
Italy	2052 (1.7)	0.0758	2022	30	0.01462
India	1838 (1.5)	0.0679	1824	14	0.00762
Spain	980 (0.1)	0.0362	975	5	0.00510
2020					
United States	10,421 (8.6)	0.1874	10,407	14	0.001343
China	8612 (7.1)	0.1549	8608	4	0.000464
Italy	5513 (4.5)	0.0991	5511	2	0.000363
India	3476 (2.9)	0.0625	3471	5	0.001438
France	2237 (1.8)	0.0402	2236	1	0.000447

Table 8. Total citations per country during the overall period and per year.

Country	Total citations, n	Average article citation
Overall time period: 2020-2021		
China	2011	0.17531
United States	550	0.03458
Italy	315	0.04164
Germany	131	0.05240
France	129	0.04087
2021		
United States	10	0.001824
China	4	0.001399
Germany	4	0.004381
Belgium	1	0.004484
France	1	0.001088
2020		
China	2007	0.23305
United States	540	0.05182
Italy	314	0.05696
France	128	0.05722
Germany	127	0.08003

Figure 15. The most productive countries during the overall period, according to authors' residences.

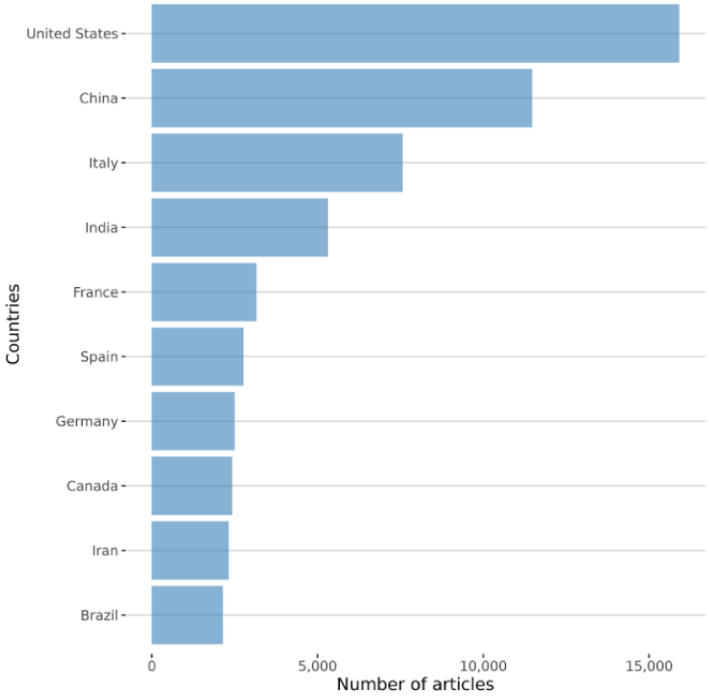
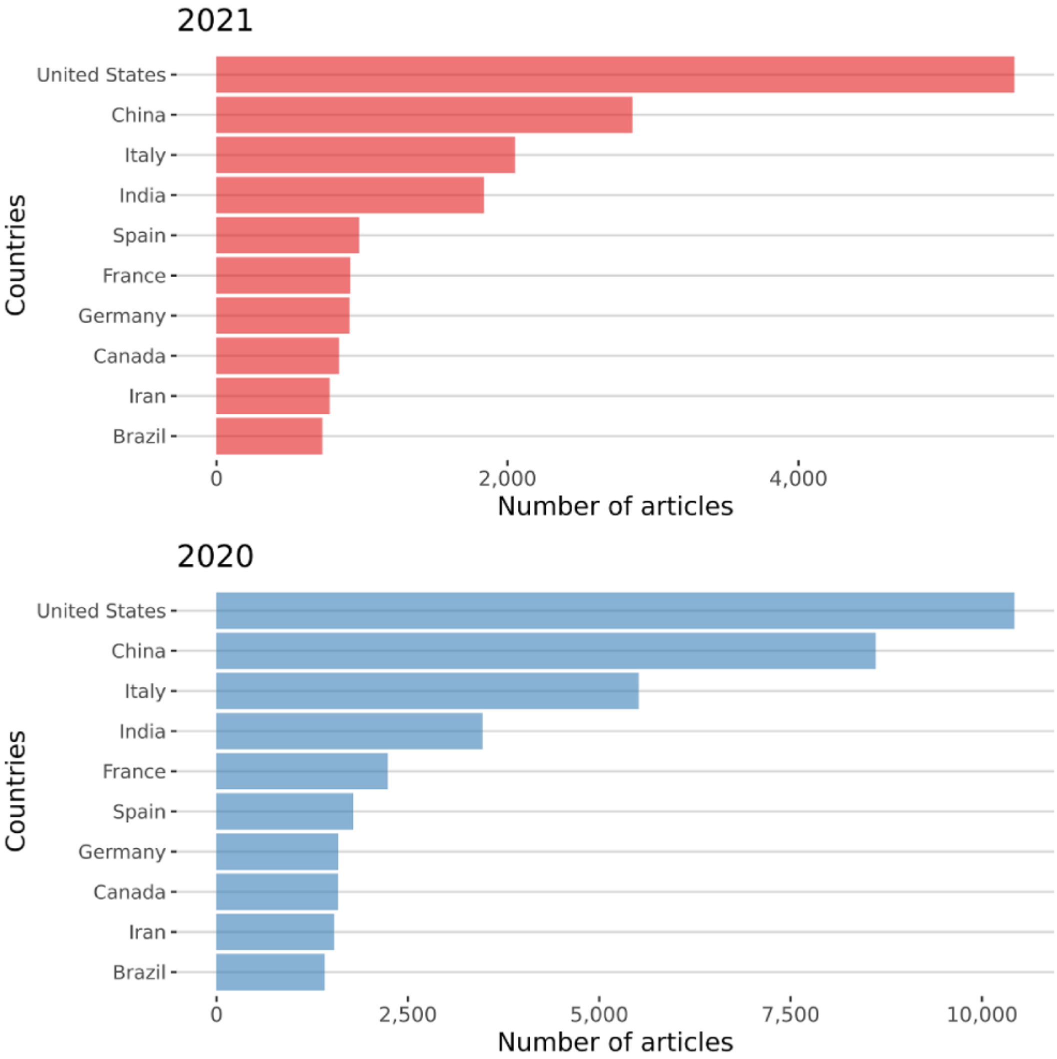


Figure 16. The most productive countries during 2021 (top) and 2020 (bottom), according to authors' residences.



We can then graph the country networks using these new measures. It is, in our opinion, an excellent showcase for public health policies and decision making. It is critical information for international health organizations, research institutions, and national governments (Figures 17-19)

Figure 17. Country collaboration networks during the overall period.

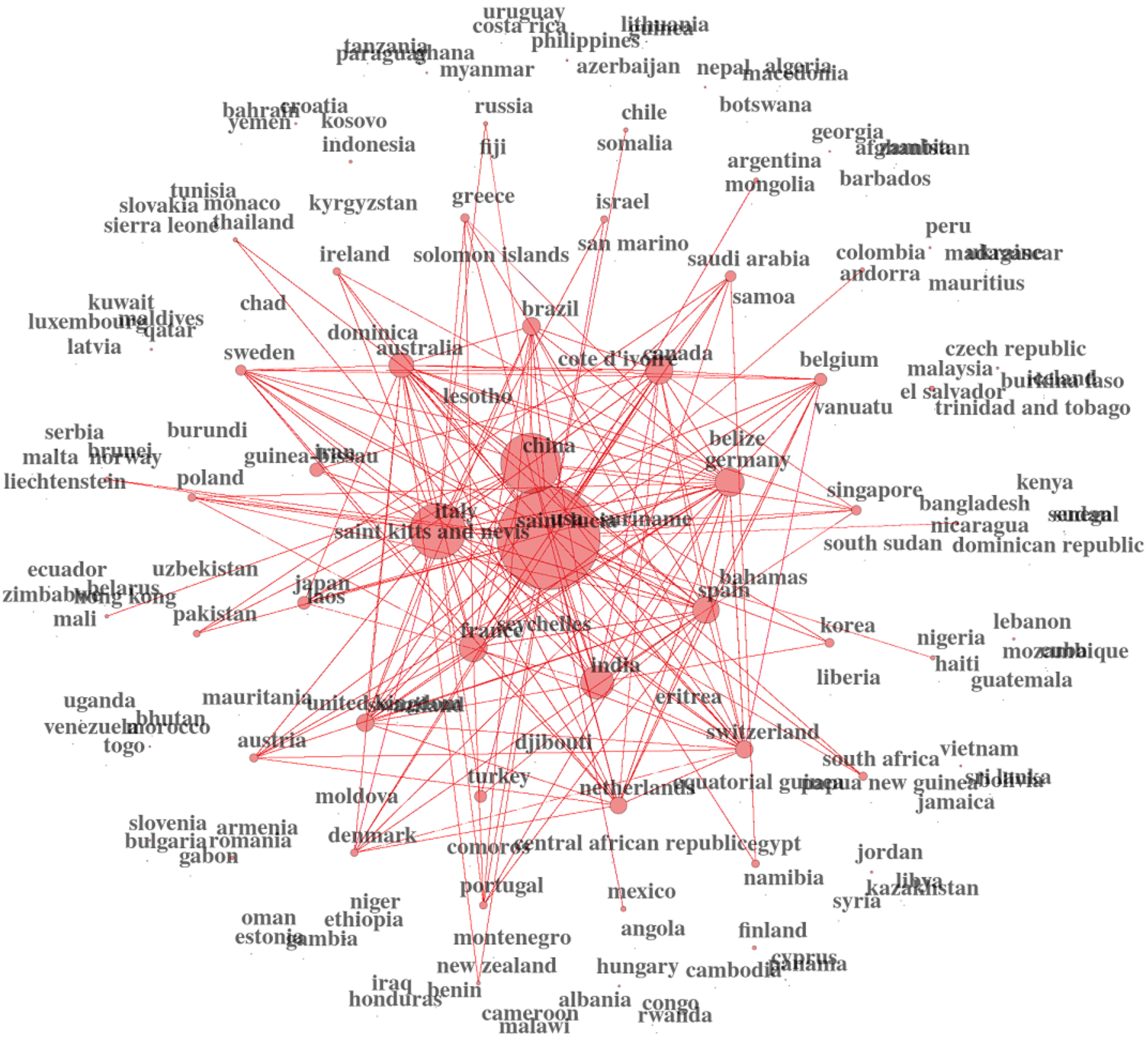


Figure 18. Country collaboration networks during 2020.

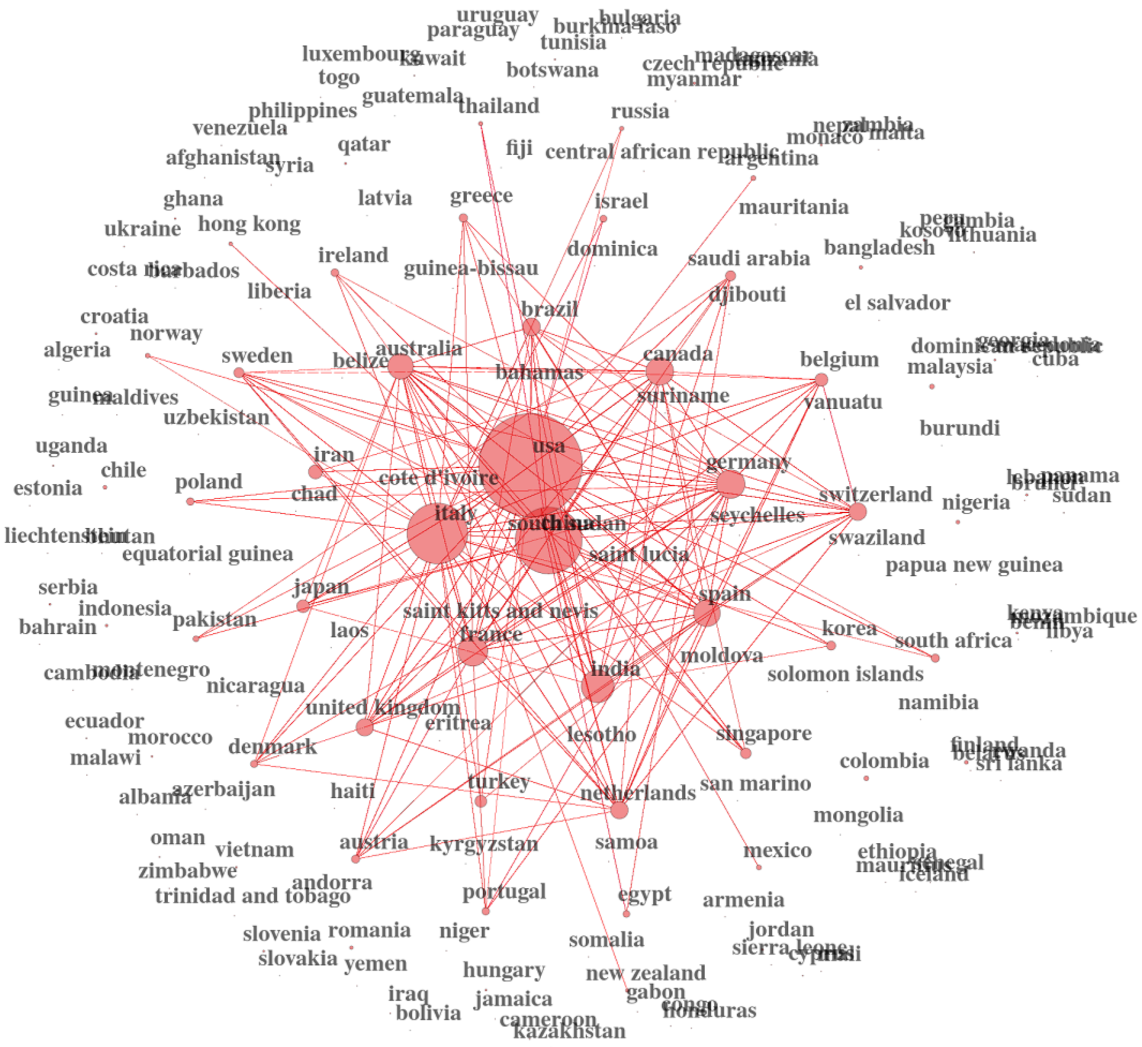
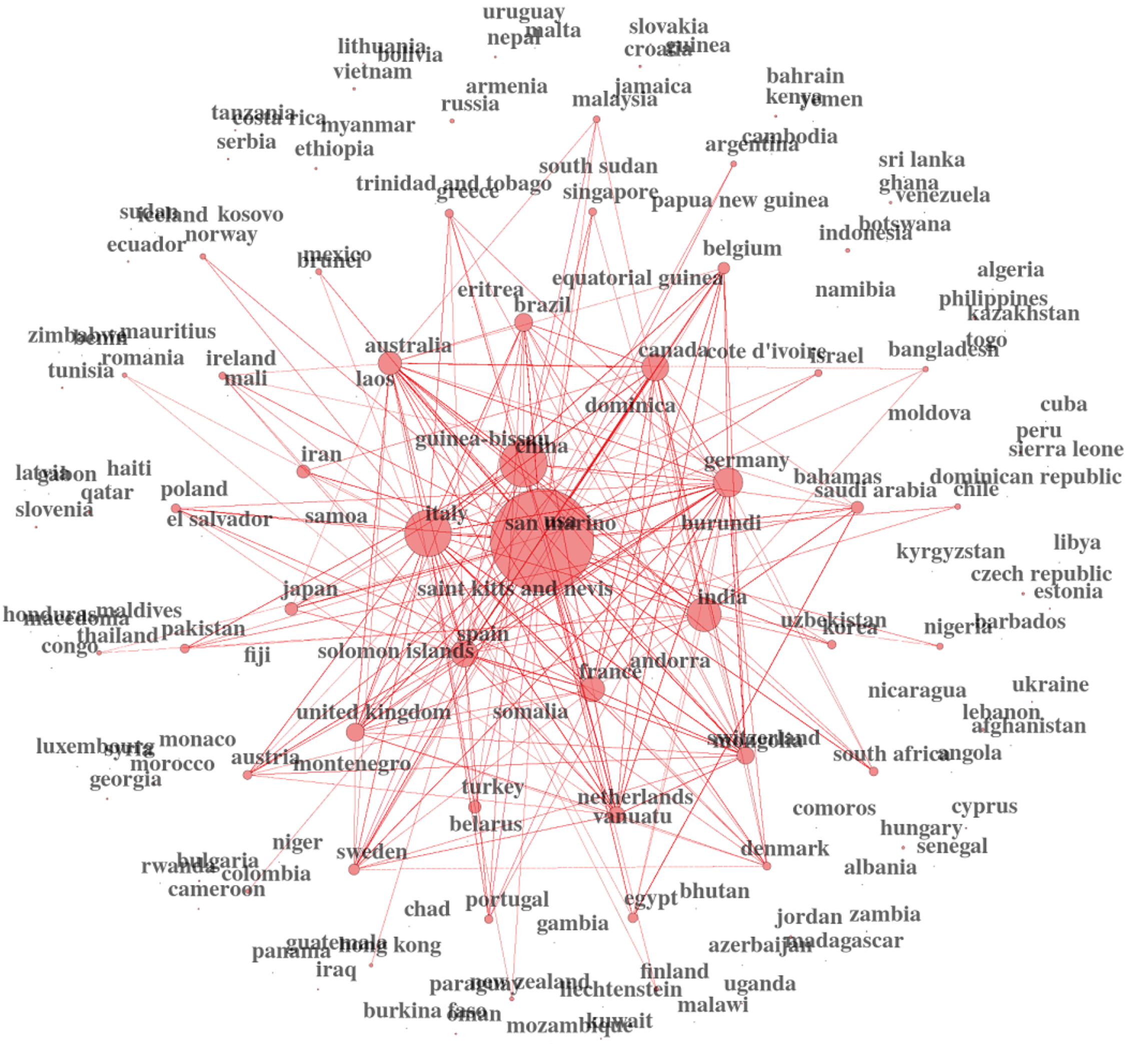


Figure 19. Country collaboration networks during 2021.



Considering the results mentioned above, the United States and China are at the forefront of academic production. Below, we also investigated the connections at the institutional level.

Co-citations of Institutional Metrics

In order to continue our social structure–oriented analysis, we made use of the collaborations that have developed among universities. We used the authors' affiliations as relevant metadata in this case, and we created a collaboration matrix to facilitate the mapping of existing links.

The network of university collaborations is also worth studying for public health policy purposes (Figures 20-22), as it indicates a strong collaboration between universities within the United States, between the United States and Canada, and between the United States and the United Kingdom.

Figure 20. University collaboration networks during the overall period.

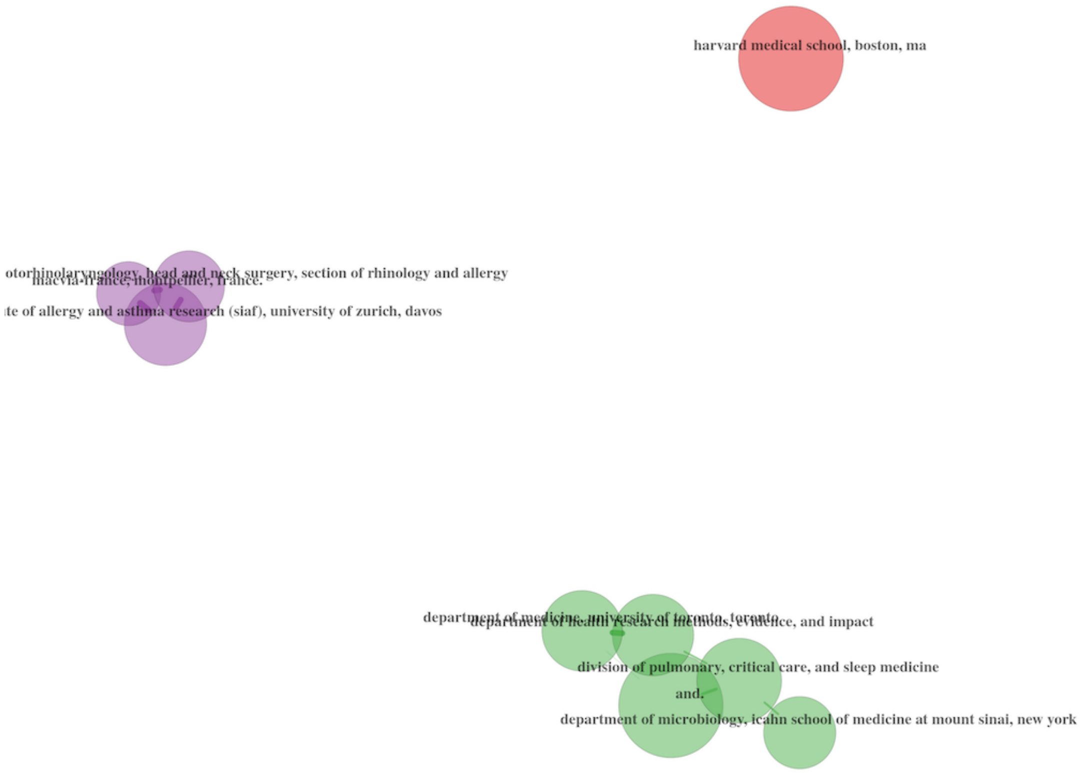

harvard medical school, boston, ma
the bmj.
otorhinolaryngology, head and neck surgery, section of rhinology and allergy
macvia-france, montpellier, france.
ıte of allergy and asthma research (siaf), university of zurich, davos
department of medicine, university of toronto, toronto
department of health research methods, evidence, and impact
division of pulmonary, critical care, and sleep medicine
and.
department of microbiology, icahn school of medicine at mount sinai, new york

Figure 21. University collaboration networks during 2020.

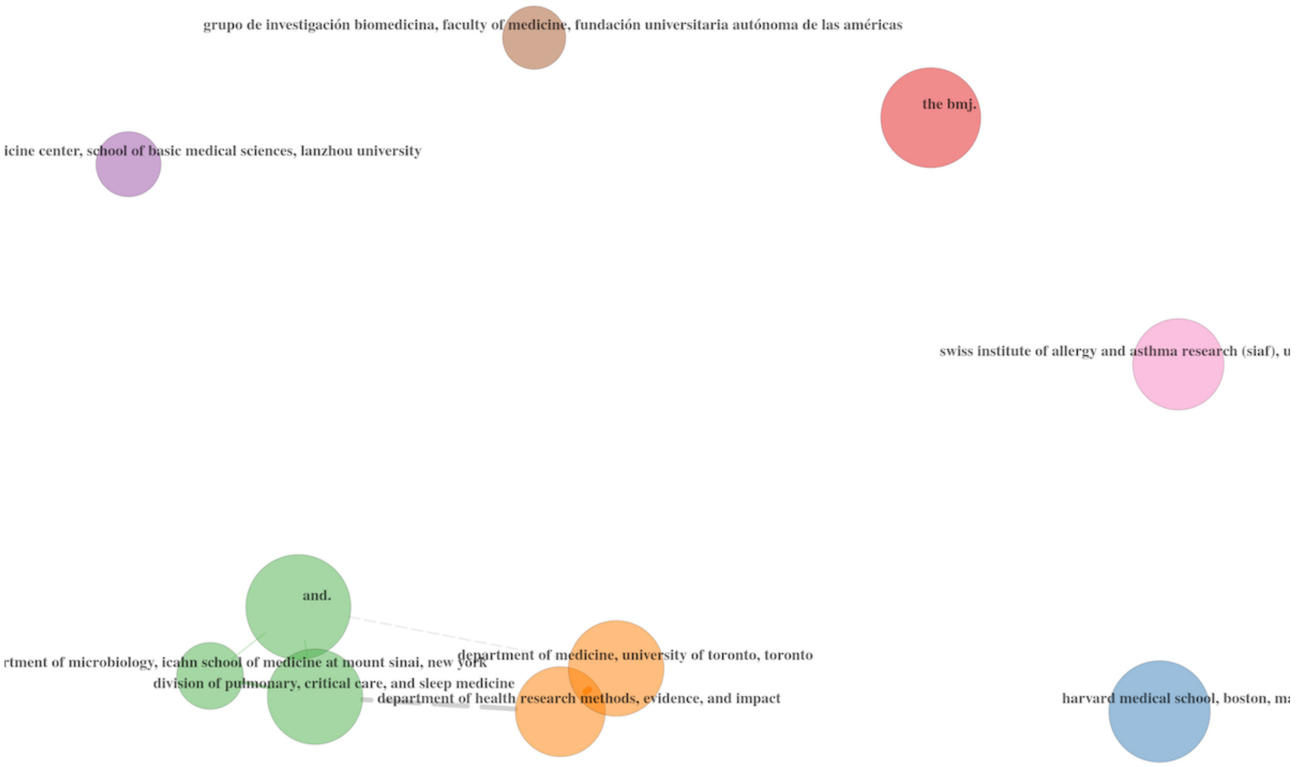
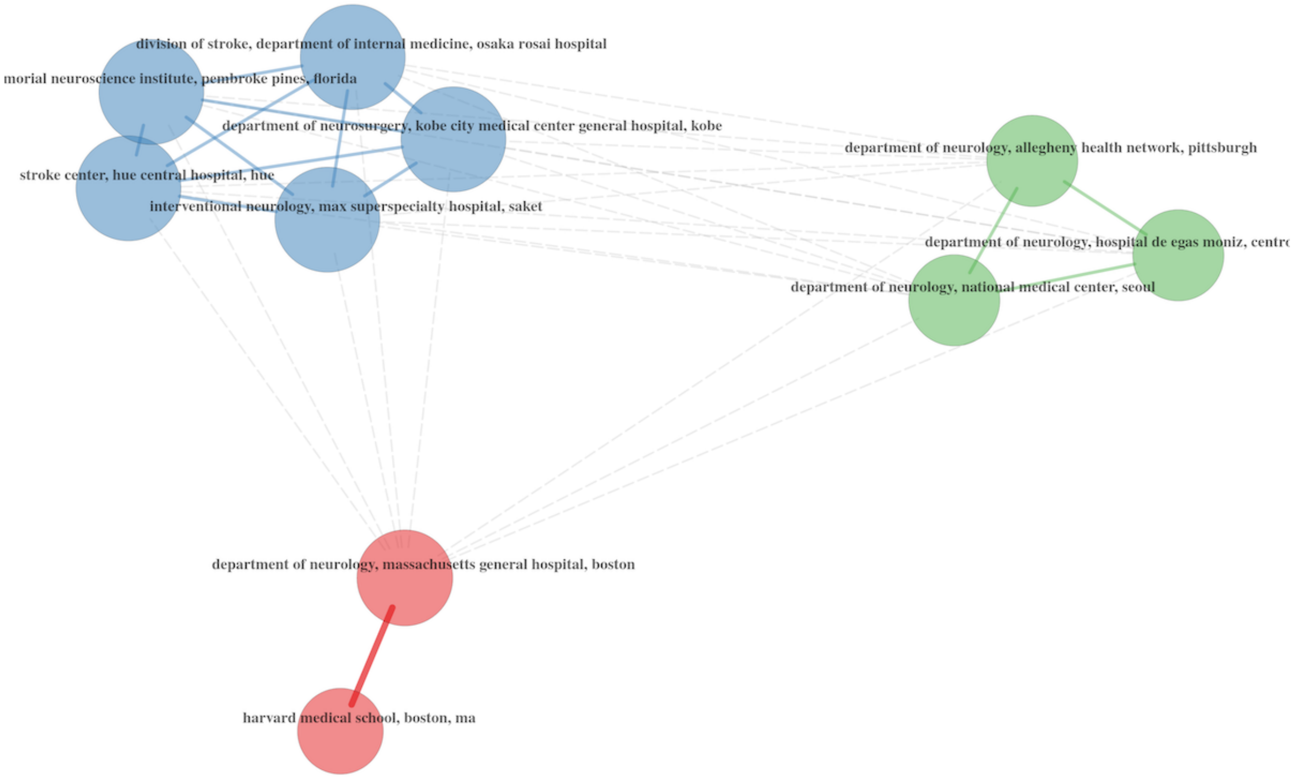

grupo de investigación biomedicina, faculty of medicine, fundación universitaria autónoma de las américas
the bmj.
icine center, school of basic medical sciences, lanzhou university
swiss institute of allergy and asthma research (siaf), u
and.
department of medicine, university of toronto, toronto
rtment of microbiology, icahn school of medicine at mount sinai, new york
division of pulmonary, critical care, and sleep medicine
department of health research methods, evidence, and impact
harvard medical school, boston, m

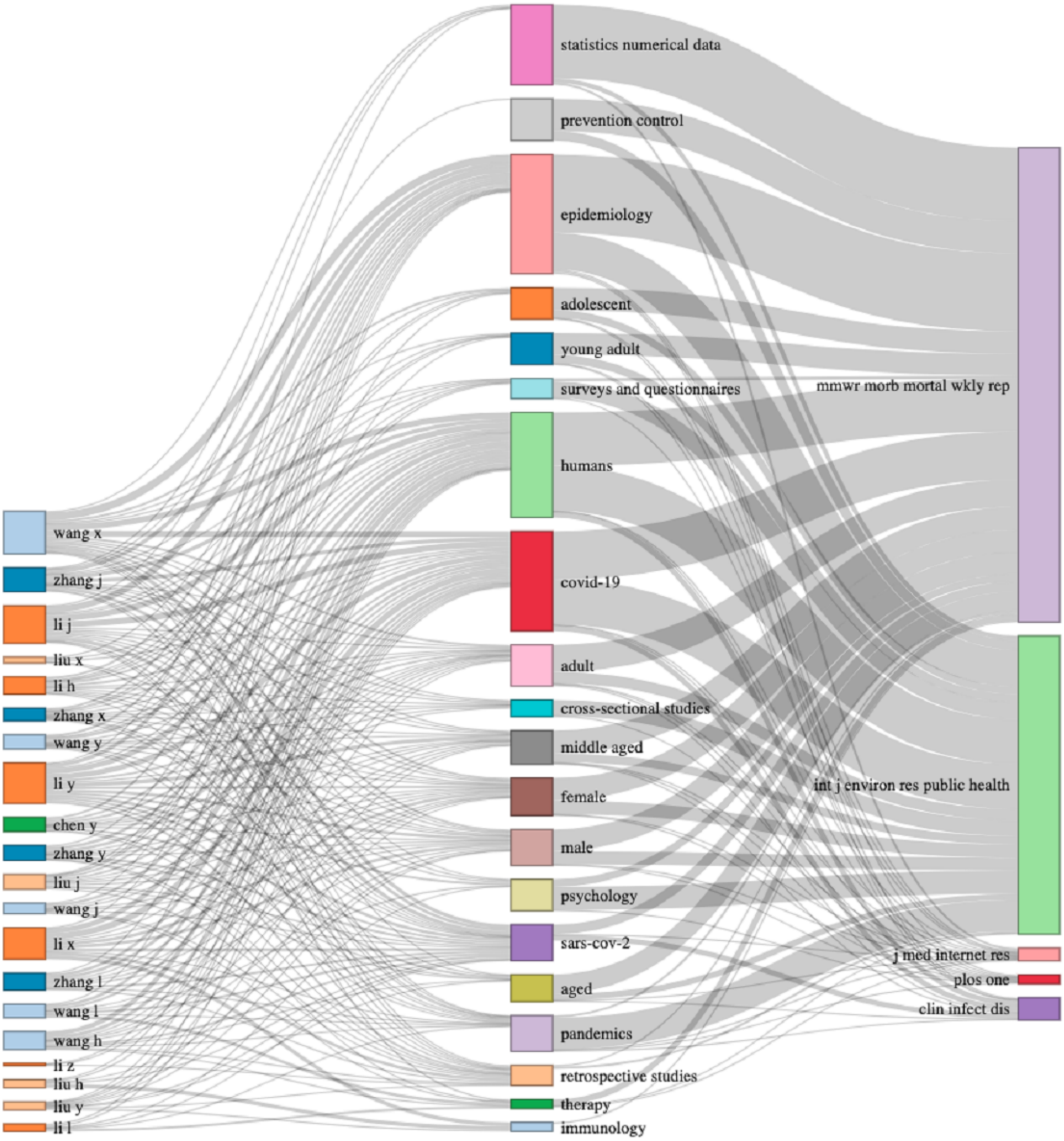
Figure 22. University collaboration networks during 2021.


division of stroke, department of internal medicine, osaka rosai hospital
morial neuroscience institute, pembroke pines, florida
department of neurosurgery, kobe city medical center general hospital, kobe
stroke center, hue central hospital, hue
interventional neurology, max superspecialty hospital, saket
department of neurology, allegheny health network, pittsburgh
department of neurology, hospital de egas moniz, centr
department of neurology, national medical center, seoul
department of neurology, massachusetts general hospital, boston
harvard medical school, boston, ma

Another point worth noting is the lack of stability between 2020 and 2021, indicating that authors from various universities preferred to collaborate on topics relevant to their research rather than replicate previous collaborations. However, we only have data for 2020 and the first half of 2021 to compare, and it would require additional research to determine whether these collaborations can be sustained over time.

To summarize, Figure 23 visualizes the major components of three fields (ie, authors, keywords, and journals) and their relationships using a so-called Sankey diagram. Particularly evident in the three fields plotted in Figure 23 are the connections between the main keywords and interest in these keywords expressed by the editors of the leading journals. We can see that the majority of the journals published articles that contained the most popular keywords suggested by the authors. Currently, there are no differentiation strategies being implemented by the publishers. Figure 23 was compiled based on 25,000 documents randomly extracted, due to the computing power limits.

Figure 23. Sankey diagram of three fields representing 2020 data: authors (left), keywords (middle), and journals (right).



Discussion

Principal Findings

We used metadata to conduct an analysis of the global research on coronaviruses. A large portion of this analysis was carried out using data science techniques, such as NLP and structured natural language analysis. It was a time-consuming and computationally intensive task. A metadata-based approach to conducting SLRs complements more traditional methods of conducting systematic reviews of the literature. There are three axes that we used to organize the literature mapping: conceptual, intellectual, and social.

When dealing with a crisis, timing is everything. Our findings were based on the transformation of text to data and then NLP analyses of the overall global research on coronaviruses. We conducted our research in order to demonstrate what we hoped would be a proof of concept. As a result, this paper falls under the umbrella term of "action research." It was our goal to demonstrate some metrics that can be applied to text-based documents, as well as how they could be applied to public health policies, with this proof of concept.

Our findings are, thus, essentially methodological and can demonstrate this approach's ability to optimize global research support. In this paper, based on data science techniques, we designed some metrics, which are static in a PDF document. Now, another powerful feature is that by using the EpiBibR data package in a research pipeline based on code, we can compile those metrics in almost real time. Indeed, all those visuals can be updated on a daily basis when the package updates itself.

In terms of actionable metrics, we have discovered that most of the research was developed in 2020 and 2021, although the first article appeared in July 1949. We also learned that the United States is the leading country in terms of scientific research on this topic. China comes second, and then individual European Union members. It was also interesting to be able to identify the international collaborations between research centers, notably between the United States, Canada, and the United Kingdom. Another interesting result was being able to capture the sizes of the research fields related to the coronaviruses, such as epidemiology, pneumology, among others.

Strengths and Limitations

Policy makers must use the most effective tools when designing public health responses in the context of the COVID-19 pandemic. Using coronaviruses as an example, this paper proposed a framework for identifying key topics and research institutions that conduct the most relevant coronavirus research.

This is especially true in the midst of what are referred to as infodemics [36]. Health policy makers may be exposed to risks associated with a lack of information, but they may also be exposed to risks associated with an overabundance of information. The quality of the information is the most important factor to consider. Indeed, one of the issues raised by WHO Director-General Tedros Ghebreyesus at the beginning of the pandemic was the “infodemic,” which is defined as the rapid spread of large volumes of information, whether true or false; the infodemic was declared on February 15, 2020 [37].

We must rely even more heavily on the contributions of the scientific community in the future. Because of advances in technology and data accessibility, policy makers today must employ the most up-to-date data science techniques in order to develop evidence-based public health policies, even more so in the COVID-19 era.

Our framework has also helped bring to light some of the limitations and biases that can be introduced into the process. These are not roadblocks, but rather concerns that a health data scientist should take into consideration. When it comes to author names, the homonymy problem serves as an excellent illustration. EDI is another aspect to consider in using those metrics. There are solutions to this problem, but they must be taken into consideration.

Another constraint is the amount of computing power required to run these machine learning routines on a large scale. National governments and international organizations, on the other hand, are not bound by this restriction in any way.

It may also be beneficial to include references from other disciplines in order to benefit from the vast number of methodologies, theories, and concepts that are available. In order to assess the spread of the disease, for example, demographers’ literature, as well as theories, would undoubtedly be relevant.

Conclusions

This is the first time that metadata have been used to analyze global research on coronaviruses. A total of 121,231 documents have been processed, resulting in a text-as-data data set. Using machine learning and NLP techniques, we have proposed a framework for public health policy makers. This framework and its metrics have the potential to assist national governments and international organizations, such as the WHO, in identifying critical global collaborations in the fight against COVID-19. It exemplifies the utility of emerging data science techniques and new modes of thought in public health.

Acknowledgments

The author expresses his gratitude to CIRANO (Centre interuniversitaire de recherche en analyse des organisations; Montreal, Canada) and Marine Leroi (CIRANO) for their support.

Conflicts of Interest

None declared.

References

1. Wu T, Hu E, Ge X, Yu G. nCov2019: An R package for studying the COVID-19 coronavirus pandemic. PeerJ 2021;9:e11421 [FREE Full text] [doi: 10.7717/peerj.11421] [Medline: 34178436]
2. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet 2020 Feb 15;395(10223):470-473 [FREE Full text] [doi: 10.1016/S0140-6736(20)30185-9] [Medline: 31986257]
3. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. Lancet 2014 Jan 11;383(9912):156-165. [doi: 10.1016/S0140-6736(13)62229-1] [Medline: 24411644]
4. Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JPA, et al. Biomedical research: Increasing value, reducing waste. Lancet 2014 Jan 11;383(9912):101-104. [doi: 10.1016/S0140-6736(13)62329-6] [Medline: 24411643]
5. Carrington PJ, Scott J, Wasserman S, editors. Models and Methods in Social Network Analysis. Cambridge, UK: Cambridge University Press; 2005.
6. Myers KR, Tham WY, Yin Y, Cohodes N, Thursby JG, Thursby MC, et al. Unequal effects of the COVID-19 pandemic on scientists. Nat Hum Behav 2020 Sep;4(9):880-883. [doi: 10.1038/s41562-020-0921-y] [Medline: 32669671]
7. Warin T. EpiBibR. GitHub. 2020. URL: https://github.com/warint/EpiBibR [accessed 2021-06-21]

8. Warin T. Global Research on Coronaviruses: An R Package. J Med Internet Res 2020 Aug 11;22(8):e19615 [FREE Full text] [doi: 10.2196/19615] [Medline: 32730218]
9. Wang L, Lo K, Chandrasekhar Y, Reas R, Yang J, Burdick D, et al. CORD-19: The COVID-19 Open Research Dataset. ArXiv. Preprint posted online on July 10, 2020 [FREE Full text]
10. Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. J Informetr 2017 Nov;11(4):959-975. [doi: 10.1016/j.joi.2017.08.007]
11. Moher D, Stewart L, Shekelle P. Establishing a new journal for systematic review products. Syst Rev 2012 Feb 09;1:1 [FREE Full text] [doi: 10.1186/2046-4053-1-1] [Medline: 22587946]
12. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. BMJ 2020 Jan 16;368:l6890 [FREE Full text] [doi: 10.1136/bmj.l6890] [Medline: 31948937]
13. Mengist W, Soromessa T, Legese G. Method for conducting systematic literature review and meta-analysis for environmental science research. MethodsX 2020;7:100777 [FREE Full text] [doi: 10.1016/j.mex.2019.100777] [Medline: 31993339]
14. Fernández del Amo I, Erkoyuncu JA, Roy R, Palmarini R, Onoufriou D. A systematic review of augmented reality content-related techniques for knowledge transfer in maintenance applications. Comput Ind 2018 Dec;103:47-71. [doi: 10.1016/j.compind.2018.08.007]
15. Lasserson TJ, Thomas J, Higgins JPT. Starting a review. In: Higgins JPT, Thomas J, editors. Cochrane Handbook for Systematic Reviews of Interventions. 2nd edition. Hoboken, NJ: John Wiley & Sons; Sep 20, 2019:3-12.
16. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: A systematic review of current approaches. Syst Rev 2015 Jan 14;4:5 [FREE Full text] [doi: 10.1186/2046-4053-4-5] [Medline: 25588314]
17. Marshall IJ, Wallace BC. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. Syst Rev 2019 Jul 11;8(1):163 [FREE Full text] [doi: 10.1186/s13643-019-1074-9] [Medline: 31296265]
18. Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying randomized controlled trials: An evaluation and practitioner's guide. Res Synth Methods 2018 Dec;9(4):602-614 [FREE Full text] [doi: 10.1002/jrsm.1287] [Medline: 29314757]
19. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. Stroudsburg, PA: Association for Computational Linguistics; 2019 Presented at: 18th BioNLP Workshop and Shared Task; August 1, 2019; Florence, Italy p. 319-327 URL: https://aclanthology.org/W19-5034.pdf [doi: 10.18653/v1/w19-5034]
20. 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] [Medline: 31501885]
21. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. 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). Stroudsburg, PA: Association for Computational Linguistics; 2019 Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 3-7, 2019; Hong Kong, China p. 3615-3620 URL: https://aclanthology.org/D19-1371.pdf [doi: 10.18653/v1/d19-1371]
22. Liang Y, Xie P. Identifying radiological findings related to COVID-19 from medical literature. ArXiv. Preprint posted online on April 4, 2020 [FREE Full text]
23. 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). Stroudsburg, PA: Association for Computational Linguistics; 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 URL: https://aclanthology.org/N19-1423.pdf [doi: 10.18653/v1/n19-1423]
24. Espinosa-Anke L, Schockaert S. SeVeN: Augmenting word embeddings with unsupervised relation vectors. In: Proceedings of the 27th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics; 2018 Presented at: 27th International Conference on Computational Linguistics; August 20-26, 2018; Sante Fe, NM p. 2653-2665 URL: https://aclanthology.org/C18-1225.pdf
25. Ahamed S, Samad M. Information mining for COVID-19 research from a large volume of scientific literature. ArXiv. Preprint posted online on April 5, 2020 [FREE Full text]
26. Koseoglu MA. Mapping the institutional collaboration network of strategic management research: 1980–2014. Scientometrics 2016 Feb 22;109(1):203-226. [doi: 10.1007/s11192-016-1894-5]
27. Grinnell M, Higgins S, Yost K, Ochuba O, Lobl M, Grimes P, et al. The proportion of male and female editors in women's health journals: A critical analysis and review of the sex gap. Int J Womens Dermatol 2020 Jan;6(1):7-12 [FREE Full text] [doi: 10.1016/j.ijwd.2019.11.005] [Medline: 32025554]

28. Thomas EG, Jayabalasingham B, Collins T, Geertzen J, Bui C, Dominici F. Gender disparities in invited commentary authorship in 2459 medical journals. JAMA Netw Open 2019 Oct 02;2(10):e1913682 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.13682] [Medline: 31642926]
29. Filardo G, da Graca B, Sass DM, Pollock BD, Smith EB, Martinez MA. Trends and comparison of female first authorship in high impact medical journals: Observational study (1994-2014). BMJ 2016 Mar 02;352:i847 [FREE Full text] [doi: 10.1136/bmj.i847] [Medline: 26935100]
30. McClelland S, Mitin T, Jagsi R, Thomas CR, Jaboin JJ. Importance of first and second authorship in assessing citation-based scholarly activity of US radiation oncology residents and subsequent choice of academic versus private practice career. J Am Coll Radiol 2018 Sep;15(9):1322-1325. [doi: 10.1016/j.jacr.2018.05.015] [Medline: 29933976]
31. Qureshi R, Lê J, Li T, Ibrahim M, Dickersin K. Gender and editorial authorship in high-impact epidemiology journals. Am J Epidemiol 2019 Dec 31;188(12):2140-2145 [FREE Full text] [doi: 10.1093/aje/kwz094] [Medline: 30995311]
32. Erren TC, Groß JV, Shaw DM, Selle B. Representation of women as authors, reviewers, editors in chief, and editorial board members at 6 general medical journals in 2010 and 2011. JAMA Intern Med 2014 Apr;174(4):633-635. [doi: 10.1001/jamainternmed.2013.14760] [Medline: 24566922]
33. Schisterman EF, Swanson CW, Lu Y, Mumford SL. The changing face of epidemiology. Epidemiology 2017;28(2):159-168. [doi: 10.1097/ede.0000000000000593]
34. Lotka AJ. The frequency distribution of scientific productivity. J Wash Acad Sci 1926 Jun 17;16(12):317-323.
35. Qureshi R, Han G, Fapohunda K, Abariga S, Wilson R, Li T. Authorship diversity among systematic reviews in eyes and vision. Syst Rev 2020 Aug 27;9(1):192 [FREE Full text] [doi: 10.1186/s13643-020-01451-1] [Medline: 32854764]
36. Tangcharoensathien V, Calleja N, Nguyen T, Purnat T, D'Agostino M, Garcia-Saiso S, et al. Framework for managing the COVID-19 infodemic: Methods and results of an online, crowdsourced WHO technical consultation. J Med Internet Res 2020 Jun 26;22(6):e19659 [FREE Full text] [doi: 10.2196/19659] [Medline: 32558655]
37. Munich Security Conference. World Health Organization. 2020 Feb 15. URL: https://www.who.int/director-general/speeches/detail/munich-security-conference [accessed 2021-11-02]

Abbreviations

2019-nCoV: novel coronavirus
AI: artificial intelligence
BERT: bidirectional encoder representations from transformers
BioBERT: bidirectional encoder representations from transformers for biomedical text mining
CIRANO: Centre interuniversitaire de recherche en analyse des organisations
CORD-19: COVID-19 Open Research Dataset
EDI: equity, diversity, and inclusion
EpiBibR: epidemiology-based bibliography for R
MERS-CoV: Middle East respiratory syndrome coronavirus
NLP: natural language processing
ORCID: Open Researcher and Contributor ID
SLR: systematic literature review
SNA: social network analysis
WHO: World Health Organization

Edited by G Eysenbach; submitted 23.06.21; peer-reviewed by A Blatch-Jones; comments to author 16.07.21; revised version received 13.08.21; accepted 27.09.21; published 30.11.21.

Please cite as:
Warin T
Global Research on Coronaviruses: Metadata-Based Analysis for Public Health Policies
JMIR Med Inform 2021;9(11):e31510
URL: https://medinform.jmir.org/2021/11/e31510
doi:10.2196/31510
PMID:34596570

Original Paper

Risk Factors Associated With Nonfatal Opioid Overdose Leading to Intensive Care Unit Admission: A Cross-sectional Study

Avijit Mitra[1], MSc; Hiba Ahsan[1], BTech; Wenjun Li[2,3], PhD; Weisong Liu[4], PhD; Robert D Kerns[5,6,7,8], PhD; Jack Tsai[9,10], PhD; William Becker[8,11], MD; David A Smelson[3,12], PsyD; Hong Yu[1,3,4,13], PhD

[1]College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA, United States

[2]Department of Public Health, University of Massachusetts Lowell, Lowell, MA, United States

[3]Center for Healthcare Organization and Implementation Research, Veterans Affairs Bedford Healthcare System, Bedford, MA, United States

[4]Department of Computer Science, University of Massachusetts Lowell, Lowell, MA, United States

[5]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, United States

[6]Department of Neurology, Yale University School of Medicine, New Haven, CT, United States

[7]Department of Psychology, Yale University School of Medicine, New Haven, CT, United States

[8]Pain Research, Informatics, Multimorbidities and Education Center, Veterans Affairs Connecticut Healthcare System, West Haven, CT, United States

[9]School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, United States

[10]National Center on Homelessness Among Veterans, United States Department of Veterans Affairs, Tampa, FL, United States

[11]Department of Internal Medicine, Yale University School of Medicine, New Haven, CT, United States

[12]Department of Psychiatry, University of Massachusetts Chan Medical School, Worcester, MA, United States

[13]Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, United States

Corresponding Author:
Hong Yu, PhD
Department of Computer Science
University of Massachusetts Lowell
1 University Avenue
Lowell, MA, 01854
United States
Phone: 1 508 612 7292
Email: Hong_Yu@uml.edu

Abstract

Background: Opioid overdose (OD) and related deaths have significantly increased in the United States over the last 2 decades. Existing studies have mostly focused on demographic and clinical risk factors in noncritical care settings. Social and behavioral determinants of health (SBDH) are infrequently coded in the electronic health record (EHR) and usually buried in unstructured EHR notes, reflecting possible gaps in clinical care and observational research. Therefore, SBDH often receive less attention despite being important risk factors for OD. Natural language processing (NLP) can alleviate this problem.

Objective: The objectives of this study were two-fold: First, we examined the usefulness of NLP for SBDH extraction from unstructured EHR text, and second, for intensive care unit (ICU) admissions, we investigated risk factors including SBDH for nonfatal OD.

Methods: We performed a cross-sectional analysis of admission data from the EHR of patients in the ICU of Beth Israel Deaconess Medical Center between 2001 and 2012. We used patient admission data and International Classification of Diseases, Ninth Revision (ICD-9) diagnoses to extract demographics, nonfatal OD, SBDH, and other clinical variables. In addition to obtaining SBDH information from the ICD codes, an NLP model was developed to extract 6 SBDH variables from EHR notes, namely, housing insecurity, unemployment, social isolation, alcohol use, smoking, and illicit drug use. We adopted a sequential forward selection process to select relevant clinical variables. Multivariable logistic regression analysis was used to evaluate the associations with nonfatal OD, and relative risks were quantified as covariate-adjusted odds ratios (aOR).

Results: The strongest association with nonfatal OD was found to be drug use disorder (aOR 8.17, 95% CI 5.44-12.27), followed by bipolar disorder (aOR 2.69, 95% CI 1.68-4.29). Among others, major depressive disorder (aOR 2.57, 95% CI 1.12-5.88), being on a Medicaid health insurance program (aOR 2.26, 95% CI 1.43-3.58), history of illicit drug use (aOR 2.09, 95% CI 1.15-3.79), and current use of illicit drugs (aOR 2.06, 95% CI 1.20-3.55) were strongly associated with increased risk of nonfatal

OD. Conversely, Blacks (aOR 0.51, 95% CI 0.28-0.94), older age groups (40-64 years: aOR 0.65, 95% CI 0.44-0.96; >64 years: aOR 0.16, 95% CI 0.08-0.34) and those with tobacco use disorder (aOR 0.53, 95% CI 0.32-0.89) or alcohol use disorder (aOR 0.64, 95% CI 0.42-1.00) had decreased risk of nonfatal OD. Moreover, 99.82% of all SBDH information was identified by the NLP model, in contrast to only 0.18% identified by the ICD codes.

Conclusions: This is the first study to analyze the risk factors for nonfatal OD in an ICU setting using NLP-extracted SBDH from EHR notes. We found several risk factors associated with nonfatal OD including SBDH. SBDH are richly described in EHR notes, supporting the importance of integrating NLP-derived SBDH into OD risk assessment. More studies in ICU settings can help health care systems better understand and respond to the opioid epidemic.

(JMIR Med Inform 2021;9(11):e32851) doi:10.2196/32851

KEYWORDS

opioids; overdose; risk factors; electronic health records; social and behavioral determinants of health; natural language processing; intensive care unit

Introduction

The opioid epidemic in the United States is one of the most severe public health emergencies in recent times, with opioid overdose (OD) deaths quadrupling from 1999 to 2019 [1]. Almost 50,000 OD-related deaths occurred in 2019 alone [2], and the estimated economic burden including opioid use disorder and fatal OD totaled US $1021 billion during 2017 [3]. The sharp rise in opioid fatalit is responsible for a decline in the US life expectancy [4] and a surge in “deaths of despair” [5]. The opioid crisis is a complex situation involving a broad range of contributing factors including social determinants of health (SDOH) [6,7].

SDOH are the conditions in which people are born, live, work, and age [8]. Adverse SDOH can affect health through various means. For example, social or familial disruptions are well-known precipitants of suicide attempt [9-11]. Behavioral determinants include alcohol consumption, tobacco usage, and use of illicit drugs, among others. Together, adverse social and behavioral determinants of health (SBDH) can be defined as those variables that can hinder an individual’s disease management and negatively impact existing medical conditions [12]. Multiple prior studies suggested strong correlations between OD and a number of SBDH [6,7,13]. Analyzing SBDH in relation to OD can help us better address the OD crisis.

Prior studies found that lack of SBDH information can significantly decrease health care quality [14,15]. Realizing the impact of SBDH on health outcomes, many prior studies focused on extracting SBDH from structured data (eg, diagnosis codes, medications) and/or unstructured data (eg, discharge summaries, progress notes) [11,12,16-18]. However, existing electronic health records (EHRs) often lack the necessary SBDH information in a structured format, undermining its use in clinical care and research settings. On the other hand, EHR notes often describe SBDH [19], for example, financial insecurity (eg, “$807 SSI and $16/month food stamps”) and risky alcohol consumption (eg, “Drinking >4 drinks on one occasion or >14 drinks per week”). In addition, EHR notes describe change of status (eg, “recently lost job” or “recently purchased a gun”) that may more precisely identify the current state of a patient. As a consequence, we can take advantage of the rich information provided by unstructured EHR notes via natural language processing (NLP) [20]. NLP has already been successfully utilized for essential information extraction from EHR text to examine various clinical problems, including opioid use and risk assessment [21,22].

With nonfatal ODs increasing, there is a growing need for critical care of these patients in the United States [23]. Although a relatively high proportion of nonfatal OD cases leads to intensive care unit (ICU) admission, little is known about the risk factors of OD for ICU admissions. [24]. This is essential to understand the severity of the opioid epidemic and anticipate critical care needs for patients with OD. There has been inadequate work on assessing risk factors associated with OD leading to ICU admission, which may be important in comprehensively preventing the public health problem of ODs.

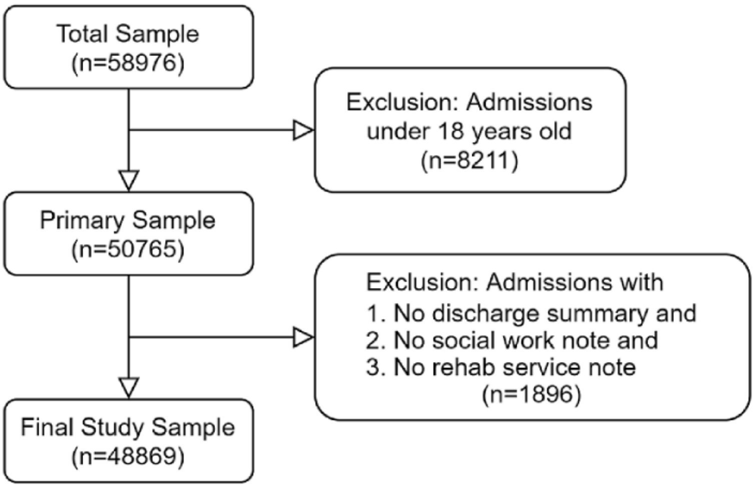
In this study, we specifically focused on the ICU setting to address the aforementioned issues. To mitigate the scarcity of structured SBDH information, we used an NLP system to automatically extract SBDH information from EHR notes and integrated that with available structured SBDH data entered upon admission. Then, we investigated the associations of various demographic, SBDH, and clinical variables with nonfatal OD for eligible ICU admissions. To date, none of the studies on OD utilized the EHR text for extracting SBDH information and few focused on the ICU setting. We bridge this gap by (1) showing that NLP systems can help extract SBDH information when structured data are inadequate and (2) identifying the risk factors that are crucial to the characterization of nonfatal OD leading to ICU admission.

Methods

Dataset

Our primary data source is MIMIC-III [25], one of the largest publicly available ICU databases encompassing 12 years of data (2001-2012) from Beth Israel Deaconess Medical Center. First, we excluded admission data from patients who were less than 18 years old at the time of admission. For inclusion, admissions were also required to have at least one note from any of these 3 categories: discharge summary, social work note, or rehabilitation service note. We selected these 3 types of notes to maximize the use of social and behavioral information for SBDH extraction: Discharge summaries are a comprehensive summary of a patient’s hospital stay, social work notes focus specifically on the social nature of a patient’s life, and rehabilitation service notes focus on improving patients' function and mobility to stabilize them for discharge. The final sample consisted of 48,869 hospital admissions from 37,361 patients. An overview of the data selection process is shown in Figure 1.

Figure 1. Data selection process.


Total Sample
(n=58976)
Exclusion: Admissions under 18 years old
(n=8211)
Primary Sample
(n=50765)
Exclusion: Admissions with
1. No discharge summary and
2. No social work note and
3. No rehab service note
(n=1896)
Final Study Sample
(n=48869)

Variables

All baseline variables were grouped into 3 categories: demographic, clinical, and SBDH. The demographic variables included age (18-39 years, 40-64 years, >64 years), gender (male or female), race/ethnicity (White, Black, Hispanic, or others), and marital status (married, divorced, widowed, single, or unknown marital status). As clinical variables, we considered drug use disorder, bipolar disorder, tobacco use disorder, major depressive disorder, alcohol use disorder, cirrhosis, chronic obstructive pulmonary disease (COPD), and renal insufficiency. This comprehensive list was made based on earlier studies related to OD [26-29], clinical judgment, and statistical analyses (see the "Statistical Analysis" section for further details). All clinical variables were detected using the International Classification of Disease, 9th Revision (ICD-9) codes from the admission diagnosis chart and included as dichotomous variables. The list of ICD-9 codes is available in Multimedia Appendix 1.

For SBDH variables, we used NLP to analyze the unstructured text data available in MIMIC-III. For each type of note, we chose the most relevant sections to extract the SBDH information: (1) discharge summaries: "Social History" sections; (2) social worker notes: "Patient/Family Assessment," "Past Addictions History," "Past Medical History" sections; (3) rehabilitation services: "Sexual and Social History" section.

We used the popular clinical NLP tool medSpaCy [30] to extract these sections from a note. We randomly chose a note, extracted the relevant sections as mentioned, and annotated for 6 categories of SBDH information. This process was repeatedly followed until we reached 1000 notes with at least one SBDH annotation. This annotated subset was later used to train a Bidirectional Encoder Representations from Transformers (BERT) model to extract SBDH at the word level. BERT [31] is a state-of-the-art language representation model that has successfully outperformed many other NLP systems across a wide range of tasks. We used the trained model to predict the SBDH information for the remaining notes. For an admission with multiple notes of the same type, we took the last note as representative of that admission as it typically includes the content of all the previous notes.

The 6 SBDH variables we chose were (1) housing insecurity, (2) unemployment, (3) social isolation, (4) alcohol use, (5) tobacco use, and (6) illicit drug use. The first 3 are social determinants and were selected based on the list of well-accepted social determinants provided by the Kaiser Family Foundation [32]. The rest were substance use–related health risk behaviors (ie, behavioral determinants) that were chosen for their clinical significance and relevance to OD. Details about the annotation process, NLP model development, and SBDH variable extraction procedures are provided in Multimedia Appendix 2.

In addition to the NLP-derived SBDH variables, we also identified social determinants from the structured data. We used the ICD-9 codes from patient diagnoses [33] to construct these 3 SBDH variables: (1) housing insecurity, (2) unemployment, and (3) social isolation. These were later integrated with the NLP-derived SBDH variables and prioritized in case of any mismatch. For example, if the NLP system detected "housing insecurity" as "No" for an admission and we obtained "Yes" from that admission's diagnoses codes, we considered "Yes" as the correct value. In the end, there were 41,669 admissions (41,669/48,869, 85.27%) with at least one SBDH variable. Table 1 illustrates the 6 SBDH variables with brief descriptions and examples. If an admission had no mention of SBDH information, SBDH variables were coded as "unknown." For instance, if an admission had no mention of patient housing status in the corresponding notes, homelessness was considered "unknown." Other than these 3 SBDH variables, we also extracted insurance provider (private, Medicaid, Medicare, other government, or self-pay) information using ICD-9 codes.

Table 1. Descriptions and examples of social and behavioral determinants of health (SBDH) variables.

SBDH Variable	Description and example
Housing insecurity	
Yes	Lack housing or stable shelter. Example: *homeless*, living with friends.
No	Has access to housing. Example: *lives* in [**location**] by himself.
Unemployment	
Yes	Patient has no source of income or lost job. Example: Patient used to work for the state lottery system, currently *unemployed*.
No	Patient has employment or some source of income. Example: He works for [**Company**].
Social isolation	
Yes	Lack of social support or community engagement. Example: Lives *alone* in [**Location**].
No	Presence of social support. Example: He is *married* and *lives* with his wife.
Alcohol use	
Current	Patient currently consumes alcohol. Example: two glasses of *wine* per night and 3 bottles over the weekend.
Former	Patient has a history of alcohol consumption. Example: He has a history of *alcohol* abuse.
None	Patient never consumed alcohol. Example: She denies any *alcohol use*.
Smoking	
Current	Patient currently smokes. Example: He *smokes one pack of cigarettes* per week.
Former	Patient has a history of tobacco usage. Example: The patient has a past history of *smoking*.
None	Patient never consumed tobacco. Example: She is a *nonsmoker*.
Illicit drug use	
Current	Patient uses non-prescribed controlled substance. Example: occasional *marijuana* use.
Former	Patient has a history of using non-prescribed controlled substance. Example: Has a h/o[a] of *cocaine* and *marijuana* abuse.
None	Patient never used non-prescribed controlled substance, e.g., cocaine, marijuana. Example: Does not drink alcohol or use recreational *drugs*.

[a] h/o: history of.

Outcome

The outcome was nonfatal OD, which was identified using ICD-9 codes [34].

Statistical Analysis

First, we performed correlation and collinearity analyses for all the variables. The correlation plot and variance inflation factor [35] did not show multicollinearity among the variables. For the clinical variables, based on earlier work and task relevance, we chose 14 comorbidities: posttraumatic stress disorder, major depressive disorder, bipolar disorder, schizophrenia, alcohol use disorder, drug use disorder, tobacco use disorder, hepatitis C, diabetes, congestive heart failure, obstructive sleep apnea, COPD, cirrhosis, and renal insufficiency. We built logistic regression models and employed the sequential forward selection procedure [36] to identify the most essential clinical variables related to OD. The final list included 8 clinical variables: drug use disorder, bipolar disorder, tobacco use disorder, major depressive disorder, alcohol use disorder, cirrhosis, COPD, and renal insufficiency.

We used a logistic regression model to examine the associations of nonfatal OD with demographic, SBDH, and clinical variables. This was assessed in terms of adjusted odds ratios (aOR) with 95% CIs. We also evaluated the crude odds ratio (OR) with 95% CIs. The statistical significance was measured at *P*<.05. Hosmer-Lemeshow test was conducted and indicated a sufficient fit for our model (χ_8=10.39; *P*=.24). All statistical analyses in this study were conducted in R (version 4.0.2).

Results

Descriptive Analysis

Table 2 presents the characteristics of our cohort (n=48,869). Our sample was comprised of mostly men (27,436/48,869, 56.14%) and white (35,058/48,869, 71.74%) adults. The majority of patients were aged 64 years or older (25,276/48,869, 51.72%). Of the clinical variables, renal insufficiency was the most prevalent (8158/48,869, 16.69%), followed by COPD (5674/48,869, 11.61%) and alcohol use disorder (4121/48,869, 8.43%). In our cohort, we observed that 7.28% (3559/48,869) of the patients were unemployed, 13.35% (6523/48,869) were socially isolated, and 0.82% (402/48,869) had housing insecurity. We found 171 (171/48,869, 0.35%) admissions with nonfatal OD.

Table 2. Prevalence of demographic, clinical, and social and behavioral determinants of health (SBDH) variables in MIMIC-III.

Variables	Overall (n=48,869)	With OD[a] (n=171)	Without OD (n=48,698)
Age[b] (years), n (%)			
<40	4715 (9.65)	62 (36.26)	4653 (9.55)
40-64	18,878 (38.63)	92 (53.80)	18,786 (38.58)
>64	25,276 (51.72)	17 (9.94)	25,259 (51.87)
Gender,[b] n (%)			
Male	27,436 (56.14)	100 (58.48)	27,336 (56.13)
Female	21,433 (43.86)	71 (41.52)	21,362 (43.87)
Race/ethnicity,[b] n (%)			
White	35,058 (71.74)	127 (74.27)	34,931 (71.73)
Black	4694 (9.61)	13 (7.60)	4681 (9.61)
Hispanic	1664 (3.40)	8 (4.68)	1656 (3.40)
Other	7453 (15.25)	23 (13.45)	7430 (15.26)
Marital status,[b] n (%)			
Married	23,378 (47.84)	42 (24.56)	23,336 (47.92)
Divorced	3664 (7.50)	22 (12.87)	3642 (7.48)
Widowed	7018 (14.36)	6 (3.51)	7012 (14.40)
Single	12,329 (25.23)	78 (45.61)	12,251 (25.16)
Unknown	2480 (5.07)	23 (13.45)	2457 (5.04)
Clinical variables,[b] n (%)			
Drug use disorder	1493 (3.06)	80 (46.78)	1413 (2.90)
Bipolar disorder	1009 (2.06)	28 (16.37)	981 (2.01)
Tobacco use disorder	3274 (6.70)	20 (11.70)	3254 (6.68)
Major depressive disorder	298 (0.61)	7 (4.09)	291 (0.60)
Alcohol use disorder	4121 (8.43)	37 (21.64)	4084 (8.39)
Cirrhosis	2431 (4.97)	19 (11.11)	2412 (4.95)
COPD[c]	5674 (11.61)	18 (10.53)	5656 (11.61)
Renal insufficiency	8158 (16.69)	12 (7.02)	8146 (16.73)
Social determinant[d]: insurance provider, n (%)			
Private	15,371 (31.45)	43 (25.15)	15,328 (31.48)
Medicaid	4307 (8.81)	60 (35.09)	4247 (8.72)
Medicare	27,365 (56.00)	48 (28.07)	27,317 (56.09)
Government (others)	1324 (2.71)	14 (8.19)	1310 (2.69)
Self-pay	502 (1.03)	6 (3.50)	496 (1.02)
Social determinant[d]: housing insecurity, n (%)			
Yes	402 (0.82)	10 (5.85)	392 (0.80)
No	27,119 (55.49)	92 (53.80)	27,027 (55.50)
Unknown	21,348 (43.69)	69 (40.35)	21,279 (43.70)
Social determinant[d]: unemployment, n (%)			
Yes	3559 (7.28)	37 (21.64)	3522 (7.22)
No	12,671 (25.93)	31 (18.13)	12,640 (25.96)

Variables	Overall (n=48,869)	With OD[a] (n=171)	Without OD (n=48,698)
Unknown	32,639 (66.79)	103 (60.23)	32,536 (66.82)
Social determinant[d]: social isolation, n (%)			
Yes	6523 (13.35)	23 (13.45)	6500 (13.35)
No	24,001 (49.11)	86 (50.29)	23,915 (49.11)
Unknown	18,345 (37.54)	62 (36.26)	18,283 (37.54)
Substance use[e]: alcohol use, n (%)			
Current	14,150 (28.96)	70 (40.94)	14,080 (28.91)
Former	2333 (4.77)	9 (5.26)	2324 (4.77)
None	15,378 (31.47)	40 (23.39)	15,338 (31.50)
Unknown	17,008 (34.80)	52 (30.41)	16,956 (34.82)
Substance use[e]: smoking, n (%)			
Current	6954 (14.23)	62 (36.26)	6892 (14.15)
Former	12,032 (24.62)	23 (13.45)	12,009 (24.66)
None	13,963 (28.57)	30 (17.54)	13,933 (28.61)
Unknown	15,920 (32.58)	56 (32.75)	15,864 (32.58)
Substance use[e]: illicit drug use, n (%)			
Current	1796 (3.67)	49 (28.65)	1747 (3.59)
Former	1362 (2.79)	26 (15.20)	1336 (2.74)
None	13,908 (28.46)	31 (18.13)	13,877 (28.50)
Unknown	31,803 (65.08)	65 (38.02)	31,738 (65.17)

[a] OD: opioid overdose.

[b] Variables extracted from structured data.

[c] COPD: chronic obstructive pulmonary disease.

[d] Variables extracted from only structured data (insurance provider) or both structured data and unstructured text notes (natural language processing [NLP]).

[e] Variables extracted from unstructured text notes (NLP).

Of the 6 NLP-derived SBDH variables, only housing insecurity, unemployment, and social isolation had associated ICD-9 diagnostic codes. Compared with their NLP-derived counterparts, these structured variables were coded infrequently. For example, using ICD-9 codes, we found 258 admissions with "housing insecurity," whereas the NLP system detected 402 admissions. For "unemployment," it was 20 for the ICD-9 codes and 10,876 for the NLP system. And more striking, for "social isolation," only 4 admissions had relevant ICD-9 codes in their diagnosis compared to 6523 admissions found by the NLP system. Due to the substantial prevalence gap, we did not compare the quality of these 2 types of SBDH variables side by side. In all, structured SBDH variables accounted for only 0.18% of the SBDH variables. This clearly shows that NLP can be useful to extract SBDH information from EHR notes when structured data are not enough. This also helps reduce bias from the use of structured data only.

Multivariable Logistic Regression Analysis

Several factors were strongly associated with nonfatal OD (Table 3). Among the demographic risk factors, Blacks (aOR 0.51, 95% CI 0.28-0.94) and older age groups (40-64 years: aOR 0.65, 95% CI 0.44-0.96; >64 years: aOR 0.16, 95% CI 0.08-0.34) had lower odds compared with White and younger patients. Among the 8 clinical variables, 5 were strong risk factors for nonfatal OD. We observed increased odds of overdose among individuals with drug use disorder (aOR 8.17, 95% CI 5.44-12.27), bipolar disorder (aOR 2.69, 95% CI 1.68-4.29), and major depressive disorder (aOR 2.57, 95% CI 1.12-5.88). Interestingly, tobacco use disorder (aOR 0.53, 95% CI 0.32-0.89) and alcohol use disorder (aOR 0.64, 95% CI 0.42-1.00) had decreased odds. Among the SBDH variables, individuals with Medicaid had increased odds compared with those with private medical insurance (aOR 2.26, 95% CI 1.43-3.58). History of (aOR 2.09, 95% CI 1.15-3.79) and current (aOR 2.06, 95% CI 1.20-3.55) use of illicit drugs were also strongly associated with the outcome.

Table 3. Multivariable logistic regression analysis for the factors associated with nonfatal opioid overdose (OD).

Variables	Crude OR[a]	95% CI	aOR[b]	95% CI
Age (years)				
<40	Ref[c]	Ref	Ref	Ref
40-64	0.37	0.27-0.51	0.65	0.44-0.96
>64	0.05	0.03-0.08	0.16	0.08-0.34
Gender				
Male	Ref	Ref	Ref	Ref
Female	0.91	0.67-1.23	1.13	0.81-1.58
Race/ethnicity				
White	Ref	Ref	Ref	Ref
Black	0.76	0.41-1.30	0.51	0.28-0.94
Hispanic	1.33	0.60-2.55	0.69	0.33-1.45
Others	0.85	0.53-1.30	0.59	0.35-0.98
Marital status				
Married	Ref	Ref	Ref	Ref
Divorced	3.36	1.97-5.57	1.56	0.89-2.74
Widowed	0.48	0.18-1.04	0.76	0.30-1.88
Single	3.54	2.44-5.19	1.03	0.65-1.61
Unknown	5.20	3.08-8.58	2.85	1.55-5.24
Clinical variables				
Drug use disorder	29.42	21.65-39.90	8.17	5.44-12.27
Bipolar disorder	9.52	6.20-14.12	2.69	1.68-4.29
Tobacco use disorder	1.85	1.12-2.88	0.53	0.32-0.89
Major depressive disorder	7.10	2.99-14.16	2.57	1.12-5.88
Alcohol use disorder	3.02	2.06-4.30	0.64	0.42-1.00
Cirrhosis	2.40	1.44-3.77	1.65	0.97-2.82
COPD[d]	7.10	2.99-14.16	1.65	0.97-2.81
Renal insufficiency	3.02	2.06-4.30	0.62	0.33-1.15
Social determinant: insurance type				
Private	Ref	Ref	Ref	Ref
Medicaid	5.04	3.41-7.50	2.26	1.43-3.58
Medicare	0.63	0.41-0.95	1.34	0.81-2.23
Government (others)	3.81	2.01-6.80	1.90	0.99-3.65
Self-paid	4.31	1.64-9.42	1.83	0.73-4.56
Social determinant: housing insecurity				
No	Ref	Ref	Ref	Ref
Yes	7.49	3.63-13.80	0.98	0.47-2.06
Unknown	0.95	0.69-1.30	0.89	0.60-1.33
Social determinant: unemployment				
No	Ref	Ref	Ref	Ref
Yes	4.28	2.66-6.95	1.10	0.65-1.87
Unknown	1.29	0.87-1.96	0.73	0.47-1.14

Variables	Crude OR[a]	95% CI	aOR[b]	95% CI
Social determinant: social isolation				
No	Ref	Ref	Ref	Ref
Yes	0.98	0.61-1.53	0.97	0.59-1.60
Unknown	0.94	0.68-1.31	1.01	0.66-1.53
Substance use: alcohol use				
None	Ref	Ref	Ref	Ref
Former	1.48	0.67-2.92	0.66	0.30-1.44
Current	1.91	1.30-2.84	1.11	0.71-1.72
Unknown	1.18	0.78-1.79	1.05	0.62-1.78
Substance use: smoking				
None	Ref	Ref	Ref	Ref
Former	0.89	0.51-1.53	0.92	0.52-1.65
Current	4.18	2.72-6.55	1.40	0.84-2.33
Unknown	1.64	1.06-2.59	1.12	0.64-1.96
Substance use: illicit drug use				
None	Ref	Ref	Ref	Ref
Former	8.71	5.12-14.7	2.09	1.15-3.79
Current	12.56	8.03-19.93	2.06	1.20-3.55
Unknown	0.92	0.60-1.42	1.05	0.65-1.70

[a] OR: odds ratio.

[b] aOR: adjusted odds ratio.

[c] Ref: Reference.

[d] COPD: chronic obstructive pulmonary disorder.

Discussion

Principal Findings

To our knowledge, this is the first study to examine the risk factors associated with nonfatal OD leading to ICU admission. In the United States, the need for characterizing critical care patients with OD is rising [23,24], and this study partially addressed that by identifying the risk factors for nonfatal OD from a large ICU database. The novelty also lies in the use of a state-of-the-art NLP system that utilized unstructured EHR notes for essential SBDH extraction due to inadequate representation from structured data. There is a growing body of literature showing that SBDH can strongly influence patient health and outcomes [12]. For example, SBDH variables have been shown to be strongly associated with suicide attempt [11], mortality [17], and mental health diagnosis [18]. The challenges here for the health care systems are to set up methods that can identify SBDH and use them at the point of care to inform clinical action [37,38]. Our work demonstrated that using NLP to detect SBDH information from EHR text can be a viable option in this regard.

According to our analysis, multiple SBDH variables were significantly associated with nonfatal OD in ICU settings. We observed that patients with economic instability (unemployed) were more likely to have an overdose, but homelessness and social isolation conferred little additional risk. Among behavioral determinants, *current* alcohol users and smokers had higher odds of overdose, whereas *former* users had decreased odds. Illicit drug use was strongly associated with nonfatal OD for both *former* and *current* users. Among clinical variables, tobacco use disorder and alcohol use disorder had strong negative associations with nonfatal OD. We hypothesize that the majority of the patients diagnosed with such disorders were already receiving additional social counseling or clinical support, which helped them build better health and behavioral practices. However, we did not have enough relevant admission data in MIMIC-III to validate this hypothesis; future research is needed to identify the reasons for this observation.

Limitations

There are several limitations of our study. EHR data are prone to variability by provider documentation and may contain incomplete SBDH information [39]. Additionally, using only ICD-9 codes to identify different medical conditions may lead to inaccurate or misleading values for the corresponding variable. However, structured data often significantly lack SBDH information (only 0.18% for this study), making an NLP-based approach a valuable integration for population studies. Finally, our data had a very low prevalence of nonfatal OD cases (171/48,869, 0.35%), and the MIMIC (ICU) database might not characterize the general outpatient/inpatient hospital setting.

While our study describes an important methodological process that can identify important SBDH factors to consider, which is a necessary first step, further research is needed on subsequent steps on how best to share and translate this information to providers so that they can effectively and actionably use the findings. As our future work, we would like to work on modeling the NLP system predictions for SBDH extraction and how they can be better tied with predictor assessment metrics (eg, OR).

Conclusions

This is the first work to evaluate the risk factors associated with nonfatal OD leading to ICU admissions. Our work concluded that data-driven NLP systems can be largely beneficial in the automatic extraction of SBDH information from unstructured EHR text data. We also showed that analyzing critical care admissions is crucial to better understand the opioid epidemic. Utilizing NLP to leverage the rich EHR notes and more epidemiological studies in critical care settings could be useful for deeper analysis of the OD crisis, leading to the development of better risk assessment tools and effective prevention systems.

Acknowledgments

We thank Minhee Sung, Jimin Kim, and Chen Kun for their valuable comments. This work was supported in part by the grant R01DA045816 from the National Institutes of Health (NIH). The contents of this paper do not represent the views of the NIH.

Conflicts of Interest

None declared.

Multimedia Appendix 1
International Classification of Disease, 9th Revision codes for clinical variables.
[DOCX File , 36 KB - medinform_v9i11e32851_app1.docx]

Multimedia Appendix 2
Natural language processing model training and evaluation.
[DOCX File , 23 KB - medinform_v9i11e32851_app2.docx]

References

1. Overdose Death Rates. National Institute on Drug Abuse (NIDA). 2017. URL: https://www.drugabuse.gov/drug-topics/trends-statistics/overdose-death-rates [accessed 2021-10-16]
2. Opioid Overdose Crisis. National Institute on Drug Abuse (NIDA). URL: https://www.drugabuse.gov/drug-topics/opioids/opioid-overdose-crisis [accessed 2021-10-16]
3. Luo F, Li M, Florence C. State-level economic costs of opioid use disorder and fatal opioid overdose - United States, 2017. MMWR Morb Mortal Wkly Rep 2021 Apr 16;70(15):541-546 [FREE Full text] [doi: 10.15585/mmwr.mm7015a1] [Medline: 33857070]
4. Dowell D, Arias E, Kochanek K, Anderson R, Guy GP, Losby JL, et al. Contribution of opioid-involved poisoning to the change in life expectancy in the United States, 2000-2015. JAMA 2017 Sep 19;318(11):1065-1067 [FREE Full text] [doi: 10.1001/jama.2017.9308] [Medline: 28975295]
5. Case A, Deaton A. Mortality and morbidity in the 21 century. Brookings Pap Econ Act 2017;2017:397-476 [FREE Full text] [doi: 10.1353/eca.2017.0005] [Medline: 29033460]
6. Dasgupta N, Beletsky L, Ciccarone D. Opioid crisis: no easy fix to its social and economic determinants. Am J Public Health 2018 Feb;108(2):182-186. [doi: 10.2105/ajph.2017.304187]
7. Volkow ND, Blanco C. The changing opioid crisis: development, challenges and opportunities. Mol Psychiatry 2021 Jan;26(1):218-233 [FREE Full text] [doi: 10.1038/s41380-020-0661-4] [Medline: 32020048]
8. Cole BL, Fielding JE. Health impact assessment: a tool to help policy makers understand health beyond health care. Annu Rev Public Health 2007;28:393-412. [doi: 10.1146/annurev.publhealth.28.083006.131942] [Medline: 17173539]
9. Kposowa AJ. Unemployment and suicide: a cohort analysis of social factors predicting suicide in the US National Longitudinal Mortality Study. Psychol Med 2001 Jan;31(1):127-138. [doi: 10.1017/s0033291799002925] [Medline: 11200951]
10. Dube SR, Anda RF, Felitti VJ, Chapman DP, Williamson DF, Giles WH. Childhood abuse, household dysfunction, and the risk of attempted suicide throughout the life span: findings from the Adverse Childhood Experiences Study. JAMA 2001 Dec 26;286(24):3089-3096. [doi: 10.1001/jama.286.24.3089] [Medline: 11754674]
11. Blosnich JR, Montgomery AE, Dichter ME, Gordon AJ, Kavalieratos D, Taylor L, et al. Social determinants and military veterans' suicide ideation and attempt: a cross-sectional analysis of electronic health record data. J Gen Intern Med 2020 Jun;35(6):1759-1767 [FREE Full text] [doi: 10.1007/s11606-019-05447-z] [Medline: 31745856]

12. Feller DJ, Bear Don't Walk Iv OJ, Zucker J, Yin MT, Gordon P, Elhadad N. Detecting social and behavioral determinants of health with structured and free-text clinical data. Appl Clin Inform 2020 Jan 04;11(1):172-181 [FREE Full text] [doi: 10.1055/s-0040-1702214] [Medline: 32131117]
13. Cantu R, Fields-Johnson D, Savannah S. Applying a social determinants of health approach to the opioid epidemic. Health Promot Pract 2020 Jul 26:1524839920943207. [doi: 10.1177/1524839920943207] [Medline: 32713219]
14. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving electronic medical records upstream: incorporating social determinants of health. Am J Prev Med 2015 Feb;48(2):215-218. [doi: 10.1016/j.amepre.2014.07.009] [Medline: 25217095]
15. Weir CR, Staggers N, Gibson B, Doing-Harris K, Barrus R, Dunlea R. A qualitative evaluation of the crucial attributes of contextual information necessary in EHR design to support patient-centered medical home care. BMC Med Inform Decis Mak 2015 Apr 16;15:30 [FREE Full text] [doi: 10.1186/s12911-015-0150-x] [Medline: 25881181]
16. Gottlieb L, Sandel M, Adler NE. Collecting and applying data on social determinants of health in health care settings. JAMA Intern Med 2013 Jun 10;173(11):1017-1020. [doi: 10.1001/jamainternmed.2013.560] [Medline: 23699778]
17. Blosnich JR, Montgomery AE, Taylor LD, Dichter ME. Adverse social factors and all-cause mortality among male and female patients receiving care in the Veterans Health Administration. Prev Med 2020 Dec;141:106272. [doi: 10.1016/j.ypmed.2020.106272] [Medline: 33022319]
18. Blosnich JR, Marsiglio MC, Dichter ME, Gao S, Gordon AJ, Shipherd JC, et al. Impact of social determinants of health on medical conditions among transgender veterans. Am J Prev Med 2017 Apr;52(4):491-498 [FREE Full text] [doi: 10.1016/j.amepre.2016.12.019] [Medline: 28161034]
19. Dorr D, Bejan CA, Pizzimenti C, Singh S, Storer M, Quinones A. Identifying patients with significant problems related to social determinants of health with natural language processing. Stud Health Technol Inform 2019 Aug 21;264:1456-1457. [doi: 10.3233/SHTI190482] [Medline: 31438179]
20. Liu F, Weng C, Yu H. Natural language processing, electronic health records, and clinical research. In: Richesson R, Andrews JE, editors. Clinical Research Informatics. New York City, NY: Springer Publishing Company; 2012:293-310.
21. Carrell DS, Cronkite D, Palmer RE, Saunders K, Gross DE, Masters ET, et al. Using natural language processing to identify problem usage of prescription opioids. Int J Med Inform 2015 Dec;84(12):1057-1064. [doi: 10.1016/j.ijmedinf.2015.09.002] [Medline: 26456569]
22. Haller IV, Renier CM, Hitz P, Palcher JA, Elliott TE. Validation of the automated diagnosis, intractability, risk, efficacy (DIRE) opioid risk assessment tool. Journal of Patient-Centered Research and Reviews 2016 Aug 15;3(3):227. [doi: 10.17294/2330-0698.1403]
23. Stevens JP, Wall MJ, Novack L, Marshall J, Hsu DJ, Howell MD. The critical care crisis of opioid overdoses in the United States. Annals ATS 2017 Dec;14(12):1803-1809. [doi: 10.1513/annalsats.201701-022oc]
24. Pfister GJ, Burkes RM, Guinn B, Steele J, Kelley RR, Wiemken TL, et al. Opioid overdose leading to intensive care unit admission: Epidemiology and outcomes. J Crit Care 2016 Oct;35:29-32. [doi: 10.1016/j.jcrc.2016.04.022] [Medline: 27481733]
25. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3(1):160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
26. Bohnert ASB, Valenstein M, Bair MJ, Ganoczy D, McCarthy JF, Ilgen MA, et al. Association between opioid prescribing patterns and opioid overdose-related deaths. JAMA 2011 Apr 06;305(13):1315-1321. [doi: 10.1001/jama.2011.370] [Medline: 21467284]
27. Dunn KM, Saunders KW, Rutter CM, Banta-Green CJ, Merrill JO, Sullivan MD, et al. Opioid prescriptions for chronic pain and overdose: a cohort study. Ann Intern Med 2010 Jan 19;152(2):85-92 [FREE Full text] [doi: 10.7326/0003-4819-152-2-201001190-00006] [Medline: 20083827]
28. Campbell CI, Bahorik AL, VanVeldhuisen P, Weisner C, Rubinstein AL, Ray GT. Use of a prescription opioid registry to examine opioid misuse and overdose in an integrated health system. Prev Med 2018 May;110:31-37 [FREE Full text] [doi: 10.1016/j.ypmed.2018.01.019] [Medline: 29410132]
29. Glanz JM, Narwaney KJ, Mueller SR, Gardner EM, Calcaterra SL, Xu S, et al. Prediction model for two-year risk of opioid overdose among patients prescribed chronic opioid therapy. J Gen Intern Med 2018 Oct 29;33(10):1646-1653 [FREE Full text] [doi: 10.1007/s11606-017-4288-3] [Medline: 29380216]
30. medspacy. GitHub. URL: https://github.com/medspacy/medspacy [accessed 2021-10-16]
31. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. 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) 2019:4171-4186 [FREE Full text] [doi: 10.18653/v1/N19-1423]
32. Heiman HJ, Artiga S. Beyond Health Care: The Role of Social Determinants in Promoting Health and Health Equity. Kaiser Family Foundation. 2018 May 10. URL: https://www.kff.org/racial-equity-and-health-policy/issue-brief/beyond-health-care-the-role-of-social-determinants-in-promoting-health-and-health-equity/ [accessed 2021-10-16]

33. Kessler RC, Bauer MS, Bishop TM, Demler OV, Dobscha SK, Gildea SM, et al. Using administrative data to predict suicide after psychiatric hospitalization in the Veterans Health Administration System. Front Psychiatry 2020 May 6;11:390 [FREE Full text] [doi: 10.3389/fpsyt.2020.00390] [Medline: 32435212]
34. Glanz J, Binswanger I, Shetterly SM, Narwaney KJ, Xu S. Association between opioid dose variability and opioid overdose among adults prescribed long-term opioid therapy. JAMA Netw Open 2019 Apr 05;2(4):e192613 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.2613] [Medline: 31002325]
35. O'Brien RM. A caution regarding rules of thumb for variance inflation factors. Qual Quant 2007 Mar 13;41(5):673-690. [doi: 10.1007/s11135-006-9018-6]
36. Ferri FJ, Pudil P, Hatef M, Kittler J. Comparative study of techniques for large-scale feature selection. Machine Intelligence and Pattern Recognition 1994;16:403-413. [doi: 10.1016/B978-0-444-81892-8.50040-7]
37. Schickedanz A, Hamity C, Rogers A, Sharp A, Jackson A. Clinician experiences and attitudes regarding screening for social determinants of health in a large integrated health system. Med Care 2019 Jun;57 Suppl 6 Suppl 2:S197-S201 [FREE Full text] [doi: 10.1097/MLR.0000000000001051] [Medline: 31095061]
38. Horwitz LI, Chang C, Arcilla HN, Knickman JR. Quantifying health systems' investment in social determinants of health, by sector, 2017-19. Health Aff (Millwood) 2020 Feb;39(2):192-198. [doi: 10.1377/hlthaff.2019.01246] [Medline: 32011928]
39. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A data quality assessment guideline for electronic health record data reuse. EGEMS (Wash DC) 2017 Sep 04;5(1):14 [FREE Full text] [doi: 10.5334/egems.218] [Medline: 29881734]

Abbreviations

aOR: adjusted odds ratio
BERT: Bidirectional Encoder Representations from Transformers
COPD: chronic obstructive pulmonary disease
EHR: Electronic health record
ICD-9: International Classification of Disease, 9th Revision
ICU: Intensive care unit
NIH: National Institutes of Health
NLP: natural language processing
OD: opioid overdose
OR: odds ratio
SBDH: social and behavioral determinants of health
SDOH: social determinants of health

Edited by G Eysenbach; submitted 12.08.21; peer-reviewed by J Coquet; comments to author 05.09.21; revised version received 23.09.21; accepted 26.09.21; published 08.11.21.

Please cite as:
Mitra A, Ahsan H, Li W, Liu W, Kerns RD, Tsai J, Becker W, Smelson DA, Yu H
Risk Factors Associated With Nonfatal Opioid Overdose Leading to Intensive Care Unit Admission: A Cross-sectional Study
JMIR Med Inform 2021;9(11):e32851
URL: https://medinform.jmir.org/2021/11/e32851
doi:10.2196/32851
PMID:34747714

Original Paper

A Neural Network Approach for Understanding Patient Experiences of Chronic Obstructive Pulmonary Disease (COPD): Retrospective, Cross-sectional Study of Social Media Content

Tobe Che Benjamin Freeman[1,2], BSc, DPhil; Raul Rodriguez-Esteban[1], PhD; Juergen Gottowik[1], PhD; Xing Yang[3], PhD; Veit Johannes Erpenbeck[4], MD; Mathias Leddin[1], PhD

[1]Roche Pharma Research and Early Development, Pharma Research and Early Development Informatics, Roche Innovation Center Basel, F. Hoffmann–La Roche Ltd, Basel, Switzerland

[2]wordup development AG, CH-8006, Zurich, Switzerland

[3]Roche Pharma Research and Early Development, Pharma Research and Early Development Informatics, Roche Innovation Center Little Falls, F. Hoffmann–La Roche Ltd, Little Falls, NJ, United States

[4]Roche Pharma Research and Early Development, Immunology, Infectious Diseases and Ophthalmology Discovery and Translational Area, Roche innovation Center Basel, F. Hoffmann–La Roche Ltd, Basel, Switzerland

Corresponding Author:
Tobe Che Benjamin Freeman, BSc, DPhil
Roche Pharma Research and Early Development
Pharma Research and Early Development Informatics
Roche Innovation Center Basel, F. Hoffmann–La Roche Ltd
Grenzacherstrasse 124
Basel, CH-4070
Switzerland
Phone: 41 793778595
Email: tobefreeman@gmail.com

Abstract

Background: The abundance of online content contributed by patients is a rich source of insight about the lived experience of disease. Patients share disease experiences with other members of the patient and caregiver community and do so using their own lexicon of words and phrases. This lexicon and the topics that are communicated using words and phrases belonging to the lexicon help us better understand disease burden. Insights from social media may ultimately guide clinical development in ways that ensure that future treatments are fit for purpose from the patient's perspective.

Objective: We sought insights into the patient experience of chronic obstructive pulmonary disease (COPD) by analyzing a substantial corpus of social media content. The corpus was sufficiently large to make manual review and manual coding all but impossible to perform in a consistent and systematic fashion. Advanced analytics were applied to the corpus content in the search for associations between symptoms and impacts across the entire text corpus.

Methods: We conducted a retrospective, cross-sectional study of 5663 posts sourced from open blogs and online forum posts published by COPD patients between February 2016 and August 2019. We applied a novel neural network approach to identify a lexicon of community words and phrases used by patients to describe their symptoms. We used this lexicon to explore the relationship between COPD symptoms and disease-related impacts.

Results: We identified a diverse lexicon of community words and phrases for COPD symptoms, including gasping, wheezy, mucus-y, and muck. These symptoms were mentioned in association with specific words and phrases for disease impact such as frightening, breathing discomfort, and difficulty exercising. Furthermore, we found an association between mucus hypersecretion and moderate disease severity, which distinguished mucus from the other main COPD symptoms, namely breathlessness and cough.

Conclusions: We demonstrated the potential of neural networks and advanced analytics to gain patient-focused insights about how each distinct COPD symptom contributes to the burden of chronic and acute respiratory illness. Using a neural network approach, we identified words and phrases for COPD symptoms that were specific to the patient community. Identifying patterns in the association between symptoms and impacts deepened our understanding of the patient experience of COPD. This approach can be readily applied to other disease areas.

(JMIR Med Inform 2021;9(11):e26272) doi:10.2196/26272

KEYWORDS

outcomes research; natural language processing; neural networks (computer); social media; exercise; sleep deprivation; social media listening; drug development

Introduction

Online content made public by patients in blogs and on forum platforms provides detailed first person accounts of the lived experience of disease [1,2]. These communications from patients use a diverse vocabulary of words and phrases for disease symptoms [3]. Online content is conveyed in the patient's own voice and is contributed in the ecological context of day-to-day life [4], namely in the sharing of experiences with other members of the patient and caregiver community. Analysis of these online communications enables a patient-centric approach to understanding disease impact.

A systematic understanding of the language used by patients to describe their symptoms has important clinical implications, not least being the need to acquire accurate patient anamneses and respond to care needs [5]. Dreisbach et al [6] note that the use of normalized medical vocabularies supports a systematic approach to identify terms for clinical and subclinical symptoms. This approach enables the identification of community terms that, while not belonging to a traditional medical lexicon, denote respiratory dysfunction unambiguously.

Many researchers use interviews, focus groups, and patient advisory boards with a goal of observing patient experiences. These approaches enable direct observation of the patient; however, they tend to be a burden to patients [7]. Moreover, interviews and focus groups are generally limited to cohorts of just a few patients, and the results are qualitative in nature.

In contrast, machine learning and related computational techniques offer a means to analyze online content at scale. Current state-of-the-art approaches using neural network architectures are being deployed to map patient community terms onto controlled medical [8] and pharmaceutical vocabularies [3]. However, these approaches are anchored in a defined lexicon of scientific terms, thus compromising patient centricity. In a patient-centric approach, our understanding of disease should instead be anchored to patients' self-reported topics [7], as observed in the ecological context of daily life [4], and not exclusively anchored to expert medical thinking, as expressed in a scientific lexicon.

We address this limitation with a novel approach based on a neural network, specifically a word embedding [9], to identify words and phrases that patients with chronic obstructive pulmonary disease (COPD) use to describe their experiences of living with the disease. Unlike traditional neural network approaches, a word embedding is not trained on any specific set of scientific keywords [10,11].

We use the word embedding to identify a diverse lexicon of hundreds of COPD-related words and phrases from the context in which words appear in a text. Next, we use that lexicon to extract all mentions of words and phrases relating to COPD symptoms and disease impacts from a large corpus of social media text. Once extracted, we can analyze the relationship between COPD symptoms and disease impacts at scale.

The quantitative analysis of this diverse community lexicon reveals insights [6] about the lived experience of COPD. These insights can contribute positively to the development of effective medical treatments that are, from the patient's perspective, fit for purpose [12].

Methods

Ethics

This work is compliant with ethical guidelines for the collection and analysis of user-generated content on open internet platforms. Data were downloaded only from open health social networking sites and communities. No information from restricted data areas has been downloaded (ie, content that requires an ID or password for access). No aggregation or enrichment of data on an individual has been performed. Extracts used for exemplary purposes were carefully paraphrased to protect the privacy of individuals.

Data Availability

All social media content included in our analysis was sourced from open social networking sites and communities. Terms and conditions apply to the availability of the original social media data. The sources used in this study can be made available upon request. Example texts shown in this manuscript have been rephrased to prevent de-anonymization of the individuals included in our analysis.

Neural Network Methodology

We trained the neural network on a corpus of 1.1 million words sourced from 22 individual blogs and online forums (Multimedia Appendix 1). We used the skip-gram negative sampling variant of the word2vec neural network algorithm described by Mikolov et al [9] to discover community words and phrases for disease symptoms. Briefly, the neural network model was trained to predict context words that appear in close proximity with symptom keywords in the corpus text.

The resulting word embedding captured semantic and syntactic features of each unique word in the text corpus. Neighboring vocabulary items in the embedding will likely share semantic and syntactic features in common. We then used cosine similarity as a metric to probe the word embedding model for words and phrases that share common meanings. This makes it possible to build and expand a lexicon of community terms for each main COPD symptom type in a systematic and repeatable manner (Table S1 in Multimedia Appendix 1).

We started our search for community words and phrases for COPD symptoms with a small seed lexicon that included breathlessness, cough, and sputum. This seed lexicon was sourced from MeSH terms from the US National Library of Medicine (NLM) [13] and from the NLM health information website for the layperson, MedlinePlus [14]. These 3 seed terms correspond to key pathophysiological manifestations of COPD, namely small airway fibrosis, emphysema, which refers to a destruction of the lungs' alveoli, and mucus hypersecretion [15-17].

We used the same approach to search for community words and phrases describing the impact of COPD on daily life. The seed terms for disease impacts include anxiety, depression, fatigue, pain, and exercise. We then scanned the entire corpus to detect posts in which COPD symptoms co-occur with mentions of disease-related impacts. Our analysis explored the relationship between specific symptoms and each of the main disease impact topics.

Results

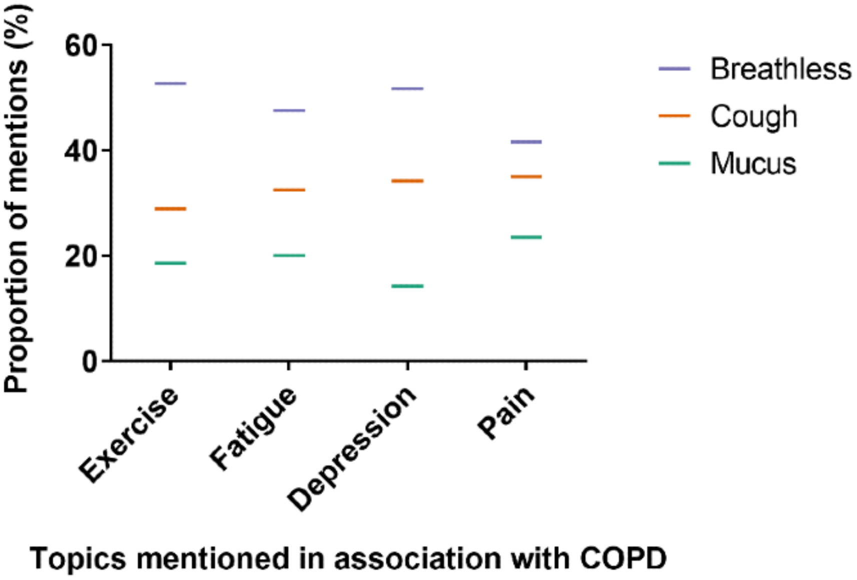
Using the cosine similarity metric to probe the word embedding model, close neighbors of the symptom seed term breathlessness included gasping, wheezy, and the phrase pursed-lip (Table S1 in Multimedia Appendix 1). The phrase pursed-lip is noteworthy as it refers to a technique, called pursed-lip breathing, used in pulmonary rehabilitation. Specifically, pursed-lip breathing is used to manage anxiety associated with breathlessness [18]. Words and phrases neighboring the seed term sputum include mucus-y, phlegm, clear mucus, and muck, as well as common misspellings of phlegm.

Probing the word embedding model with the seed term exercise, we found walk and the phrases low impact and difficulty exercising (Table S1 in Multimedia Appendix 1). These community terms are, as we might expect, for a relatively aged and exercise-limited patient cohort [19]. Manual inspection of individual excerpts from the corpus featuring symptom keywords further confirmed the relevance of these keywords (Table S4 in Multimedia Appendix 1).

Summing the number of mentions corresponding to each symptom lexicon across the entire corpus (Table S2 in Multimedia Appendix 1), the breathlessness lexicon was mentioned most frequently (mentioned in 10.49% [413/3938] of posts), followed by the lexicon for cough (270/3938, 6.86%) and, finally, mucus hypersecretion (159/3938, 4.04%).

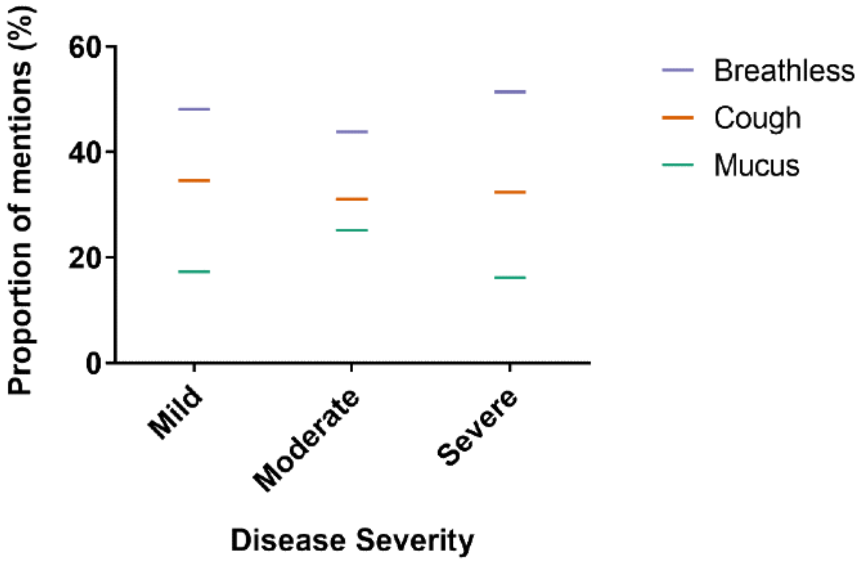
Leveraging these distinct lexicons of symptoms and disease impacts (Table S3 in Multimedia Appendix 1), we were able to explore the relationship between specific symptoms and each of the main disease-impact topics. Figure 1 examines posts in which COPD symptoms co-occurred with mentions of disease-related impacts. The analysis shows that breathlessness was the symptom most frequently mentioned in association with the 4 main topics and impacts considered. The most frequent disease impact associated with COPD symptoms was fatigue, followed closely by self-reports of anxiety and depression.

Figure 1. Topics co-occurring with symptom mentions in the same post.



Breathlessness and cough followed a broadly similar trend, while the trend in the co-occurrence between mucus and the 3 disease severity levels was distinctive (Figure 2). The co-occurrence between mucus and mild severity was lower than that between mucus and moderate disease severity, inverting the relationships observed for breathless and for cough. Taken together, it was apparent that there was an association between mucus and moderate disease severity that distinguished mucus from the symptoms breathlessness and cough.

Figure 2. Relationship between the symptoms of chronic obstructive pulmonary disease and disease severity.



By applying principal component analysis (PCA), we visualized semantic relationships [20,21] between each symptom lexicon and a mapping of the psychological salience of these symptoms. PCA arranged data points corresponding to individual words and phrases on a 2D map [20] (see Multimedia Appendix 1 for further details). Our PCA results showed that words and phrases belonging to the 3 symptom lexicons were arranged in 3 distinct clusters on this map (Figure S1 in Multimedia Appendix 1).

By adding a lexicon of affective states such as feel depressed and be embarrassing to the PCA map, we could explore the psychological salience of these symptoms. The lexicon of affective states also appeared as a distinct cluster on the map and was positioned closest to the cough symptom cluster. The mucus cluster was displaced further away from the cluster of affective states than the cough cluster. Note, however, that the cough and mucus clusters were aligned along a single axis with respect to the cluster of affective states.

Discussion

Principal Findings

Our findings demonstrate the potential to deploy advanced analytics in the search for disease-related insights from hundreds of patients and many thousands of self-reports published online.

By probing a word embedding model trained on a corpus of online content contributed by COPD patients, we found a lexicon of community terms expressing a broad range of topics and meanings (Table S1 in Multimedia Appendix 1). Many terms found this way were related to COPD in a direct and intuitive fashion. And some terms revealed associations with unexpected, yet highly relevant topics (eg, pursed-lip) [18]. This term relates to the pursed-lip technique for managing anxiety associated with breathlessness.

The finding that breathlessness was the most frequently mentioned symptom accorded with medical consensus. As stated by the internationally recognized guidelines of the Global Initiative for Chronic Obstructive Lung Disease (GOLD) [15,22], a decline in lung capacity, in combination with other disease-specific symptoms [23,24], forms the basis of a clinical diagnosis of COPD, and measurements of lung function and lung volumes are used to monitor disease progression [17].

In agreement with recent social media studies of COPD patients, our results highlight mucus hypersecretion as an important COPD symptom [25,26]. Compared with breathlessness and cough, mucus terms co-occur with mentions of moderate disease and co-occur less often with mild or severe disease. Similarly, when compared with breathlessness and cough, mucus symptoms were mentioned relatively less frequently when patients reported affective impacts of COPD such as depression.

These distinct associations relating to mucus hypersecretion were corroborated by a novel analysis using PCA to map the psychological salience of the 3 COPD symptoms. Relative to breathlessness and cough, mucus symptoms were mapped furthest from the affective impacts of COPD, suggesting that mucus has the weakest association with perceived affective impacts of the disease.

Mucus symptoms were mentioned at less than half the frequency that breathlessness was mentioned in the corpus. This finding is consistent with the GOLD report and reports indicating that not all COPD patients experience mucus hypersecretion as a symptom of their disease and that mucin concentrations are lower in COPD versus other obstructive lung diseases like cystic fibrosis or bronchiectasis [27]. And yet mucus hypersecretion is an important clinical factor in COPD. For example, mucus symptoms can motivate patients to take timely action against life-threatening respiratory infections [28]. Hypersecretion also drives cough symptoms and expectoration [15].

Without these advanced analytics, our insights about mucus symptoms would have been obscured by the overall dominance of breathlessness and cough symptoms mentioned in the corpus. Examining the co-occurrences between symptoms and disease impacts informed a deeper understanding of disease burden. The approach was able to quickly and accurately identify patient populations whose experience was especially impacted by a particular symptom, adding greater potential for personalization.

This approach can ultimately guide clinical development in ways that ensure that future treatments are fit for purpose from the patient's perspective [12] and from the perspective of patients' perceived treatment needs.

Limitations

The forum content we included in the corpus had been posted anonymously and so we were unable to verify any bias arising from the demographics of forum contributors. Beyond the general guidance posted online by forum moderators, we could not explore biases introduced by a moderator removing posts from the forum.

We can expect a degree of clinical inaccuracy in the contributions posted by individuals who may not have formal medical training. Furthermore, the anonymity of social media makes it all but impossible to determine whether a post is authored by a genuine patient or caregiver or by someone merely posing as one. Taken together, any clinical interpretations we make from social media must take these uncertainties into account. However, because every post was manually reviewed, obviously fraudulent content from bots, scammers, and marketers was eliminated.

Despite limitations, the societal benefits that may be gained from large scale analysis of social media content are substantial, as researchers Gleibs et al [29] and Golder et al [30] have noted. The research community should ideally work closely with patients and health care advocates to ensure that people can continue to contribute to online forums and other social media platforms in a way that protects their privacy and ensures they are safe from potentially harmful misinformation.

Conclusions

Using a novel neural network approach, we demonstrate how online content can be a rich source of insights about the lived experience of COPD. Our findings demonstrate the potential of neural networks to gain a quantitative, patient-focused understanding about how each distinct COPD symptom contributes to the burden of chronic and acute respiratory illness. This approach can be readily applied to other disease areas in which there exists sufficient online content contributed by patients and caregivers.

Acknowledgments

The authors wish to acknowledge the management of Roche Pharma Research and Early Development. We would also like to thank our colleagues working at F Hoffmann–La Roche Ltd in regulatory affairs and data science, particularly Venus So, and, last but not least, the medical experts who kindly reviewed our findings.

Authors' Contributions

TF and RRE authored the paper. TF conducted the analysis. JG and XY sourced and prepared the corpus of content for downstream analysis. VJE and ML contributed to the analysis plan and manuscript writing.

Conflicts of Interest

VJE is an employee of F. Hoffmann—La Roche Ltd and holds stocks. JG and RRE are employees of F. Hoffmann—La Roche Ltd.

Multimedia Appendix 1
Further advanced analyses.
[DOCX File , 118 KB - medinform_v9i11e26272_app1.docx]

References

1. Gijsen V, Maddux M, Lavertu A, Gonzalez-Hernandez G, Ram N, Reeves B, et al. #Science: the potential and the challenges of utilizing social media and other electronic communication platforms in health care. Clin Transl Sci 2020 Jan;13(1):26-30 [FREE Full text] [doi: 10.1111/cts.12687] [Medline: 31392837]
2. Limsopatham N, Collier N. Normalising medical concepts in social media texts by learning semantic representation. Proc 54th Ann Mtg Assoc Computational Linguistics 2016:1014-1023. [doi: 10.18653/v1/P16-1096]
3. Tutubalina E, Miftahutdinov Z, Nikolenko S, Malykh V. Medical concept normalization in social media posts with recurrent neural networks. J Biomed Inform 2018 Aug;84:93-102 [FREE Full text] [doi: 10.1016/j.jbi.2018.06.006] [Medline: 29906585]
4. Taylor KI, Staunton H, Lipsmeier F, Nobbs D, Lindemann M. Outcome measures based on digital health technology sensor data: data- and patient-centric approaches. NPJ Digit Med 2020 Jul 23;3(1):97. [doi: 10.1038/s41746-020-0305-8] [Medline: 32715091]
5. van Rosse F, de Bruijne M, Suurmond J, Essink-Bot M, Wagner C. Language barriers and patient safety risks in hospital care. A mixed methods study. Int J Nurs Stud 2016 Feb;54:45-53. [doi: 10.1016/j.ijnurstu.2015.03.012] [Medline: 25840899]

6. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform 2019 May;125:37-46 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.02.008] [Medline: 30914179]
7. Humphrey L, Willgoss T, Trigg A, Meysner S, Kane M, Dickinson S, et al. A comparison of three methods to generate a conceptual understanding of a disease based on the patients' perspective. J Patient Rep Outcomes 2017;1(1):9 [FREE Full text] [doi: 10.1186/s41687-017-0013-6] [Medline: 29757313]
8. Weissenbacher D, Sarker A, Klein A, O'Connor K, Magge A, Gonzalez-Hernandez G. Deep neural networks ensemble for detecting medication mentions in tweets. J Am Med Inform Assoc 2019 Dec 01;26(12):1618-1626 [FREE Full text] [doi: 10.1093/jamia/ocz156] [Medline: 31562510]
9. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. ArXiv. Preprint posted online on January 16, 2013 2013:1 [FREE Full text] [doi: 10.3126/jiee.v3i1.34327]
10. Mikolov T. Distributed representations of words and phrases and their compositionality. ArXiv. Preprint posted online on October 16, 2013 2013:1 [FREE Full text] [doi: 10.5860/choice.189890]
11. Goldberg Y, Levy O. word2vec explained: deriving Mikolov et al's negative-sampling word-embedding method. ArXiv. Preprint posted online on February 15, 2014 2014:1 [FREE Full text]
12. Patient-focused drug development: methods to identify what is important to patients guidance for industry. Food and Drug Administration. 2019. URL: https://www.fda.gov/media/131230/download [accessed 2021-10-08]
13. NLM MeSH: chronic obstructive pulmonary disease. URL: https://www.ncbi.nlm.nih.gov/mesh?Db=mesh&Cmd=DetailsSearch&Term=%22Pulmonary+Disease,+Chronic+Obstructive%22%5BMeSH+Terms%5D [accessed 2021-10-08]
14. NLM MedlinePlus: chronic obstructive pulmonary disease. URL: https://medlineplus.gov/copd.html [accessed 2021-10-08]
15. Global Initiative for Chronic Obstructive Lung Disease (GOLD). URL: https://goldcopd.org/ [accessed 2021-10-08]
16. Kessler R, Partridge MR, Miravitlles M, Cazzola M, Vogelmeier C, Leynaud D, et al. Symptom variability in patients with severe COPD: a pan-European cross-sectional study. Eur Respir J 2011 Feb;37(2):264-272 [FREE Full text] [doi: 10.1183/09031936.00051110] [Medline: 21115606]
17. Miravitlles M, Worth H, Soler Cataluña JJ, Price D, De Benedetto F, Roche N, et al. Observational study to characterise 24-hour COPD symptoms and their relationship with patient-reported outcomes: results from the ASSESS study. Respir Res 2014 Oct 21;15:122 [FREE Full text] [doi: 10.1186/s12931-014-0122-1] [Medline: 25331383]
18. Ubolnuar N, Tantisuwat A, Thaveeratitham P, Lertmaharit S, Kruapanich C, Chimpalee J, et al. Effects of pursed-lip breathing and forward trunk lean postures on total and compartmental lung volumes and ventilation in patients with mild to moderate chronic obstructive pulmonary disease: an observational study. Medicine (Baltimore) 2020 Dec 18;99(51):e23646 [FREE Full text] [doi: 10.1097/MD.0000000000023646] [Medline: 33371099]
19. Elliott MW, Adams L, Cockcroft A, MacRae KD, Murphy K, Guz A. The language of breathlessness. Use of verbal descriptors by patients with cardiopulmonary disease. Am Rev Respir Dis 1991 Oct;144(4):826-832. [doi: 10.1164/ajrccm/144.4.826] [Medline: 1928956]
20. Hie B, Zhong ED, Berger B, Bryson B. Learning the language of viral evolution and escape. Science 2021 Jan 15;371(6526):284-288. [doi: 10.1126/science.abd7331] [Medline: 33446556]
21. Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 2019 Jul 3;571(7763):95-98. [doi: 10.1038/s41586-019-1335-8] [Medline: 31270483]
22. 2021 GOLD reports: 2021 global strategy for prevention, diagnosis, and management of COPD. URL: https://goldcopd.org/wp-content/uploads/2020/11/GOLD-REPORT-2021-v1.1-25Nov20_WMV.pdf [accessed 2021-10-08]
23. Barnes PJ, Burney PGJ, Silverman EK, Celli BR, Vestbo J, Wedzicha JA, et al. Chronic obstructive pulmonary disease. Nat Rev Dis Primers 2015 Dec 03;1:15076. [doi: 10.1038/nrdp.2015.76] [Medline: 27189863]
24. Agustí A, Hogg JC. Update on the pathogenesis of chronic obstructive pulmonary disease. N Engl J Med 2019 Sep 26;381(13):1248-1256. [doi: 10.1056/nejmra1900475]
25. Cook NS, Kostikas K, Gruenberger J, Shah B, Pathak P, Kaur VP, et al. Patients' perspectives on COPD: findings from a social media listening study. ERJ Open Res 2019 Feb 10;5(1):00128-02018 [FREE Full text] [doi: 10.1183/23120541.00128-2018] [Medline: 30775374]
26. Patalano F, Gutzwiller FS, Shah B, Kumari C, Cook NS. Gathering structured patient insight to drive the PRO strategy in COPD: patient-centric drug development from theory to practice. Adv Ther 2020 Jan 09;37(1):17-26 [FREE Full text] [doi: 10.1007/s12325-019-01134-x] [Medline: 31707715]
27. Ghosh A, Boucher RC, Tarran R. Airway hydration and COPD. Cell Mol Life Sci 2015 Oct;72(19):3637-3652 [FREE Full text] [doi: 10.1007/s00018-015-1946-7] [Medline: 26068443]
28. Hewitt R, Farne H, Ritchie A, Luke E, Johnston SL, Mallia P. The role of viral infections in exacerbations of chronic obstructive pulmonary disease and asthma. Ther Adv Respir Dis 2016 Apr;10(2):158-174 [FREE Full text] [doi: 10.1177/1753465815618113] [Medline: 26611907]
29. Gleibs IH. Turning virtual public spaces into laboratories: thoughts on conducting online field studies using social network sites. Anal Soc Iss Public Pol 2014 Jan 22;14(1):352-370. [doi: 10.1111/asap.12036]

30. Golder S, Ahmed S, Norman G, Booth A. Attitudes toward the ethics of research using social media: a systematic review. J Med Internet Res 2017 Jun 06;19(6):e195 [FREE Full text] [doi: 10.2196/jmir.7082] [Medline: 28588006]

Abbreviations

COPD: chronic obstructive pulmonary disease
GOLD: Global Initiative for Chronic Obstructive Lung Disease
NLM: National Library of Medicine
PCA: principal component analysis

Edited by R Kukafka, G Eysenbach; submitted 04.12.20; peer-reviewed by S Elliott, Y Kwan; comments to author 18.01.21; revised version received 18.04.21; accepted 20.09.21; published 11.11.21.

Please cite as:
Freeman TCB, Rodriguez-Esteban R, Gottowik J, Yang X, Erpenbeck VJ, Leddin M
A Neural Network Approach for Understanding Patient Experiences of Chronic Obstructive Pulmonary Disease (COPD): Retrospective, Cross-sectional Study of Social Media Content
JMIR Med Inform 2021;9(11):e26272
URL: https://medinform.jmir.org/2021/11/e26272
doi:10.2196/26272
PMID:34762056

Original Paper

The Evolution of Rumors on a Closed Social Networking Platform During COVID-19: Algorithm Development and Content Study

Andrea W Wang[1], MSc; Jo-Yu Lan[2], BEng; Ming-Hung Wang[3], PhD; Chihhao Yu[1], MFA

[1]Information Operations Research Group, Taipei, Taiwan

[2]Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan

[3]Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan

Corresponding Author:
Chihhao Yu, MFA
Information Operations Research Group
7F-13, No. 103, Sec. 1, Fuxing S. Rd., Da'an Dist.
Taipei, 106
Taiwan
Phone: 886 933 263 989
Email: chihhao@iorg.tw

Abstract

Background: In 2020, the COVID-19 pandemic put the world in a crisis regarding both physical and psychological health. Simultaneously, a myriad of unverified information flowed on social media and online outlets. The situation was so severe that the World Health Organization identified it as an infodemic in February 2020.

Objective: The aim of this study was to examine the propagation patterns and textual transformation of COVID-19–related rumors on a closed social media platform.

Methods: We obtained a data set of suspicious text messages collected on Taiwan's most popular instant messaging platform, LINE, between January and July 2020. We proposed a classification-based clustering algorithm that could efficiently cluster messages into groups, with each group representing a rumor. For ease of understanding, a group is referred to as a "rumor group." Messages in a rumor group could be identical or could have limited textual differences between them. Therefore, each message in a rumor group is a form of the rumor.

Results: A total of 936 rumor groups with at least 10 messages each were discovered among 114,124 text messages collected from LINE. Among 936 rumors, 396 (42.3%) were related to COVID-19. Of the 396 COVID-19–related rumors, 134 (33.8%) had been fact-checked by the International Fact-Checking Network–certified agencies in Taiwan and determined to be false or misleading. By studying the prevalence of simplified Chinese characters or phrases in the messages that originated in China, we found that COVID-19–related messages, compared to non–COVID-19–related messages, were more likely to have been written by non-Taiwanese users. The association was statistically significant, with $P<.001$, as determined by the chi-square independence test. The qualitative investigations of the three most popular COVID-19 rumors revealed that key authoritative figures, mostly medical personnel, were often misquoted in the messages. In addition, these rumors resurfaced multiple times after being fact-checked, usually preceded by major societal events or textual transformations.

Conclusions: To fight the infodemic, it is crucial that we first understand why and how a rumor becomes popular. While social media has given rise to an unprecedented number of unverified rumors, it also provides a unique opportunity for us to study the propagation of rumors and their interactions with society. Therefore, we must put more effort into these areas.

(JMIR Med Inform 2021;9(11):e30467) doi:10.2196/30467

KEYWORDS

COVID-19; rumors; rumor diffusion; rumor propagation; social listening; infodemic; social media; closed platform; natural language processing; machine learning; unsupervised learning; computers and society

Introduction

Online social media has democratized content. By creating a direct path from content producers to consumers, the power of production and sharing of information has been redistributed from limited parties to general populations. However, social media platforms have also given rise to the proliferation of misinformation and enabled the fast dissemination of unverified rumors [1-3]. In 2020, the COVID-19 pandemic put the world in a crisis regarding both physical and psychological health. A myriad of unverified information flowed on social media. Rumors and claims of erroneous health practices even interfered with the control of COVID-19 in various parts of the world [4,5]. The World Health Organization (WHO) identified this situation as an infodemic in February 2020 [6], indicating its seriousness.

Previous studies revealed that people relied on social media to gather COVID-19 information and guidelines [7,8]. Efforts have, thus, been put into studies examining true and false rumors on social media [9-11]. For example, Cinelli et al [9] compared feedback to reliable and questionable COVID-19 information across five platforms, including Twitter, YouTube, and Gab. Gallotti et al [10] looked at how much unreliable COVID-19 information Twitter users were exposed to across countries.

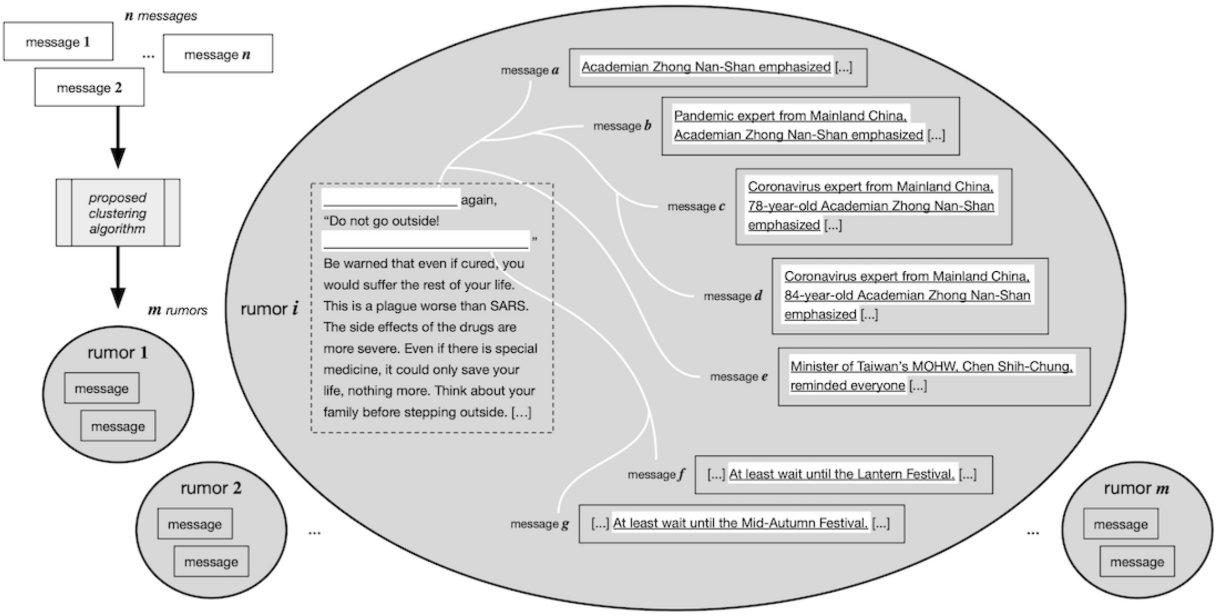
Machine learning and deep learning techniques have been employed to study COVID-19 posts on social media, with much of the focus on topic modeling, sentiment analysis, and misinformation detection [12-25]. Both sentiment analysis and misinformation detection are supervised classification problems. Many studies have employed the Valence Aware Dictionary and Sentiment Reasoner (VADER) model or long short-term memory (LSTM) for sentiment analysis and ensemble machine learning models, such as Extreme Gradient Boosting (XGBoost), for misinformation detection [13-15,17-19,23,25]. Topic modeling, on the other hand, is an unsupervised clustering method. Among topic modeling studies, latent Dirichlet allocation (LDA) was the most widely used algorithm [12,15,17,19,21,22,24], and other favorites included k-means clustering [14,16]. For example, Chandrasekaran et al [15] utilized LDA to extract 26 topics among 13.9 million English COVID-19 Twitter posts. Then they adopted the VADER model to compute sentiment scores for each topic. Jelodar et al [19] employed LDA to extract topics from 560,000 COVID-19 Twitter posts and then used the LSTM neural network to identify the sentiments of the posts. Kwok et al [21] employed LDA to extract topics and Stanford University's CoreNLP (natural language processing) to study the sentiments of Twitter posts regarding COVID-19 vaccinations from Australian Twitter accounts. Also, Chen et al [16] compared the COVID-19 discussions on Twitter and Weibo using t-distributed stochastic neighbor embedding dimensionality reduction with the k-means clustering algorithm to extract topics.

Despite the instructive knowledge provided by the aforementioned machine learning studies, there are two identifiable gaps. First, most studies concentrated on *public* social media platforms, with the majority using Twitter as the data source [12-19,21,22,24,25]. Investigations on closed social media platforms, such as WhatsApp, WeChat, Telegram, or LINE, remain extremely scarce. Secondly, most studies looked at posts via their high-level theme, such as "misconceptions and complaints about COVID-19 control" [21], "psychological stress" [17], or "government response" [15]. There were limited efforts put into the study of individual narratives or rumors under a high-level theme, for example, rumors such as "protect yourself from coronavirus by putting bleach in your body" and "check for COVID-19 by holding your breath for 10 seconds or longer" under the theme of "erroneous health practices."

While high-level themes and sentiments can give us an overview of the public discourse, the capability to efficiently identify individual narratives would be extremely helpful for picking up trending rumors and claims. Discussions on social media platforms are most likely not independent from each other. Thus, simply looking at billions of individual messages is not effective for identifying what rumors are receiving attention. Therefore, there is an apparent need for an efficient way to group and extract the narratives to recognize the popular ones.

Recognizing the limitations from previous studies and to solve the aforementioned problem, our goal was to use machine learning to identify individual COVID-19 rumors from a pool of social media messages, as shown in Figure 1. After identifying the rumors, we then investigated the propagation patterns and textual transformation of those rumors on a closed platform. To achieve this, we proposed a classification-based clustering algorithm to efficiently group tens of thousands of messages according to the similarity of messages. Then, we applied the algorithm to the suspicious messages on LINE, a popular messaging platform in Taiwan. Furthermore, according to the clustering results, we investigated how the messages evolved from temporal and cultural perspectives during the pandemic. To the best of our knowledge, this is the first study to examine COVID-19 rumor diffusion on a closed platform.

Figure 1. A graphical depiction of this study's goal in using machine learning to extract individual rumors from a pool of social media messages. MOHW: Ministry of Health and Welfare.


n messages
message 1
...
message n
message 2
proposed clustering algorithm
m rumors
rumor 1
message
message
rumor 2
message
message
...
rumor i
________ again,
"Do not go outside!
________"
Be warned that even if cured, you would suffer the rest of your life.
This is a plague worse than SARS.
The side effects of the drugs are more severe. Even if there is special medicine, it could only save your life, nothing more. Think about your family before stepping outside. [...]
message a
Academian Zhong Nan-Shan emphasized [...]
message b
Pandemic expert from Mainland China, Academian Zhong Nan-Shan emphasized [...]
message c
Coronavirus expert from Mainland China, 78-year-old Academian Zhong Nan-Shan emphasized [...]
message d
Coronavirus expert from Mainland China, 84-year-old Academian Zhong Nan-Shan emphasized [...]
message e
Minister of Taiwan's MOHW, Chen Shih-Chung, reminded everyone [...]
message f
[...] At least wait until the Lantern Festival. [...]
message g
[...] At least wait until the Mid-Autumn Festival. [...]
...
rumor m
message
message

Methods

Data Collection

LINE is an instant messaging platform. According to the 2018 Taiwan Communication Survey (TCS), 98.5% of people in Taiwan used LINE as their primary messaging tool [26], making it the most popular closed messaging platform. In light of the increasing amount of unreliable information being exchanged through LINE, fact-checking agencies or groups, such as the Taiwan FactCheck Center, Cofacts, or MyGoPen, have developed LINE chatbots for users to voluntarily forward suspicious messages. These chatbots archive the messages and check them against their existing databases to reply with the fact-checked results.

We obtained a data set of suspicious messages forwarded by LINE users to a fact-checking LINE bot between January and July 2020. The data set included messages related to COVID-19 as well as other topics.

Along with the text content of each reported message, we also obtained the report time of each message and a unique identifier for the LINE user that reported the message. The user identifiers we received were scrambled; therefore, it was not possible for us to use the identifiers to attribute any reported message back to any actual LINE user.

Data Preprocessing

After obtaining the text messages, we preprocessed them using the following steps. First, we removed all characters that were neither simplified Chinese nor traditional Chinese. Second, we tokenized each message using the Jieba library [27] in Python (version 3.7; Python Software Foundation) and then removed tokens that were Chinese stop words from the token list. To focus on longer messages, we only kept messages with at least 20 tokens from our data set. Finally, the CountVectorizer module from Python's scikit-learn package [28] was used to create a binary word vector for each message.

Clustering Messages Into Rumor Groups by the Classification-Based Clustering Algorithm

In order to determine what messages belonged to the same rumor, we needed to define distance between messages. We wanted two messages, A and B, to be close to each other if the overlapping text between the two constituted the majority of both messages. When the overlapping text makes up the majority of A but not B, it signals that message A only constitutes a portion of B, meaning that B is likely a combination of several other rumors. In this situation, A and B should be in different groups; therefore, we would like the distance between them to be larger. Based on this idea, we defined the distance between two messages, A and B, to be as follows:

where *tok*(·) is the set of tokens of one message and |·| denotes the number of elements in a set.

While most work relied on the LDA or k-means algorithm to separate messages into groups, both algorithms required a predefined number of final groups. That is, the users need to tell the algorithm how many groups to separate the messages into before being applied. Even though what we wanted to discover was how many narratives, or rumors, there were in all the messages by comparing the distance (equation 1) among all messages, such a requirement contradicted our needs. Hierarchical agglomerative clustering (HAC), on the other hand, starts by merging messages closer to one another into clusters and then iteratively merging closer clusters together until the distance between each cluster exceeds a predefined threshold. That is, instead of predefining a specific number of final groups like in LDA or k-means clustering, HAC determines the number of groups based on a predefined distance threshold. In addition, HAC has the advantage of accepting self-defined distance metrics. Therefore, HAC was the clustering algorithm that fitted our needs.

However, HAC can be quite slow and memory consuming. It suffers with large data sets, especially in the case of social media messages. Therefore, we devised a classification-based clustering algorithm, one that combined the k-nearest neighbors (KNN) algorithm with HAC, to efficiently perform the clustering task. The idea was to randomly select a portion of messages on which to perform HAC; the result was then used to train a KNN algorithm. The trained KNN algorithm was subsequently used to predict the rest of the messages. A detailed algorithm is outlined in Textbox 1, and a flowchart of the algorithm is presented in Figure 2. The experimentation details are outlined in Multimedia Appendix 1, and we demonstrate the efficiency and effectiveness of this algorithm in the following subsection. The algorithm was implemented with the KNeighborsClassifiers and AgglomerativeClustering modules from the Python library scikit-learn [28]. The library gensim (version 3.8.3) [29] was also used in experiments to implement the LDA model for comparisons. We released the code to implement the model in a GitHub repository [30].

Textbox 1. The classification-based clustering algorithm (hierarchical agglomerative clustering plus k-nearest neighbors algorithm).

Notation:

1. $(A)_j$: j^{th} element of set A.

Input:

1. D: the set of all documents to be grouped.
2. D^T: the set of tokenized documents. The order is preserved as D.
3. Train portion u: a number >0 and ≤1.
4. Distance threshold λ: a number >0 and ≤1. Throughout this paper, we set $\lambda=0.6$.

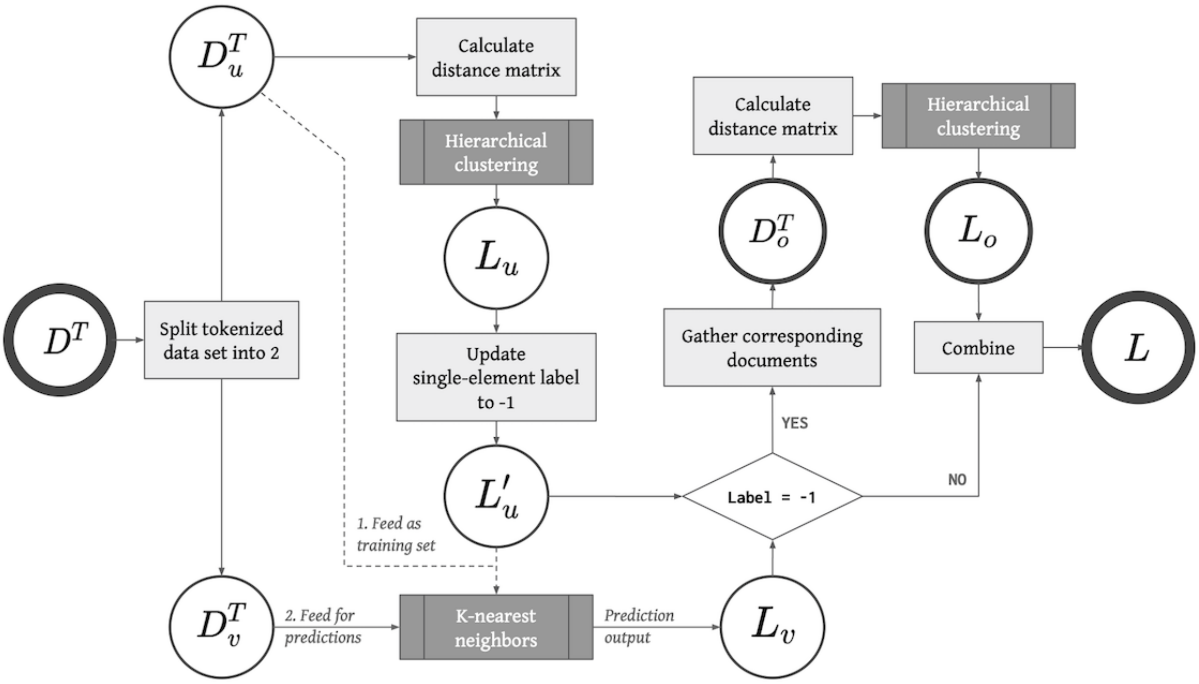
Algorithm:

1. Select $u \times |D^T|$ elements from D^T, denoted as $D^T{}_u$, and the rest not selected as set $D^T{}_v$.
2. Construct distance matrix M for $D^T{}_u$, where $M_{ij} = d((D^T{}_u)_i, (D^T{}_u)_j)$ by equation 1. Note that M is symmetric.
3. Feed M into hierarchical clustering with a distance threshold of λ. We will get back a sequence of labels L_u, where $(L_u)_i$ is the label of element $(D^T{}_u)_i$. Elements with the same label are in the same cluster. Since the label itself does not carry meaning, manipulate them so they are all nonnegative whole numbers.
4. For each unique label x in L_u, if $|\{ k \mid k = x\ \forall\ k \in L_u \}| = 1$, then replace the value of x with −1. Denote the updated label set as L'_u.
5. Train a k-nearest neighbors classifier K using the training set $(D^T{}_u, L'_u)$. Then use K to predict the labels of $D^T{}_v$. Denote the prediction as L_v.
6. Construct L by combining L'_u and L_v, where $(L)_i$ is the label of $(D^T)_i$.
7. Construct $D^T{}_o = \{d_i \mid \text{Label}\ (d_i) = -1\ \forall\ d_i \in D^T\}$.
8. Redo steps 2 and 3 for $D^T{}_o$. Denote the output as L_o. Make sure the values of L_o do not overlap with the values of L from step 6.
9. Update L from step 6 with L_o.

Algorithm output:

A list of labels L, where $(L)_i$ denotes the label of document $(D)_i$. Note that the value of the label itself does not carry any meaning. However, elements in D^T with the same label belong to the same group.

Figure 2. Flowchart for the classification-based clustering algorithm (hierarchical agglomerative clustering + k-nearest neighbors algorithm).



Comparing the Classification-Based Clustering Algorithm With Other Popular Algorithms

From Figure 3, we can see that the classification-based clustering algorithm, the HAC+KNN model, greatly reduced the runtime compared to using only HAC, especially when the train portion value *u* was less than 0.60. Furthermore, such a significant gain in speed did not compromise the clustering results. With the HAC model's results as the gold standard to compare with, the precision values (Figure 4), recall values (Figure 5), and *F* scores (Figure 6) from the HAC+KNN model remained greater than 99% when the train portion *u* was not lower than 0.40. The results demonstrated that the HAC+KNN model's assignments of groups were complete, as measured by recall, and the use of KNN did not introduce too many errors in each group, as measured by precision.

Figure 3. Speed comparison with 95% CIs between HAC and HAC+KNN across different levels of train portion *u*. HAC: hierarchical agglomerative clustering; KNN: k-nearest neighbors.

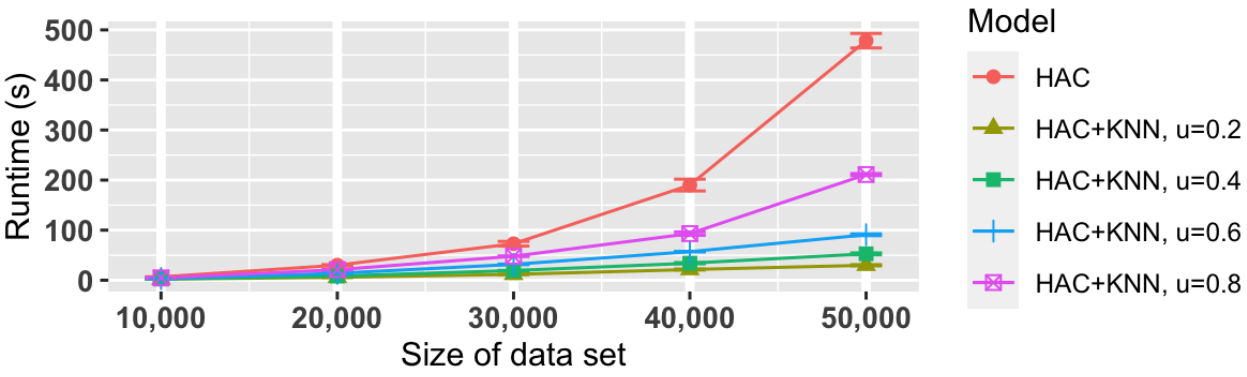


Figure 4. Precision values and 95% CIs (whiskers) of the HAC+KNN algorithm across different data set sizes and train portion *u*. HAC: hierarchical agglomerative clustering; KNN: k-nearest neighbors.

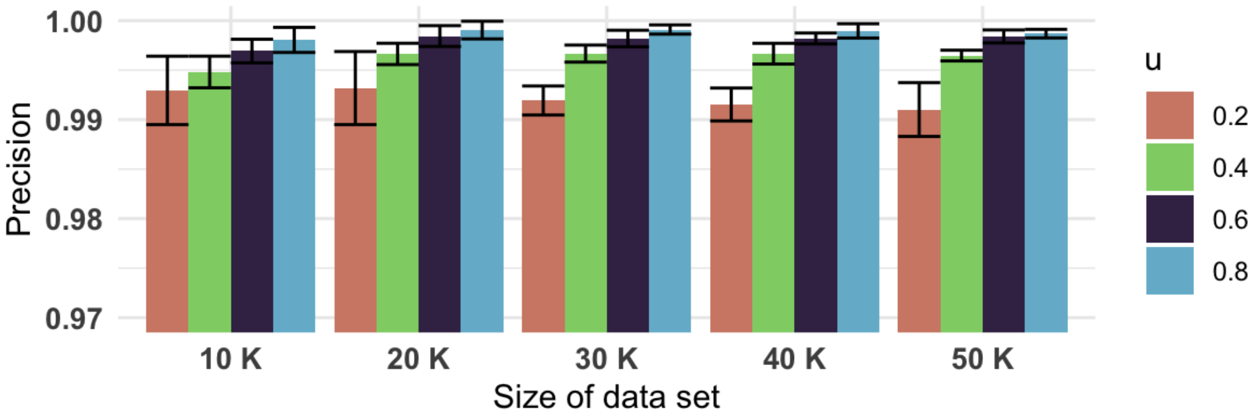


Figure 5. Recall values and 95% CIs (whiskers) of the HAC+KNN algorithm across different data set sizes and train portion *u*. HAC: hierarchical agglomerative clustering; KNN: k-nearest neighbors.

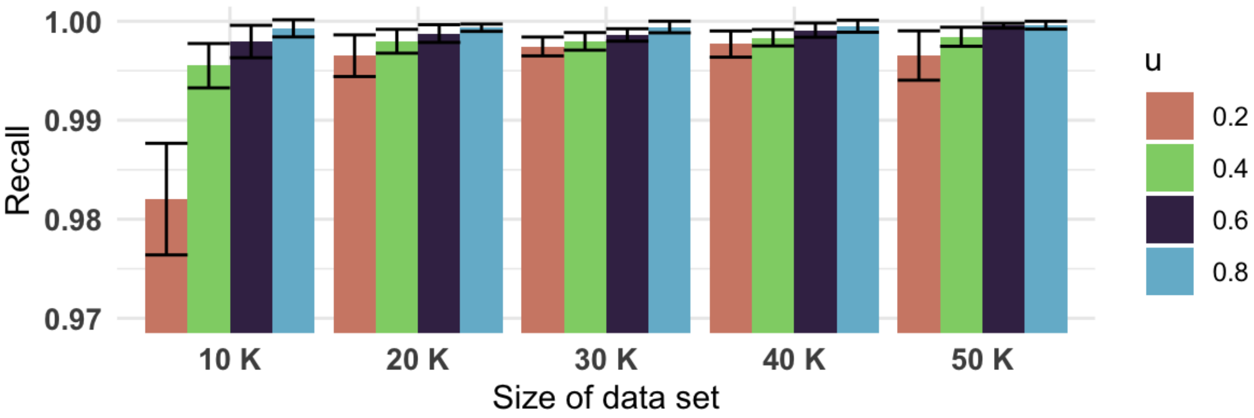
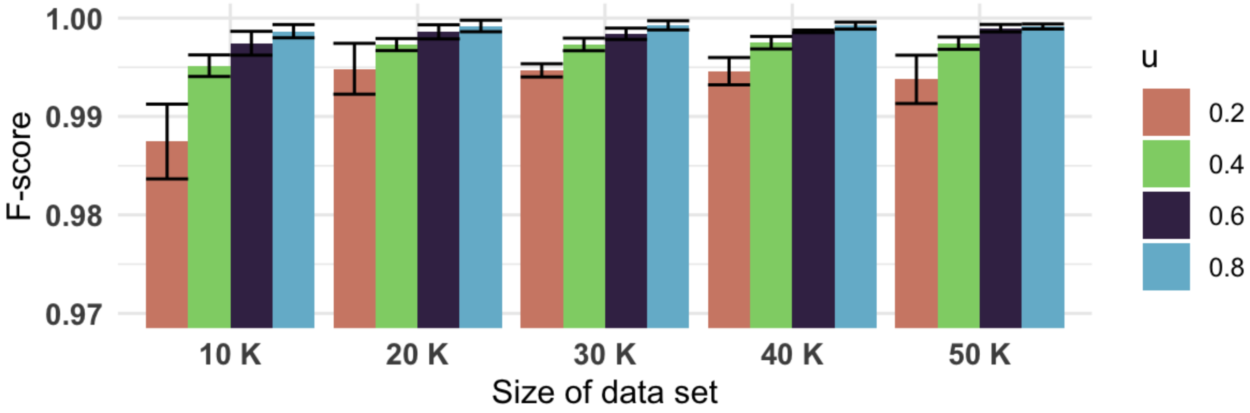


Figure 6. *F* scores and 95% CIs (whiskers) of the HAC+KNN algorithm across different data set sizes and train portion *u*. HAC: hierarchical agglomerative clustering; KNN: k-nearest neighbors.



We observed that the runtime of the k-means clustering was 10 times slower than that of the HAC algorithm, and the LDA model's runtime was the slowest among all models (Table 1). In addition, the precision of the LDA model was very low, meaning that predicted groups had many false positives. While the precision of the k-means model was comparable to that of the HAC+KNN model, recall was only 73%. This showed that the k-means model missed out on many messages.

Table 1. Performance comparisons between HAC, HAC+KNN, LDA, and k-means models for data sets with 10,000 messages.

Model	Runtime (seconds), mean (SD)	Precision, mean (SD)	Recall, mean (SD)	*F* score, mean (SD)
HAC[a]	6.594 (0.245)	N/A[b]	N/A	N/A
HAC+KNN[c] (*u*=0.2)	2.172 (0.097)	0.993 (0.003)	0.982 (0.005)	0.986 (0.004)
HAC+KNN (*u*=0.4)	2.502 (0.023)	0.995 (0.001)	0.996 (0.002)	0.995 (0.001)
HAC+KNN (*u*=0.6)	3.418 (0.071)	0.997 (0.001)	0.998 (0.001)	0.997 (0.001)
HAC+KNN (*u*=0.8)	4.697 (0.146)	0.998 (0.001)	0.999 (0.001)	0.999 (0.001)
LDA[d]	1788.981 (62.444)	0.624 (0.029)	0.939 (0.006)	0.704 (0.023)
K-means	41.143 (1.334)	0.993 (0.002)	0.734 (0.011)	0.823 (0.010)

[a] HAC: hierarchical agglomerative clustering.

[b] N/A: not applicable, because model does not include the parameter *u*.

[c] KNN: k-nearest neighbors.

[d] LDA: latent Dirichlet allocation.

Determining Whether a Rumor Is Related to COVID-19

A rumor group contains many messages. To determine if a rumor group is related to COVID-19, we first identified how many messages in the group contained any of the COVID-19 keywords from the list that we put together (Textbox 2). Next, rumor groups with more than 60% of messages containing COVID-19–related keywords were passed to the authors to decide if such a rumor was really about COVID-19. If a rumor was deemed COVID-19–related, then all messages in the group were also deemed COVID-19–related, regardless of whether that message itself contained the keywords. Recognizing COVID-19–relatedness by close neighbors of each message is a more inclusive approach, as there were messages without the keywords that were obviously related to the pandemic; see Multimedia Appendix 2 as an example.

Textbox 2. A list of 33 COVID-19–related keywords.

指揮中心, 奎寧, 急性呼吸道感染, 新型病毒, 疫情, 口罩, 負壓, 抗疫, 陽性, 新型冠狀病毒, 潛伏期, 李文亮, 纖維化, 自主管理, 群聚, 隔離, 確診, 武漢, 譚德塞, 陰性, 新冠, 染疫, 武肺, 封城, 肺炎, 自主健康管理, 防疫, 冠狀, 家庭感染, covid, ibuprofen, 2019-ncov, coronavirus

Results

Data Set

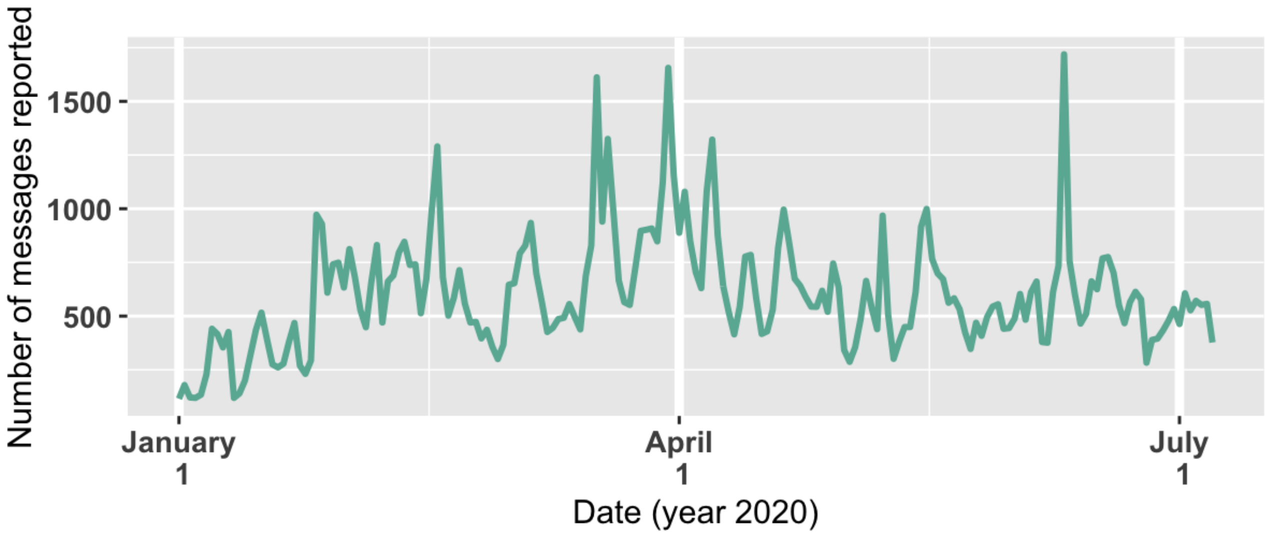
Our data set, after preprocessing, contained 114,124 messages. The character distribution is presented in Table 2, and the number of messages reported per date is shown in Figure 7.

Table 2. Breakdown of characters in the data set of 114,124 suspicious messages.

Statistic	Type of character				
	All	Chinese	Digit	Alphabetical	Others[a]
Minimum, n	24	24	0	0	0
Median (IQR)	233 (333)	145 (225)	7 (17)	2 (22)	38 (79)
Maximum, n	10,012	8132	3252	7014	5532

[a] This category includes characters such as punctuation marks and emojis.

Figure 7. Distribution of 114,124 messages by report dates.



Rumor Group Overview

By using the HAC+KNN algorithm, 114,124 messages were separated into 12,260 rumor groups. A total of 8529 rumor groups had only 1 message. Therefore, the rest of the 105,595 messages were separated into 3731 rumor groups. There were 936 rumor groups with at least 10 messages, with the largest one having 2546 messages. We present the statistics of the rumor group sizes in Table 3.

Table 3. Statistics of the rumor group sizes.

Minimum number of messages per rumor group	Messages				
	Mean (SD)	Maximum, n	3rd quartile, n	2nd quartile, n	Total, n
1	9.309 (71)	2546	2	1	114,124
2	28.302 (126.907)	2546	10	3	105,595
10	102.96 (238.31)	2546	75	27	96,373

Among 936 rumor groups with at least 10 messages, we identified 396 (42.3%) that were related to COVID-19; these consisted of a total of 42,829 messages. Among 396 COVID-19–related rumor groups, 134 (33.8%) were deemed false or misleading by either the Taiwan FactCheck Center or MyGoPen, two International Fact-Checking Network (IFCN)–certified fact-checking agencies in Taiwan.

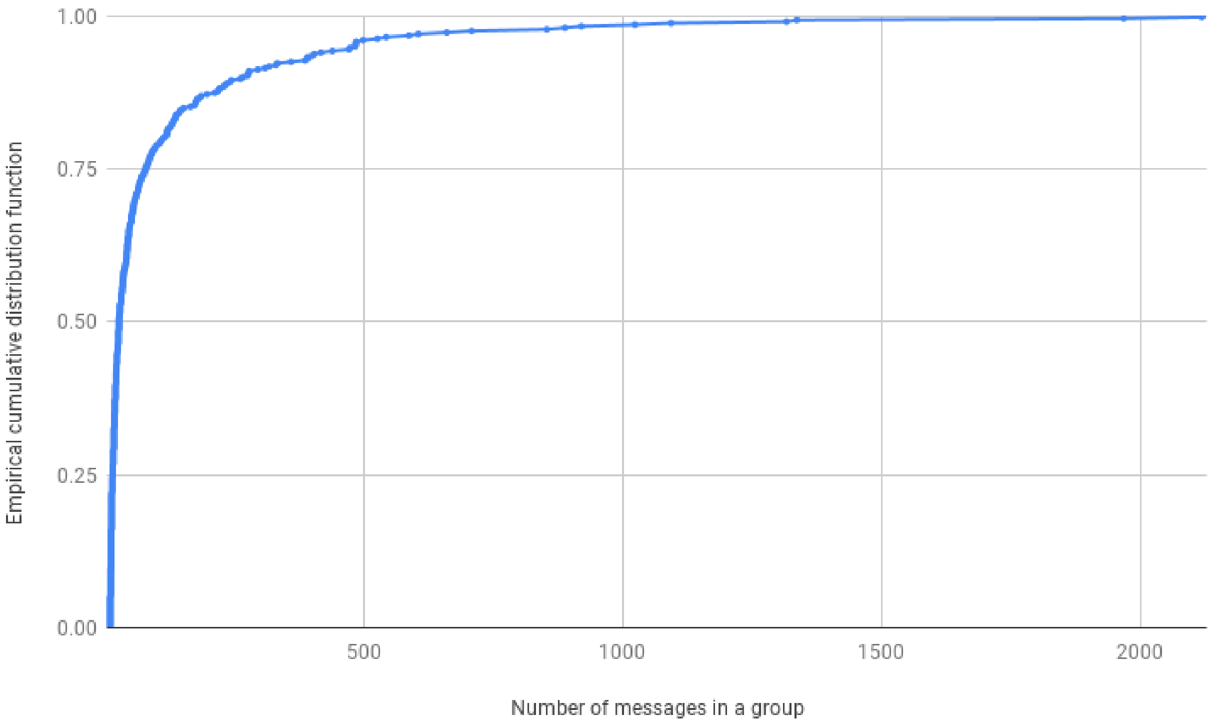
After recognizing many messages containing simplified Chinese characters or phrases originating from China, we compared the prevalence of those characters and phrases between COVID-19–related and non–COVID-19–related messages. Compared to non–COVID-19–related messages, the pool of COVID-19–related messages had significantly more messages using simplified Chinese characters or phrases that originated from China (Table 4). The association was significant as determined by the chi-square independence test with Yates' continuity correction (χ^2_1=1088.0, n=96,373; *P*<.001).

Table 4. Contingency table of COVID-19–relatedness using simplified Chinese characters or phrases originating from China.

Message type	Messages with simplified Chinese characters or phrases originating from China, n		Total messages, n
	Yes	No	
Related to COVID-19	16,957	25,872	42,829
Not related to COVID-19	15,776	37,768	53,544
Total	32,733	63,640	96,373

The COVID-19–related rumor group sizes had a very long-tailed distribution (Figure 8). Most of the rumor groups only contained a few messages. In fact, only 15 rumor groups contained more than 1000 messages. In the following subsection, we discuss how we qualitatively analyzed the three COVID-19 rumor groups with the largest number of messages.

Figure 8. Empirical cumulative distribution function of the number of messages in COVID-19–related rumor groups.



Case Studies of the Three Largest COVID-19–Related Rumor Groups

Overview

We qualitatively analyzed the three rumor groups with the largest number of messages among the 936 COVID-19–related rumor groups. In fact, a total of 7523 messages from the three rumor groups made up 17.6% of all 42,829 COVID-19–related messages.

To study the interactions of the rumors' popularity with society, we picked out some major societal events, as shown in Table 5. While there were multiple important events regarding the pandemic every day, we picked out incidents that were the first occurrences.

Table 5. Major societal events related to COVID-19.

Date (year 2020)	Events
February 9	• First asymptomatic laboratory-confirmed COVID-19 case in Taiwan
February 15	• First COVID-19 death case in Taiwan
February 21	• Passengers on Diamond Princess cruise ship returned to Taiwan
March 11	• COVID-19 declared a global pandemic by the World Health Organization
March 18	• The director of the CECC[a], Chen Shih-Chung, went to the Legislative Yuan for interpellation about the pandemic for the first time • A total of 100 confirmed cases was reached; single-day confirmed cases hit record high for 3 consecutive days
March 26	• The CECC released the first report on the analysis of confirmed cases in Taiwan
March 30	• First death case in Taiwan's first hospital cluster infection
April 1	• The day before a 4-day long weekend • First day of mask requirement on public transportation
April 5	• The last day of a 4-day long weekend

[a] CECC: Central Epidemic Command Center.

Case 1: Do Not Go Outside!

The rumor content for Case 1 is presented in Textbox 3. This rumor first appeared in the data set on February 2, 2020. Over the course of 3.5 months, there were a total of 2119 messages reported. The reported messages went viral at least four times: they peaked on February 22 with 80 messages, they peaked on March 16 with 68 messages, they reached the highest number on April 2 with 205 messages, and they peaked on April 7 with 197 messages (Figure 9). During this period, we observed several content changes (Table 6).

Textbox 3. Content of Case 1 rumor.

English translation:

Academian Zhong Nan-Shan emphasized again, “Do not go outside! At least wait until the Lantern Festival.” Be warned that even if cured, you would suffer the rest of your life. This is a plague worse than SARS. The side effects of the drugs are more severe. Even if there is special medicine, it could only save your life, nothing more. Think about your family before stepping outside...This is a war, not a game...No one is an outsider in this war...Please share it with others. By Zhong Nan-Shan.

Original content:

鐘南山院士再次強調：別出門，元宵後，再看疫情控制情況！警告：一旦染上，就算治癒了，後遺症也會拖累後半生！這場瘟疫比17年前的非典更嚴重，用的藥副作用更大。如果出了特效藥，也只能保命，僅此而已！出門前想想你的家人，別連累家人，能不出門就不出門，大家一起轉發吧！這是一場戰役，不是兒戲，收起你盲目的自信和僥倖心理，也收起你事不關己高高掛起的態度，在這場戰役中沒有局外人！在家！在家！在家！不要點贊！求轉發——鐘南山

Figure 9. The number of Case 1 (ie, "Do not go outside!") messages reported by date. The number peaked on April 2 with 205 messages, when messages started misquoting the Central Epidemic Command Center (CECC) director. There were also a large number of reports after a 4-day long weekend on April 6 with 166 messages and on April 7 with 197 messages. Refer to Table 5 for major societal events.

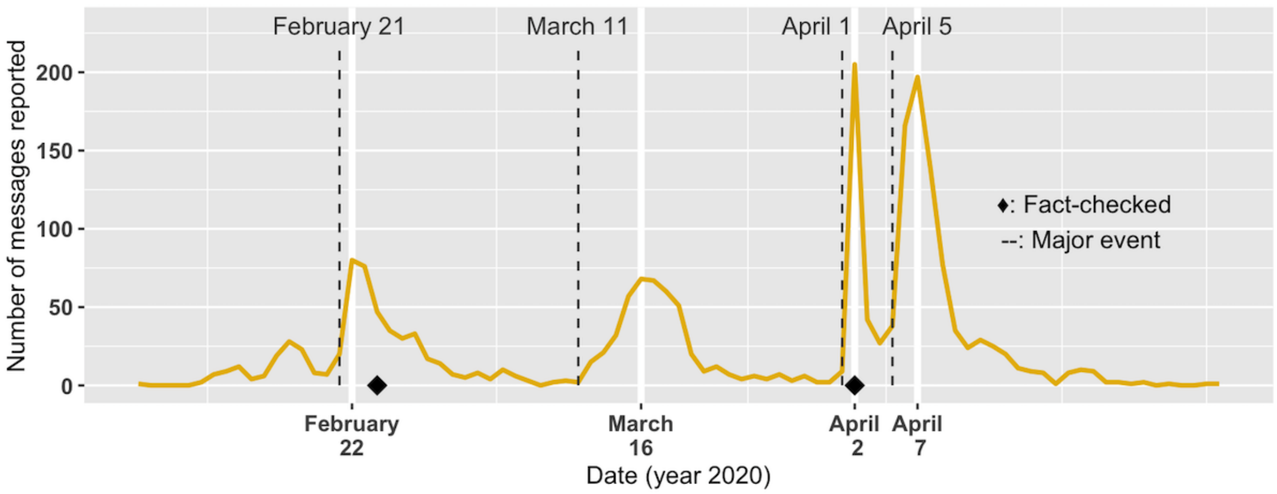


Table 6. Change log for Case 1 rumor content.

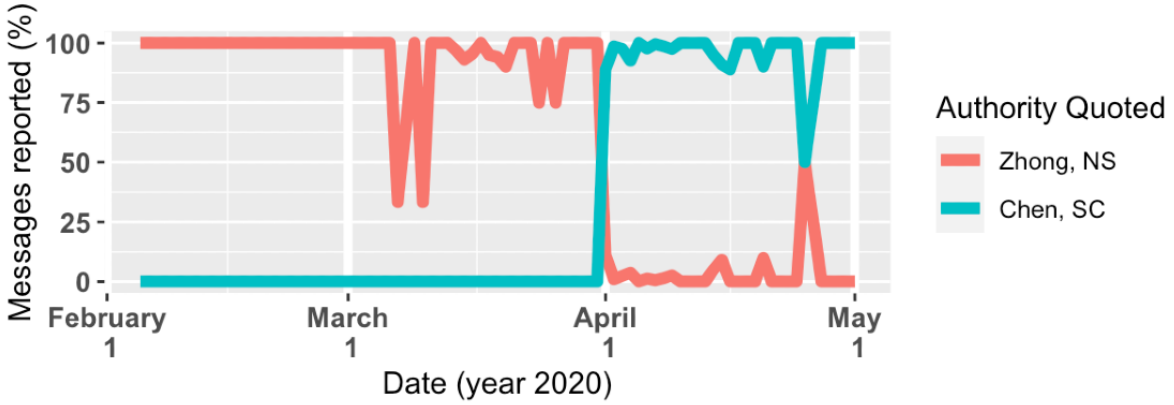
Date (year 2020)	English translation (original)	
	Previous content	New content
February 17	Academian Zhong Nan-Shan emphasized (鍾南山院士再次強調)	Pandemic expert from Mainland China, Academian Zhong Nan-Shan emphasized (大陸防疫專家鍾南山院士再次強調)
February 18	Academian Zhong Nan-Shan emphasized (鍾南山院士再次強調)	Coronavirus expert from Mainland China, 78-year-old Academian Zhong Nan-Shan emphasized (大陸，冠狀病毒專家鐘南山78歲院士再次強調)
February 27	Academian Zhong Nan-Shan emphasized (鍾南山院士再次強調)	Coronavirus expert from Mainland China, 84-year-old Academian Zhong Nan-Shan emphasized (大陸，冠狀病毒專家鐘南山84歲院士再次強調)
April 1	Academian Zhong Nan-Shan emphasized (鍾南山院士再次強調)	Minister of Taiwan's MOHW[a], Chen Shih-Chung, reminded everyone (台灣衛福部長陳時中提醒大家)
February 18	Do not go outside! At least wait until the Lantern Festival. (別出門，元宵後，再看疫情控制情況)	Do not go outside! At least wait until the Dragon Boat Festival. (別出門，端午節過後，再看疫情控制情況)

[a] MOHW: Ministry of Health and Welfare.

First, the time-sensitive information in the messages evolved. In early February, most messages mentioned "Lantern Festival," which took place on February 8, 2020. However, from February 18 onward, there were messages that replaced "Lantern Festival" with "March." Then, after March 10, most messages included "Dragon Boat Festival," which took place on June 25, 2020.

Second, among 2119 reported messages, 2095 (98.9%) falsely quoted authority. Zhong Nan-Shan, the leader of China's National Health Commission's expert panel for investigating the COVID-19 outbreak in China, and Chen Shih-Chung, the director of the Central Epidemic Command Center (CECC)—the two most popular misquoted targets—showed up in 975 (46.5%) and 1117 (53.3%) messages, respectively. Efforts were made to emphasize the authoritativeness of the quoted party as well. For example, titles for Zhong Nan-Shan became longer, from "Expert in Pandemic from Mainland China" and "Expert in Coronavirus" to "Expert in Coronavirus from Mainland China, 78-year-old Academian Zhong Nan-Shan." Starting from April 1, 2020, every reported message had Zhong replaced with Chen Shih-Chung (Figure 10). As the Minister of the Ministry of Health and Welfare (MOHW) and director of Taiwan's CECC, Chen's popularity skyrocketed during the pandemic through his daily press conferences.

Figure 10. Chen Shih-Chung replaced Zhong Nan-Shan as the most quoted party in the Case 1 rumor after April 1, 2020.



Due to the prevalence of this message spreading on the web and closed platforms, the MOHW and the CECC both sent out a press release [31] on April 2, 2020, reminding the public that this was misinformation. Nevertheless, this did not stop another viral spread of the same message at the end of a 4-day long weekend holiday in Taiwan, where crowds were seen at every tourist attraction on the island. For days, people worried that the long weekend would lead to another outbreak of the pandemic, providing an explanation as to why the message bearing the key topic "do not go out" would become a big hit.

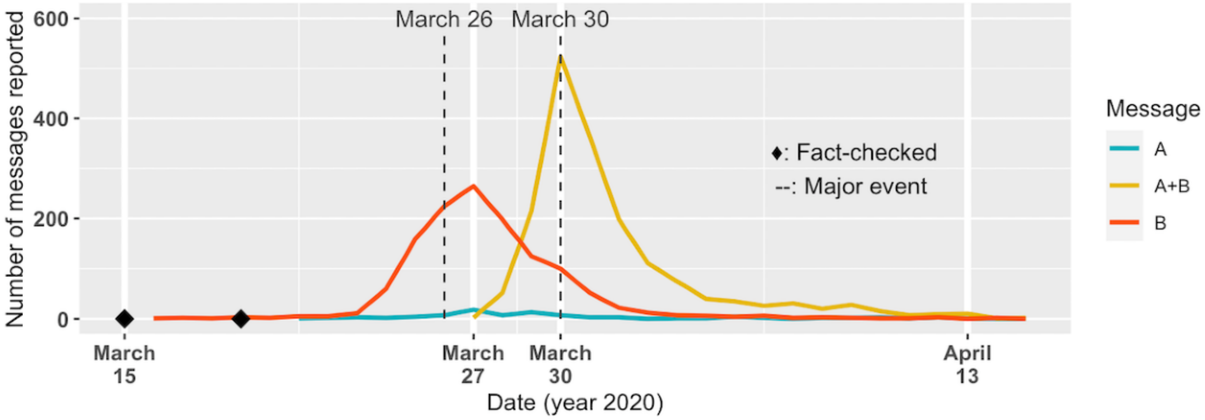
Case 2: Drinking Salt Water Can Prevent the Spread of COVID-19

This rumor promoted drinking salt water to prevent COVID-19. Interestingly, this rumor was actually the combination of two individual rumors (Table 7). Message B had a peak on March 27 with 265 messages, and Message A+B received the most attention on March 30 with 523 messages (Figure 11).

Table 7. Content of Case 2 rumor.

Message	English translation	Original content
A	This is 100% accurate...Why did we see a huge decline of confirmed cases in China during the last few days? They simply forced their citizens to rinse mouths with salted water three times a day and then drink water for 5 minutes. The virus would attack throats before the lungs, and when getting in touch with salted water, the virus would die or get destroyed in lungs. This is the only way to prevent the spread of COVID-19. There is no need to buy medicine as there is nothing effective on the market.	這是100%準確的信息... 為什麼中國過去幾天大大減少了感染人數？他們只是簡單地強迫他們的人民每天漱口3次鹽水.完成後，喝水5分鐘.因為該病毒只能在喉嚨中侵襲，然後再侵襲肺部，當受到鹽水侵襲時，該病毒會死亡或從胃中流下來並在胃中銷毀，這是預防冠狀病毒流行的唯一方法.市場上沒有藥品，所以不要購買.
B	Before reaching the lungs, the novel coronavirus would survive in throats for 4 days. At this stage, people would experience sore throats and start coughing. If one can drink as much warm water with salt and vinegar as they can, the virus could be destroyed. Share this information to save people’s lives.	新冠肺炎在還沒有來到肺部之前，它會在喉嚨部位存活4天.在這個時候,人們會開始咳嗽及喉嚨痛.如果他能儘量喝多溫開水及鹽巴或醋,就能消滅病菌.儘快把此訊息轉達一下，因爲你會救他人一命！
A+B	Why did Mainland China show a huge decline of confirmed cases over the last few days? Besides wearing masks and washing hands, they simply rinse mouths with salted water three times a day and then drink water for 5 minutes...Dr Wang of Tung Hospital stated that the novel coronavirus would survive in throats for 4 days before reaching the lungs...If one can drink as much warm water with salt and vinegar as they can, the virus could be destroyed...	為什麼中國大陸過去幾天大大減少了感染人數？除了戴口罩勤洗手外，他們只是簡單地每天漱口3次鹽水.完成後，喝水5分鐘...新冠肺炎在還沒有來到肺部之前，它會在喉嚨部位存活4天...如果他能儘量喝多溫開水及鹽巴或醋，就能消滅病菌...

Figure 11. The number of Case 2 (ie, "Drinking salt water can prevent the spread of COVID-19") messages reported by date. The rumor had been fact-checked rather early; however, the information still received widespread attention. Message B peaked on March 27 with 265 messages, and the combined message peaked on March 30 with 523 messages. Refer to Table 5 for major societal events. Refer to Table 7 for contents of Message A, B, and A+B.



Among 3283 reported messages, 3093 (94.2%) misquoted medical professionals. The most popular misquoted parties were Dr Wang of Tung Hospital and the director of the Veteran Hospital, each seen in 2340 (71.3%) and 753 (22.9%) messages, respectively.

Drinking salt water to prevent COVID-19 was a popular false claim about COVID-19 internationally. This rumor was fact-checked several times in March by Taiwan’s fact-checking agencies [32,33], and even the WHO had fact-checked a similar claim about rinsing noses with saline [34]. However, this did not stop this piece of misinformation from receiving attention (Figure 11). In fact, several translations of the combined rumor (ie, Message A+B) were observed in April. The translations included English, Indonesian, Filipino, and Tibetan.

The lifespan of this “drink salt water” rumor was rather long. One famous fact-checking platform in Taiwan, MyGoPen, released an article to disprove this false medical advice again in October 2020 [35], 7 months after it was first seen in our data set.

Case 3: This Is a Critical Period; Here Are Some Suggestions

This rumor mentioned that Taiwan “entered a critical period of the pandemic” and provided a list of suggested measures for people to follow (Textbox 4). Some of the suggestions made sense in terms of personal hygiene, while others were without basis. This rumor first appeared in the data set on February 6 and included a total of 2121 messages. Over the 1.5 months of its most popular period, it went viral at least three times: February 10 with 120 messages, February 17 with 394 messages, and March 19 with 543 messages (Figure 12).

Textbox 4. Content of the Case 3 rumor.

English translation:

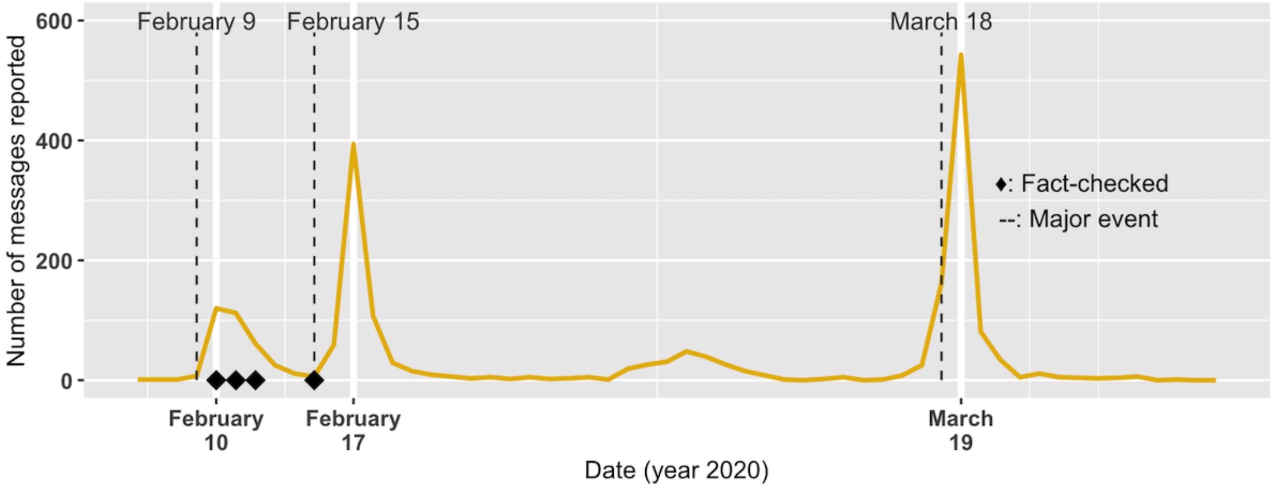
10 days from now, Taiwan will be in a critical period to combat COVID-19. Here are some suggested measures.

1. Strictly prohibit going to public places. 2. Take out from restaurants. 3. Eat outside in open spaces. 4. Wash your hands the right way (extremely important). 5. When taking the subway or bus, choose the seats at the front half of the vehicle. 6. Do not wear contact lenses. 7. Eat warm food and more vegetables. 8. Avoid constipation. 9. Drink warm water. 10. Do not visit hair salons. 11. Hang the clothes you're wearing outside for two hours the first moment you get home. 12. Do not wear jewelry. 13. Wash your hands immediately after touching cash or coins. Put coins you just received inside a plastic bag for one day before using them. 14. Do not use a colleague's phone when working. If you have to, disinfect before using. 15. Avoid taking public transportation during rush hour. 16. Do not visit night markets or traditional markets. 17. Exercise. 18. Avoid going to the gym.

Original content:

今天開始10天，台灣正式進入武漢肺炎関鍵期。建議如下: 1.嚴禁進入公共場所. 2.用餐儘量將食物外帶.3.用餐環境儘量在外. 4.正確方式的洗手(特別重要). 5.坐捷運(公車)，選擇在車前頭. 6.避免戴隱形眼鏡 7.吃熱食,避開生凉食物,多吃蔬菜 8.保持腸胃暢通. 9.多喝溫水. 10.暫停去髮廊. 11.穿過的衣服(外套,長褲),回家先單獨吊在外2小時 12.暫停戴首飾. 13. 一有接觸錢幣,一定要洗手,剛拿進來的錢幣,先單獨放在塑膠袋中,一天後,才拿出來. 14.在公司不要使用別人的電話筒.電話筒需消毒.15. 避開巔峰時間坐車. 16.不去傳統市場及夜市. 17.適當的運動.18. 暫停進入健身房.

Figure 12. The number of Case 3 (ie, "This is a critical period; here are some suggestions") messages reported by date. The rumor was fact-checked several times in early February. However, higher peaks were still seen later on February 17 with 394 messages and, after a month, on March 19 with 543 messages. Refer to Table 5 for major societal events.



Among 2121 reported messages, 1778 (83.8%) misquoted authorities as *endorsing* the rumor. The Taiwan Medical Association and the CECC director, Chen Shih-Chung, were the most misquoted parties, each seen in 1637 (77.2%) and 393 (18.5%) messages, respectively (Figure 13). A major revision of the rumor appeared on February 12 (Table 8), 6 days after the first message. In the revision, the original 18 bullets were pruned to 14, removing the ones that were perhaps more ridiculous or hard to follow. Strong words were also modified to a gentler tone. The Taiwan Medical Association, the most misquoted party, also first appeared in the message.

Figure 13. The Taiwan Medical Association (TMA) was quoted in almost every message in this rumor group, even though the TMA released a statement on February 12, 2020, saying that they did not endorse the material. Later, after Chen Shih-Chung went to Legislative Yuan on March 18, the same rumor started misquoting him.

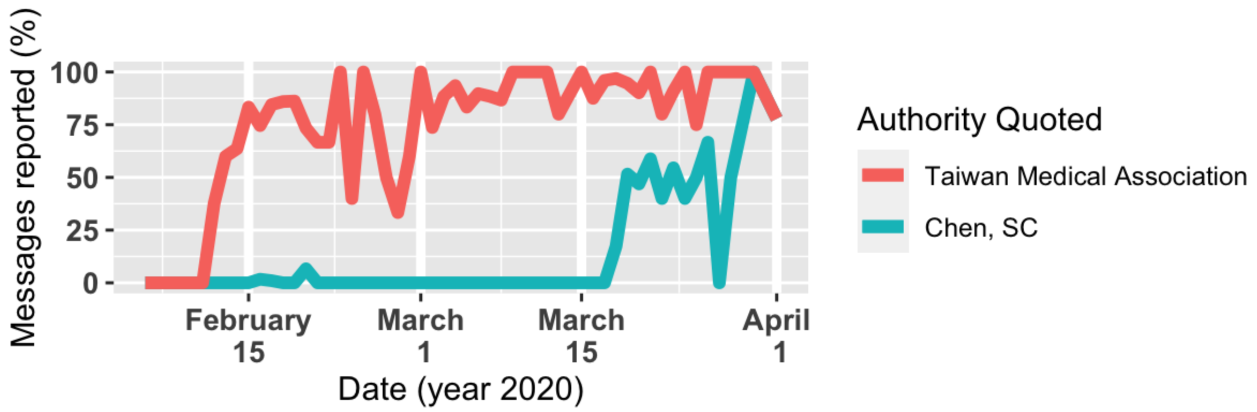


Table 8. Change log for the Case 3 rumor content.

Date	English translation (original)	
	Previous content	New content
February 12, 2020		
	• Strictly prohibit going to public places. (嚴禁進入公共場所.)	• Reduce going to public places. (減少進入公共場所.)
	• Eat outside in open spaces. (用餐環境儘量在外.) • When taking the subway or bus, choose the seats at the front half of the vehicle. (坐捷運(公車)，選擇在車前頭.) • Do not visit hair salons. (暫停去髮廊.) • Do not visit night markets or traditional markets. (不去傳統市場及夜市.)	• Content was deleted
	• No previous content	• Regards from the Taiwan Medical Association. (醫師全聯會關心您.)
March 18, 2020		
	• 10 days from now, Taiwan will be in a critical period to combat COVID-19. Here are some suggested measures... (今天起10天，台灣正式進入武漢肺炎関鍵期，建議如下...)	• 10 days from now, Taiwan will be in a critical period to combat COVID-19 (explained by Chen Shih-Chung in the Legislative Yuan on March 18, 2020). Here are some suggested measures... (今天起10天，台灣正式進入武漢肺炎関鍵期，(3/18陳時中立法院說明) 建議如下...)
	• Regards from the Taiwan Medical Association. (醫師全聯會關心您.)	• Content was deleted

After almost a month with only a few messages circulating (Figure 12), on March 18, the CECC director, Chen Shih-Chung, went to the Legislative Yuan (similar to the US Congress) for interpellation about the pandemic. Chen started to be quoted in messages on the same day, making the “suggested measures” look like they were said by Chen during his interpellation (Table 8). The next day, on March 19, the reported message count skyrocketed to the highest peak. Of the 543 messages reported on March 19, 280 (51.2%) misquoted Chen.

Several fact-checking agencies published reports pointing out the falsity of the message [36-38] between February 10 and 15 (Figure 12). The Taiwan Medical Association, which was misquoted in 1637 out of 2121 (77.2%) messages, also released a clarifying statement on February 12 [39], stating explicitly that they did not endorse the material. However, similar to what we observed in the previous two cases, such fact-checking efforts did not prevent the rumor from getting widespread attention later. Rather, societal events might have played a larger role in the popularity of the rumor. For example, the spike on February 17 (Figure 12) was preceded by the first COVID-19 death case and a local cluster in Taiwan. A taxi driver tested positive for the virus and died the same day on February 15. Over the next few days, four of the driver’s family members also tested positive, forming the first local cluster of COVID-19 infection in Taiwan. The highest spike on March 19 was preceded by the CECC director’s interpellation in the Legislative Yuan, the event after which the messages started misquoting the director.

Discussion

Principal Findings

First, we demonstrated that by using a combination of HAC and KNN algorithms, we could efficiently separate a large number of social media text messages into fine-grained narratives, or *rumors*. The addition of the KNN classification algorithm enabled the speedup and, at the same time, achieved near-equivalent results compared to using HAC alone. Hence, this classification-based clustering algorithm could enable future large-scale studies of rumor transformation with social media post content.

We identified 396 rumors related to COVID-19 from the pool of 114,124 suspicious messages collected from the LINE platform between January and July 2020. Among the COVID-19–related rumors, more than one-third were deemed false or misleading by IFCN-certified fact-checking agencies in Taiwan. Compared to non–COVID-19–related messages, COVID-19–related messages were more likely to contain simplified Chinese characters or phrases originating from China. The association was statistically significant. As the official language in Taiwan is traditional Chinese, the result suggested that COVID-19–related messages were more likely to have originated from non-Taiwanese users than the non–COVID-19–related messages.

We qualitatively investigated three COVID-19–related rumors with the highest number of messages and observed several commonalities among these highly popular rumors. First, a significant number of messages from all three rumor groups misquoted key authoritative figures. Given the nature of the pandemic, the authorities were usually medical personnel. At times, a change in the quoted authority figures signaled a paradigm shift, indicating whom the public looked up to, for example, from Zhong Nan-Shan to Chen Shih-Chung. At other times, the quoted party did not seem to make any sense. For example, Dr Wang in Case 2 was in fact an orthopedist, a specialty not directly related to COVID-19. Second, in all three rumors, we observed spikes in reported messages even after several fact-checking agencies released reports that deemed the content false or misleading. Echoing the findings of Wood and Porter [40], the current practice of fact-checking did not seem to effectively stop the false information from getting widespread attention later. In fact, by identifying major societal events preceding each resurfacing peak, we asserted that resurfacing patterns were more influenced by major societal events and textual transformation. However, each peak of popularity would not last long, and it was often without good explanation about how one wave of attention ended.

Our work offers several insights into the landscape of misinformation in a closed platform as well as the behaviors of some popular COVID-19 rumors. These characteristics could serve as rules to discover possible false information as early detection mechanisms. Although we identified these characteristics manually in this study, it is quite possible to employ techniques such as NLP to automatically recognize these textual changes in the future, making it possible to have an automatic early warning system of possible misinformation before fact-checking efforts by professionals.

Comparison With Prior Work

Our work adds to the limited collection of COVID-19 infodemic studies in closed platforms [41]. Compared with other rumor diffusion studies, such as the study of 17 political rumors by Shin et al [42], this work provided an efficient machine learning algorithm that could enable large-scale rumor evolution studies on social media platforms in the future. In comparison to other machine learning applications in COVID-19 infodemic studies, this work focused on fine-grained narratives, or *rumors*, rather than high-level topics, in order to study individual rumor propagation. To the best of our knowledge, this is the first study to examine rumor diffusion and propagation patterns of COVID-19 misinformation on a closed platform.

Limitations

This study had several limitations. First, the data were collected by LINE users' reports. Therefore, it was impossible to infer the true distribution of messages without making some assumptions. For example, if there was more health-related misinformation in our data, it did not necessarily translate to more health-related rumors circulating in the platform. In fact, it could also be that people were more alert and skeptical of health-related information. Second, we only looked at text messages. Therefore, information distributed visually or in audio form was not covered. Lastly, our algorithm for grouping messages does not work well with short texts.

Conclusions

While social media may give rise to an unprecedented number of unverified rumors, it also provides a unique opportunity to study rumor propagation. In fact, to combat the infodemic, we need to first understand how and why some rumors became popular. In our studies, we proposed an algorithm that enables the research community to perform large-scale studies on the evolution of text messages at the rumor level rather than at the topic level. Moreover, we showed textual commonalities in widespread rumors in Taiwan during COVID-19. We also showed that the attention one rumor received was connected to major societal events and content changes. To the best of our knowledge, this is one of the few studies that has examined COVID-19 misinformation on a closed messaging platform and the first to examine the textual evolution of COVID-19–related rumors during their propagation. We hope that this will further spark more studies in rumor propagation patterns as an effort to fight the infodemic.

Acknowledgments

We cited a survey study where the data were collected by a research project of the TCS [26]. The TCS project is supported by the Ministry of Science and Technology of the Republic of China. We appreciate the assistance by the institute in providing data.

The views expressed herein are the authors' own. This work was supported by the Information Operations Research Group, which was funded by the Institute for War and Peace Reporting and the Taiwan Foundation for Democracy. The funders played no role in the design, implementation, or reporting of this research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Experiment setup for comparing algorithms.
[DOCX File , 1097 KB - medinform_v9i11e30467_app1.docx]

Multimedia Appendix 2
An example message without any of the COVID-19 keywords, but that could be identified as COVID-19–related by a close neighbor.
[DOCX File , 705 KB - medinform_v9i11e30467_app2.docx]

References

1. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, et al. The spreading of misinformation online. Proc Natl Acad Sci U S A 2016 Jan 19;113(3):554-559 [FREE Full text] [doi: 10.1073/pnas.1517441113] [Medline: 26729863]
2. Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, et al. The science of fake news. Science 2018 Mar 09;359(6380):1094-1096. [doi: 10.1126/science.aao2998] [Medline: 29590025]
3. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018 Mar 09;359(6380):1146-1151. [doi: 10.1126/science.aap9559] [Medline: 29590045]
4. Abdoli A. Gossip, rumors, and the COVID-19 crisis. Disaster Med Public Health Prep 2020 Aug;14(4):e29-e30 [FREE Full text] [doi: 10.1017/dmp.2020.272] [Medline: 32713376]
5. Tasnim S, Hossain MM, Mazumder H. Impact of rumors and misinformation on COVID-19 in social media. J Prev Med Public Health 2020 May;53(3):171-174 [FREE Full text] [doi: 10.3961/jpmph.20.094] [Medline: 32498140]
6. Novel Coronavirus (2019-nCoV): Situation Report - 13. Geneva, Switzerland: World Health Organization; 2020 Feb 02. URL: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf [accessed 2021-06-01]
7. Mubeen SM, Kamal S, Kamal S, Balkhi F. Knowledge and awareness regarding spread and prevention of COVID-19 among the young adults of Karachi. J Pak Med Assoc 2020 May;70(Suppl 3)(5):S169-S174. [doi: 10.5455/JPMA.40] [Medline: 32515406]
8. Mat Dawi N, Namazi H, Hwang HJ, Ismail S, Maresova P, Krejcar O. Attitude toward protective behavior engagement during COVID-19 pandemic in Malaysia: The role of e-government and social media. Front Public Health 2021;9:609716 [FREE Full text] [doi: 10.3389/fpubh.2021.609716] [Medline: 33732677]
9. Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, et al. The COVID-19 social media infodemic. Sci Rep 2020 Oct 06;10(1):16598 [FREE Full text] [doi: 10.1038/s41598-020-73510-5] [Medline: 33024152]
10. Gallotti R, Valle F, Castaldo N, Sacco P, De Domenico M. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. Nat Hum Behav 2020 Dec;4(12):1285-1293. [doi: 10.1038/s41562-020-00994-6] [Medline: 33122812]
11. Pulido CM, Villarejo-Carballido B, Redondo-Sama G, Gómez A. COVID-19 infodemic: More retweets for science-based information on coronavirus than for false information. Int Sociol 2020 Apr 15;35(4):377-392. [doi: 10.1177/0268580920914755]
12. 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] [Medline: 32287039]
13. Al-Rakhami MS, Al-Amri AM. Lies kill, facts save: Detecting COVID-19 misinformation in Twitter. IEEE Access 2020;8:155961-155970. [doi: 10.1109/access.2020.3019600]
14. Alsudias L, Rayson P. COVID-19 and Arabic Twitter: How can Arab world governments and public health organizations learn from social media? In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. Stroudsburg, PA: Association for Computational Linguistics; 2020 Presented at: 1st Workshop on NLP for COVID-19 at ACL 2020; July 9-10, 2020; Virtual URL: https://aclanthology.org/2020.nlpcovid19-acl.16.pdf
15. Chandrasekaran R, Mehta V, Valkunde T, Moustakas E. Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study. J Med Internet Res 2020 Oct 23;22(10):e22624 [FREE Full text] [doi: 10.2196/22624] [Medline: 33006937]
16. Chen S, Zhou L, Song Y, Xu Q, Wang P, Wang K, et al. A novel machine learning framework for comparison of viral COVID-19-related Sina Weibo and Twitter posts: Workflow development and content analysis. J Med Internet Res 2021 Jan 06;23(1):e24889 [FREE Full text] [doi: 10.2196/24889] [Medline: 33326408]

17. Hung M, Lauren E, Hon ES, Birmingham WC, Xu J, Su S, et al. Social network analysis of COVID-19 sentiments: Application of artificial intelligence. J Med Internet Res 2020 Aug 18;22(8):e22590 [FREE Full text] [doi: 10.2196/22590] [Medline: 32750001]
18. Imran AS, Daudpota SM, Kastrati Z, Batra R. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. IEEE Access 2020;8:181074-181090. [doi: 10.1109/access.2020.3027350]
19. Jelodar H, Wang Y, Orji R, Huang S. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. IEEE J Biomed Health Inform 2020 Oct;24(10):2733-2742. [doi: 10.1109/JBHI.2020.3001216] [Medline: 32750931]
20. Jo W, Lee J, Park J, Kim Y. Online information exchange and anxiety spread in the early stage of the novel coronavirus (COVID-19) outbreak in South Korea: Structural topic model and network analysis. J Med Internet Res 2020 Jun 02;22(6):e19455 [FREE Full text] [doi: 10.2196/19455] [Medline: 32463367]
21. Kwok SWH, Vadde SK, Wang G. Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: Machine learning analysis. J Med Internet Res 2021 May 19;23(5):e26953 [FREE Full text] [doi: 10.2196/26953] [Medline: 33886492]
22. Satu MS, Khan MI, Mahmud M, Uddin S, Summers MA, Quinn JM, et al. TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets. Knowl Based Syst 2021 Aug 17;226:107126 [FREE Full text] [doi: 10.1016/j.knosys.2021.107126] [Medline: 33972817]
23. Shi A, Qu Z, Jia Q, Lyu C. Rumor detection of COVID-19 pandemic on online social networks. In: Proceedings of the IEEE/ACM Symposium on Edge Computing. 2020 Presented at: IEEE/ACM Symposium on Edge Computing; November 12-14, 2020; San Jose, CA p. 376-281. [doi: 10.1109/SEC50012.2020.00055]
24. Xue J, Chen J, Hu R, Chen C, Zheng C, Su Y, et al. Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach. J Med Internet Res 2020 Nov 25;22(11):e20550 [FREE Full text] [doi: 10.2196/20550] [Medline: 33119535]
25. Zhang C, Xu S, Li Z, Hu S. Understanding concerns, sentiments, and disparities among population groups during the COVID-19 pandemic via Twitter data mining: Large-scale cross-sectional study. J Med Internet Res 2021 Mar 05;23(3):e26482 [FREE Full text] [doi: 10.2196/26482] [Medline: 33617460]
26. Chang C. The 2018 Taiwan Communication Survey (Phase Two, Year Two): Media Use and Social Implications. Taipei, Taiwan: Taiwan Communication Survey; 2020. URL: https://srda.sinica.edu.tw/datasearch_detail.php?id=3053 [accessed 2021-11-15]
27. Junyi S. Jieba. GitHub. URL: https://github.com/fxsjy/jieba [accessed 2021-05-20]
28. 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]
29. Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of the LREC Workshop on New Challenges for NLP Frameworks. 2010 Presented at: LREC Workshop on New Challenges for NLP Frameworks; May 22, 2010; La Valleta, Malta p. 46-50 URL: https://is.muni.cz/publication/884893/lrec2010-rehurek-sojka.pdf [doi: 10.13140/2.1.2393.1847]
30. LINE-rumor-clustering. GitHub. URL: https://github.com/iorg-tw/LINE-rumor-clustering [accessed 2021-10-26]
31. "Director Chen said do not go out before Dragon Boat Festival" is false information. Ministry of Health and Welfare. URL: https://www.mohw.gov.tw/cp-4633-52577-1.html [accessed 2021-05-20]
32. "Coronavirus will stay in your throat for four days" is a false image and a false rumor!. MyGoPen. URL: https://www.mygopen.com/2020/03/gargling-eliminate-coronavirus.html [accessed 2021-05-20]
33. "Drinking warm water with salt and vinegar could eradicate coronavirus" is false information. Taiwan FactCheck Center. URL: https://tfc-taiwan.org.tw/articles/3207 [accessed 2021-05-20]
34. Coronavirus disease (COVID-19) advice for the public: Mythbusters. FACT: Rinsing your nose with saline does NOT prevent COVID-19. World Health Organization. 2021 May 05. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters#saline [accessed 2021-05-20]
35. "Drink salt water to prevent coronavirus" is a misleading false rumor. MyGoPen. URL: https://www.mygopen.com/2020/10/salt-water.html [accessed 2021-05-20]
36. "10 days from today, Taiwan enters the critical period of COVID-19" is misleading information. MyGoPen. URL: https://www.mygopen.com/2020/02/10-key.html [accessed 2021-05-20]
37. Rumor has it that "10 days from today, Taiwan will enter the critical period of COVID-19", so what date is today? Rumor & Truth. 2020. URL: https://www.rumtoast.com/12842 [accessed 2021-05-20]
38. Clarifying "From Taiwan Medical Association, 10 days from today, Taiwan will enter the critical period of COVID-19". Taiwan FactCheck Center. 2020. URL: https://tfc-taiwan.org.tw/articles/2547 [accessed 2021-05-20]
39. COVID-19: Clarification statement of the All-Union Federation. Taiwan Medical Association. 2020 Feb 12. URL: https://www.tma.tw/meeting/meeting_info04.asp?/9112.html [accessed 2021-05-20]
40. Wood T, Porter E. The elusive backfire effect: Mass attitudes' steadfast factual adherence. Polit Behav 2018 Jan 16;41(1):135-163. [doi: 10.1007/s11109-018-9443-y]

41. Ng LHX, Loke JY. Analyzing public opinion and misinformation in a COVID-19 Telegram group chat. IEEE Internet Comput 2021 Mar 1;25(2):84-91. [doi: 10.1109/mic.2020.3040516]

42. Shin J, Jian L, Driscoll K, Bar F. The diffusion of misinformation on social media: Temporal pattern, message, and source. Comput Human Behav 2018 Jun;83:278-287. [doi: 10.1016/j.chb.2018.02.008]

Abbreviations

CECC: Central Epidemic Command Center
HAC: hierarchical agglomerative clustering
IFCN: International Fact-Checking Network
KNN: k-nearest neighbors
LDA: latent Dirichlet allocation
MOHW: Ministry of Health and Welfare
NLP: natural language processing
TCS: Taiwan Communication Survey
VADER: Valence Aware Dictionary and Sentiment Reasoner
WHO: World Health Organization
XGBoost: Extreme Gradient Boosting

Edited by C Lovis; submitted 21.05.21; peer-reviewed by C Zhang, S Lalmuanawma; comments to author 07.06.21; revised version received 29.06.21; accepted 10.09.21; published 23.11.21.

Please cite as:
Wang AW, Lan JY, Wang MH, Yu C
The Evolution of Rumors on a Closed Social Networking Platform During COVID-19: Algorithm Development and Content Study
JMIR Med Inform 2021;9(11):e30467
URL: https://medinform.jmir.org/2021/11/e30467
doi:10.2196/30467
PMID:34623954

Original Paper

Assessing the Value of Unsupervised Clustering in Predicting Persistent High Health Care Utilizers: Retrospective Analysis of Insurance Claims Data

Raghav Ramachandran[1], PhD; Michael J McShea[1], MSc; Stephanie N Howson[1], MSc; Howard S Burkom[1], PhD; Hsien-Yen Chang[2], PhD; Jonathan P Weiner[2], DrPH; Hadi Kharrazi[2], MD, PhD

[1]Applied Physics Laboratory, Johns Hopkins University, Baltimore, MD, United States

[2]Center for Population Health Information Technology, Department of Health Policy and Management, Johns Hopkins School of Public Health, Baltimore, MD, United States

Corresponding Author:
Hadi Kharrazi, MD, PhD
Center for Population Health Information Technology
Department of Health Policy and Management
Johns Hopkins School of Public Health
Baltimore, MD
United States
Phone: 1 4432878264
Email: kharrazi@jhu.edu

Abstract

Background: A high proportion of health care services are persistently utilized by a small subpopulation of patients. To improve clinical outcomes while reducing costs and utilization, population health management programs often provide targeted interventions to patients who may become persistent high users/utilizers (PHUs). Enhanced prediction and management of PHUs can improve health care system efficiencies and improve the overall quality of patient care.

Objective: The aim of this study was to detect key classes of diseases and medications among the study population and to assess the predictive value of these classes in identifying PHUs.

Methods: This study was a retrospective analysis of insurance claims data of patients from the Johns Hopkins Health Care system. We defined a PHU as a patient incurring health care costs in the top 20% of all patients' costs for 4 consecutive 6-month periods. We used 2013 claims data to predict PHU status in 2014-2015. We applied latent class analysis (LCA), an unsupervised clustering approach, to identify patient subgroups with similar diagnostic and medication patterns to differentiate variations in health care utilization across PHUs. Logistic regression models were then built to predict PHUs in the full population and in select subpopulations. Predictors included LCA membership probabilities, demographic covariates, and health utilization covariates. Predictive powers of the regression models were assessed and compared using standard metrics.

Results: We identified 164,221 patients with continuous enrollment between 2013 and 2015. The mean study population age was 19.7 years, 55.9% were women, 3.3% had ≥1 hospitalization, and 19.1% had 10+ outpatient visits in 2013. A total of 8359 (5.09%) patients were identified as PHUs in both 2014 and 2015. The LCA performed optimally when assigning patients to four probability disease/medication classes. Given the feedback provided by clinical experts, we further divided the population into four diagnostic groups for sensitivity analysis: acute upper respiratory infection (URI) (n=53,232; 4.6% PHUs), mental health (n=34,456; 12.8% PHUs), otitis media (n=24,992; 4.5% PHUs), and musculoskeletal (n=24,799; 15.5% PHUs). For the regression models predicting PHUs in the full population, the F1-score classification metric was lower using a parsimonious model that included LCA categories (F1=38.62%) compared to that of a complex risk stratification model with a full set of predictors (F1=48.20%). However, the LCA-enabled simple models were comparable to the complex model when predicting PHUs in the mental health and musculoskeletal subpopulations (F1-scores of 48.69% and 48.15%, respectively). F1-scores were lower than that of the complex model when the LCA-enabled models were limited to the otitis media and acute URI subpopulations (45.77% and 43.05%, respectively).

Conclusions: Our study illustrates the value of LCA in identifying subgroups of patients with similar patterns of diagnoses and medications. Our results show that LCA-derived classes can simplify predictive models of PHUs without compromising predictive accuracy. Future studies should investigate the value of LCA-derived classes for predicting PHUs in other health care settings.

(JMIR Med Inform 2021;9(11):e31442) doi:10.2196/31442

KEYWORDS

persistent high users; persistent high utilizers; latent class analysis; comorbidity patterns; utilization prediction; unsupervised clustering; population health analytics; health care; prediction models; health care services; health care costs

Introduction

A small segment of the patient population utilizes a high volume of health care services [1,2]. Population health management programs often aim to identify high-utilizing subpopulations and provide them with appropriate preventative interventions to reduce undesired health outcomes while lowering utilization [2,3]. Reducing unnecessary health care utilization such as avoidable inpatient admissions enables more effective use of health care resources across the patient population, hence improving the overall health of the managed population [2-4].

Population health programs are often managed by insurers and health care providers [2,5]. Traditionally, health care payers use insurance claims to identify members/enrollees with high rates of utilization. Health care providers are increasingly using electronic health records (EHRs) to identify high-utilizing patients [6,7]. Payers and providers routinely apply established risk stratification techniques against their data to predict the members/patients who will become a high utilizer in the short term (eg, 30 days to 12 months) [8-11]. However, predicting who will continuously remain a high utilizer in the long term (eg, 24 months or more) has proven to be a challenging task for population health risk stratification [12].

Persistent high users/utilizers (PHUs) are patients who have a high utilization rate over an extended period (eg, a patient whose annual costs are in the top 20% of all patients' costs over 4 consecutive 6-month periods) [1,13]. Recent studies have taken several approaches to characterizing PHUs, including the frequency and type of utilization, total costs, and number of chronic conditions [1,8-13]. Despite the variety of terminologies used for PHUs (eg, high-cost high-need, super-utilizers), population health analysts have typically faced barriers in extracting the common probability classes of diagnoses and medications for PHUs to improve the management of health care resources in specific subpopulations [13,14].

PHUs constitute a small percentage of the patient population [1]. PHUs of a health system may present a different mix of comorbidities and medications compared with those of PHUs in other health systems [8-14]. The variability of the underlying probabilities of PHUs' diseases and medications across different settings complicates the use of traditional approaches for identifying PHUs from groupings of diagnostic codes. Considering this diversity of conditions, the manual grouping of diagnostic and medication codes by clinical experts will not only be burdensome to compile for a given health system but also impractical to use elsewhere [1-3]. Automated clustering/grouping techniques can be a valuable alternative to characterizing PHUs for a specific health system patient subpopulation [15-19]. Automated groupings of health care utilization patterns can also enhance the prediction of PHUs through traditional analytical methods such as logistic regression [15].

To address the difficulties of identifying common patterns of comorbidities among PHUs, in this study, we implemented an unsupervised clustering methodology, latent class analysis (LCA) [20], to semiautomatically classify PHU patients by a limited number of probability classes of characteristic comorbidities and medications. We then used the LCA classes along with a few demographic and health system factors to predict PHU status for each member of the total study population and a selected set of patient subpopulations. We finally compared our LCA-enabled predictive model with a sophisticated (but more complex) risk stratification model that uses several demographic, clinical, and medication factors to predict PHU status.

Methods

Overall Aims and Definitions

The overall goal of our study was to identify subpopulations of PHUs where changes in care delivery could reduce the risk of high utilization. Our analysis aimed to automate the extraction of common probabilistic patterns of comorbidities and medications for PHUs, and then use such information to improve the prediction of PHUs among the study population as well as specific diagnostic subpopulations.

We defined a PHU as an individual whose medical charges remained in the top 20% of the highest health care costs for 4 consecutive 6-month periods (ie, total of 2 years after the base period) [1]. Health care costs were defined as the sum of hospital inpatient, outpatient department, emergency department (ED), and professional and pharmacy costs covered by the insurer and the patient's out-of-pocket costs [1,6].

Data Source and Preparation

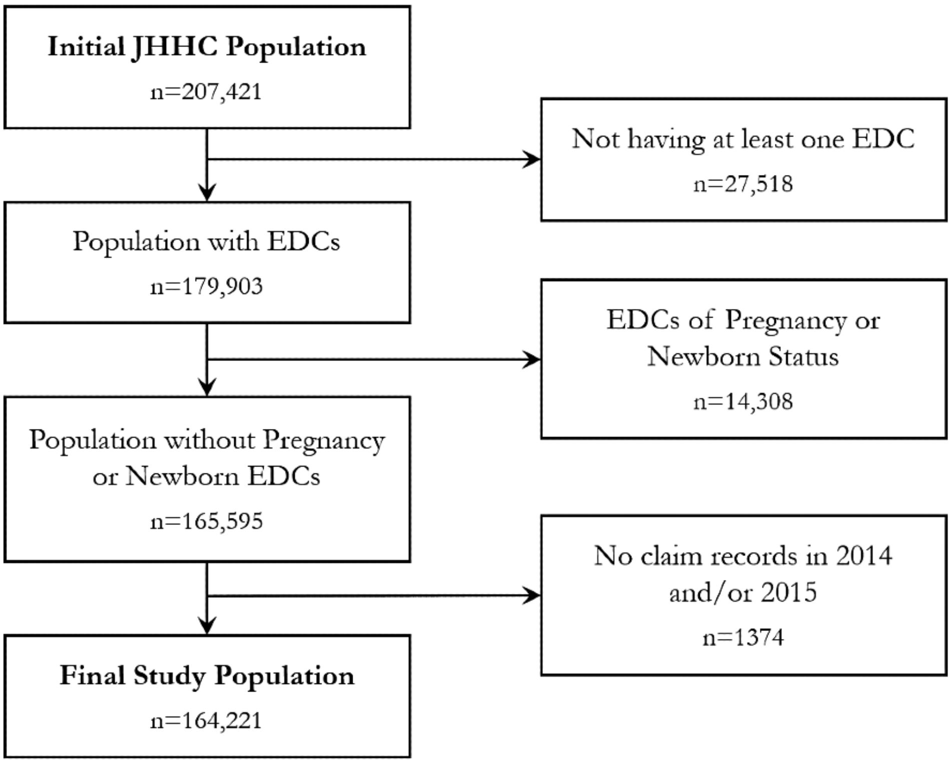
We performed a retrospective analysis of the Johns Hopkins Health Care (JHHC) insurance claims data captured between 2013 and 2015. JHHC provides health insurance to a variety of enrollees, including Medicaid and employer-based members. JHHC enrollees can also seek care outside of the Johns Hopkins health system. We applied the Johns Hopkins Adjusted Clinical Groups (ACG) software to the claims data to generate additional health care utilization variables consistent with previous PHU analyses [1,21]. We categorized the diagnostic codes into higher-level diagnosis groupings defined by the ACG methodology as expanded diagnostic clusters (EDCs), and grouped the medication data into ACG prescription-defined morbidity groups (RxMGs) [21]. EDCs and RxMGs, which are extensively validated and routinely used for risk stratification [1,6], were used in our analysis as the base diagnosis and medication categories, respectively.

Study Population

Our initial sample population included 207,421 patients with at least one JHHC claims record in 2013 and at least 2 years of continuous JHHC enrollment between 2013 and 2015 (Figure 1). Following the CONSORT (Consolidated Standards of Reporting Trials) statement [22], we first excluded 27,518 patients with missing EDC diagnosis codes since EDCs were used to identify clusters of patients within the population. Next, we excluded 14,308 patients with pregnancy or newborn EDC codes since high costs typical of pregnancy complications differ from those that distinguish PHUs. Finally, we excluded an additional 1374 patients without JHHC claims in 2014-2015 since data in 2013 were used to predict PHUs in 2014 and 2015. The final study population included 164,221 patients (Figure 1).

To explore the sensitivity of our approach, we further divided the study population into four distinct diagnostic-driven subpopulations. These subpopulations were chosen based on the frequency of the underlying EDC data and were validated by two clinicians. The clinicians reviewed the combination of EDCs and asserted their practical use in clinical settings. These subpopulations were identified as: (1) otitis media (n=24,992 patients), (2) mental health (n=34,456), (3) musculoskeletal signs and symptoms (n=24,799), and (4) acute upper respiratory infection (URI; n=53,232).

Figure 1. Selection process of the study population. JHHC: Johns Hopkins Health Care; EDC: expanded diagnostic cluster.


Initial JHHC Population
n=207,421
Not having at least one EDC
n=27,518
Population with EDCs
n=179,903
EDCs of Pregnancy or Newborn Status
n=14,308
Population without Pregnancy or Newborn EDCs
n=165,595
No claim records in 2014 and/or 2015
n=1374
Final Study Population
n=164,221

Predictors and Outcome

The full study population and each subpopulation contained several predictor variables and the outcome variable. Predictors (ie, independent variables) included demographics, EDCs, Rx-MGs, and other health utilization variables (eg, hospitalization, care coordination) generated by the ACG system. Many of these predictors, including all EDCs and Rx-MGs, are categorical variables [21].

The outcome of interest, a binary variable, was whether or not a patient became a PHU after the base year (ie, being in the top 20% of the highest health care costs over 4 consecutive 6-month periods from 2014 to 2015). The outcome variable was calculated separately in the full population and in each of the diagnostic subpopulations (eg, a patient might be considered a PHU in a subpopulation but not in the full population).

Statistical Approach

Unsupervised Clustering to Identify Diagnoses Clusters

LCA was performed on the full study population and on each subpopulation separately to identify "phenotypes" (ie, classes) of disease subtypes [20]. LCA is an unsupervised data-driven clustering technique that identifies unobserved subtypes (latent classes) within a population based on probability theory. A key assumption in LCA is that conditional independence (ie, latent class membership) explains all of the shared variance across variables [20].

The main parameters generated by LCA are the probabilities of latent class membership for each individual (ie, each patient in the mental health subpopulation; n=34,456) and the class-specific probabilities of observing each binary variable (eg, tobacco use EDC among mental health patients). These probabilities distinguish LCA from binning techniques in which each individual (eg, patient) is merely assigned a probability of belonging to an unobserved/latent class (eg, representing a specific pattern of comorbidities) based on a well-established statistical theory [20].

LCA creates latent classes that optimize minimizing the variance across individuals within each class while maximizing the variance between individuals in different classes. Moreover, LCA is a person-centered approach, does not make distributional assumptions, and works well with categorical data, making it particularly applicable to subtype identification of patients using diagnostic data such as EDCs [20].

LCA models with a varied number of latent classes (2 to 6 classes) were constructed using EDC, Rx-MG, and selected patient-level resource utilization variables. For both the full population and the select subpopulations, 4-class models were chosen because they provided the right balance between optimal model fit and interpretability of the classes. Although models with more classes (eg, 5- and 6-class models) might fit the data slightly better, the interpretation of the classes becomes less clear, and often classes may differ only across a few variables. In other words, the gain in fit is not sufficient to overcome the decline in interpretability that comes from adding too many classes to the model. Additionally, LCA models with more than 6 classes did not improve the standard fit metrics, explained a very small proportion of patients, and had limited mathematical convergence, and were therefore not considered in this study.

LCA fit was measured using G^2, Akaike information criterion (AIC), and Bayesian information criterion (BIC) metrics; lower values of G^2, AIC, and BIC imply a better fit [23,24]. Similar to standard regression techniques, LCA uses maximum-likelihood estimation to determine its model parameters. The goal of maximum-likelihood estimation is to maximize the probability (likelihood, L) that the process described by the model produced the observed data: $G^2=–2\times\log(L)$, $\text{AIC}=–2\times\log(L)+2\times k$, and $\text{BIC}=–2\times\log(L)+k\times\log(N)$, where k is the number of estimated model parameters and N is the sample size. Since L is maximized to achieve the best fit to the data, $–2\times\log(L)$ is also minimized, and thus lower G^2, AIC, and BIC values indicate a better model fit. For a large sample where $\log(N)>2$, AIC tends to favor more complex models (ie, more model parameters) over BIC [23,24].

LCA does not bin each individual into a class but rather calculates the probability that an individual's characteristics most closely match those of the other individuals in each class. Classes are constructed to maximize similarity of individuals' characteristics within a class and dissimilarity of individuals across classes. For example, in this study, the LCA methodology generated four different class probabilities for each patient representing the similarity of the patient's comorbidities (ie, mix of EDCs and RxMGs) to comorbidities of patients in each LCA-derived class of the entire study population.

Logistic Regression Modeling to Predict PHUs

Once the classes were constructed via LCA and health utilization characteristics of the classes were graphically compared, we trained logistic regression models to predict PHUs in both the full population and in each subpopulation using the following variables: (1-3) latent class membership probabilities for 3 of the 4 classes (the class with the lowest chronic EDC/RxMG probabilities was chosen to be the reference class); (4) gender (male; reference=female); (5-9) race (Black, Asian, Hispanic, other, missing; reference=White); (10) medical and pharmacy coverage in 2013; (11) Medicaid eligibility; (12) number of acute care inpatient days; (13) number of acute care inpatient stays; (14) presence of frailty conditions; and (15-16) likely or possibly experiencing care coordination issues (yes/no). Variables 12 to 16 were generated by the ACG system [21] using the JHHC medical claims data.

We also used the ACG system's internal risk stratification functions (ie, embedded models) to predict PHU status in the full population [21]. The ACG system implements a complex model that uses over 300 variables (eg, demographics, all EDCs, all RxMGs, and dozens of health system variables) to predict health care utilization such as inpatient admissions, ED visits, and overall medical or pharmacy costs. Predictive performance of all regression models was assessed and compared using sensitivity, predictive positive value (PPV), and the F1-score.

All analyses, including the descriptive analysis of the full population and all subpopulations, were performed in R (v3.5.1). We used R's basic packages for the LCA clustering [25] and logistic regression predictions.

Results

Descriptive Analyses

Descriptive statistics for the full population are summarized in Table 1. Overall, approximately 5% of the full population were identified as PHUs. The average age of PHUs was more than twice that of the non-PHU population. The percentage of males was smaller among PHUs than among non-PHUs. As expected, a larger percentage of PHUs had one or more inpatient or outpatient visits compared to non-PHUs (18.7% vs 2.5% for inpatient visits and 99.7% vs 97.3% for outpatient visits, respectively). Similar descriptive statistics were generated for each of the four diagnostic subpopulations (see Multimedia Appendix 1-4).

Table 1. Characteristics of the study populations.

Characteristic	Overall study population (N=164,221)	Non-PHU[a] population (n=155,862)	PHU population (n=8359)
Age group (years), n (%)			
0-17	100,811 (61.4)	99,352 (63.7)	1459 (17.5)
18-64	62,396 (38.0)	55,666 (35.7)	6730 (80.5)
65+	1014 (0.6)	844 (0.5)	170 (2.0)
Age (years), mean (SD)	19.79 (17.43)	18.79 (16.82)	38.51 (18.01)
Male, n (%)	72,418 (44.1)	69,683 (44.7)	2735 (32.7)
Race, n (%)			
White	41,219 (25.1)	38,762 (24.9)	2,457 (29.4)
Black	53,872 (32.8)	50,993 (32.7)	2,879 (34.4)
Other[b]	149 (0.1)	143 (0.1)	6 (0.1)
Missing[c]	68,981 (42.0)	65,964 (42.3)	3017 (36.1)
Inpatient visits, n (%)			
0	158,763 (96.7)	151,971 (97.5)	6792 (81.3)
1-5	5,366 (3.3)	3,866 (2.5)	1500 (17.9)
6-10	74 (<0.1)	20 (<0.1)	54 (0.6)
11+	18 (<0.1)	5 (<0.1)	13 (0.2)
Outpatient visits, n (%)			
0	3,690 (2.2)	3,663 (2.4)	27 (0.3)
1-5	95,372 (58.1)	94,138 (60.4)	1234 (14.8)
6-10	33,745 (20.5)	32,317 (20.7)	1428 (17.1)
11+	31,414 (19.1)	25,744 (16.5)	5670 (67.8)

[a] PHU: persistent high users.

[b] “Other”describes members of known race/ethnicity not equal to Asian, Hispanic, White, or Black.

[c] “Missing” describes members with empty values for race.

Latent Class (Cluster) Analyses

LCA models with 2 to 6 classes were trained using the full population to identify the optimal number of classes. The fit statistics for these models were then calculated and compared for the full population (Table 2). The 4-class models were chosen for both the full population and subpopulations as they optimally balanced good model fit with interpretability of the classes (see Multimedia Appendix 5). The LCA’s 4 classes represented probability patterns of diseases and medications that were deemed to be optimal and interpretable for identifying subgroups of patients within the full sample and in each of the diagnostic subpopulations.

A model with the lowest AIC tends to be more complex if it is not the same as the model with the lowest BIC [23]. Thus, we selected the 4-class LCA model since it fit the data better than the 2- and 3-class models, and the classes were more interpretable than those in the 5- and 6-class models (Table 2). Additionally, AIC and BIC metrics can be compared only across nested models (ie, when the terms in one model are a subset of the terms in the other model). As a result, AIC and BIC measures should not be compared across different study subpopulations (Multimedia Appendix 5).

The LCA models were run with 178 different EDCs and RxMGs on the full population and with the same EDCs/RxMGs on the diagnostic subpopulations, excluding the EDCs used to define the subpopulations. Examining all EDCs/RxMGs in our 4-class LCA models, excluding the EDCs used to define our subpopulation, led us to very similar descriptions of each class. A caveat to this observation is that many EDCs/RxMGs had very low or very high probabilities of being observed in all classes and hence were not useful for distinguishing among classes.

Each LCA class contained item-response probabilities for each of the EDC/RxMG codes; however, for only a few of the EDC/RxMG codes, the probability was ≥0.4 in every class. Figure 2 depicts the EDC/RxMG codes that reached the threshold of 0.4 within the full population across all classes. Within the figure, the selected EDC categories that made the threshold are shown along the x-axis and their (item-response) probabilities are shown on the y-axis. The color shading indicates the four different LCA classes, which have different levels of probabilities across different EDCs. Only items with a maximum difference in probability of 0.4 (40%) or greater across pairs of classes are shown for simplicity. Classes 1, 3, and 4 represent people with moderate, high, and low likelihoods of EDCs, respectively. Class 2 is associated with higher probabilities of infections.

The selected subtype characteristics from the LCA and fractions of patients assigned to each subtype were also explored for each of the four diagnostic subpopulations (Figures 3-6). For example, within the full study population, 21.2% of the patients were attributed to class 1 (Figure 2). However, 13.2%, 14.9%, 30.0%, and 46.2% of the patients were in class 1 for the otitis media (Figure 3), mental health (Figure 4), musculoskeletal (Figure 5), and acute URI (Figure 6) subpopulations, respectively. In Figures 3 to 6, only items with a maximum difference in probability of 0.4 (40%) or greater across pairs of classes are shown for simplicity. In Figure 3, classes 1, 2, and 3 represent people with moderate, low, and high (particularly chronic conditions) likelihoods of EDCs, respectively, whereas class 4 is associated with higher probabilities of infections (eg, URI) and fever. In Figure 4, classes 1 and 3 represent people with high and low likelihoods of EDCs, respectively, whereas class 2 is associated primarily with a high likelihood of minor infections and class 4 represents people with moderate likelihoods of infections and pain. In Figure 5, classes 1, 3, and 4 represent people with moderate, low, and high likelihoods of EDCs, respectively, whereas class 2 is associated primarily with a high likelihood of minor infections. In Figure 6, classes 1, 3, and 4 represent people with low, moderate, and high likelihoods of EDCs, respectively, whereas class 2 is associated primarily with a high likelihood of airway hyperactivity.

Only a handful of EDCs clearly distinguished the four classes in each LCA model (full population and the diagnostic subpopulations). In the full population and in most of the diagnostic subpopulations, three of these classes were associated with uniformly high, moderate, or low probabilities of the EDCs. The remaining class was characterized primarily by a high likelihood of minor infections, pain, or respiratory diagnoses (Figures 2-6).

Table 2. Model fit statistics for latent class analysis models with 2 to 6 classes (N=164,221).

Model	G^2[a]	AIC[b]	BIC[c]
2-class model	5,487,702	9,113,315	9,116,888
3-class model	5,213,964	8,839,935	8,845,300
4-class model	5,088,223	8,714,552	8,721,708
5-class model	4,934,192	8,560,878	8,569,826
6-class model	4,874,634	8,501,679	8,512,419

[a] G^2: likelihood ratio/deviance statistic.

[b] AIC: Akaike information criterion.

[c] BIC: Bayesian information criterion.

Figure 2. Latent class item-response probabilities for the full population (N=164,221).

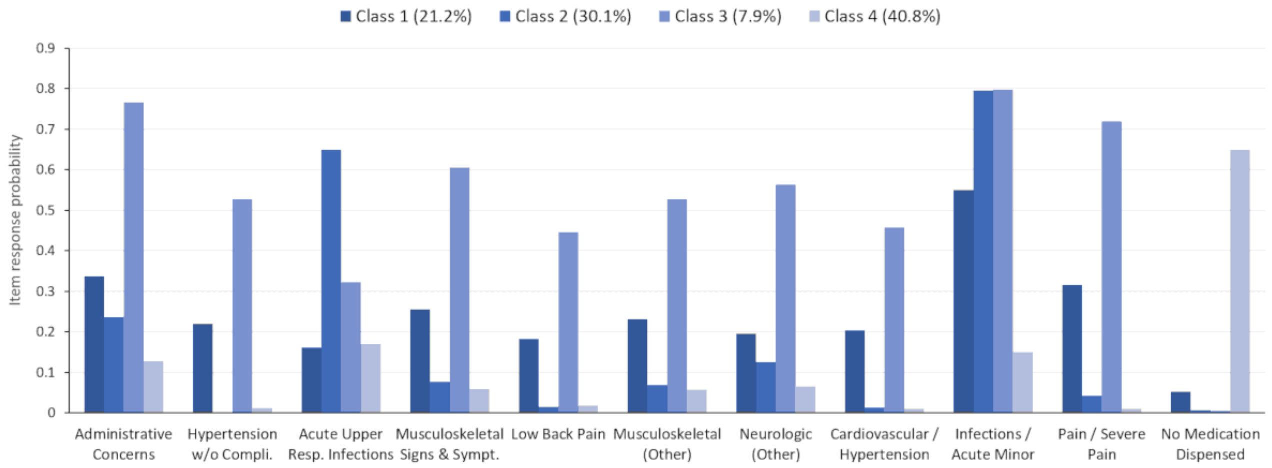


Figure 3. Latent class item-response probabilities for the otitis media subpopulation (n=24,992).

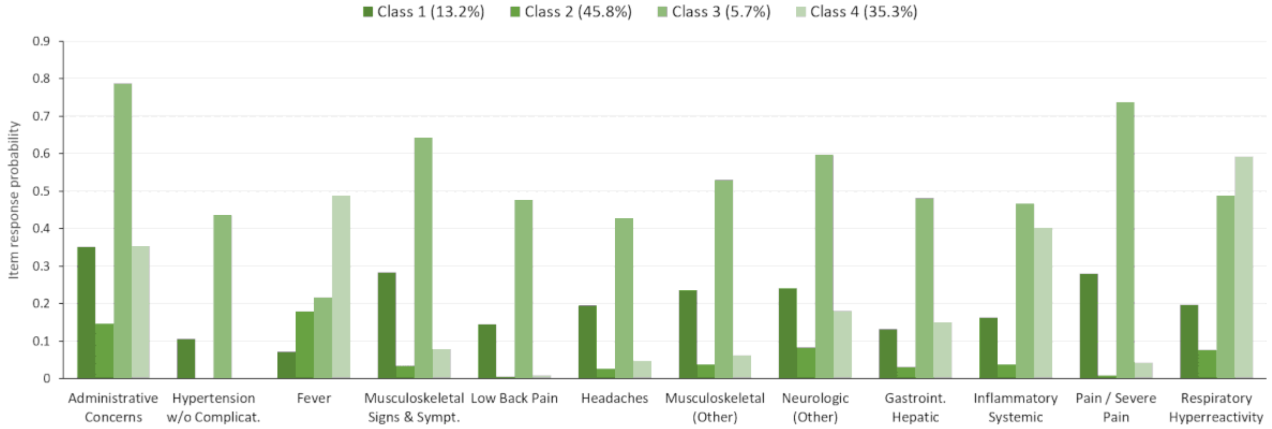


Figure 4. Latent class item-response probabilities for the mental health subpopulation (n=34,456).

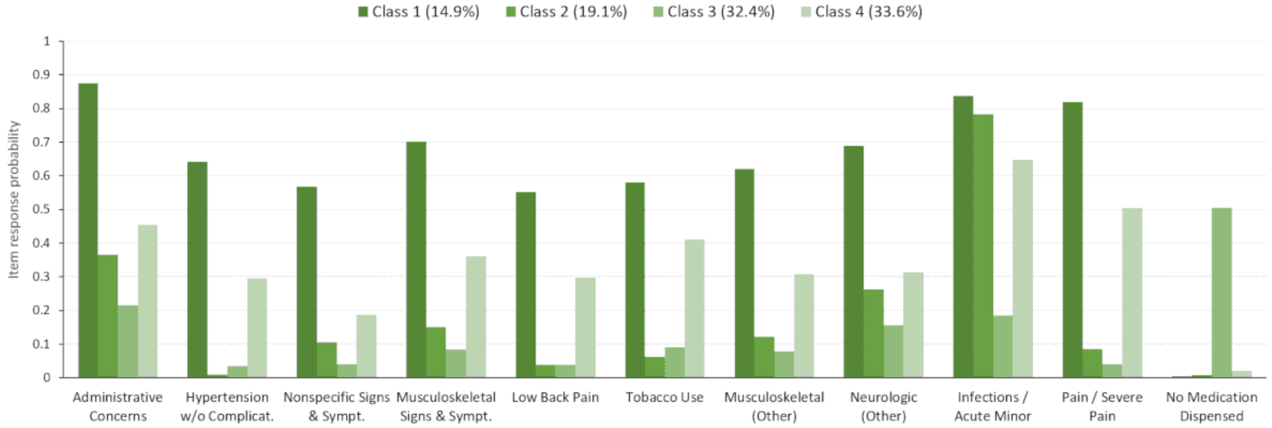


Figure 5. Latent class item-response probabilities for the musculoskeletal subpopulation (n=24,799).

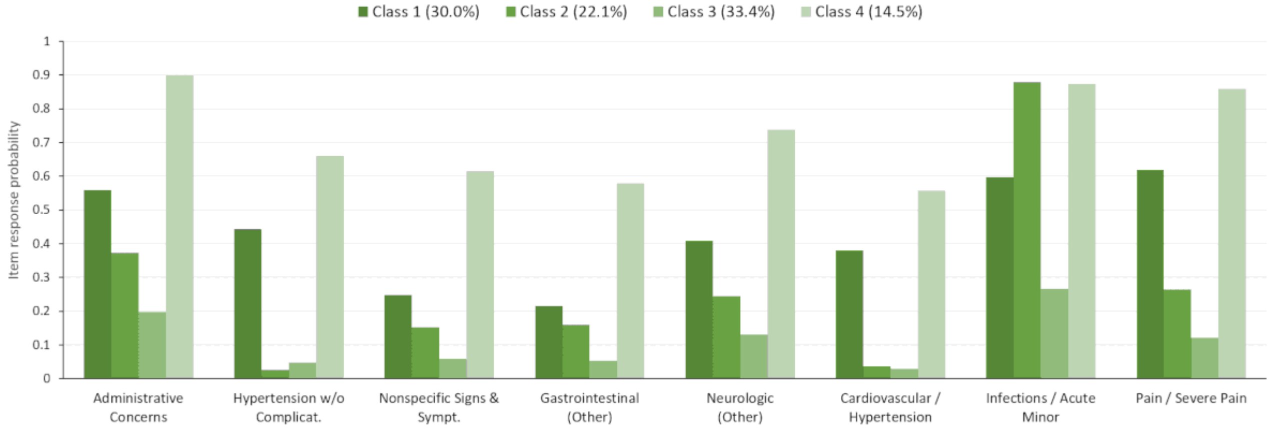
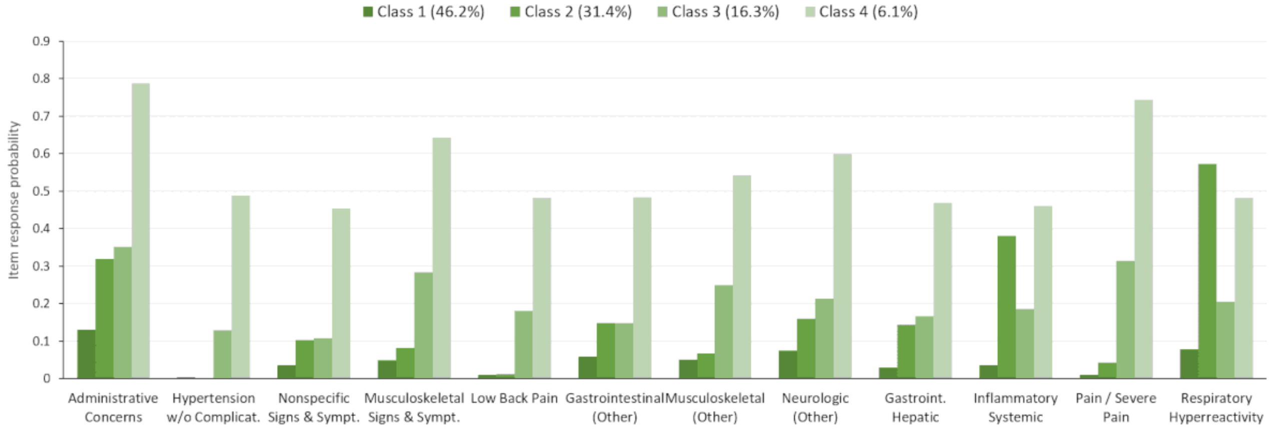


Figure 6. Latent class item-response probabilities for the acute upper respiratory infection subpopulation (n=53,232).



PHU Predictive Modeling (Logistic Regression)

Logistic regression models were developed for the full population and for each subpopulation to predict PHUs from latent class membership probabilities along with demographic and health utilization characteristics of each patient. These models were trained on a randomly selected sample of 80% of the patients in the full population/subpopulation and were evaluated on a test data set with the other 20% of patients. Classification metrics for each of these models (Table 3) revealed that PHU predictions are more accurate within subpopulations that have a high prevalence of PHUs. For example, the F1-score reached 38.6 in the LCA-enabled regression models predicting PHUs in the full population, whereas the F1-score reached 45.8, 48.7, 48.1, and 43.0 among the otitis media, mental health, musculoskeletal, and acute URI subpopulations, respectively. Although the musculoskeletal subpopulation had the highest percentage of PHUs (Table 3), the regression model for the mental health subpopulation performed the best in terms of the sensitivity and F1-score (62.4 and 48.7 vs 55.1 and 48.1, respectively).

The LCA-enabled regression model for the full population performed modestly lower than the ACG model (ie, F1-score 38.6 vs 48.2); however, the LCA-enabled model had fewer predictors (16 variables) than the ACG model (≥300 variables). The F1-scores of the LCA-enabled regression models in the subpopulations were comparable to the F1-score of the complex ACG model in predicting PHUs in the full population (ie, F1-scores ranging from 43.0 to 48.7 vs 48.2). Since the specificity, sensitivity, PPV, and F1-score were calculated for specific thresholds, only one estimate was calculated for each of those metrics (ie, the 95% CI was not applicable).

Table 3. Comparing classification metrics for predicting persistent high user/utilizer (PHU) status.

Metric	Full study population (N=164,221)		Otitis media (n=24,992)	Mental health (n=34,456)	MSK[a] (n=24,799)	acute URI[b] (n=53,232)
	ACG[c]	LCA-LRM[d]	LCA-LRM	LCA-LRM	LCA-LRM	LCA-LRM
PPV[e] (%)	48.60	38.53	44.40	39.91	42.74	41.28
Sensitivity (%)	47.90	38.72	47.23	62.43	55.14	44.99
F1-score (%)	48.20	38.62	45.77	48.69	48.15	43.05
Percentile (threshold)	95th (0.33)	95th (0.33)	95th (0.18)	80th (0.25)	95th (0.53)	95th (0.23)
PHUs (%)	5.1	5.1	4.5	12.8	15.5	4.6

[a] MSK: musculoskeletal.

[b] URI: upper respiratory infection.

[c] ACG: Adjusted Clinical Groups; latent class analysis results not included in the model.

[d] LCA-LRM: latent class analysis-logistic regression model; latent class probabilities included as predictors in the model.

[e] PPV: positive predictive value.

Odds ratios (ORs) of the LCA-enabled regression models predicting PHUs in the full population and in each of the diagnostic subpopulations were calculated separately (Multimedia Appendices 6-10). In all LCA-enabled regression models, the class probabilities were statistically significant in predicting PHUs and resulted in the highest ORs of 22.3, 6.0, and 135.3 for classes 1, 2, and 3 in the full population model, respectively. Other predictors were either not statistically significant (eg, sex, inpatient hospitalization days) or, if significant, had a small effect size (ie, ORs ranging between 0.4 and 3.0). Being Asian or Hispanic, having medical or pharmacy insurance coverage, and being on Medicaid were protective against PHUs (ie, ORs of 0.77, 0.41, 0.85, and 0.69, respectively), while being Black, having a high count of inpatient stays, holding frailty conditions, and likely or possibly experiencing care coordination issues were associated with

PHUs (ie, ORs of 1.18, 1.25, 1.14, 1.68, and 3.07, respectively). These findings highlight some of the demographic and health care factors associated with a higher or lower likelihood of being a PHU.

Discussion

Principal Findings

PHUs are defined as the patient population who stay in the highest deciles of health care costs and/or utilization for multiple years [1,8-15]. Predicting PHUs is a challenge as their underlying mix of comorbidities and medications may differ across settings [12,13]. To address this analytic gap and improve the efficiency of grouping underlying conditions of PHUs, we applied LCA, a novel unsupervised clustering approach, to the JHHC's insurance claims data to identify classes of high-utilizing patients with similar probabilities for different sets of diseases and medications. We then explored the value of the LCA classes for predicting which patients, within the full population or specific subpopulations, will become PHUs using a simple parsimonious regression model, and then compared its predictions to those of a more detailed complex predictive model.

Our study demonstrated the use of nontraditional statistical clustering methods such as LCA to facilitate the automated development of diagnostic and medication probability classes that can be effectively used in traditional logistic regression models to predict PHUs, without the need for complex predictive models. Two of our study findings specifically support the use of LCA in predicting PHUs. First, the F1-score of the LCA-enabled logistic regression was comparable to that of the complex predictive model despite having a fraction of the variable predictors (16 vs ≥300 variables). Second, the ORs of the LCA-derived classes were much higher (ranging from 22 to 135) than those of the other variables (ranging from 0.4 to 3.0) used in the logistic regressions. Therefore, LCA can be an efficient (ie, unsupervised process requires minimal manual effort), effective (ie, high ORs in the predictive models), and usable (ie, avoiding complex predictive models) method for predicting PHUs in different settings.

The mix of LCA classes may differ among PHUs of different health systems. For example, our study population of 164,221 patients included 130,711 members enrolled in a special Medicaid insurance plan (ie, Johns Hopkins Priority Partners) targeting mothers and children. Thus, as 79.6% of the study population were enrolled in this Medicaid program, the average age of the full population was close to 20 years. Consequently, the most common EDCs for three of the four diagnostic subpopulations included pediatric conditions such as ear problems [26], which led our clinical experts to categorize one of the subpopulations as otitis media. In addition, the fact that one of the diagnostic subpopulations was identified as "mental health" reflects the reported association of higher health care costs for children with mental health conditions [27], which made this subpopulation particularly relevant to our study of PHUs.

Comparison With Prior Work

A few prior studies have explored the use of LCA and other classifying techniques to improve the prediction of PHUs. One study focused on US older and middle-aged patients and grouped them using the Medical Expenditure Panel Survey data set to explore high to moderate utilization rates [16]. Due to the older demographic of their population, the study found age, unemployment, insurance status, and number of chronic conditions and medications as key clustering factors. Two separate studies in Singapore applied LCA to segment populations into different utilization classes [18,19]. Their first study focused on primary health care patients enrolled in governmental insurance programs, and found that a specific class with metabolic diseases and multiorgan complications had the highest hospital admissions and ED visits [18]. Their second study focused on patients enrolled in the government-sponsored hospital-to-home transitional care program, and found that patients with frailty and cognitive impairment had the highest hospital readmission rate [19]. Another study in the United States further explored the use of LCA grouping for improving the prediction of superutilizers; however, that study was limited to veterans experiencing homelessness [15]. Veterans who were in an LCA group representing older, male, White, unmarried, and disabled patients proved most likely to be superutilizers. However, none of these studies explored the Medicaid population (with a high percentage of pediatric patients), assessed the LCA classes in separate diagnostic subpopulations in addition to the full population, or compared the value of LCA classes in predicting PHUs compared to a standard/complex utilization prediction model.

Practical Implications

Health care providers increasingly use risk stratification tools to manage their patient populations. However, providers often do not have access to insurance claims data and use local EHRs to risk stratify patients and predict PHUs [6,7,28]. Despite the advances in using unique EHR data in improving risk prediction [29-34], quality issues render EHR data challenging to use in complex predictive models of utilization [35-38]. Using an unsupervised methodology to classify underlying diagnostic and medications can enable providers to surmount some of these deficiencies and improve the prediction of PHUs using EHR data [37]. Furthermore, LCA and similar classification approaches can help providers to better understand the unique needs of their underlying patient populations and to better target their population health interventions [39]. Nonetheless, fully automating the LCA classes, and excluding clinical feedback in the process, may result in identifying subpopulations that may not provide a meaningful clinical context for targeted care management.

Limitations

Our study has several limitations. First, the results of our LCA approach, and the improvement of the PHU prediction, may not generalize to other populations (eg, older adults, Medicare), settings (eg, inpatient only), or data sources (eg, EHRs). Future research should explore the use of LCA in new populations and settings using alternate data sources. Second, our specific definition for PHU (ie, percentile of cost and time period) may not fit all populations. The risk stratification research community should offer a harmonized definition of PHU so that various research findings on PHUs can be compared effectively to establish generalizable evidence. Third, results of the logistic regression should be interpreted with caution as race and ethnicity are likely to be closely linked to differences in health care coverage and quality rather than being directly related to PHU [40,41]. Fourth, although the LCA approach automates the classification of the populations, clinical feedback is still key to produce useful results. Hence, the LCA process may become more complex to incorporate in clinical settings compared to the traditional regression models such as ACGs [1,21]. Finally, our selection of the diagnostic subpopulations was based on subjective feedback provided by clinical experts. Future research should examine a mix of qualitative and quantitative methods to normalize and expedite this process. Moreover, with even ideal classification of high-cost health care users, effective operational use of these classes in clinical and operational settings remains to be determined.

Conclusion

A small percentage of patients use most of the health care services continuously over extended periods. We used LCA, an unsupervised clustering approach, to automate the process of extracting classes of comorbidity and medication probabilities for individual patients that can be effectively used in predicting PHUs. The latent classes highlight broad differences in health care utilization patterns among groups of people, while also providing a way to condense critical information into a smaller set of variables to simplify the PHU prediction model and improve its interpretability. From a care management perspective, the LCA and PHU prediction models provide care managers with insights on specific resource utilization variables that are strongly associated with PHU. Future studies should investigate the value of LCA-derived classes for predicting PHUs in other health care settings with potentially different underlying populations.

Acknowledgments

We acknowledge the contributions of Sheri Maxim, Jonathan Thornhill, Jason Lee, Hong Kan, and Thomas Richards to this project. This project was funded by the Johns Hopkins APL’s National Health Mission Area (NHMA) Independent Research and Development (IRAD) program.

Authors' Contributions

HK and MM codirected the research project. RR and SH analyzed the data. HC provided analytical insight and calculated claims costs. HK, MM, HB, and JW reviewed and interpreted the results. HK, RR, and MM drafted the manuscript. All authors reviewed and contributed to the final manuscript. HK prepared the manuscript for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Descriptive statistics for the otitis media subpopulation (n=24,992).

[DOC File , 40 KB - medinform_v9i11e31442_app1.doc]

Multimedia Appendix 2

Descriptive statistics for the mental health subpopulation (n=34,456).

[DOC File , 41 KB - medinform_v9i11e31442_app2.doc]

Multimedia Appendix 3

Descriptive statistics for the musculoskeletal subpopulation (n=24,799).

[DOC File , 42 KB - medinform_v9i11e31442_app3.doc]

Multimedia Appendix 4

Descriptive statistics for acute upper respiratory tract infection (URI) subpopulation (n=53,232).

[DOC File , 41 KB - medinform_v9i11e31442_app4.doc]

Multimedia Appendix 5

Model fit statistics for latent class analysis (LCA) in diagnostic subpopulations.

[DOC File , 30 KB - medinform_v9i11e31442_app5.doc]

Multimedia Appendix 6

Odds ratios of predictors in the LCA-enabled logistic regression model predicting persistent health care users/utilizers (PHUs) in the full population (N=164,221).
[DOC File , 43 KB - medinform_v9i11e31442_app6.doc]

Multimedia Appendix 7
Logistic regression odds ratios for the otitis media subpopulation.
[DOC File , 57 KB - medinform_v9i11e31442_app7.doc]

Multimedia Appendix 8
Logistic regression odds ratios for the mental health subpopulation.
[DOC File , 58 KB - medinform_v9i11e31442_app8.doc]

Multimedia Appendix 9
Logistic regression odds ratios for the musculoskeletal subpopulation.
[DOC File , 57 KB - medinform_v9i11e31442_app9.doc]

Multimedia Appendix 10
Logistic regression odds ratios for the acute upper respiratory infection subpopulation.
[DOC File , 56 KB - medinform_v9i11e31442_app10.doc]

References

1. Chang H, Boyd CM, Leff B, Lemke KW, Bodycombe DP, Weiner JP. Identifying consistent high-cost users in a health plan: comparison of alternative prediction models. Med Care 2016 Sep;54(9):852-859. [doi: 10.1097/MLR.0000000000000566] [Medline: 27326548]
2. Iezzoni LI, editor. Risk adjustment for measuring health care outcomes. Chicago, IL: Health Administration Press; 2012.
3. Kharrazi H, Gamache R, Weiner J. Role of informatics in bridging public and population health. In: Magnuson JA, Dixon BE, editors. Public health informatics and information systems. London, UK: Springer; 2020:59-79.
4. Gamache R, Kharrazi H, Weiner JP. Public and population health informatics: the bridging of big data to benefit communities. Yearb Med Inform 2018 Aug;27(1):199-206 [FREE Full text] [doi: 10.1055/s-0038-1667081] [Medline: 30157524]
5. Kharrazi H, Lehmann H. Role of population health informatics in understanding data, information and knowledge. In: Joshi A, Thorpe L, Waldron L, editors. Population health informatics: driving evidence-based solutions into practice. Burlington, MA: Jones and Bartlett Learning; 2017:61-86.
6. Kharrazi H, Chi W, Chang H, Richards TM, Gallagher JM, Knudson SM, et al. Comparing population-based risk-stratification model performance using demographic, diagnosis and medication data extracted from outpatient electronic health records versus administrative claims. Med Care 2017 Aug;55(8):789-796. [doi: 10.1097/MLR.0000000000000754] [Medline: 28598890]
7. Kharrazi H, Weiner JP. A practical comparison between the predictive power of population-based risk stratification models using data from electronic health records versus administrative claims: setting a baseline for future EHR-derived risk stratification models. Med Care 2018 Feb;56(2):202-203. [doi: 10.1097/MLR.0000000000000849] [Medline: 29200132]
8. Ng SH, Rahman N, Ang IYH, Sridharan S, Ramachandran S, Wang DD, et al. Characterization of high healthcare utilizer groups using administrative data from an electronic medical record database. BMC Health Serv Res 2019 Jul 05;19(1):452 [FREE Full text] [doi: 10.1186/s12913-019-4239-2] [Medline: 31277649]
9. Sterling S, Chi F, Weisner C, Grant R, Pruzansky A, Bui S, et al. Association of behavioral health factors and social determinants of health with high and persistently high healthcare costs. Prev Med Rep 2018 Sep;11:154-159 [FREE Full text] [doi: 10.1016/j.pmedr.2018.06.017] [Medline: 30003015]
10. Hwang W, LaClair M, Camacho F, Paz H. Persistent high utilization in a privately insured population. Am J Manag Care 2015 Apr;21(4):309-316 [FREE Full text] [Medline: 26014469]
11. Kim YJ, Park H. Improving prediction of high-cost health care users with medical check-up data. Big Data 2019 Sep;7(3):163-175. [doi: 10.1089/big.2018.0096] [Medline: 31246499]
12. Lee NS, Whitman N, Vakharia N, Taksler GB, Rothberg MB. High-cost patients: hot-spotters don't explain the half of it. J Gen Intern Med 2017 Jan;32(1):28-34 [FREE Full text] [doi: 10.1007/s11606-016-3790-3] [Medline: 27480529]
13. Guilcher SJT, Bronskill SE, Guan J, Wodchis WP. Who are the high-cost users? A method for person-centred attribution of health care spending. PLoS One 2016;11(3):e0149179 [FREE Full text] [doi: 10.1371/journal.pone.0149179] [Medline: 26937955]
14. Wodchis WP, Austin PC, Henry DA. A 3-year study of high-cost users of health care. CMAJ 2016 Feb 16;188(3):182-188 [FREE Full text] [doi: 10.1503/cmaj.150064] [Medline: 26755672]

15. Szymkowiak D, Montgomery AE, Johnson EE, Manning T, O'Toole TP. Persistent super-utilization of acute care services among subgroups of veterans experiencing homelessness. Med Care 2017 Oct;55(10):893-900. [doi: 10.1097/MLR.0000000000000796] [Medline: 28863030]
16. Zayas CE, He Z, Yuan J, Maldonado-Molina M, Hogan W, Modave F, et al. Examining healthcare utilization patterns of elderly middle-aged adults in the United States. Proc Int Fla AI Res Soc Conf 2016 May;2016:361-366 [FREE Full text] [Medline: 27430035]
17. Hu J, Wang F, Sun J, Sorrentino R, Ebadollahi S. A healthcare utilization analysis framework for hot spotting and contextual anomaly detection. AMIA Annu Symp Proc 2012;2012:360-369 [FREE Full text] [Medline: 23304306]
18. Yan S, Seng BJJ, Kwan YH, Tan CS, Quah JHM, Thumboo J, et al. Identifying heterogeneous health profiles of primary care utilizers and their differential healthcare utilization and mortality - a retrospective cohort study. BMC Fam Pract 2019 Apr 23;20(1):54 [FREE Full text] [doi: 10.1186/s12875-019-0939-2] [Medline: 31014231]
19. Ng SCW, Kwan YH, Yan S, Tan CS, Low LL. The heterogeneous health state profiles of high-risk healthcare utilizers and their longitudinal hospital readmission and mortality patterns. BMC Health Serv Res 2019 Dec 04;19(1):931 [FREE Full text] [doi: 10.1186/s12913-019-4769-7] [Medline: 31801537]
20. Hagenaars JA, McCutcheon AL, editors. Applied latent class analysis. Cambridge, UK: Cambridge University Press; 2002.
21. The Johns Hopkins ACGs System, Version 12. Johns Hopkins School of Public Health. 2019. URL: https://www.hopkinsacg.org/ [accessed 2021-08-07]
22. Begg C. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996 Aug 28;276(8):637-639. [doi: 10.1001/jama.1996.03540080059030]
23. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr 1974 Dec;19(6):716-723. [doi: 10.1109/tac.1974.1100705]
24. Schwarz G. Estimating the Dimension of a Model. Ann Statist 1978 Mar 1;6(2):461-464. [doi: 10.1214/aos/1176344136]
25. Linzer DA, Lewis JB. poLCA: an R package for polytomous variable latent class analysis. J Stat Soft 2011;42(10):1-29 [FREE Full text] [doi: 10.18637/jss.v042.i10]
26. Gotcsik M. Otitis media. In: Elzouki AY, Harfi HA, Nazer HM, Stapleton FB, Oh W, Whitley RJ, editors. Textbook of clinical pediatrics. Berlin Heidelberg: Springer-Verlag; 2012:863-871.
27. Perrin JM, Asarnow JR, Stancin T, Melek SP, Fritz GK. Mental health conditions and health care payments for children with chronic medical conditions. Acad Pediatr 2019;19(1):44-50. [doi: 10.1016/j.acap.2018.10.001] [Medline: 30315948]
28. Kharrazi H, Gonzalez CP, Lowe KB, Huerta TR, Ford EW. Forecasting the maturation of electronic health record functions among US hospitals: retrospective analysis and predictive model. J Med Internet Res 2018 Aug 07;20(8):e10458 [FREE Full text] [doi: 10.2196/10458] [Medline: 30087090]
29. Chang H, Richards TM, Shermock KM, Elder Dalpoas S, J Kan H, Alexander GC, et al. Evaluating the impact of prescription fill rates on risk stratification model performance. Med Care 2017 Dec;55(12):1052-1060. [doi: 10.1097/MLR.0000000000000825] [Medline: 29036011]
30. Kan HJ, Kharrazi H, Leff B, Boyd C, Davison A, Chang H, et al. Defining and assessing geriatric risk factors and associated health care utilization among older adults using claims and electronic health records. Med Care 2018 Mar;56(3):233-239. [doi: 10.1097/MLR.0000000000000865] [Medline: 29438193]
31. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The value of unstructured electronic health record data in geriatric syndrome case identification. J Am Geriatr Soc 2018 Aug;66(8):1499-1507. [doi: 10.1111/jgs.15411] [Medline: 29972595]
32. Lemke KW, Gudzune KA, Kharrazi H, Weiner JP. Assessing markers from ambulatory laboratory tests for predicting high-risk patients. Am J Manag Care 2018 Jun 01;24(6):e190-e195 [FREE Full text] [Medline: 29939509]
33. Kharrazi H, Chang H, Heins SE, Weiner JP, Gudzune KA. Assessing the impact of body mass index information on the performance of risk adjustment models in predicting health care costs and utilization. Med Care 2018 Dec;56(12):1042-1050 [FREE Full text] [doi: 10.1097/MLR.0000000000001001] [Medline: 30339574]
34. Chang H, Kan HJ, Shermock KM, Alexander GC, Weiner JP, Kharrazi H. Integrating e-prescribing and pharmacy claims data for predictive modeling: comparing costs and utilization of health plan members who fill their initial medications with those who do not. J Manag Care Spec Pharm 2020 Oct;26(10):1282-1290. [doi: 10.18553/jmcp.2020.26.10.1282] [Medline: 32996394]
35. Kharrazi H, Wang C, Scharfstein D. Prospective EHR-based clinical trials: the challenge of missing data. J Gen Intern Med 2014 Jul;29(7):976-978 [FREE Full text] [doi: 10.1007/s11606-014-2883-0] [Medline: 24839057]
36. Ma X, Jung C, Chang H, Richards TM, Kharrazi H. Assessing the population-level correlation of medication regimen complexity and adherence indices using electronic health records and insurance claims. J Manag Care Spec Pharm 2020 Jul;26(7):860-871. [doi: 10.18553/jmcp.2020.26.7.860] [Medline: 32584680]
37. Kan HJ, Kharrazi H, Chang H, Bodycombe D, Lemke K, Weiner JP. Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. PLoS One 2019;14(3):e0213258 [FREE Full text] [doi: 10.1371/journal.pone.0213258] [Medline: 30840682]

38. Hu Z, Du D. A new analytical framework for missing data imputation and classification with uncertainty: Missing data imputation and heart failure readmission prediction. PLoS One 2020;15(9):e0237724 [FREE Full text] [doi: 10.1371/journal.pone.0237724] [Medline: 32956366]
39. Pandya CJ, Chang H, Kharrazi H. Electronic health record-based risk stratification: a potential key ingredient to achieving value-based care. Popul Health Manag 2021 Jun 14:online ahead of print. [doi: 10.1089/pop.2021.0131] [Medline: 34129398]
40. Chang H, Hatef E, Ma X, Weiner JP, Kharrazi H. Impact of area deprivation index on the performance of claims-based risk-adjustment models in predicting health care costs and utilization. Popul Health Manag 2021 Jun;24(3):403-411. [doi: 10.1089/pop.2020.0135] [Medline: 33434448]
41. Hatef E, Ma X, Rouhizadeh M, Singh G, Weiner JP, Kharrazi H. Assessing the impact of social needs and social determinants of health on health care utilization: using patient- and community-level data. Popul Health Manag 2021 Apr;24(2):222-230. [doi: 10.1089/pop.2020.0043] [Medline: 32598228]

Abbreviations

ACG: Adjusted Clinical Groups
AIC: Akaike information criterion
BIC: Bayesian information criterion
CONSORT: Consolidated Standards of Reporting Trials
ED: emergency department
EDC: expanded diagnostic cluster
EHR: electronic health record
JHHC: Johns Hopkins Health Care
LCA: latent class analysis
OR: odds ratio
PHU: persistent high user/utilizer
PPV: positive predictive value
RxMG: prescription-defined morbidity groups
URI: upper respiratory infection

Edited by C Lovis; submitted 21.06.21; peer-reviewed by N Lee; comments to author 20.07.21; revised version received 26.07.21; accepted 30.09.21; published 25.11.21.

Please cite as:
Ramachandran R, McShea MJ, Howson SN, Burkom HS, Chang HY, Weiner JP, Kharrazi H
Assessing the Value of Unsupervised Clustering in Predicting Persistent High Health Care Utilizers: Retrospective Analysis of Insurance Claims Data
JMIR Med Inform 2021;9(11):e31442
URL: https://medinform.jmir.org/2021/11/e31442
doi:10.2196/31442
PMID:34592712

Original Paper

Incorporating Domain Knowledge Into Language Models by Using Graph Convolutional Networks for Assessing Semantic Textual Similarity: Model Development and Performance Comparison

David Chang[1], DPhil; Eric Lin[2], MD; Cynthia Brandt[1,3,4], MPH, MD; Richard Andrew Taylor[1,3], MD

[1]Yale Center for Medical Informatics, Yale University, New Haven, CT, United States

[2]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, United States

[3]Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT, United States

[4]West Haven Campus, Veterans Affairs Connecticut Healthcare System, West Haven, CT, United States

Corresponding Author:
Richard Andrew Taylor, MD
Yale Center for Medical Informatics
Yale University
Suite 501
300 George St
New Haven, CT
United States
Phone: 1 2037854058
Email: richard.taylor@yale.edu

Abstract

Background: Although electronic health record systems have facilitated clinical documentation in health care, they have also introduced new challenges, such as the proliferation of redundant information through the use of copy and paste commands or templates. One approach to trimming down bloated clinical documentation and improving clinical summarization is to identify highly similar text snippets with the goal of removing such text.

Objective: We developed a natural language processing system for the task of assessing clinical semantic textual similarity. The system assigns scores to pairs of clinical text snippets based on their clinical semantic similarity.

Methods: We leveraged recent advances in natural language processing and graph representation learning to create a model that combines linguistic and domain knowledge information from the MedSTS data set to assess clinical semantic textual similarity. We used bidirectional encoder representation from transformers (BERT)–based models as text encoders for the sentence pairs in the data set and graph convolutional networks (GCNs) as graph encoders for corresponding concept graphs that were constructed based on the sentences. We also explored techniques, including data augmentation, ensembling, and knowledge distillation, to improve the model's performance, as measured by the Pearson correlation coefficient (*r*).

Results: Fine-tuning the BERT_base and ClinicalBERT models on the MedSTS data set provided a strong baseline (Pearson correlation coefficients: 0.842 and 0.848, respectively) compared to those of the previous year's submissions. Our data augmentation techniques yielded moderate gains in performance, and adding a GCN-based graph encoder to incorporate the concept graphs also boosted performance, especially when the node features were initialized with pretrained knowledge graph embeddings of the concepts (*r*=0.868). As expected, ensembling improved performance, and performing multisource ensembling by using different language model variants, conducting knowledge distillation with the multisource ensemble model, and taking a final ensemble of the distilled models further improved the system's performance (Pearson correlation coefficients: 0.875, 0.878, and 0.882, respectively).

Conclusions: This study presents a system for the MedSTS clinical semantic textual similarity benchmark task, which was created by combining BERT-based text encoders and GCN-based graph encoders in order to incorporate domain knowledge into the natural language processing pipeline. We also experimented with other techniques involving data augmentation, pretrained concept embeddings, ensembling, and knowledge distillation to further increase our system's performance. Although the task and its benchmark data set are in the early stages of development, this study, as well as the results of the competition, demonstrates the potential of modern language model–based systems to detect redundant information in clinical notes.

(JMIR Med Inform 2021;9(11):e23101) doi:10.2196/23101

KEYWORDS

natural language processing; graph neural networks; National NLP Clinical Challenges; bidirectional encoder representation from transformers

Introduction

Electronic health records (EHRs) have introduced efficiencies in clinical documentation via the automatic insertion of commonly used documentation phrases and the use of the copy and paste command, which copies the content of one day's notes into that of the next day's notes, but at the same time, these tools have resulted in notes becoming increasingly bloated with sometimes outdated, irrelevant, and even erroneous information [1]. To trim down bloated clinical documentation, one approach of interest is to identify highly similar text snippets for the goal of removing such text. Wang et al [2,3] created the MedSTS data set—a clinical analogue of the natural language understanding benchmark task of assessing semantic textual similarity (STS)—to be a resource for this line of study. In this paper, we show the model, as well as subsequent improvements, that was used in the August 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) Consortium semantic similarity shared task challenge [2], which featured the MedSTS data set.

In the broader natural language processing (NLP) community, STS assessment is a task in which the similarity of semantic meanings and content among natural language texts is calculated [3], and at the time of its release in late 2018, the bidirectional encoder representation from transformers (BERT) language model had the best published performance on the commonly used general English STS Benchmark (STS-B) data set [4]. For the MedSTS data set, it was shown that a BERT model that was fine-tuned to the biomedical domain also outperformed most prior state-of-the-art models [5]. The first iteration of the MedSTS challenge in 2018 (ie, prior to the release of BERT) saw 4 submissions involving the mixed use of traditional machine learning models, like random forests, and more recent deep learning architectures, like recurrent neural networks and convolutional neural networks. The 2019 MedSTS challenge saw over 30 submissions, and the majority of these submissions used BERT in some capacity. The increased number of submissions, as well as the increased average performance of submissions, can be attributed in large part to the recent progress in the development of language models, of which BERT is a popular example.

Despite such advances, researchers have noted that although language models demonstrate a small degree of commonsense reasoning and basic knowledge, such models are very limited in terms of their ability to generate factually correct text or even recall explicit facts from training data [6]. The attempts to mitigate these shortcomings of language models have often involved the use of graph representation learning techniques [7-9], which provide a natural way for working with knowledge in the form of graphs.

Recent progress in graph representation learning has given rise to 2 promising classes of methods that can be used in conjunction with NLP models to incorporate knowledge (either domain knowledge or commonsense knowledge)—graph convolutional networks (GCNs) [10] and knowledge graph embeddings (KGEs) [11].

GCNs generalize the notion of convolution from images to graph-structured data, thereby enabling the application of deep learning techniques on graphs. KGE methods are used to encode entities (nodes) and relationships (edges) in a knowledge graph into dense vector representations, much like word embeddings. KGEs provide a way of obtaining embeddings of concepts, and GCNs provide a natural way of using that information in the context of graph-based learning. For instance, GCNs can be used to initialize node features with pretrained KGEs.

In this study, we leveraged these recent advances in NLP and graph representation learning to develop a more knowledge-aware approach to assessing the MedSTS benchmark data set. We further investigated the benefits of other techniques, such as data augmentation, multisource ensembling, and knowledge distillation, and they resulted in competitive performance values for the 2019 n2c2/OHNLP Consortium semantic similarity shared task challenge.

Methods

Data Set

MedSTS is a data set of sentence pairs that were gathered from the clinical EHRs at Mayo Clinic. Deidentified sentences were selected based on their frequency of appearance and an assumption that frequently appearing sentences tend to contain less protected health information. Sentence pairings were arranged so that they had at least some degree of surface-level similarity. This was based on a combination of surface lexical similarity metrics. Broadly speaking, sentences generally fell into the following four categories: signs and symptoms, disorders, procedures, and medications. Further details are discussed in the original MedSTS paper [3]. For the 2019 n2c2/OHNLP competition and this study, a subset of annotated sentence pairs was examined; of the 2054 sentence pairs in this subset, 1652 (80.4%) were included in the training set, and 412 (20.1%) were included in the test set [2]. This subset was independently scored by 2 medical experts for semantic similarity. A 6-point (range: 0-5) rubric was provided to the annotators; 0 denotes complete dissimilarity, 1 indicates that 2 sentences are topically related but are otherwise not equivalent, and 5 represents complete similarity. The agreement between the two annotators received a weighted Cohen κ score of 0.67. The average of the two scores served as the gold standard against which STS systems would be evaluated [3].

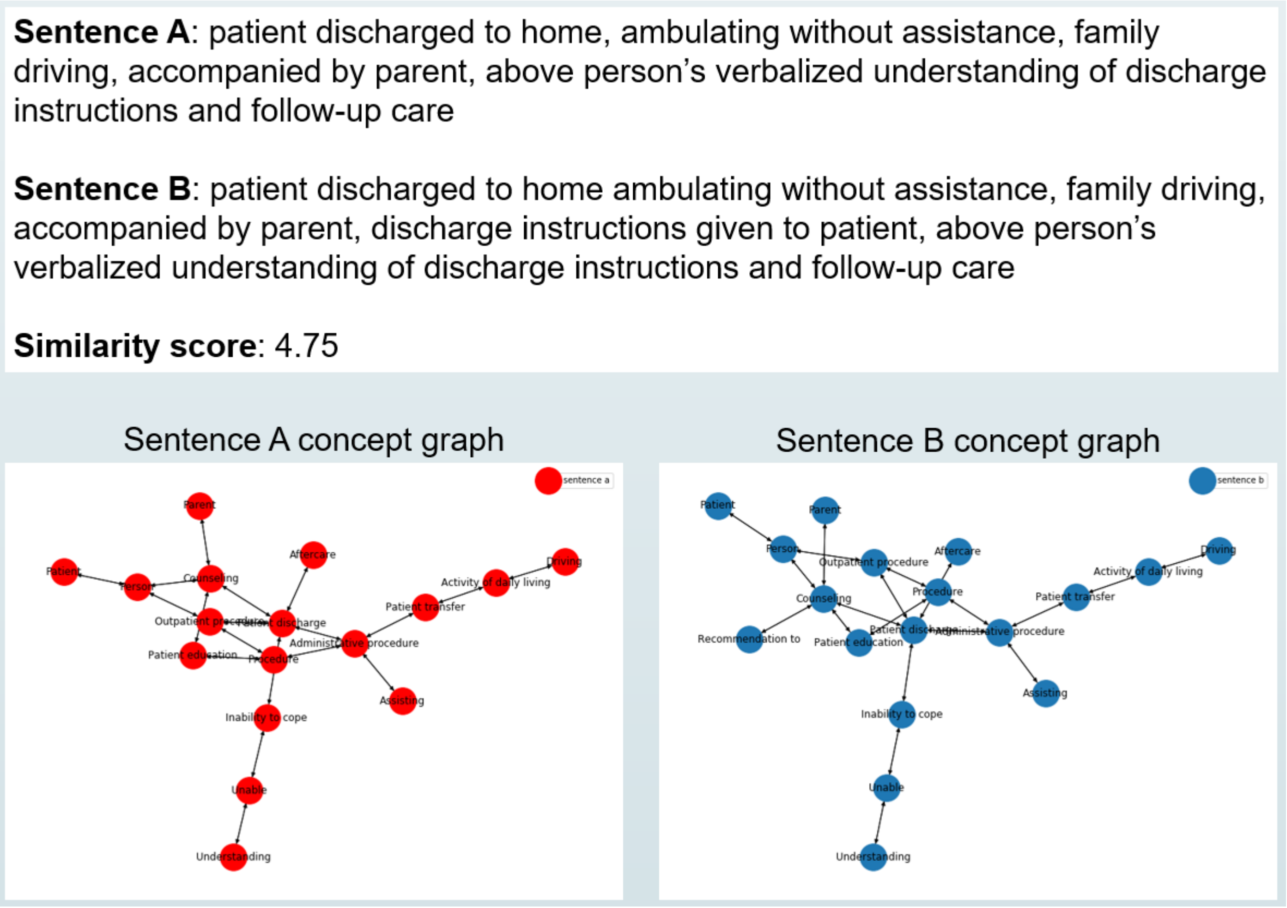
Concept Graph Construction

For each sentence in the MedSTS data set, we constructed a corresponding concept graph to represent the domain knowledge aspect of the data set. The concept graphs consisted of concepts that were tagged with a domain-specific tagger called *MetaMap* [12] and were mapped to a specified medical terminology. The idea was that such a graph would provide an additional representation of data containing explicit domain knowledge in the form of mapped concepts and their connections.

The Unified Medical Language System (UMLS) [13] is an important resource in biomedical and health care research that integrates many health and biomedical vocabularies and terminologies under a unified, interoperable system. MetaMap [12] is a widely used NLP tool that maps concepts in biomedical and clinical text to the UMLS Metathesaurus. We applied MetaMap on the MedSTS data set to extract biomedical and clinical entities that belonged to the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) terminology of the UMLS. Thus, for each sentence, we obtained a corresponding list of extracted concepts, their concept unique identifiers, and semantic type information.

We then constructed a graph of SNOMED CT terminology from the raw UMLS files by using the concepts (MRCONSO.RRF files) as nodes and by using the relationships (MRREL.RRF files) among them as edges. For simplicity, we only considered the connectivity information among the concepts and left the semantic information in the relation types for future work. Once we had a full SNOMED CT graph, we induced subgraphs for each sentence from MedSTS by taking the shortest paths between the concepts that were extracted from the sentences. More concretely, this was done by using the shortest path method via the Dijkstra algorithm in the Networkx [14] library. Although there are many possible ways of constructing such sentence graphs, we decided to use the simple and heuristic shortest path method to obtain a connected graph that represents each sentence. Examples of such concept graphs, along with their original sentences, are shown in (Figure 1).

Figure 1. An example sentence pair, its similarity score, and a visualization of the corresponding concept graphs constructed from the concepts in the sentences.



Sentence A: patient discharged to home, ambulating without assistance, family driving, accompanied by parent, above person's verbalized understanding of discharge instructions and follow-up care

Sentence B: patient discharged to home ambulating without assistance, family driving, accompanied by parent, discharge instructions given to patient, above person's verbalized understanding of discharge instructions and follow-up care

Similarity score: 4.75

Data Augmentation

Given the small size of the data set, we decided to augment it by including additional domain knowledge from the MetaMap output files. Notably, there were 2 pieces of information that we chose to use—the preferred name of the mapped concept in the source terminology and the semantic type of the concept within the UMLS Semantic Network. The preferred name of a mapped concept can often be the same as how the concept appears in the text, but the preferred name sometimes provides potentially valuable information in the form of synonyms or abbreviation expansions. For example, in the text snippet "the patient was taken to the pacu in stable condition," the term *pacu* is mapped to the UMLS concept *postoperative anesthesia care unit (PACU)*, thereby providing the full description of the abbreviated term. The strings of the preferred names of mapped concepts were simply appended to the original sentences in the data set. Likewise, the semantic types of the mapped concepts (eg, *Health Care Related Organization* for the term *pacu*) were appended to the original sentences. Another method we used was doubling the data set size by simply feeding the model a copy of the data set that included sentences formatted in the reverse order (ie, “sentence2:sentence1”). This yielded slightly better results than those that were obtained by simply doubling the number of training epochs, suggesting that feeding the model the reverse copy of the data set might have given it more explicit hints that the task was agnostic to the order of the sentences. Although the data augmentation techniques we used were simple and yielded moderate improvements in performance, a recent paper [15] provides more interesting approaches to data augmentation. In the paper [15], the authors used back-translation and performed segment reordering to augment the MedSTS data set.

The BERT Model

The BERT model is a widely used NLP model that is part of the recently emerging class of language models that use transformers [16] as the building blocks. The BERT model stacks multiple layers of transformer-based modules that primarily use the multiheaded self-attention mechanism to encode text into dense embeddings. The model is trained by using the masked language modeling objective and the next sentence prediction objective, and pretrained models for BERT (and other similar models) are readily available on the HuggingFace Transformers library [17]. Shortly after the BERT model dominated the general NLP field, several variations of the BERT model that were adapted to the biomedical and clinical domains also became available [5,18,19]. These domain-adapted versions of the BERT model were trained on some combination of the Medical Information Mart for Intensive Care version 3 [20], PubMed [21], and PubMed Central [22] databases, and these versions have been shown to outperform the original BERT model in several clinical NLP tasks, suggesting that they are more appropriate for working with clinical text data sets like MedSTS.

The GCN Method

Kipf et al [10] contributed to the popularization of graph neural networks by providing an efficient implementation method for GCNs and demonstrating their effectiveness in analyzing several benchmark graph data sets for graph classification, node classification, and link prediction. Variants of GCNs were soon applied successfully to various domains and problems, including the modeling of interactions in physical systems [23], drug-drug interactions [24], and text classification [25]. GCNs have become a popular deep learning model for working with graph-structured data, and we used GCNs to encode the concept graphs.

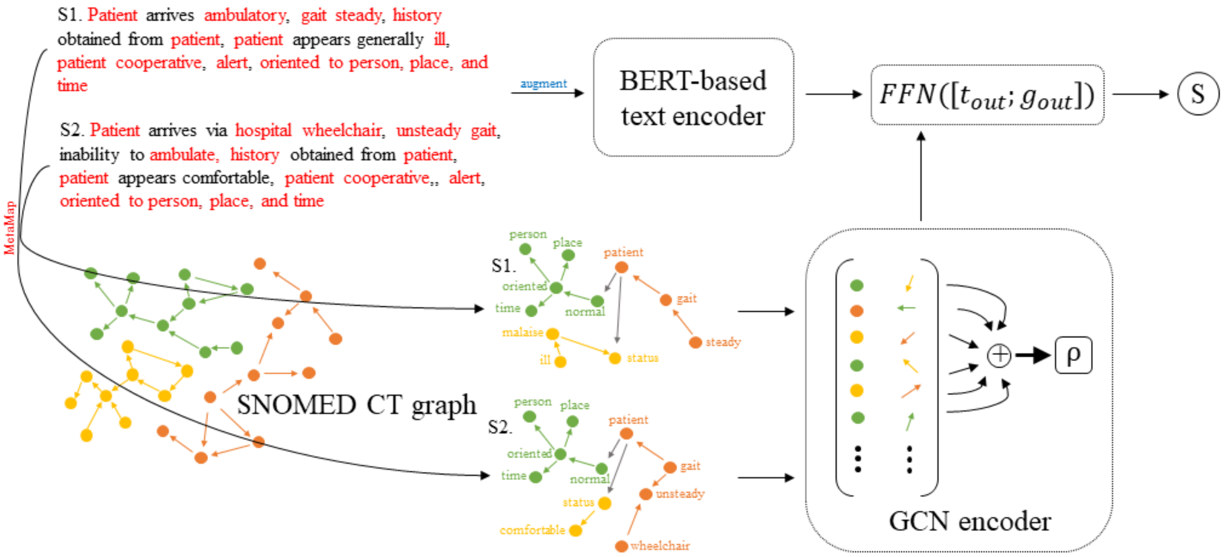
KGE Methods

KGEs are a relatively novel class of methods for learning dense vector representations of entities and relations in multi-relational, heterogeneous knowledge graphs. Essentially, a KGE model maps entities and relations to embedding spaces by using a predefined scoring function. Due to their growing popularity and the availability of implementation methods, KGEs have recently been applied to various domains, including biomedical knowledge graphs [26]. Chang et al [26] showed that using KGEs for learning concept embeddings from medical terminologies and knowledge graphs is arguably a more principled and effective approach than using previous methods based on skip-gram–based models like Cui2Vec [27] or network embedding–based models like Snomed2Vec [28]. Although we initially used Cui2Vec for our entity vectors at the time of submission, we later used SNOMED CT KGEs after they became available in recent months.

Augmenting BERT With KGEs for MedSTS

We combined the components of GCNs and KGEs into a single model in the following way: we used a BERT-based model as our text encoder for the sentence pairs in MedSTS, used a GCN-based model as our graph encoder for the concept graphs that corresponded to the sentence pairs, initialized the node embeddings in the graphs by using pretrained SNOMED CT KGEs, concatenated the outputs of the text and graph encoders, and passed the final concatenated vector to a fully connected layer to obtain a semantic similarity score. We also tested the benefits of using the SNOMED CT KGEs by comparing this method to random initialization and initialization with Cui2Vec embeddings. A visualization of the pipeline is shown in Figure 2.

Figure 2. A simplified diagram of our pipeline. We passed the sentences through MetaMap to extract concepts belonging to the SNOMED CT and induced concept graphs by using the relationships among the terminology. We then passed the augmented sentence pairs to the text encoder and passed the concept graphs to the graph encoder. The outputs from the encoders were concatenated and passed to a fully connected layer to obtain an S. BERT: bidirectional encoder representation from transformers; FFN: feed-forward network; GCN: graph convolutional network; S: similarity score; S1: sentence 1; S2: sentence 2; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.


S1. Patient arrives ambulatory, gait steady, history obtained from patient, patient appears generally ill, patient cooperative, alert, oriented to person, place, and time
S2. Patient arrives via hospital wheelchair, unsteady gait, inability to ambulate, history obtained from patient, patient appears comfortable, patient cooperative,, alert, oriented to person, place, and time
augment
MetaMap
BERT-based text encoder
$FFN([t_{out}; g_{out}])$
S
SNOMED CT graph
S1.
person
place
patient
oriented
time
normal
gait
malaise
steady
ill
status
S2.
person
place
patient
oriented
time
normal
gait
status
unsteady
comfortable
wheelchair
ρ
GCN encoder

Ensemble and Knowledge Distillation

After training our model, we took an ensemble to further improve the model's performance. In accordance with Xu et al [29], we performed multisource ensembling with the following variants of BERT: BERT_base [4], SciBERT [30], ClinicalBERT [18], multi-task deep neural networks (MT-DNNs) [31], and BlueBERT [32]. Afterward, we performed knowledge distillation—an effective model compression method in which a smaller model is trained to mimic a larger model (ie, the ensemble). We used the predictions of the multisource ensemble model as soft labels in a teacher bounded regression loss function, in accordance with Chen et al [33], to train more individual models and obtain a final ensemble of the knowledge-distilled models.

Results

We split the provided training set of MedSTS into 1313 training examples and 329 validation examples and reported the Pearson correlation coefficient for the held-out test set of 412 examples. The Pearson correlation coefficient was the chosen metric for the competition. We used the HuggingFace Transformers library for implementations related to language models, and we used PyTorch Geometric [34] for implementations of GCNs. Many of the default training and fine-tuning hyperparameters were used, while the following hyperparameters were tuned on the validation set: a learning rate of $1e^{-4}$ for BERT-based models (chosen from $5e^{-5}$, $1e^{-4}$, and $5e^{-4}$), a learning rate of $1e^{-3}$ for GCNs (chosen from $1e^{-2}$, $1e^{-3}$, and $1e^{-4}$), and 4 epochs (chosen from 3, 4, and 5 epochs).

Table 1 shows the contributions of the different components in the pipeline. Simply using the off-the-shelf BERT_base model and fine-tuning it on MedSTS yielded higher performance values compared to those of the 2018 submissions. Using ClinicalBERT and using our previously described data augmentation technique each yielded moderate gains.

Table 1. The results for the base model and each model version (an additional component was added to each system). The columns under Pearson correlation analysis show the scores for the test set (all) and the four subsets of the test set, which included sentences regarding patients' conditions or statuses (status), patients' education or interactions (education), patients' medications (meds), and miscellaneous topics (miscellaneous).

Model name	Pearson correlation analysis				
	All, *r*	Status, *r*	Education, *r*	Meds, *r*	Miscellaneous, *r*
BERT_base	0.842	0.643	0.721	0.522	0.414
ClinicalBERT	0.848	0.662	0.735	0.541	0.425
ClinicalBERT-DA[a]	0.855	0.671	0.737	0.553	0.432
ClinicalBERT-DA + GCN_rand	0.861	0.675	0.742	0.532	0.427
ClinicalBERT-DA + GCN_cui2vec	0.863	0.682	0.753	0.536	0.442
ClinicalBERT-DA + GCN_snomedkge	0.868	0.693	0.761	0.562	0.463

[a]The ClinicalBERT-DA model refers to the ClinicalBERT model after data augmentation.

Adding a graph encoder, in addition to our other modifications, to incorporate the concept graphs resulted in minor improvements when the node embeddings were either initialized randomly or initialized with pretrained Cui2Vec embeddings. However, using SNOMED CT KGEs as the node features in the GCN resulted in an increase in performance, that is, an increase of 1.3% above the performance of ClinicalBERT (ie, after data augmentation), suggesting that SNOMED CT KGEs served as better starting representations of the concepts. It is worth noting that since the BERT-based text encoder is initialized with a pretrained checkpoint, it might be especially important to initialize the graph encoder with decent pretrained embeddings to allow the graph encoder to "catch up" with the text encoder. We called this best performing setting *ClinicalBERT_all*.

We also manually categorized the sentence pairs into the following four categories: sentences related to patients' conditions and statuses (status), education or interactions with patients (education), medications (meds), and miscellaneous or clearly dissimilar topics (miscellaneous). The columns in Table 1 (those under *Pearson correlation analysis*) show the scores for the test set (all) and for the four categories described. Sentence pairs in the status and education categories received relatively higher scores, as expected, since many of the sentences and text snippets in these categories often repeated. Specifically, text snippets beginning with "patient arrives...," "discussed the risks...," or "identified illness as a learning need..." recurred noticeably in these two categories. Further, the medication and miscellaneous categories received relatively low correlation scores. For the miscellaneous category, this was expected, since many of the sentence pairs in this category were more difficult for the model to learn due to their greater variability. For the medication category, the gold-standard scores assigned by the annotators proved to be rather inconsistent and challenging to predict, even upon manual review by a medical expert.

Table 2 shows the results for ensembling and knowledge distillation. First, we took the ensemble of 10 ClinicalBERT_all models with slightly varied hyperparameters and saw a moderate increase in performance, as expected of ensembles. Second, in accordance with Xu et al [29], we took an ensemble of 10 models consisting of a variety of model types (BERT_base, SciBERT, ClinicalBERT, MT-DNNs, and BlueBERT), along with the graph encoder, based on their validation performance and saw a slight improvement. Finally, by using a teacher bounded regression loss function [33], we used the outputs of the multisource ensemble model as soft labels to train more best-setting models of different types and took an ensemble consisting of 10 such knowledge-distilled models for slight performance gain.

Table 2. Results for the ensembling of the best performing models from (ClinicalBERT_all), the ensembling of multiple language models (LMs; each with a graph convolutional network), and the ensembling of knowledge-distilled (KD) multisource ensembles.

Ensemble type	Performance, %
Ensemble of ClinicalBERT_all	87.5
Ensemble with multiple LMs	87.8
Ensemble of KD models	88.2
IBM-N2C2[a]	90.1

[a]The best performing model from the IBM team at the time of the competition was included for reference.

Discussion

Main Findings

We implemented a list of techniques in our pipeline for the clinical MedSTS benchmark task and reported slight to moderate improvements in performance for each technique. Using a pretrained, off-the-shelf, BERT-based model and fine-tuning it alone served as a strong baseline that outperformed all pre-BERT systems in the task. We found that our data augmentation technique helped slightly, but again, Wang et al

[15] has provided more interesting and effective data augmentation approaches for MedSTS.

Adding a graph encoder to incorporate concept graphs into the pipeline yielded decent gains, especially when the graph encoder was initialized by using pretrained SNOMED CT KGEs. We stress that since the graph encoder was trained jointly with a pretrained text encoder, it is important to consider providing the graph encoder with pretrained embeddings as well, so that it does not fall too far behind in training.

As expected, ensembling leads to improved performance. Further improvements can be achieved by using language models from different sources as well as by performing knowledge distillation, which can be followed by the ensembling of the distilled models.

We also attempted to use several other techniques that did not yield any performance gains. First, we tried multi-task learning by using different general and clinical domain NLP data sets, including the Medical Natural Language Inference [35], Recognizing Question Entailment [36], and English STS-B [37] data sets, following an implementation of multi-task learning for MT-DNNs, but this approach did not result in any improvements and substantially increased the training time. Second, we tried manually annotating the MedSTS data for different sentence categories (medication, status, education, and miscellaneous). This was done as an auxiliary classification task (also an example of multi-task learning), but this did not lead to noticeable gains in performance. Lastly, we tried experimenting with different variants of GCNs, but we found that training multiple types of graph neural networks jointly with a large language model was difficult in terms of hyperparameter tuning and decided to limit our analysis to basic GCNs.

Limitations of the Method

Although the results show that the strategies for data augmentation and the incorporation of domain knowledge through concept embeddings and GCNs do confer some benefit, we address some of the limitations in this section.

The data augmentation techniques we used involved including additional textual and semantic information from the MetaMap output and reversing the sentence order to double the data set size. There are many other potential data augmentation techniques in the general NLP field that could be useful. Notably, Wang et al [15] recently performed segment reordering and back-translation to substantially improve their model's performance on a task.

As for the pretrained concept embeddings and GCNs, combining them with a large pretrained language model is still largely experimental. This can be improved by using recent developments in the field of graph representation learning, such as graph attention networks [38] and graph matching networks [39].

Limitations of the Data Set

Both the positive and negative findings should be considered with caution due to the abundance of the potential ways of implementing each component as well as the size and quality of the data set, which was relatively smaller and of lower quality compared to data sets in mainstream, nonclinical NLP domains that have less complicated access to labeled data.

After working closely with the data set for several months, we noticed that certain sentence pairs had large irregularities in terms of their scores from the two annotators of the data set. This was the most notable in the sentence pairs that discussed medications; often, these sentence pairs described the prescribing of medications to patients and differed in terms of dosing or drug class. At one level of categorization, the similarity of a sentence pair related to prescribing could be seen as high, regardless of the medication class or dosing. At another level of categorization, it appeared that several such pairs were noted to be of low similarity when the medications or dosing regimens differed. This discrepancy in scoring also seemed to differ depending on the drug classes being mentioned. Without knowing which annotator was responsible for a given score, it is difficult to speak conclusively, but we speculate that certain drug classes were of greater salience to each annotator. As an example, someone with a specialty in a mental health may subjectively perceive 2 different psychiatric medications of different classes to be quite different but view cardiology drugs to be subjectively more similar. In contrast, an individual in the field of cardiology may perceive various cardiology drugs as being different but may perceive drugs in the psychiatric medications category overall as being more similar. Such differences in perspectives may also be influenced by aspects of an annotator's practice, such as whether their practice occurs in inpatient settings, outpatient settings, the operating room, or the medical clinic.

Many of the scoring irregularities may have been related to the nature of the task of rating subjective similarity. One approach to mitigating annotator bias, as discussed in the original MedSTS paper [3], is to increase the number of annotators and set the average score as the gold standard. For example, in the English STS-B, 5 annotators were used for each sentence, and annotators were limited to a certain number of sentence pairs that they could annotate [37]. Although such an approach can be prohibitively expensive due to the need to hire enough medical annotators and be very cumbersome to implement for clinical text due to patient privacy protections, another approach for the case of having few annotators could be to reveal potentially biasing factors toward annotation, such as clinical background, or to assign an annotator ID to each score. Stating the biases or allowing teams to model the annotator biases may help with understanding scoring irregularities that may be difficult to resolve without the use of specifically tailored algorithm designs or features, which require specific domain knowledge to adapt to unique annotator biases.

Despite our concerns with the fundamental difficulty of objectively rating subjective semantic similarity, the high Pearson correlation coefficient achieved by our model suggests that the task is still largely tractable. MedSTS also remains one of the few, if not only, publicly available data sets for studying clinical STS in EHRs. We hope that our suggestions may introduce additional strategies for modeling the variance from subjective elements and provide some insights to future data set annotation processes for this important yet challenging problem.

Conclusions

As participants of the 2019 n2c2/OHNLP shared task challenge, we developed a system for the clinical MedSTS benchmark task by combining BERT-based text encoders and GCN-based graph encoders in order to incorporate domain knowledge into the NLP pipeline. We also experimented with other techniques involving data augmentation, pretrained concept embeddings, ensembling, and knowledge distillation to further increase our model's performance. Although our results lagged behind those of the top scoring model at the n2c2 workshop, the incorporation of domain knowledge into deep learning NLP models via graph-based methods was a new advance in clinical NLP. We highlight our concerns about the impact of specific difficulties with subjective semantic similarities in data set annotation, but overall, we believe that clinical semantic similarity remains an important topic of study, and continued work on the MedSTS benchmark—one of the few clinical STS data sets available—will yield advances in processing valuable unstructured data in EHRs. The MedSTS data set should continue to be improved and enlarged through the further careful annotation of the original pool of sentence pairs, and future work should explore novel methods that can effectively leverage both linguistic and domain knowledge.

Acknowledgments

This project was supported by the National Institutes of Health Training Grant (grant 5T15LM007056-33), the National Institute of Mental Health Training Grant (grant 2R25MH071584-12), and the Veterans Affairs Boston Medical Informatics Fellowship.

Conflicts of Interest

None declared.

References

1. Hirschtick RE. A piece of my mind. Copy-and-paste. JAMA 2006 May 24;295(20):2335-2336. [doi: 10.1001/jama.295.20.2335] [Medline: 16720812]
2. Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H. The 2019 n2c2/OHNLP track on clinical semantic textual similarity: overview. JMIR Med Inform 2020 Nov 27;8(11):e23375 [FREE Full text] [doi: 10.2196/23375] [Medline: 33245291]
3. Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, et al. MedSTS: a resource for clinical semantic textual similarity. Lang Resour Eval 2018 Oct 24;54:57-72. [doi: 10.1007/s10579-018-9431-1]
4. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online on May 24, 2019. [FREE Full text]
5. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. 2019 Aug Presented at: The 18th BioNLP Workshop and Shared Task; August 1, 2019; Florence, Italy p. 58-65 URL: https://aclanthology.org/W19-5006.pdf [doi: 10.18653/v1/w19-5006]
6. Logan IV RL, Liu NF, Peters ME, Gardner M, Singh S. Barack's wife Hillary: Using knowledge-graphs for fact-aware language modeling. 2019 Jul Presented at: The 57th Annual Meeting of the Association for Computational Linguistics; July 28 to August 2, 2019; Florence, Italy p. 5962-5971 URL: https://aclanthology.org/P19-1598.pdf [doi: 10.18653/v1/p19-1598]
7. Peters ME, Neumann M, Logan IV RL, Schwartz R, Joshi V, Singh S, et al. Knowledge enhanced contextual word representations. 2019 Nov Presented at: The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 3-7, 2019; Hong Kong, China p. 43-54 URL: https://aclanthology.org/D19-1005.pdf [doi: 10.18653/v1/d19-1005]
8. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: Enhanced language representation with informative entities. 2019 Jul Presented at: The 57th Annual Meeting of the Association for Computational Linguistics; July 28 to August 2, 2019; Florence, Italy p. 1441-1451 URL: https://aclanthology.org/P19-1139.pdf [doi: 10.18653/v1/p19-1139]
9. Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, Deng H, et al. K-BERT: Enabling language representation with knowledge graph. In: Proc Conf AAAI Artif Intell. 2020 Feb Presented at: The Thirty-Fourth AAAI Conference on Artificial Intelligence; February 7-12, 2020; New York, New York, USA p. 2901-2908. [doi: 10.1609/aaai.v34i03.5681]
10. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv Preprint posted online on February 22, 2017. [FREE Full text]
11. Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: A survey. Knowl Based Syst 2018 Jul 01;151:78-94. [doi: 10.1016/j.knosys.2018.03.022]
12. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21 [FREE Full text] [Medline: 11825149]
13. 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] [Medline: 14681409]

14. Hagberg A, Swart P, Schult D. Exploring network structure, dynamics, and function using NetworkX. 2008 Presented at: The 7th Python in Science Conference (SciPy2008); August 21, 2008; Pasadena, California URL: https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-08-05495 [doi: 10.2172/425288]
15. Wang Y, Liu F, Verspoor K, Baldwin T. Evaluating the utility of model configurations and data augmentation on clinical semantic textual similarity. 2020 Jul Presented at: The 19th SIGBioMed Workshop on Biomedical Language Processing; July 9, 2020; Online URL: https://aclanthology.org/2020.bionlp-1.11.pdf [doi: 10.18653/v1/2020.bionlp-1.11]
16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017 Dec Presented at: The 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, California, USA p. 6000-6010.
17. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-art natural language processing. 2020 Oct Presented at: The 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; November 16-20, 2020; Online p. 38-45 URL: https://aclanthology.org/2020.emnlp-demos.6.pdf [doi: 10.18653/v1/2020.emnlp-demos.6]
18. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. 2019 Jun Presented at: The 2nd Clinical Natural Language Processing Workshop; June 7, 2019; Minneapolis, Minnesota p. 72-78 URL: https://aclanthology.org/W19-1909.pdf [doi: 10.18653/v1/w19-1909]
19. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
20. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, 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] [Medline: 27219127]
21. PubMed. National Library of Medicine. URL: https://pubmed.ncbi.nlm.nih.gov/ [accessed 2021-11-09]
22. Home - PMC - NCBI. National Center for Biotechnology Information. URL: https://www.ncbi.nlm.nih.gov/pmc/ [accessed 2021-11-09]
23. Kipf T, Fetaya E, Wang KC, Welling M, Zemel R. Neural relational inference for interacting systems. arXiv Preprint posted online on June 6, 2018. [FREE Full text]
24. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 2018 Jul 01;34(13):i457-i466 [FREE Full text] [doi: 10.1093/bioinformatics/bty294] [Medline: 29949996]
25. Yao L, Mao C, Luo Y. Graph convolutional networks for text classification. In: Proc Conf AAAI Artif Intell. 2019 Presented at: The Thirty-Third AAAI Conference on Artificial Intelligence; January 27 to February 1, 2019; Honolulu, Hawaii, USA p. 7370-7377. [doi: 10.1609/aaai.v33i01.33017370]
26. Chang D, Balazevic I, Allen C, Chawla D, Brandt C, Taylor RA. Benchmark and best practices for biomedical knowledge graph embeddings. 2020 Jul Presented at: The 19th SIGBioMed Workshop on Biomedical Language Processing; July 9, 2020; Online p. 167-176 URL: https://aclanthology.org/2020.bionlp-1.18.pdf [doi: 10.18653/v1/2020.bionlp-1.18]
27. Beam AL, Kompa B, Schmaltz A, Fried I, Weber G, Palmer N, et al. Clinical concept embeddings learned from massive sources of multimodal medical data. arXiv Preprint posted online on August 20, 2019. [FREE Full text] [doi: 10.1142/9789811215636_0027]
28. Agarwal K, Eftimov T, Addanki R, Choudhury S, Tamang S, Rallo R. Snomed2Vec: Random walk and Poincaré embeddings of a clinical knowledge base for healthcare analytics. arXiv Preprint posted online on July 19, 2019. [FREE Full text]
29. Xu Y, Liu X, Li C, Poon H, Gao J. DoubleTransfer at MEDIQA 2019: Multi-source transfer learning for natural language understanding in the medical domain. 2019 Aug Presented at: The 18th BioNLP Workshop and Shared Task; August 1, 2019; Florence, Italy p. 399-405 URL: https://aclanthology.org/W19-5042.pdf [doi: 10.18653/v1/w19-5042]
30. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. 2019 Nov Presented at: The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 3-7, 2019; Hong Kong, China p. 3615-3620 URL: https://aclanthology.org/D19-1371.pdf [doi: 10.18653/v1/d19-1371]
31. Liu X, He P, Chen W, Gao J. Multi-task deep neural networks for natural language understanding. 2019 Jul Presented at: The 57th Annual Meeting of the Association for Computational Linguistics; July 28 to August 2, 2019; Florence, Italy p. 4487-4496 URL: https://aclanthology.org/P19-1441.pdf [doi: 10.18653/v1/p19-1441]
32. Peng Y, Chen Q, Lu Z. An empirical study of multi-task learning on BERT for biomedical text mining. 2020 Jul Presented at: The 19th SIGBioMed Workshop on Biomedical Language Processing; July 9, 2020; Online p. 205-214 URL: https://aclanthology.org/2020.bionlp-1.22.pdf [doi: 10.18653/v1/2020.bionlp-1.22]
33. Chen G, Choi W, Yu X, Han T, Chandraker M. Learning efficient object detection models with knowledge distillation. 2017 Dec Presented at: The 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, California, USA p. 742-751.
34. Fey M, Lenssen JE. Fast graph representation learning with PyTorch Geometric. arXiv Preprint posted online on April 25, 2019. [FREE Full text]

35. Romanov A, Shivade C. Lessons from natural language inference in the clinical domain. 2018 Presented at: The 2018 Conference on Empirical Methods in Natural Language Processing; October 31 to November 4, 2018; Brussels, Belgium p. 1586-1596 URL: https://aclanthology.org/D18-1187.pdf [doi: 10.18653/v1/d18-1187]
36. Abacha AB, Shivade C, Demner-Fushman D. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. 2019 Aug Presented at: The 18th BioNLP Workshop and Shared Task; August 1, 2019; Florence, Italy URL: https://aclanthology.org/W19-5039.pdf [doi: 10.18653/v1/w19-5039]
37. Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. 2017 Aug Presented at: The 11th International Workshop on Semantic Evaluation (SemEval-2017); August 3-4, 2017; Vancouver, Canada p. 1-14 URL: https://aclanthology.org/S17-2001.pdf [doi: 10.18653/v1/s17-2001]
38. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv Preprint posted online on February 4, 2018. [FREE Full text]
39. Li Y, Gu C, Dullien T, Vinyals O, Kohli P. Graph matching networks for learning the similarity of graph structured objects. 2019 Jun Presented at: The 36th International Conference on Machine Learning; June 10-15, 2019; Long Beach, California URL: http://proceedings.mlr.press/v97/li19d/li19d.pdf [doi: 10.1093/oso/9780198788348.003.0005]

Abbreviations

BERT: bidirectional encoder representation from transformers
EHR: electronic health records
GCN: graph convolutional network
KGE: knowledge graph embedding
MT-DNN: multi-task deep neural network
n2c2: National NLP Clinical Challenges
NLP: natural language processing
OHNLP: Open Health Natural Language Processing Consortium
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
STS: semantic textual similarity
STS-B: Semantic Textual Similarity Benchmark
UMLS: Unified Medical Language System

Edited by Y Wang; submitted 31.07.20; peer-reviewed by A Trifan, B Dandala, M Torii; comments to author 22.10.20; revised version received 15.12.20; accepted 14.01.21; published 26.11.21.

Please cite as:
Chang D, Lin E, Brandt C, Taylor RA
Incorporating Domain Knowledge Into Language Models by Using Graph Convolutional Networks for Assessing Semantic Textual Similarity: Model Development and Performance Comparison
JMIR Med Inform 2021;9(11):e23101
URL: https://medinform.jmir.org/2021/11/e23101
doi:10.2196/23101
PMID:34842531

Original Paper

A Pipeline to Understand Emerging Illness Via Social Media Data Analysis: Case Study on Breast Implant Illness

Vishal Dey[1], BSc; Peter Krasniak[2], MD; Minh Nguyen[2], MD; Clara Lee[2], MD; Xia Ning[1,2,3], PhD

[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, United States

[2]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, United States

[3]Translational Data Analytics Institute, The Ohio State University, Columbus, OH, United States

Corresponding Author:
Xia Ning, PhD
Department of Biomedical Informatics
The Ohio State University
1800 Cannon Drive
Columbus, OH, 43210
United States
Phone: 1 6143662287
Email: ning.104@osu.edu

Abstract

Background: A new illness can come to public attention through social media before it is medically defined, formally documented, or systematically studied. One example is a condition known as breast implant illness (BII), which has been extensively discussed on social media, although it is vaguely defined in the medical literature.

Objective: The objective of this study is to construct a data analysis pipeline to understand emerging illnesses using social media data and to apply the pipeline to understand the key attributes of BII.

Methods: We constructed a pipeline of social media data analysis using natural language processing and topic modeling. Mentions related to signs, symptoms, diseases, disorders, and medical procedures were extracted from social media data using the clinical Text Analysis and Knowledge Extraction System. We mapped the mentions to standard medical concepts and then summarized these mapped concepts as topics using latent Dirichlet allocation. Finally, we applied this pipeline to understand BII from several BII-dedicated social media sites.

Results: Our pipeline identified topics related to toxicity, cancer, and mental health issues that were highly associated with BII. Our pipeline also showed that cancers, autoimmune disorders, and mental health problems were emerging concerns associated with breast implants, based on social media discussions. Furthermore, the pipeline identified mentions such as rupture, infection, pain, and fatigue as common self-reported issues among the public, as well as concerns about toxicity from silicone implants.

Conclusions: Our study could inspire future studies on the suggested symptoms and factors of BII. Our study provides the first analysis and derived knowledge of BII from social media using natural language processing techniques and demonstrates the potential of using social media information to better understand similar emerging illnesses.

(JMIR Med Inform 2021;9(11):e29768) doi:10.2196/29768

KEYWORDS

breast implant illness; social media; natural language processing; topic modeling

Introduction

Background

The ubiquity of social media has resulted in early descriptions of new and evolving diseases on social media platforms before they can be systematically studied [1-7], particularly during the era of the medical internet [8-14]. Social media users increasingly turn to platforms such as Twitter (Twitter Inc), Facebook (Facebook Inc), and YouTube (Google LLC) to share personal experiences, including diseases and illnesses they have experienced, or to seek support and resources, such as health and medical resources. Recent studies have shown the potential of social media in the detection of mental illness and depression [15-17] and in the early detection of food-borne illnesses [18-20] and other infectious diseases [2,21-24]. Furthermore, several studies have demonstrated social media as an effective tool to disseminate information regarding symptoms, personal well-being, and public health resources during multiple influenza outbreaks [25-28]. During the early stages of COVID-19, studies [4,29,30] analyzed posts on Sina Weibo (Weibo Corporation)—a major Chinese microblogging site—to characterize patient symptoms and public concerns in multiple provinces of China. From the analysis of Weibo (Weibo Corporation) posts, Huang et al [30] concluded that most of the affected patients were older persons, with fever as the most common symptom. These studies demonstrate that public social media data can be leveraged to better understand emerging illnesses and to accommodate prompt responses.

One new illness we studied in this manuscript was breast implant illness (BII). Breast implants have gained popularity over the last 20 years [31]. During this period, more than 400,000 women have undergone breast augmentation or postmastectomy surgeries every year in the United States [32]. There was a 4% increase in the number of breast augmentation procedures between 2017 and 2018, and a 6% increase in breast implant removal procedures occurred over the same period [32]. Concerns about the safety of breast implants have also arisen [33-38] and persisted [39-45]. However, although a causal link between breast implants and systemic diseases has not been definitively shown, a phenomenon called *breast implant illness*, which attributes systemic symptoms to breast implants, has emerged [46]. Unlike other new medical illnesses, however, BII has been reported minimally in the medical literature, being primarily limited to social media [11,47-50]. For example, a recent analysis [49] demonstrated increasing public interest in BII based on Twitter and Google Trends data from February 2018 to February 2019. To summarize the key symptoms, diseases, and disorders defining BII, several cohort studies [51,52] have analyzed patient-reported outcomes before and after breast explant surgeries. These studies showed some potential relationships between explant surgeries and the improvement of specific symptoms in the patient population. Unfortunately, these studies were not definitive because of their limited study design secondary to their lack of control groups, data collection bias, and lack of randomization. The lack of medical knowledge about BII makes it difficult to define the condition, and therefore, it is nearly impossible to conduct rigorous epidemiological or clinical studies. BII is just one disease process for which the lack of medical knowledge is apparent, but there are many other new illnesses for which this is the case. Any initial knowledge that is supported by sufficient social media data would be meaningful as a reference for formal studies in the future, and thus, the techniques to discover such knowledge are highly required.

Objectives

To identify and summarize the key attributes of a new illness, in this study, we constructed a data analysis pipeline for the social media data analysis of BII. The pipeline incorporated natural language processing (NLP) and topic modeling methods. Our primary objective is to derive novel knowledge about BII, a medical condition that has not yet been systematically studied and defined in the medical literature, by constructing a data analysis pipeline and applying the pipeline to social media data. As medical knowledge and literature on BII have not been established and the related concepts are not well defined or accepted, using social media data to understand emerging issues could be a meaningful starting point. We applied this pipeline to better understand the symptoms and signs associated with BII. To the best of our knowledge, this study is the first to use social media data to derive the knowledge of BII from social media. This demonstrates the potential of using social media information to better understand the conditions that have primarily been reported on social media. It also establishes the effectiveness of our pipeline and its potential application to understand other new illnesses. In the following discussion, we have described our analysis pipeline in the context of BII. However, our pipeline is not specific to BII and is applicable to other illnesses as well.

Methods

Data

We collected and used data from select social media websites. These websites were selected because they were dedicated to BII discussions and information and were focused on user groups with interest in BII. Often, dedicated social media websites (eg, forums and Twitter pages) are available for a particular illness or disease. For example, some dedicated websites [53-55] contain the stories and experiences of patients fighting different cancers, some [56,57] contain posts and stories of users experiencing chronic pain and illness, and others [58-60] contain stories and experiences from COVID-19 survivors. The social media sources used in our study were as follows:

- BII [61]: This was a dedicated public website with articles on BII-related topics and offered resources related to implant and explant procedures, etc. This website also allowed individuals to post their experiences and concerns about breast implants and related health issues. We extracted individual posts from the website (up to May 10, 2019), and the resulting data set was referred to as BIIweb.
- Healing BII [62]: This website contained information on postimplant disorders, postexplant healing, breast implant safety, etc. The discussion board of this website had multiple posts and comments on symptoms, signs, etc, which are experienced by individuals with a breast implant or by those who have undergone an explant. The data set extracted from the discussion board of this website (up to May 10, 2019) was referred to as HealingBII.
- Instagram posts about BII [63]: This website contained a collection of publicly available Instagram posts that used *breastimplantillness* as a hashtag. We extracted the associated texts for each Instagram post with a timestamp between January 10, 2012, and September 4, 2019. The data set extracted from this site was referred to as IG-BII.

All the comments and posts from the 3 websites were included in the corresponding data sets. Table 1 presents a summary of the social media data collected. The BIIweb data set had only 187 posts (where each post on average has 129 words, SD 124) but these were larger (larger average length of posts in words) on average than those in the other 2 data sets. HealingBII was the second largest data set, with 1920 posts, each with 85 words on average (l_{avg}) (SD 107). IG-BII was the largest data set, with 28,987 posts and 123 words per post on average (SD 113).

Table 1. Statistical summary of social media data analyzed.

Data set	Posts[a] (n=31,094), n (%)	l_{max}[b]	l_{min}[c]	l_{avg}[d], mean (SD)	Words[e], n (%)
BIIweb	187 (0.6)	669	3	129 (124)	24,191 (0.64)
HealingBII	1920 (6.17)	1330	1	85 (107)	165,090 (4.38)
IG-BII	28,987 (93.22)	515	1	123 (113)	3,581,081 (94.98)

[a] Posts: the number of posts and comments in the respective data sets.

[b] l_{max}: the minimum length of a post in words.

[c] l_{min}: the maximum length of a post in words.

[d] l_{avg}: the average length of posts in words.

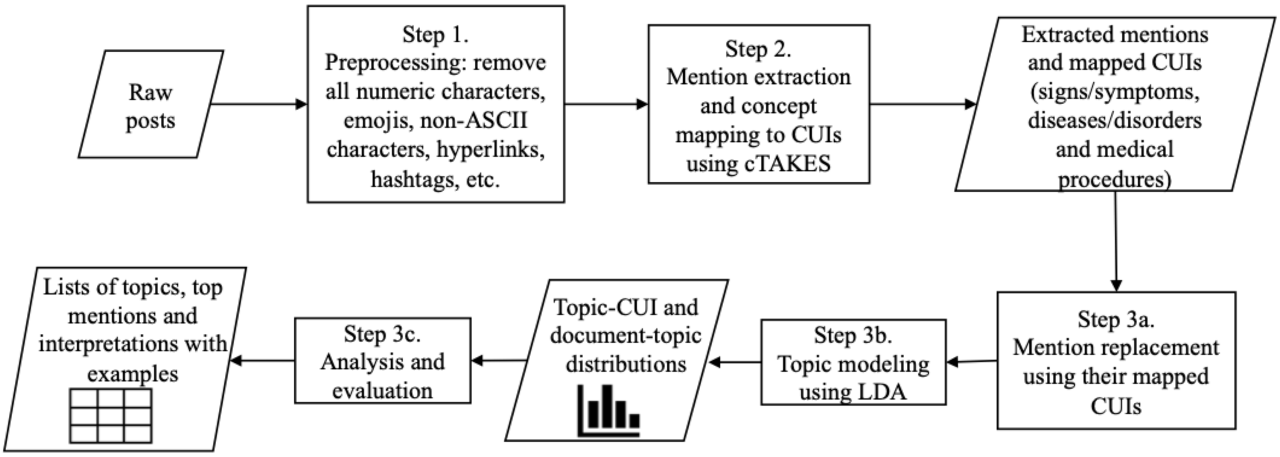
[e] Words: the total number of words in the respective data sets.

The Pipeline

Overview

Figure 1 shows an overview of the pipeline. We extracted major topics of interest primarily related to symptoms, diseases, and medical procedures from our data sets through the following 3 steps. Each of the steps will be discussed in detail later. The first step involved data preprocessing. We removed all stop words, numeric characters, hyperlinks, hashtags, etc, and converted the remaining characters into lowercase. The second step was of mention extraction and concept mapping. We extracted mentions related to signs, symptoms, diseases, disorders, and medical procedures using the clinical Text Analysis and Knowledge Extraction System (cTAKES) [64]. The extracted mentions were further mapped to standard medical concepts represented by concept unique identifiers (CUIs) in the unified medical language system (UMLS) [65] ontology. The third step involved topic modeling. We summarized the mapped concepts to topics using latent Dirichlet allocation (LDA) [66]. LDA is a probabilistic generative model for topic modeling. It represents each document as a mixture of latent topics, where each topic is modeled as a distribution over words. This modeling consisted of 3 stages: (1) mention replacement, (2) topic modeling using LDA, and (3) analysis and evaluation. In mention replacement, we replaced each extracted mention in the posts with its mapped CUIs and discarded all other words in the posts. We have discussed this step in more detail in the section *Topic modeling*. Then, in topic modeling using LDA, given the corpus of mapped CUIs, LDA generates document-topic and topics-CUI probability distributions. We have discussed this step in more detail in the section *Topic modeling*. Finally, during our analysis and evaluation, we further analyzed these distributions to derive a list of topics using the most representative mentions and summarized the extracted mentions for each data set. We have discussed this step in more detail in the section *Results: LDA topics*.

Figure 1. Pipeline for breast implant illness social media analysis. ASCII: American standard code for information interchange; CUI: concept unique identifier; LDA: latent Dirichlet allocation; cTAKES: clinical Text Analysis and Knowledge Extraction System.


Raw posts
Step 1. Preprocessing: remove all numeric characters, emojis, non-ASCII characters, hyperlinks, hashtags, etc.
Step 2. Mention extraction and concept mapping to CUIs using cTAKES
Extracted mentions and mapped CUIs (signs/symptoms, diseases/disorders and medical procedures)
Step 3a. Mention replacement using their mapped CUIs
Step 3b. Topic modeling using LDA
Topic-CUI and document-topic distributions
Step 3c. Analysis and evaluation
Lists of topics, top mentions and interpretations with examples

Data Preprocessing

We used the Natural Language Toolkit tokenizer [67] to tokenize the raw text for each data set. Out of the obtained tokens, we removed the stop-words (most frequently occurring, function words such as conjunctions, prepositions, determiners, etc) using the Natural Language Toolkit English stop-words list. As stop-words carried little or no information on our topics of interest in BII, they could be safely removed, as is typically done in NLP. We also removed all the numeric characters, emojis, non–American Standard Code for Information Interchange (ASCII) characters, hyperlinks, hashtags, and Instagram handles using regular expression matching and converted all the remaining tokens into lower cases to unify different cases for downstream processing.

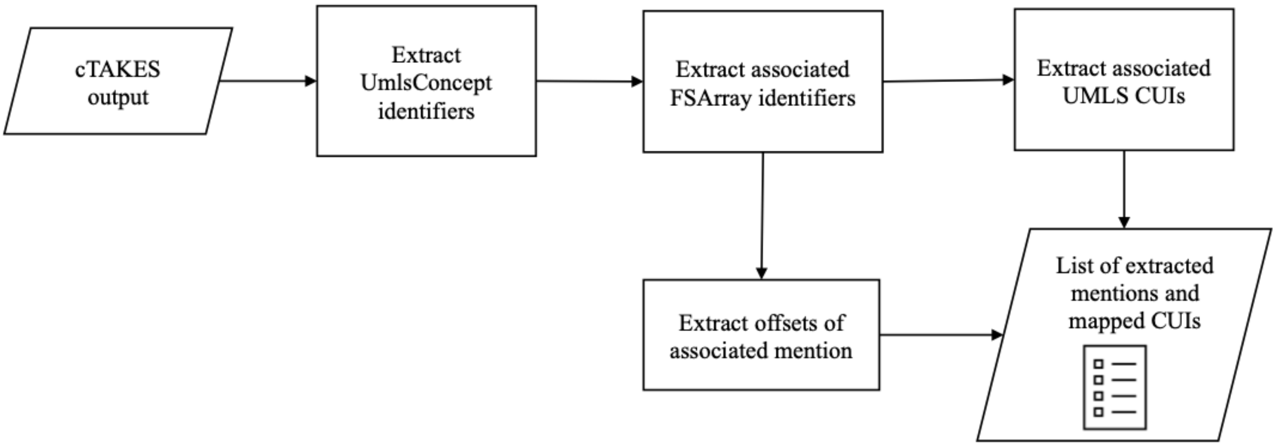
Mention Extraction and Concept Mapping

Mention extraction refers to the extraction of words or phrases that convey a medical concept. We used the cTAKES tool for mention extraction. The cTAKES tool is an open-source NLP tool for clinical information extraction from unstructured clinical texts. cTAKES extracts mentions (ie, words or phrases that convey a medical concept) from posts and maps these mentions to standard medical concepts. In doing so, it also categorizes each extracted mention into one of 5 cTAKES categories: sign, symptom, disease, disorder, medication, procedure, and anatomy; that is, while cTAKES extracts mentions, it also automatically classifies the mentions into one of the 5 categories. For example, in the sentence “Over the years, my tinnitus has become worse to almost debilitating levels,” cTAKES extracts *tinnitus* as a mention of sign and symptom category. Below, we discuss how to configure the cTAKES in detail.

We used the fast-dictionary-lookup annotator in cTAKES to extract mentions from the processed data. This annotator identifies and extracts mentions in texts and normalizes them into CUIs in the UMLS standard medical ontology. This normalization of extracted mentions into CUIs is referred to as concept mapping. Each CUI in the UMLS ontology uniquely identifies a medical concept. Hence, we represented extracted mentions using the standard medical concepts of CUIs that cTAKES maps the mentions to. We configured the annotator to use an exact string match and to use the all-term-persistence property. Thus, the annotator could retain all terms, irrespective of the semantic properties of each term. For example, for the phrase *back pain*, the annotator would annotate the generic term *pain* as well as the precise term *back pain*. We chose to use the all-term-persistence property to retain maximum information with respect to precise and generic medical concepts. Finally, the annotator stored the generated annotations in XML Metadata Interchange (XMI) files.

To obtain the annotations in a human-readable format from the XMI files, we performed the following steps (Figure 2). We used a custom interpreter to process the XMI files produced by cTAKES and to obtain mappings between mentions and CUIs from cTAKES. We first searched for *UmlsConcept* XML identifiers in the XMI files, where each *UmlsConcept* XML identifier is generally grouped under the *FSArray*, and each *FSArray* is associated with a single ontology concept and the category of the concept. Each concept is assigned one category out of 5 cTAKES categories: sign, symptom, disease, disorder, medication, procedure, and anatomy. Each ontology concept is further associated with a UMLS CUI and an *ontologyConceptArr* identifier. It must be noted that a mention can be mapped to multiple CUIs. For example, the mention *allergic reaction* is categorized as sign and symptom but mapped to 2 different CUIs: *C1527304* and *C0020517*. Then, we extracted the ontology concepts that describe any of these categories: diseases, disorders, signs, symptoms, and medical procedures. Finally, we used the *begin* and *end* markers associated with each *ontologyConceptArr* identifier to obtain the position of the annotated mention in the input post. In this work, we were only interested in the first 3 categories (ie, sign, symptom, disease, disorder, and procedure) to understand BII-related issues. Hence, we only used the mentions categorized into either of these 3 categories.

Figure 2. Pipeline for obtaining annotations out of Clinical text analysis and knowledge extraction system. cTAKES: clinical Text Analysis and Knowledge Extraction System; CUI: concept unique identifier; UMLS: unified medical language system.


cTAKES output
Extract UmlsConcept identifiers
Extract associated FSArray identifiers
Extract associated UMLS CUIs
Extract offsets of associated mention
List of extracted mentions and mapped CUIs

Topic Modeling

To conduct topic modeling, we processed the posts as follows: we substituted each mention in the posts with its mapped CUIs and discarded all other words in the posts, which were considered as nonmedical concepts by cTAKES or were not among the 3 categories of interest. If a mention was mapped to multiple CUIs, we replaced it with multiple CUIs. If multiple mentions were mapped to the same CUI, we replaced all such mentions with the CUI. In this way, each post was represented as a bag-of-CUI, instead of a collection of mentions, as the input to the topic modeling and our vocabulary consisted of CUIs. Upon topic modeling, we interpreted the topic-CUI distribution to derive the topics.

We used LDA [66] to learn the topic distributions of each post and the CUI distributions of each topic. LDA is a generative probabilistic model for modeling topics within a document corpus. LDA models each document in the corpus as a mixture of latent topics, where each topic is modeled as a distribution over words in all documents. LDA derives the optimal distributions by maximizing the likelihood of observing the corpus, following perspective distributions. A brief description of LDA is provided in Multimedia Appendix 1 [66]. In our experiments, a bag-of-CUIs generated as described above was used as a document in LDA, and the CUIs were words in the document. We used the lda-c software [68], which is a very efficient implementation of the LDA method, to conduct topic modeling.

When LDA is used in topic modeling for general documents (eg, news, scientific literature), words and their frequencies in the documents are used. However, in our analysis, we aimed to understand the medical concepts related to BII from social media texts. Different words may indicate the same medical concepts. For example, joint aches, painful joints, arthralgia, and aching joints all indicate joint pain and are associated with a single medical concept represented by a single CUI. Therefore, instead of using words, we used medical concepts, represented by CUIs, in our LDA analysis. Because multiple words indicating the same medical concept can be mapped to the same CUI, using CUIs can also aggregate and strengthen the information from multiple words, compared with using words, which may be sparse and thus not easy to learn topics from.

Results

cTAKES Annotations

Table 2 presents the summary statistics for the annotated mentions and their CUIs mapped by cTAKES. In BIIweb, cTAKES extracted 2186 mentions and mapped them to 475 unique CUIs. In HealingBII, cTAKES extracted 11,080 mentions and mapped them to 1177 unique CUIs. In the largest data set IG-BII, cTAKES extracted 5530 unique mentions and mapped them to 2871 unique CUIs. Note that the same mention can be mapped to multiple CUIs and can have multiple categories (each CUI has only one category). For example, the mention *flashes* is mapped to 2 different CUIs and then 2 different categories: diseases and medical procedures. Table 2 presents the statistics for each category of extracted mentions. For each data set, most of the extracted mentions were categorized as signs and symptoms by cTAKES.

Table 2. Statistical summary of annotations of the clinical Text Analysis and Knowledge Extraction System.

Data set	cwords[a]	annots[b]	maps[c]	M[d]	C[e]	M/C[f]	C/M[g]	S[h]	D[i]	P[j]
BIIweb	24,034	2186	661	640	475	1.39	1.03	385	149	106
HealingBII	163,352	11,080	1740	1685	1177	1.48	1.03	891	503	292
IG-BII	3,116,966	185,339	5694	5530	2871	1.98	1.03	3049	1549	932

[a] cwords: the total number of words recognized by the clinical Text Analysis and Knowledge Extraction System.

[b] annots: the total number of extracted mentions belonging to the 3 semantic types (ie, signs, symptoms, diseases, disorders, and medical procedures).

[c] maps: the number of unique mention–concept unique identifier mappings.

[d] M: the number of unique extracted mentions.

[e] C: the number of unique mapped concept unique identifiers.

[f] M/C: the average number of extracted mentions mapped to a given concept unique identifier.

[g] C/M: the average number of concept unique identifiers mapped to an extracted mention.

[h] S: the number of unique extracted mentions mapped to the signs and symptoms category.

[i] D: the number of unique extracted mentions that are mapped to the diseases and disorders category.

[j] P: the number of unique extracted mentions mapped to the medical procedures category.

To determine if cTAKES can sufficiently extract relevant mentions, we performed a manual annotation and compared the 2 lists of extracted mentions: one from using cTAKES and the other from using manual annotation. We randomly sampled 50 posts from each of the 3 data sets and manually annotated these posts. Upon manual annotation, we extracted mentions (words or phrases) that conveyed the concerns and experiences of social media users involving BII-related symptoms, diseases, and medical procedures. For a random sample of 50 posts (l_{avg}=134.18) from BIIweb, we obtained a total of 575 mentions from using manual annotation, and 637 mentions using cTAKES; there were 479 common mentions. Each mention was associated with a post identifier and a character offset. A mention was considered to belong to both lists if it occurred in both lists with the same post identifier and character offset. We found that 83.3% (479/575) of manually annotated mentions were covered by cTAKES. This high coverage demonstrates that cTAKES can capture most of the relevant medical concepts. In contrast, 75.2% (479/637) of the annotated mentions by cTAKES were covered by manual annotation. This further demonstrates that most of the annotated mentions of cTAKES can be confirmed by manual annotation. Similarly, for a random sample of 50 posts (l_{avg}=80.02) from HealingBII, 69.5% (194/279) of manually annotated mentions were covered by cTAKES; 70.3% (194/276) of mentions annotated by cTAKES were confirmed by manual annotation. For a random sample of 50 posts (l_{avg}=121.00) from IG-BII, the corresponding values were 75.2% (182/242) and 64.3% (182/283), respectively. According to the high overlap in the results between manual annotation and cTAKES across multiple data sets used in our study, it is reasonable to assume that cTAKES is a decent surrogate of manual annotation for BII study through social media data.

LDA Topics

To identify the best topic models, we used a grid search to identify the best parameter values for the Dirichlet prior $\alpha \in \{0.01,0.05,0.1,0.5,1,1.5,2,5,10,15,20,25\}$ and the number of topics K ∊ {3,4,5,10,15,20}. To evaluate topic models, we analyzed each LDA topic modeling result for every combination of α and K values corresponding to low perplexity scores [66,69,70].

For each topic modeling result, we analyzed the document-topic and topic-CUI probability distributions to derive topics and their respective top 10 representative mentions. The top 10 representative mentions for a given topic were the most frequent mentions, corresponding to the top 10 CUIs, with the highest probabilities of belonging to the topic. Multiple mentions could be mapped to a given CUI (Table 2). We only presented the most frequent mention because all mentions mapped to the same CUI had similar semantics. We further evaluated the quality of topic modeling based on how well the derived topics summarized the most representative mentions. We analyzed each LDA topic modeling result for every combination of α and K and chose the one where the derived topics were distinct and best summarized the most representative mentions. Finally, we identified distinct and meaningful topics using (1) K=4 and α=10 for BIIweb, (2) K=5 and α=10 for HealingBII, and (3) K=5 and α=1.5 for IG-BII. We observed that with higher K values, the most representative mentions were similar across the topics. Hence, the derived topics were not distinct and were difficult to interpret.

Tables 3-5 present the top 10 representative mentions, the frequencies of CUIs corresponding to the mentions (in %), and the interpretations of the topics indicated by the mentions (eg, common signs and symptoms). Note that the frequencies of CUIs are among all the posts, not only in those posts with the highest probability belonging to a certain topic. We presented these frequencies because each post had a certain probability of belonging to a certain topic, and thus frequencies among all posts would better represent the topic information across all the posts. These tables also present examples of posts that have a high probability of belonging to the respective topic. In the examples, the mentions that had high probabilities of belonging to the corresponding topics are italicized. Note that we used CUIs in LDA to derive the topic and word distributions (as discussed in the section Methods—Topic modeling), but we have presented the most frequent mentions (with clear semantics) that were mapped to the respective CUIs (which are identifiers without semantics) in these tables. The mentions in these tables were sorted based on the probabilities of their corresponding CUIs belonging to the respective topics. Please note that these probabilities have not been presented in the tables (they are not the frequencies presented in the tables). Therefore, each topic was represented by its most representative mentions, and thus, summarized such mentions. For example, we interpreted a topic as pain and other signs if there were a significant number of mentions related to pain, such as neck pain, chest pain, and headache. Please note that the topics have not been sorted, and the first columns in Tables 3 to 5 are nominal identifiers. Below, we have discussed the topics derived from LDA for BIIweb and HealingBII data sets from the original posts. Note that 2 topics can still share the same representative mention with different probabilities in the LDA.

Table 3. Derived topics in BIIweb.

Topic	Top 10 mentions	Interpretation
1	• Testing (2.34); illness (4.46); problem (2.82); work (1.17); swollen (0.78); drains (0.61); feel common (2.51); fatigue (1.82); exhausted (0.39); sensitivity (0.95) • Example: "I had silicone implants done 5 years ago, three years ago after going to the doctor with extreme *fatigue*[a] (I was sleeping 14-16 hours a day and was still *exhausted*)"	Common signs and symptoms
2	• Breast implant (6.80); removal (1.30); cancer (0.95); autoimmune (0.95); infection (0.87); scleroderma (0.39); pain (3.68); diagnosis (0.30); alcl (0.30); breast cancer (0.30) • Example: "I had stage 4 breast *cancer* and had chemo and radiation. I tried to have my *breast implants* removed due to *pain*...Then I had an acute *infection* occur a month and a half after they put the new implants in and they were forced to perform an emergency *removal* of the newer implants. I have had all the symptoms of breast implant illness—even after their removal."	Diseases or disorders
3	• Breast implant (6.80); illness (4.46); toxicity (1.17); foreign body (0.87); heal (0.78); support (0.65); rupture (0.52); cancer (0.95); awareness (0.35); inflammation (0.56) • Example: "...I never had a problem until 2006 at which time I thought something had happened however, my surgeon said I must have just pulled a muscle and that the *implants* seemed fine. Now that surgeon is old and the shop is closed up. I have been suffering for the past 13 years with arthritis, fatigue, brain fog, *inflammation*, hormone imbalances, and adrenal fatigue..."	Toxicity
4	• Pain (3.68); feel (2.51); fatigue (1.82); back pain (0.87); illness (4.46); joint pain (0.56); worse (0.65); anxiety (0.52); ear ringing (0.39); headache (0.39) • Example: "It wasn't until 2017 where I started to experience *anxiety* and panic attacks (which I didn't know I was having at the time). With that, along came crazy *headaches*, feeling dizzy, sick, light-headed, and my right eye would always be swollen and never knew why."	Pain and stress-related disorders

[a] The mentions in the examples that had high probabilities of belonging to the corresponding topics are italicized.

Table 4. Derived topics in HealingBII.

Topic	Top 10 mentions	Interpretation
1	• Rupture (1.34); supported (0.87); read (1.17); suffering (0.87); happy (0.6); mastectomy (0.46); work (0.96); scare (0.77); reconstruction (0.41); mri (0.72) • Example: “Double *mastectomy*[a] in 2015. *Reconstruction* process with expanders then permanent 1000 ml saline implants in early 2016. After that was 9 procedures, a hysterectomy and now MANY health problems.”	Surgeries and procedures
2	• Pain (3.91); joint pain (0.79); fatigued (0.96); ailment (4.70); removal (0.84); hair loss (0.52); headache (0.47); muscle ache (0.34); rash (0.39); infection (0.84) • Example: “In addition to the neuromuscular spasms and *pain*, I’ve suffered with incapacitating chronic *fatigue*, BRAIN FOG and confusion (yes, even while driving), loss of vision and hearing, vertigo, mysterious skin *rashes, hair loss, migraines*...”	Pain and other signs
3	• Problem (2.64); cancer (0.90); autoimmune (0.57); breast cancer (0.38); scars (0.35); treatment (0.43); diagnose (0.29); autoimmune disorder (0.27); lupus (0.29); arthritis (0.26) • Example: “I had capsules form on both breasts from about 2010. I got sick with BII symptoms from 2005 with lots of infections required intravenous and oral antibiotics. My environmental and drug allergies got worse, onset of *arthritis*, skin rashes, *autoimmune* symptoms, started growing low grade *cancers*...”	Cancer and other disorders
4	• Breast implant (3.85); ailment (4.70); toxicity (3.05); healing (1.56); capsulectomy (0.64); infection (0.84); inflammation (0.39); detoxification (0.32); foreign object (0.25); bleed (0.23) • Example: “Some women with silicone *toxicity* have bruising and *bleeding* problems. If I was you, I would try and have the lymph node localized and checked for silicone and removed if it is contaminated beyond detoxing much like a silicone granuloma is removed.”	Toxicity
5	• Emotion (3.70); think (2.26); feel (0.84); normal (0.65); anxiety (0.50); ill (0.61); sensation (0.33); tired (0.28); sores (0.27); depression (0.33) • Example: “Even more heartbreaking and discouraging, has been the *emotional* pain of not being able to freely play with her on the floor due to hip and knee pain, along with leg and foot spasms...but I struggle with many *feelings* of failure as a wife and mother due to physical limitations.”	Mental health

[a] Italic text indicates the mentions in the examples that had high probability of belonging to the corresponding topics.

Table 5. Derived topics in IG-BII.

Topic	Top 10 mentions	Interpretation
1	• Heal (1.46); working (0.90); weighted (1.05); able (0.99); rest (0.37); stress (0.29); exercise (0.28); therapeutic (0.35); sleep (0.36); run (0.23) • Example: "It's been 14 months since my explant. The journey to *healing*[a] hasn't been an easy one due to setbacks and relapses but better than daily anaphylaxis from getting cold, food, smells, crying, *exercise* and *stress*, then add angina attacks from anaphylaxis."	Physical health
2	• Malignancy (1.10); removal (0.96); scar (0.75); capsulectomy (0.68); rupture (0.43); ciactrice (0.43); alcl (0.41); augmentation (0.37); lymphoma (0.35); removal of implants (0.29) • Example: "The ugly side of breast implants. It's not a matter of IF you will get sick...it's WHEN. implants leak toxic heavy metals without rupture It's called a gel bleed. Women with implants are 3 times more likely to develop brain, lung and *lymphatic cancer* than women with implants."	Cancer and medical procedures
3	• Loving (2.43); happiness (2.11); emotion (1.64); think (1.05); feel (0.87); scare (0.55); confidence (0.35); tired (0.38); emotional (0.27); sensation (0.33) • Example: "I was *scared* of looking incomplete. After much deep, inner work on myself, I realized that my worth wasn't dependent on what I looked like or how big my chest was. I realized that true *happiness* would come from 100% acceptance of what and who I was"	Mental health
4	• Breast implant (7.21); ailment (5.67); toxicity (1.67); aware (0.96); felt worse (0.36); test (0.64); foreign body (0.45); alone (0.33); suffering (0.21); complication (0.20) • Example: "...We get *toxic* from the chemical makeup of the silicone, the *toxic* chemicals that are released when the shell degrades, sick from rupture and sometimes mold."	Toxicity
5	• Pain (2.52); inflammatory reaction (0.89); fatigue (0.83); anxiousness (0.72); allergy (0.43); depression (0.37); joint pain (0.33); autoimmune disorder (0.32); swell (0.43); infection (0.31) • Example: "For three years, doctors have been unable to diagnose or explain upper body weakness, hand *pain*, and general *inflammation*. I have suffered from periods of high *inflammation*, debilitating *fatigue*, migraines, inability to lose weight, insomnia, low libido, body and *joint pain*, hair loss, dry skin, dry eyes, brain fog, etc."	Common disorders

[a] Italic text indicates the mentions in the examples that had high probability of belonging to the corresponding topics.

Table 3 presents the topics in the data set BIIweb data set. Although BIIweb was the smallest the data set (Table 1), we were still able to identify 4 distinct topics with the most representative mentions, namely, fatigue, infection, toxicity, and anxiety. Table 4 presents the topics in the data set HealingBII, which shared some common topics and representative mentions with those in BIIweb. For example, pain, cancer, and toxicity were common across these 2 data sets. However, a focused topic unique to HealingBII was surgeries and procedures, where people (mostly patients) discuss the procedures among themselves and share their related experiences. Another unique topic in HealingBII was mental health.

In addition to physical symptoms, individuals reported significant emotional and mental difficulties, such as depression, and expressed serious symptoms on social media. Table 5 presents the topics in the data set IG-BII data set. IG-BII was the largest data set (Table 1) and had significantly more posts than the other two. We observed that cancers, mental health, and toxicity emerged as significant topics in this large data set, consistent with those in HealingBII. In IG-BII, people also discussed their recovery process from the issues or events associated with BII. We identified from these 3 data sets frequent mentions of rupture, pains, and fatigue. We also identified mentions of cancer, lupus, and autoimmune disorders.

Please note that Table 3 contains 4 topics for BIIweb, but Tables 4 and 5 contain 5 topics for HealingBII and IG-BII, respectively. This is because the number of topics was determined by how distinct the topics were, not by the prespecified number of topics.

Table 6 presents the top 10 representative mentions, the frequencies of CUIs corresponding to the mentions (in %), and interpretations of the topics on the unified data set, combining all 3 data sets BIIweb, HealingBII, and IG-BII. We obtained a unified data set by combining all the posts from the 3 data sets into one corpus. To perform topic modeling, we processed the posts in the unified data set in the same way as we processed the posts in the individual data sets (discussed in the section Methods—Topic modeling). Upon topic modeling, we identified 5 distinct topics using K=5 and α=1.5. We observed that physical health, cancers, mental health, toxicity, and common disorders emerged as significant topics in the unified data set, consistent with those in IG-BII. This was because IG-BII was the largest data set out of the three and comprised 93.22% (28,987/31,094) of the unified data set. We also identified common concerns such as pain, allergy, depression, weight gain, cancer, inflammation, and toxicity issues from the individual and unified data sets. This implies that the above-mentioned factors were frequently associated with BII.

Table 6. Derived topics in the unified data set.

Topic	Top 10 mentions	Interpretation
1	• Working (1.45); ate (0.92); weight (0.79); runs (0.40); thinking (2.68); exercise (0.25); talk (0.50); walking (0.35); nutrition (0.15); move (0.28); • Example: "...I'm now healthier than I have been in the last 7 years of my life!...I explanted in Feb of 2018, a few months after explant, I gained my *weight*[a] back and found a love for true self care and *working* out."	Physical health
2	• Illnesses (4.45); cancer (0.87); ruptures (0.77); removal (0.76); awareness (0.73); suffers (0.83); capsulectomy (0.54); autoimmune (0.52); breast augmentation (0.30); augmentation (0.28); • Example: "I was diagnosed with breast *cancer* at the young age of 30 and ended up with a double mastectomy as part of that process...now 10 years later I have just 15 weeks ago had my implants removed. They had *ruptured*, were toxic and giving me health issues"	Cancer and medical procedures
3	• Feel (5.94); loved (2.97); thinking (2.68); happier (1.64); feelings (1.47); afraid (0.66); confidence (0.27); support (0.79); able (0.77); alive (0.17); • Example: "When I found out I was sick and I had to tear apart my body to get better I never thought I'd be happy with myself again. I am 4 weeks post op and *feeling* more happy and healthy than ever. I was worried I'd never be *loved* again."	Mental health
4	• Heal (2.26); scars (0.58); scarred (0.33); drain (0.26); toxic (1.97); sights (1.25); inflammation (0.68); bulge (0.36); tenderness (0.20); red (0.15); damage (0.16); • Example: "I was so worried about how *red* and raised up my *scars* were...then they got really inflamed, sore and raised up around 3 weeks and i was really stressed over it. then overnight the *inflammation* and redness went down..."	Common signs, symptoms, and toxicity
5	• Pain (2.09); tired all the time (0.69); anxiety (0.57); joint pain (0.46); alopecia (0.39); weight gain (0.37); allergies (0.35); depression (0.29); pain back (0.23); headache (0.22) • Example: "Before I had the explant, I had many unexplained symptoms (brain fog, *joint pain*, back and neck pain, *tired all the time*, psoriasis, afib, just to mention a few) since I awoke from surgery I have had absolutely no neck, back, or joint pain."	Common disorders

[a] Italic text indicates the mentions in the examples that had high probability of belonging to the corresponding topics.

Table 7 presents the percentage of posts per topic, where a post *d* is considered to belong to a topic z if among all topics that *d* has, z has the highest probability. Although the distributions are not completely consistent across data sets, toxicity remained a notable topic among all data sets. This indicates that these were common issues that were significantly associated with BII. In addition, pain, cancer, mental health, and other disorders were also associated with breast implants.

Table 7. Distribution of posts among the topics.

Data set and topics	Posts, n (%)
BIIweb	
Common signs and symptoms	62 (33.2)
Diseases or disorders	28 (15)
Toxicity	50 (26.7)
Pain and stress-related disorders	47 (25.1)
HealingBII	
Surgeries and procedures	713 (37.1)
Pain and other signs	221 (11.5)
Cancer and other disorders	221 (11.5)
Toxicity	505 (26.3)
Mental health	260 (13.6)
IG-BII	
Physical health	11,299 (39)
Cancer and medical procedures	3890 (13.4)
Mental health	4879 (16.8)
Toxicity	5415 (18.7)
Common disorders	3504 (12.1)
Unified	
Physical health	4760 (15.3)
Cancer and medical procedures	10,637 (34.2)
Mental health	7954 (25.6)
Common signs, symptoms, and toxicity	4030 (13)
Common disorders	3713 (11.9)

Discussion

Principal Findings

To understand the signs, symptoms, and diseases or disorders associated with BII, a condition reported primarily on social media rather than in medical reports, we collected social media posts and analyzed them using NLP and topic modeling. We extracted mentions related to signs, symptoms, diseases, disorders, and medical procedures using cTAKES, mapped them to standard medical concepts, and summarized the mapped concepts to topics using LDA. We found that mentions such as rupture, infection, inflammation, pain, and fatigue were common self-reported issues. We also found that mental health–related concerns such as stress, anxiety, and depression, as well as diseases such as cancers and autoimmune disorders, were common concerns. The cTAKES was able to extract medication and anatomy information as well, but they were not used in our LDA analysis, given that the objective of our study was not to study the medications used or the anatomy related to BII.

In our method, we relied on cTAKES and the rich UMLS dictionary to extract all relevant mentions, including their lexical variants (synonyms, abbreviations, paraphrases). To determine if cTAKES could sufficiently extract relevant mentions, we performed a manual annotation to extract all the relevant mentions and compared them with the extracted mentions from cTAKES. We found that cTAKES could sufficiently capture relevant medical concepts and was comparable with manual annotation. It is worth noting that we did not evaluate the performance of our mention extraction module on all the posts of each data set, which is typically performed using precision and recall metrics when there are ground-truth labels associated with each mention. However, in order to have such labels, careful manual annotations based on domain knowledge of BII are required. Unfortunately, such domain knowledge on complications, symptoms, and other issues associated with or caused by BII were not fully available. Our goal in this study is to provide useful information from social media data that could complement our current knowledge. Therefore, in this preliminary study, we used all annotated mentions, assuming that cTAKES enabled high-quality annotations.

Strengths and Limitations

We acknowledge that cTAKES might not have been able to extract all relevant mentions from our social media data sets. This is because cTAKES was originally designed for extraction of medical entities from clinical notes, which have very different wording and writing styles compared with social media data. As social media data comprise informal phrases, short ambiguous texts, emoticons, and a wide range of lexical variants corresponding to a single concept, cTAKES might not work flawlessly on social media data, although we observed reasonable output from cTAKES. We also observed that cTAKES often associated a single mention with multiple CUIs belonging to the same category. We think this was because of the presence of multiple mappings for a given mention in the UMLS metathesaurus. Regardless, the extracted mentions and the mapping of mentions to UMLS CUIs, as generated by cTAKES, were used for topic modeling without any manual verification or evaluation. In the future, we will develop a detailed guideline to further evaluate the extracted mentions before using them in topic modeling.

Our study had some limitations. First, LDA is an unsupervised learning technique in which the number of topics (K) is assumed to be known a priori. However, it is difficult to accurately estimate K for a given data set. In our study, we used a grid search to obtain different K values. Even without full domain knowledge, it remains nontrivial to evaluate the LDA results for each K value. In our study, we selected the topics based on α and K values. We did not use perplexity [66,69,70], a widely used metric in topic modeling, to select the topics, because as studied in the literature (eg, Chang et al [71]), perplexity often does not correlate well with topic interpretability; in our case, the lowest perplexity did not always enable intuitive or meaningful topics. In the future, we will develop more rigorous ways to select the number of topics and evaluate the topic modeling results. In this study, we did not conduct a sentiment analysis of the posts to understand the positive or negative opinions expressed in the posts. We plan to include this process before topic modeling to generate a cleaner data set for topic modeling.

It is worth noting that social media data could be of variable quality (eg, misspelling, misconception, and biased opinions), particularly compared with medical literature data. Anyone can post on social media, and so the derived content may be from individuals who may have other implant-specific issues such as capsular contracture or implant infection. Thus, understanding the diseases, disorders, symptoms, signs, etc, associated with a drug, disease, or medical procedure from social media data would always be at risk from confounders or errors. However, given that the medical knowledge and literature on BII have not been well established, and the related concepts are not well defined or well accepted, using social media data to understand emerging issues could be a meaningful starting point. Still, any findings from social media data would require a rigorous evaluation and validation based on medical and biological knowledge, experiments, clinical practice, etc. In addition, we have only analyzed 3, though the most relevant and prolific websites dedicated to BII discussions. A more comprehensive analysis of social media data on a much larger scale would be beneficial to better understand BII in a larger, diverse population. Sentiment analysis of social media data could be another valuable analysis to enable more insights into the health experiences of users or patients and their emotions or feelings. We will consider sentiment analysis in our future research when BII is better understood, and we can accurately annotate social media data.

Conclusions

This study has important implications for future methodological and clinical research. Future methodological research on NLP could include causality inference between BII and symptom and sign mentions from social media to understand their relations, etc. Our findings could provide the relevant domains for clinical research studies seeking to develop measures of BII and to identify its causes. More specifically, our results can provide a patient-derived definition of BII, which can be useful to clinicians treating patients with BII concerns to use this patient-centered language. Our methods and informatics strategies applied in this study would also provide working examples for analyzing other emerging but not well-defined illnesses from social media data.

Our analysis of social media data identified mentions such as rupture, infection, inflammation, pain, and fatigue, which were common self-reported issues on social media sites dedicated to BII. In addition, our analysis showed that a significant number of user comments and posts were also concerned with mental and physical health and toxicity issues after having breast implants. The findings from our study could be used to further the scientific study of BII, as well as the care of patients presenting with the described symptoms, by allowing clinicians to develop a patient-centered language to better approach the patients with concerns. Our study provides the first analysis and derived knowledge of BII from social media using NLP techniques and demonstrates the potential of using social media information to better understand emerging illnesses.

Acknowledgments

XN conceived the research, obtained funding for the research, and supervised VD; PK, MN, and CL provided substantial medical background and insights; Vishal Dey and Xia Ning conducted the research, including data curation, methodology design and implementation, and analysis; Vishal Dey drafted the original manuscript; Vishal Dey and Xia Ning conducted the manuscript editing; PK, MN, and CL reviewed the manuscript and provided constructive suggestions and feedback.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A brief description of latent Dirichlet allocation.

[DOCX File , 53 KB - medinform_v9i11e29768_app1.docx]

References

1. Barros JM, Duggan J, Rebholz-Schuhmann D. The application of internet-based sources for public health surveillance (Infoveillance): systematic review. J Med Internet Res 2020 Mar 13;22(3):e13680 [FREE Full text] [doi: 10.2196/13680] [Medline: 32167477]
2. Schillinger D, Chittamuru D, Ramírez AS. From "Infodemics" to health promotion: a novel framework for the role of social media in public health. Am J Public Health 2020 Sep;110(9):1393-1396. [doi: 10.2105/AJPH.2020.305746] [Medline: 32552021]
3. Li D, Chaudhary H, Zhang Z. Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining. Int J Environ Res Public Health 2020 Jul 10;17(14):4988 [FREE Full text] [doi: 10.3390/ijerph17144988] [Medline: 32664388]
4. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in mainland china: observational infoveillance study. J Med Internet Res 2020 May 28;22(5):e19421 [FREE Full text] [doi: 10.2196/19421] [Medline: 32452804]
5. Aiello AE, Renson A, Zivich PN. Social media- and internet-based disease surveillance for public health. Annu Rev Public Health 2020 Apr 02;41:101-118. [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
6. Li C, Chen LJ, Chen X, Zhang M, Pang CP, Chen H. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. Euro Surveill 2020 Mar;25(10):2000199 [FREE Full text] [doi: 10.2807/1560-7917.ES.2020.25.10.2000199] [Medline: 32183935]
7. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb 19;457(7232):1012-1014. [doi: 10.1038/nature07634] [Medline: 19020500]
8. Naslund JA, Grande SW, Aschbrenner KA, Elwyn G. Naturally occurring peer support through social media: the experiences of individuals with severe mental illness using YouTube. PLoS One 2014;9(10):e110171 [FREE Full text] [doi: 10.1371/journal.pone.0110171] [Medline: 25333470]
9. Foufi V, Timakum T, Gaudet-Blavignac C, Lovis C, Song M. Mining of textual health information from reddit: analysis of chronic diseases with extracted entities and their relations. J Med Internet Res 2019 Jun 13;21(6):e12876 [FREE Full text] [doi: 10.2196/12876] [Medline: 31199327]
10. Wakamiya S, Morita M, Kano Y, Ohkuma T, Aramaki E. Tweet classification toward twitter-based disease surveillance: new data, methods, and evaluations. J Med Internet Res 2019 Feb 20;21(2):e12783. [doi: 10.2196/12783] [Medline: 30785407]
11. Attai DJ, Cowher MS, Al-Hamadani M, Schoger JM, Staley AC, Landercasper J. Twitter social media is an effective tool for breast cancer patient education and support: patient-reported outcomes by survey. J Med Internet Res 2015;17(7):e188 [FREE Full text] [doi: 10.2196/jmir.4721] [Medline: 26228234]
12. Osadchiy V, Mills JN, Eleswarapu SV. Understanding patient anxieties in the social media era: qualitative analysis and natural language processing of an online male infertility community. J Med Internet Res 2020 Mar 10;22(3):e16728 [FREE Full text] [doi: 10.2196/16728] [Medline: 32154785]
13. Nobles AL, Leas EC, Althouse BM, Dredze M, Longhurst CA, Smith DM, et al. Requests for diagnoses of sexually transmitted diseases on a social media platform. J Am Med Assoc 2019 Nov 05;322(17):1712-1713. [doi: 10.1001/jama.2019.14390] [Medline: 31688875]
14. Kahlor L, Mackert M. Perceptions of infertility information and support sources among female patients who access the internet. Fertil Steril 2009 Jan;91(1):83-90. [doi: 10.1016/j.fertnstert.2007.11.005] [Medline: 18243181]
15. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci 2017 Dec;18:43-49. [doi: 10.1016/j.cobeha.2017.07.005]
16. Karmen C, Hsiung RC, Wetter T. Screening internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods. Comput Methods Programs Biomed 2015 Jun;120(1):27-36. [doi: 10.1016/j.cmpb.2015.03.008] [Medline: 25891366]
17. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform 2019 May;125:37-46 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.02.008] [Medline: 30914179]
18. Chapman B, Raymond B, Powell D. Potential of social media as a tool to combat foodborne illness. Perspect Public Health 2014 Jul;134(4):225-230. [doi: 10.1177/1757913914538015] [Medline: 24990140]
19. Harris JK, Mansour R, Choucair B, Olson J, Nissen C, Bhatt J, Centers for Disease ControlPrevention. Health department use of social media to identify foodborne illness - Chicago, Illinois, 2013-2014. MMWR Morb Mortal Wkly Rep 2014 Aug 15;63(32):681-685 [FREE Full text] [Medline: 25121710]
20. Casas J, Mugellini E, Abou K. Early detection of foodborne illnesses in social media. In: Proceedings of the 2nd International Conference on Human Interaction and Emerging Technologies: Future Applications (IHIET-AI 2020). Lausanne, Switzerland: Springer; 2020 Presented at: 2nd International Conference on Human Interaction and Emerging Technologies: Future Applications (IHIET-AI 2020); April 23-25, 2020; Lausanne, Switzerland p. 415-420. [doi: 10.1007/978-3-030-44267-5_62]

21. McGough SF, Brownstein JS, Hawkins JB, Santillana M. Forecasting zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data. PLoS Negl Trop Dis 2017 Jan;11(1):e0005295 [FREE Full text] [doi: 10.1371/journal.pntd.0005295] [Medline: 28085877]
22. Zhao J, Han H, Zhong B, Xie W, Chen Y, Zhi M. Health information on social media helps mitigate Crohn's disease symptoms and improves patients' clinical course. Comput Hum Behav 2021 Feb;115:106588. [doi: 10.1016/j.chb.2020.106588]
23. Pandrekar S, Chen X, Gopalkrishna G, Srivastava A, Saltz M, Saltz J, et al. Social media based analysis of opioid epidemic using Reddit. In: AMIA Annu Symp Proc. 2018 Presented at: AMIA Annual Symposium; November 3-7, 2018; San Francisco, CA p. 867-876 URL: http://europepmc.org/abstract/MED/30815129
24. Marques-Toledo CD, Degener CM, Vinhal L, Coelho G, Meira W, Codeço CT, et al. Dengue prediction by the web: tweets are a useful tool for estimating and forecasting Dengue at country and city level. PLoS Negl Trop Dis 2017 Jul;11(7):e0005729 [FREE Full text] [doi: 10.1371/journal.pntd.0005729] [Medline: 28719659]
25. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One 2011;6(5):e19467 [FREE Full text] [doi: 10.1371/journal.pone.0019467] [Medline: 21573238]
26. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One 2013;8(12):e83672 [FREE Full text] [doi: 10.1371/journal.pone.0083672] [Medline: 24349542]
27. Klembczyk JJ, Jalalpour M, Levin S, Washington RE, Pines JM, Rothman RE, et al. Google flu trends spatial variability validated against emergency department influenza-related visits. J Med Internet Res 2016;18(6):e175 [FREE Full text] [doi: 10.2196/jmir.5585] [Medline: 27354313]
28. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res 2009;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
29. Li J, Xu Q, Cuomo R, Purushothaman V, Mackey T. Data mining and content analysis of Chinese social media platform Weibo during early COVID-19 outbreak: a retrospective observational infoveillance study. JMIR Public Health Surveill 2020 Apr 14;6(2):e18700 [FREE Full text] [doi: 10.2196/18700] [Medline: 32293582]
30. Huang C, Xu X, Cai Y, Ge Q, Zeng G, Li X, et al. Mining the characteristics of COVID-19 patients in China: analysis of social media posts. J Med Internet Res 2020 May 17;22(5):e19087 [FREE Full text] [doi: 10.2196/19087] [Medline: 32401210]
31. 2019 Plastic Surgery Statistics Report. American Society of Plastic Surgeons (ASPS). 2019. URL: https://www.plasticsurgery.org/news/plastic-surgery-statistics?sub=2019+Plastic+Surgery+Statistics [accessed 2021-03-29]
32. 2018 National Plastic Surgery Statistics. American Society of Plastic Surgeons (ASPS). 2018. URL: https://www.plasticsurgery.org/documents/News/Statistics/2018/plastic-surgery-statistics-report-2018.pdf [accessed 2019-07-18]
33. Balk EM, Earley A, Avendano EA, Raman G. Long-term health outcomes in women with silicone gel breast implants: a systematic review. Ann Intern Med 2016 Feb 02;164(3):164-175. [doi: 10.7326/M15-1169] [Medline: 26550776]
34. Watad A, Rosenberg V, Tiosano S, Tervaert JW, Yavne Y, Shoenfeld Y, et al. Silicone breast implants and the risk of autoimmune/rheumatic disorders: a real-world analysis. Int J Epidemiol 2018 Dec 01;47(6):1846-1854. [doi: 10.1093/ije/dyy217] [Medline: 30329056]
35. Labadie JG, Korta DZ, Barton N, Mesinkovska NA. Cutaneous hypersensitivity-like reactions associated with breast implants: a review. Dermatol Surg 2018 Mar;44(3):323-329. [doi: 10.1097/DSS.0000000000001448] [Medline: 29293108]
36. Calobrace MB, Stevens WG, Capizzi PJ, Cohen R, Godinez T, Beckstrand M. Risk factor analysis for capsular contracture: a 10-year sientra study using round, smooth, and textured implants for breast augmentation. Plast Reconstr Surg 2018 Apr;141(4S):20-28. [doi: 10.1097/PRS.0000000000004351] [Medline: 29595715]
37. Rohrich RJ, Kaplan J, Dayan E. Silicone implant illness: science versus myth? Plast Reconstr Surg 2019;144(1):98-109. [doi: 10.1097/prs.0000000000005710]
38. Coroneos C, Selber J, Offodile A, Butler C, Clemens M. US FDA breast implant postapproval studies: long-term outcomes in 99,993 patients. Ann Surg 2019 Jan;269(1):30-36. [doi: 10.1097/SLA.0000000000002990] [Medline: 30222598]
39. Gabriel SE, O'Fallon WM, Kurland LT, Beard CM, Woods JE, Melton LJ. Risk of connective-tissue diseases and other disorders after breast implantation. N Engl J Med 1994 Jun 16;330(24):1697-1702. [doi: 10.1056/NEJM199406163302401] [Medline: 8190133]
40. Peters W, Smith D, Fornasier V, Lugowski S, Ibanez D. An outcome analysis of 100 women after explantation of silicone gel breast implants. Ann Plast Surg 1997 Jul;39(1):9-19. [doi: 10.1097/00000637-199707000-00002] [Medline: 9229086]
41. Janowsky EC, Kupper LL, Hulka BS. Meta-analyses of the relation between silicone breast implants and the risk of connective-tissue diseases. N Engl J Med 2000 Mar 16;342(11):781-790. [doi: 10.1056/NEJM200003163421105] [Medline: 10717013]
42. Rohrich RJ, Kenkel JM, Adams WP, Beran S, Conner WC. A prospective analysis of patients undergoing silicone breast implant explantation. Plast Reconstr Surg 2000 Jun;105(7):2529-2538. [doi: 10.1097/00006534-200006000-00036] [Medline: 10845310]

43. Nahabedian MY, Tsangaris T, Momen B, Manson PN. Infectious complications following breast reconstruction with expanders and implants. Plast Reconstr Surg 2003 Aug;112(2):467-476. [doi: 10.1097/01.PRS.0000070727.02992.54] [Medline: 12900604]
44. Siggelkow W, Klosterhalfen B, Klinge U, Rath W, Faridi A. Analysis of local complications following explantation of silicone breast implants. Breast 2004 Apr;13(2):122-128. [doi: 10.1016/j.breast.2003.08.003] [Medline: 15019692]
45. Lee I, Cook NR, Shadick NA, Pereira E, Buring JE. Prospective cohort study of breast implants and the risk of connective-tissue diseases. Int J Epidemiol 2011 Feb;40(1):230-238 [FREE Full text] [doi: 10.1093/ije/dyq164] [Medline: 20943932]
46. Tang SY, Israel JS, Afifi AM. Breast implant illness: symptoms, patient concerns, and the power of social media. Plast Reconstr Surg 2017 Nov;140(5):765-766. [doi: 10.1097/PRS.0000000000003785] [Medline: 28753149]
47. Tang SY, Israel JS, Poore SO, Afifi AM. Facebook facts: breast reconstruction patient-reported outcomes using social media. Plast Reconstr Surg 2018 May;141(5):1106-1113. [doi: 10.1097/PRS.0000000000004275] [Medline: 29697604]
48. Magnusson MR, Cooter RD, Rakhorst H, McGuire PA, Adams WP, Deva AK. Breast implant illness: a way forward. Plast Reconstr Surg 2019 Mar;143(3S):74-81. [doi: 10.1097/PRS.0000000000005573] [Medline: 30817559]
49. Adidharma W, Latack KR, Colohan SM, Morrison SD, Cederna PS. Breast implant illness: are social media and the internet worrying patients sick? Plast Reconstr Surg 2020 Jan;145(1):225-227. [doi: 10.1097/PRS.0000000000006361] [Medline: 31625990]
50. Keane G, Chi D, Ha A, Myckatyn T. En bloc capsulectomy for breast implant illness: a social media phenomenon? Aesth Surg J 2021;41(4):448-459. [doi: 10.1093/asj/sjaa203]
51. Wee CE, Younis J, Isbester K, Smith A, Wangler B, Sarode AL, et al. Understanding breast implant illness, before and after explantation: a patient-reported outcomes study. Ann Plast Surg 2020 Jul;85(S1 Suppl 1):82-86 [FREE Full text] [doi: 10.1097/SAP.0000000000002446] [Medline: 32530850]
52. Lee M, Ponraja G, McLeod K, Chong S. Breast implant illness: a biofilm hypothesis. Plast Reconstr Surg Glob Open 2020 Apr;8(4):e2755. [doi: 10.1097/GOX.0000000000002755] [Medline: 32440423]
53. Blog - Cancer.net. URL: https://www.cancer.net/blog [accessed 2021-01-05]
54. Blog - Living Beyond Breast Cancer. URL: https://www.lbbc.org/blog [accessed 2021-01-05]
55. Stories from patients with breast, lung and other cancers. Cancer Treatment Centers of America. URL: https://www.cancercenter.com/patient-stories [accessed 2021-01-05]
56. Chronic illness. Mighty Well Archives. URL: https://blog.mighty-well.com/category/chronic-illness/ [accessed 2021-01-05]
57. Practical pain management - symptoms, causes, treatments, medications for chronic pain. Remedy Health Media, LLC. URL: https://www.practicalpainmanagement.com/patients [accessed 2021-01-05]
58. Resources: for patients. Body Politic. URL: https://www.wearebodypolitic.com/resources [accessed 2021-01-05]
59. Coronavirus blog team. Medium. URL: https://medium.com/@coronavirus_blog_team [accessed 2021-01-05]
60. COVID-19 patient stories. Johns Hopkins Medicine. URL: https://www.hopkinsmedicine.org/coronavirus/patient-stories/ [accessed 2021-01-05]
61. About breast implant illness. Breast Implant Illness. URL: https://www.breastimplantillness.com/symptoms/ [accessed 2019-05-10]
62. Breast implant illness - symptoms, explant, surgeons, detox. Healing Breast Implant Illness. URL: https://healingbreastimplantillness.com [accessed 2019-05-10]
63. #breastimplantillness hashtag on Instagram. Instagram. URL: https://www.instagram.com/explore/tags/breastimplantillness [accessed 2019-09-05]
64. 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]
65. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):267-270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
66. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993-1022. [doi: 10.1016/B978-0-12-411519-4.00006-9]
67. Loper E, Bird S. NLTK: the Natural Language Toolkit. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. USA: Association for Computational Linguistics; 2002 Presented at: ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics; July 7, 2002; Philadelphia Pennsylvania p. 63-70. [doi: 10.3115/1118108.1118117]
68. Blei DM. C implementation of variational EM for latent Dirichlet Allocation (LDA). Github. 2013. URL: https://github.com/blei-lab/lda-c [accessed 2019-07-02]
69. Hofmann T. Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International Acm Sigir Conference on Research and Development in Information Retrieval. United States: Association for Computing Machinery; 1999 Presented at: SIGIR99: 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; August 15 - 19, 1999; Berkeley California USA p. 50-57. [doi: 10.1145/312624.312649]

70. Blei D, Lafferty J. Correlated topic models. In: Proceedings of the Advances in Neural Information Processing Systems. Cambridge, United States: MIT Press; 2006 Presented at: Advances in Neural Information Processing Systems; December 4-7, 2006; Vancouver, Canada. [doi: 10.5555/2976248.2976267]

71. Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei D. Reading tea leaves: how humans interpret topic models. In: Proceedings of the Advances in Neural Information Processing Systems. United States: Curran Associates Inc; 2009 Presented at: Advances in Neural Information Processing Systems; December 2009; Vancouver, Canada. [doi: 10.5555/2984093.2984126]

Abbreviations

ASCII: American Standard Code for Information Interchange
BII: breast implant illness
cTAKES: clinical Text Analysis and Knowledge Extraction System
CUI: concept unique identifier
LDA: latent Dirichlet allocation
NLP: natural language processing
UMLS: unified medical language system
XMI: XML metadata interchange

Edited by C Lovis; submitted 19.04.21; peer-reviewed by A Chen, V Foufi, S Wakamiya, T Timakum; comments to author 05.06.21; revised version received 31.07.21; accepted 23.09.21; published 29.11.21.

Please cite as:
Dey V, Krasniak P, Nguyen M, Lee C, Ning X
A Pipeline to Understand Emerging Illness Via Social Media Data Analysis: Case Study on Breast Implant Illness
JMIR Med Inform 2021;9(11):e29768
URL: https://medinform.jmir.org/2021/11/e29768
doi:10.2196/29768
PMID:34847064

Original Paper

Detecting Adverse Drug Events Through the Chronological Relationship Between the Medication Period and the Presence of Adverse Reactions From Electronic Medical Record Systems: Observational Study

Kei Teramoto[1,2*], MS; Toshihiro Takeda[1*], MD, PhD; Naoki Mihara[3], MD, PhD; Yoshie Shimai[1], PhD; Shirou Manabe[1], MS; Shigeki Kuwata[4], PhD; Hiroshi Kondoh[2], MD, PhD; Yasushi Matsumura[1], MD, PhD

[1]Department of Medical Informatics, Graduate School of Medicine, Osaka University, Suita, Japan

[2]Division of Medical Informatics, Tottori University Hospital, Yonago, Japan

[3]Department of Medical Informatics, National Cancer Center Hospital, Tokyo, Japan

[4]Department of Clinical Information Management, Nara City Hospital, Nara, Japan

[*]these authors contributed equally

Corresponding Author:
Toshihiro Takeda, MD, PhD
Department of Medical Informatics
Graduate School of Medicine
Osaka University
2-2 Yamada-oka
Suita, 565 0871
Japan
Phone: 81 6 6879 5900
Email: ttakeda@hp-info.med.osaka-u.ac.jp

Abstract

Background: Medicines may cause various adverse reactions. An enormous amount of money and effort is spent investigating adverse drug events (ADEs) in clinical trials and postmarketing surveillance. Real-world data from multiple electronic medical records (EMRs) can make it easy to understand the ADEs that occur in actual patients.

Objective: In this study, we generated a patient medication history database from physician orders recorded in EMRs, which allowed the period of medication to be clearly identified.

Methods: We developed a method for detecting ADEs based on the chronological relationship between the presence of an adverse event and the medication period. To verify our method, we detected ADEs with alanine aminotransferase elevation in patients receiving aspirin, clopidogrel, and ticlopidine. The accuracy of the detection was evaluated with a chart review and by comparison with the Roussel Uclaf Causality Assessment Method (RUCAM), which is a standard method for detecting drug-induced liver injury.

Results: The calculated rates of ADE with ALT elevation in patients receiving aspirin, clopidogrel, and ticlopidine were 3.33% (868/26,059 patients), 3.70% (188/5076 patients), and 5.69% (226/3974 patients), respectively, which were in line with the rates of previous reports. We reviewed the medical records of the patients in whom ADEs were detected. Our method accurately predicted ADEs in 90% (27/30patients) treated with aspirin, 100% (9/9 patients) treated with clopidogrel, and 100% (4/4 patients) treated with ticlopidine. Only 3 ADEs that were detected by the RUCAM were not detected by our method.

Conclusions: These findings demonstrate that the present method is effective for detecting ADEs based on EMR data.

(JMIR Med Inform 2021;9(11):e28763) doi:10.2196/28763

KEYWORDS

real world data; electronic medical record; adverse drug event

Introduction

The investigation of adverse events in clinical trials and postmarketing surveillance requires an enormous amount of money and effort [1-3]. As clinical trials are performed with limited numbers of participants and limited investigation periods, they do not always clearly identify the full range of possible adverse events [4-6]. Although postmarketing surveillance, which is executed by specialized agencies in many countries, has focused on gathering information on adverse drug events (ADEs), the identification of ADEs in actual clinical settings remains insufficient due to its dependence upon voluntary reporting [7-11]. The introduction of electronic medical records (EMRs) by many hospitals has allowed for the secondary use of EMR data from multiple hospitals [12-15]. This enables a greater understanding of the ADEs that occur in actual patients without the costs associated with the traditional methods of determining the incidence of adverse events.

The occurrence of ADEs can be detected based on the chronological relationship between the presence of the adverse event and the medication period. The key data for the detection of an ADE are the date when a patient started to take the medicine and the date on which the medication was discontinued. It is not easy to accurately determine the medication period based on patient records because the medication data obtained from EMRs are based on a computer physician order entry (CPOE) system in which prescription orders are created for each prescription. In the clinical setting, physicians usually consider the amount of remaining medicine due to missed doses or overlapping previous prescriptions when they are preparing the prescription order. In the present study, we developed a medication history database in which both the start and end dates of medication were determined by combining the prescription order data according to the estimated amount of remaining medicine. To verify our ADE detection method, we focused on identifying ADEs with alanine aminotransferase (ALT) elevation using the medication history database and the serum ALT values obtained from the EMR. The accuracy of the detection of ADEs was examined by a review of medical records and by comparison with the Roussel Uclaf Causality Assessment Method (RUCAM), which is a standard method for detecting drug-induced liver injury (DILI) [16-25].

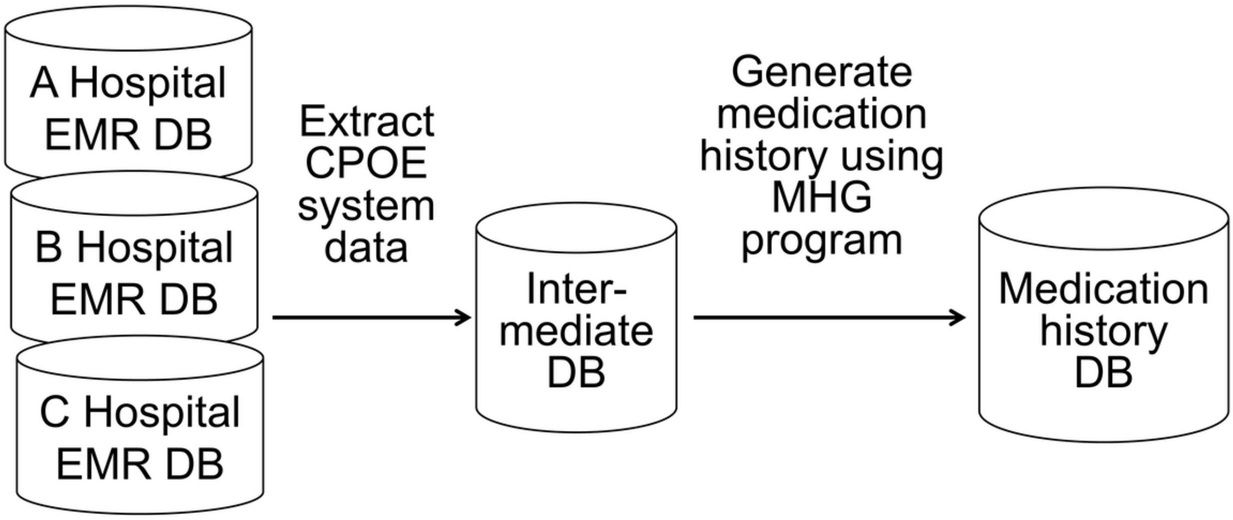
Methods

Experimental Environment

This study was performed in accordance with the World Medical Association Declaration of Helsinki, and the study protocol was approved by the institutional review board of the Osaka University Hospital (OUH), National Cerebral and Cardiovascular Center (NCVC), and Tottori University Hospital (TUH). This study was an observational study and did not obtain individual informed consent from the participants included in the study. However, the study protocol was posted on our webpage, giving the study participants an opportunity to opt out.

Because each CPOE system has its own database, the systems have different structures. We first developed an intermediate database to unify the database structure. The data from the original CPOE database were transferred to this intermediate database. We then generated the medication history by applying a medication history generation (MHG) program to the intermediate database (Figure 1). The medication history generation program was developed with Microsoft Visual Basic for Applications 7.0, and Microsoft Access 2010 was used for the intermediate database. Both the program and the database were installed on a laptop PC (Intel Core i7-2640M CPU; 8 GB of memory) with the Microsoft Windows operating system.

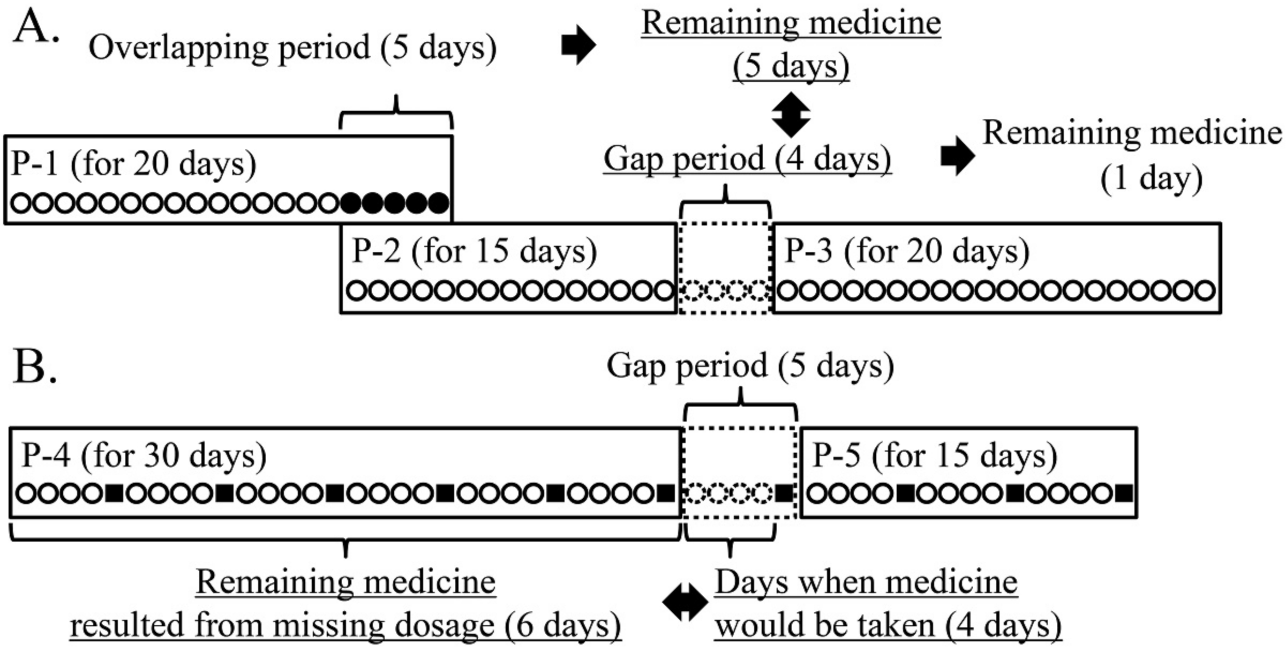
Figure 1. Procedure for generating the medication history. The data were extracted with an individually customized Structured Query Language from each CPOE database in the different medical facilities and transferred to an intermediate database. The MHG program was applied to the data in the intermediate database to generate the medication history. DB: database; CPOE: computer physician order entry; EMR: electronic medical record; MHG: medication history generation.



Generation of the Medication History

The medication history includes the start and the end dates of medication for each medicine prescribed to a patient. To construct the medication history database, the CPOE records were combined with consideration to the remaining medicine (Figure 2).

Figure 2. Process to generate a medication history. A. Generation of medication history with consideration to the overlapped period and gap period. The prescription order records (P-1, P-2, and P-3) were combined if the calculated remaining medicine was more than that needed for the days of the gap period. B. The medication history was generated under consideration of missing doses, assuming that missing doses occur once in 5 days. The prescription order records (P-4 and P-5) were combined if the amount of a remaining medicine was more than that needed for days of the gap period. Open circles indicate the days on which the patient took the medicine. Closed circles indicate the days in which the prescription orders overlapped. Closed squares indicate the days of missing doses.


A.
Overlapping period (5 days)
Remaining medicine (5 days)
Gap period (4 days)
Remaining medicine (1 day)
P-1 (for 20 days)
P-2 (for 15 days)
P-3 (for 20 days)
B.
Gap period (5 days)
P-4 (for 30 days)
P-5 (for 15 days)
Remaining medicine resulted from missing dosage (6 days)
Days when medicine would be taken (4 days)

First, the CPOE records for each medicine taken by an individual patient were extracted and combined sequentially from the oldest record to the newest record. As shown in Figure 2A, in cases where the last day of prescription 1 (P-1) was after the first day of P-2, the first day of P-1 was set as the start date of medication while the last day of P-2 was set as the end date, and the amount of the remaining medicine was estimated. In the case that a gap period lay between the last day of P-2 and the first day of P-3, if the amount of the remaining medicine was not less than the amount of medicine that would have been consumed during the gap period, P-2 and P-3 were combined.

We estimated the amount of the remaining medicines due to noncompliance by the patient, assuming that the rate of missed doses was constant. Accordingly, we set an unused medicine index (UMI), which indicated the rate of missing doses as a ratio of the period in which patients actually took the medicine to the prescription period. The amount of remaining medicine due to missing doses in P-4 was calculated (Figure 2B). If the amount of the remaining medicine was greater than that needed for the days during the gap period, P-4 and P-5 were combined. Figure 2 shows the algorithm for generating the medication history database.

The appropriate UMI value was determined by generating the medication history records for 9 medicines (pravastatin, cilostazol, isosorbide, nifedipine, ursodeoxycholic acid, rebamipide, amlodipine, aspirin, and methylcobalamin), which were considered to be long-term prescriptions. We set the UMI value as 1.4 because unnatural short-term gap periods tended to be observed when the UMI was <1.3 and because gap periods of a few months (considered to be the cessation of medicine) tended to be combined when the UMI was >1.5. To evaluate the validity of the UMI, we randomly selected 9 patients who had been treated for cardiovascular disease for >5 years and generated 725 medication history records. The medication history records were reviewed by the chief physicians, resulting in 98% of the records being considered appropriate.

Detection of ADEs with Serum ALT Elevation

We detected the occurrence of ADEs based on the chronological relationship between the presence of the adverse event and the medication period. In this study, we focused on identifying ADEs with ALT elevation, which is known to reflect hepatocellular injury-type DILI. The elevation of ALT was selected because, in the RUCAM, the severity of hepatocellular injury-type DILI is defined by serum ALT. The ALT values were obtained from the laboratory test data in the EMR database. The criteria for the diagnosis of ADE with ALT elevation are shown in Table 1. ADEs with ALT elevation were detected during the medication period, and those with a decrease in the ALT level were detected after the cessation of the medication (criteria 1 and 2). If the elevated value decreased during the medication period, then the medicine was considered not to be causative; thus, it was excluded as a cause of ADE (criterion 3). Because it was difficult to distinguish an ALT elevation caused by a previous liver injury, viral hepatitis, or an operation from hepatocellular injury-type DILI, we excluded patients with any of these factors (criterion 4-III).

Table 1. Criteria for the diagnosis of hepatocellular injury with ALT elevation.

Criterion	Criterion details
Inclusion criteria	
Elevation of ALT[a] after initiation of medication	Peak ALT[b]> ULN[c] of ALT and Peak ALT ≥ ALT (before start of medication)[d] × 2
Decrease of ALT after cessation of medication	ALT (after cessation of medication)[e] < Max ALT [f] × 0.5 or ALT (after cessation of medication) < ALT (ULN)
Exclusion criteria	
Decrease of ALT during the medication period but after the day of peak ALT	ALT (during medication period)[g] < peak ALT × 0.25 or ALT (during medication period) < ULN of ALT
Liver injury induced by nondrug causes	Previous liver injury[h], viral hepatitis[i], or surgical operation[j]

[a] ALT: alanine aminotransferase.

[b] Highest ALT value within 90 days from the start of medication.

[c] ULN: upper limit of normal.

[d] ALT value on the last day before the initiation of medication.

[e] Lowest ALT value within 30 days after the cessation of medication.

[f] Maximum ALT value during the medication period.

[g] Lowest ALT value during the medication period from the day of peak ALT to within 30 days from the date of medication cessation.

[h] Patients whose electronic medical records showed the following diseases (International Classification of Disease code 10): alcohol dependence (F10), liver disease (K70-K77), and gallbladder and bible duct disease (K80-K87).

[i] Patients whose electronic medical records showed positive results in the following laboratory blood tests: viral hepatitis A, B, and C (immunoglobulin M antibody to hepatitis A virus antigen, hepatitis B surface antigen, hepatitis C virus core antigen); cytomegalovirus; and Epstein-Barr virus.

[j] Patients whose EMRs indicated that they had undergone surgery within 14 days before the day of peak ALT.

The Study Population and the Target Medicines

In the present study, EMR data were obtained from 3 medical facilities: OUH, NCVC, and TUH. These medical institutions have independent EMR systems. In the study period, the data from a total of 1,587,939 patients were registered, and the total number of CPOE records was 37,935,783 (an average of 23.9 records per patient; Table 2).

Table 2. The medical facilities in the present study.

Characteristic	OUH[a]	NCVC[b]	TUH[c]
Manufacturer	NEC Corp.	NEC Corp.	IBM Corp.
CPOE[d] database model	Oracle	Oracle	Database 2
Data range (mm/dd/yy)	04/01/00-12/01/12	04/01/00-02/01/14	0/01/03-09/01/13
Patients, n	1,028,852	251,143	307,944
CPOE records, n	20,447,443	8,128,059	9,360,281

[a] OUH: Osaka University Hospital.

[b] NCVC: National Cerebral and Cardiovascular Center.

[c] TUH: Tottori University Hospital.

[d] CPOE; computer physician order entry.

The target medicines were aspirin, clopidogrel, and ticlopidine. These are antiplatelet drugs that have been reported to cause hepatocellular injury-type DILI [26-28]. Earlier studies have suggested that clopidogrel is associated with a lower risk of hepatocellular injury-type DILI in comparison to ticlopidine [29].

The Rates of ADE With ALT Elevation With Each Target Medicine

To calculate the rates of ADE with ALT elevation that occurred with each medicine, we counted the number of patients who met the diagnostic criteria (Table 1). The severity of ADE with ALT elevation was categorized according to the maximum ALT value as mild elevation (maximum ALT ≥40 IU/L), moderate elevation (maximum ALT ≥80 IU/L), and severe elevation (maximum ALT ≥200 IU/L). The rate of ADEs with ALT elevation was calculated by dividing the number of ADE patients by the number of patients who took the targeted medicine, and the ALT values were tested at least 3 times (before, during, and after the medication period).

Evaluating Results That Were Indicative of ADE With ALT Elevation.

We selected the patients with moderate and severe ALT elevation (maximum ALT ≥80 IU/L) whose medical records were recorded electronically at OUH and TUH and checked the progress notes recorded from 3 days before to 3 days after the date of the peak ALT value. The numbers of medical records subjected to review for each of the drugs were as follows: aspirin (n=83), clopidogrel (n=29), and ticlopidine (n=8). These records were used to determine whether or not the elevation of ALT was due to an ADE. The ADE cases were categorized into 3 groups: (1) ADE caused by the targeted medicine, (2) ADE caused by a concomitant medicine, and (3) offending medicine not identified.

Comparison of the Detection of ADE With ALT Elevation Between Our Proposed Method and the RUCAM

The RUCAM is the standard method for detecting DILI. The RUCAM uses a 5-stage scoring system to assess the possibility of DILI by classifying the condition as hepatocellular, cholestatic, or mixed based on the laboratory test data and clinical data.

We compared the accuracy of detecting hepatocellular-type ADE between our method and the RUCAM. Patients with ALT levels of >200 were included in the analysis (10,608 patients from OUH and 5464 patients from TUH).

The primary screening was performed to select hepatocellular-type ADE for the RUCAM. The screening criterion was as follows: ALT level >200 and (ALT/upper limit of normal/(alkaline phosphatase/upper limit of normal)>5 within 90 days of the first day of using the verified medication. Next, we determined the RUCAM score based on a review of medical records. Probable and highly probable scores according to the RUCAM system were classified as hepatocellular-type DILI in this study.

Statistical Analysis

Multiple comparisons were performed using the Ryan method, and the Fisher exact test was used to compare the rates of ADE. *P* values of <0.05 were considered to indicate statistical significance. All statistical analyses were performed using the R software version 3.1.2 (The R Foundation for Statistical Computing).

Results

Table 3 shows a summary of the medication history records for the target medicines that were generated by our system. Aspirin was the most frequently used medication in our study population. The numbers of patients who were treated with clopidogrel and ticlopidine were approximately equal. The CPOE records were combined into a single medication history record in 8.80% (58,873/668,765), 13.81% (12,224/88,520 patients), and 8.51% (8654/104,003) of the patients treated with aspirin, clopidogrel, and ticlopidine, respectively, which indicated that the medication histories were correctly generated.

Table 3. The medication histories generated for the target medicines (N=1,587,939).

Values	Aspirin	Clopidogrel	Ticlopidine
Patients, n (%[a])	40,938 (2.58)	10,263 (0.65)	6224 (0.39)
CPOE[b] records, n	668,765	88,520	104,003
CPOE records per patient, mean	16.3	8.6	16.7
Medication history records, n	58,873	12,224	8,854
Medication history records per patient, mean	1.4	1.2	1.4

[a] Percentage of the study population treated with the target medicine/electronic medical record–registered population (1,5879,939 patients).

[b] CPOE: computer physician order entry.

The rate of ADEs with ALT elevation among patients who received ticlopidine was significantly higher than that among patients who received the other 2 medicines (Table 4). The rates of ADE with ALT elevation in patients who received aspirin and clopidogrel did not differ to a statistically significant extent. The rates of severe ALT elevation with each of the target medicines showed the same tendency.

We reviewed the medical records of the patients in whom an ADE with ALT elevation was detected by our system (Table 5). The number of records subjected to review for each of the drugs was 83 for aspirin, 29 for clopidogrel, and 8 for ticlopidine. The number of records in which the cause of liver injury was described was 30 for aspirin, 9 for clopidogrel, and 4 for ticlopidine. Among these, the number of records in which an ADE with ALT elevation was diagnosed was 27 (90%) for aspirin, 9 (100%) for clopidogrel, and 4 (100%) for ticlopidine. These findings demonstrated that the method of the present study was appropriate for detecting ADE with ALT elevation. However, the causative medicines of ADEs with ALT elevation described in the medical records were not only the target medicine but also concomitant medicines. There were cases in which the offending medicine was not specified. In the cases in which the concomitant medicine was described as the causative medicine of an ADE with ALT elevation, the target medicine was also thought to be a candidate based on the chronological pattern of the medication period and ALT elevation. This may be due to physicians suspecting an ADE and then discontinuing all of the possible causative medicines.

Table 4. The rates of adverse drug events with ALT elevation.

Patients	Aspirin	Clopidogrel	Ticlopidine
Target patient distribution, n	26,059	5076	3974
DILI[a] patients			
(MAX[b] ALT[c] >ULN[c])	868 (3.33%)	188 (3.70%)	226 (5.69%)[e]
MAX ALT ≥ 80 IU/L	341 (0.95%)	69 (0.93%)	83 (1.43%)[e]
MAX ALT ≥ 200 IU/L	93 (0.36%)	22 (0.43%)	26 (0.65%)[f]

[a] DILI: drug-induced liver injury.

[b] MAX: maximum.

[c] ALT: alanine aminotransferase.

[d] ULN: upper limit of normal.

[e] *P*<.001 vs other groups.

[f] *P*<.001 vs Aspirin.

Table 5. Evaluation by review of medical records.

Medical record values	Aspirin	Clopidogrel	Ticlopidine
ADEs[a] with ALT[b] elevation, n	27	9	4
Caused by target medicine	8	6	1
Caused by concomitant medicine	11	1	2
Offending medicine not specified	8	2	1
Other causes of liver injury, n	3	0	0
Total, n	30	9	4

[a] ADE: adverse drug event.

[b] ALT: alanine aminotransferase.

The number of patients diagnosed with hepatocellular-type ADE with our proposed method and the RUCAM are shown in Table 6. The first RUCAM screening identified 10 patients at OUH and 39 patients at TUH as candidates of hepatocellular-type ADE. The number of candidate patients was very few at OUH because the testing rate of alkaline phosphatase (ALP) within 90 days from starting the medication was very low (882/16,735, 5.26%) for OUH. As a result, none of the patients were suspected as hepatocellular-type ADE at OUH, while 51 patients were suspected as hepatocellular-type ADE by our method. On the other hand, the rate of ALP testing within 90 days from starting the medication was not low at TUH (6692/9097, 73.56%). At TUH, 11 patients were detected as DILI by both our method and the RUCAM. Two patients were detected as hepatocellular-type ADE only by our method, and both patients were thought to be hepatocellular-type ADE by the review of medical records. Three patients were not detected as hepatocellular-type ADE by our method because the ALT levels of these patients did not recover within 30 days of termination of the medication (within 33 days, 40 days, and 45 days, respectively).

Table 6. ADE with alanine aminotransferase level elevation detection results by RUCAM and the proposed method.

Values	Aspirin		Clopidogrel		Ticlopidine		Total	
	OUH[a]	TUH[b]	OUH	TUH	OUH	TUH	OUH	TUH
Target patients	7611	4002	1266	951	1731	511	10,608	5464
RUCAM[c]								
First[d] screening	5	28	3	9	2	2	10	39
ADE[e,f]	0	10	0	3	0	1	0	14
ADE medication history[g]	26	10	7	2	18	1	51	13

[a] OUH: Osaka University Hospital.

[b] TUH: Tottori University Hospital.

[c] RUCAM: Roussel Uclaf Causality Assessment Method.

[d] Alanine aminotransferase level >200 and (alkaline phosphatase /200)/(alanine aminotransferase/40) <5.

[e] ADE: adverse drug event.

[f] The number of patients diagnosed with "probable" suspected of drug-induced liver injury or with a degree greater than "probable" by the RUCAM (alanine aminotransferase ALT level >200) includes first screening patients.

[g] The number of patients diagnosed with an ADE by the proposed method (alanine aminotransferase level >200).

Discussion

Principal Findings

Accurate demonstration of the start and end dates of a medication period is important in pharmacoepidemiologic research. However, the CPOE records in EMRs cannot clearly demonstrate the total duration of the medication period. In the present study, we generated a medication history database from the CPOE databases of 3 hospitals and systematically diagnosed ADEs with ALT elevation according to the chronological relationship between the changes in ALT values and the duration of medication using a medication history database. Because the medication history database can be applied not only to the detection of ADEs but also to crossover studies that compare drug efficacy in the same patients, it can become a basis for pharmacoepidemiologic research.

The comparison of the RUCAM and our method revealed that the rates of ALT and ALP testing influenced the accuracy of the RUCAM in the detection of ADEs. In a prospective study, laboratory test data can be obtained according to a research plan. However, in a retrospective study, missing data often become problematic. Scoring in the RUCAM requires information such as the use of concomitant medications, drug risk information, the presence or absence of a rechallenge, and the history of alcohol consumption. This information is not registered as structured data in EMRs. In this study, a review of medical records was needed to determine the score for the RUCAM. In contrast, our method used only standardized data, such as laboratory test data, prescription data, disease name data, and surgical data. For this reason, our method is applicable to the detection of ADEs in a retrospective analysis of big data generated by EMRs.

The population characteristics greatly affect the rate of adverse events. In clinical trials, the incidence of adverse events may be accurate because blood testing is routinely performed in all patients. On the other hand, in observational studies, the timing of blood testing differs for each patient. There may be great differences in the rates of adverse events depending on how the study population is defined. A previous clinical study in Japan reported that the rates of serious liver injury among patients receiving ticlopidine and clopidogrel were 13.6% (129/948) and 5.1% (115/2261), respectively [30,31]. However, these studies had different study populations, and caution must be exercised when interpreting the comparison of the rates of adverse events. The present method determined the rates of adverse events for some medicines under the same conditions for ticlopidine (188/5076 ,3.70%) and clopidogrel (226/3974, 5.69%); thus, this method could be used to compare the risk of adverse events between medicines (ticlopidine therapy is associated with a greater risk of developing ADEs in comparison to clopidogrel).

When physicians suspect an ADE with ALT elevation, all of the medicines that might have caused the ADE are likely to be discontinued. Thus, it was difficult to differentiate the causative medicine from the concomitant medicines using our method. Our method demonstrated the maximum rate of ADEs with ALT elevation induced by a targeted medicine, assuming that the targeted medicine was the causative medicine in all cases.

Although aspirin has been reported as a cause of liver injury, the rate in Asian populations remains unclear. According to the clopidogrel versus aspirin in patients at risk of ischemic events (CAPRIE) Steering Committee report, the rates of liver injury in patients receiving aspirin and clopidogrel were 2.97% (285/9599) and 3.15% (302/9586), respectively, which are in line with the rates obtained in the present study (aspirin: 868/26,059, 3.33%; clopidogre: 188/5076, 3.70%) [32]. The rates of severe liver injury in the same report were 0.19% for aspirin (93/5076, 0.36% in this study) and 0.11% for clopidogrel (22/5076, 0.43% in this study). Similar to our study, the rates of severe liver injury did not differ between patients using aspirin and those using clopidogrel.

Even though the absolute risk of a medicine is difficult to estimate, our method can estimate the upper limit of the risk. Furthermore, for some medicines, our method can estimate the risk of for ADE with ALT elevation one at a time under the same conditions, and the risk can be compared among different medicines.

Limitations

In this study, we used the medication history database created from CPOE records to detect DILI, but we did not detect all cases of DILI. First, we focused on elevated serum ALT levels. Elevated serum ALT can capture hepatocellular-type DILI, but it may not detect cholestatic-type DILI, which is characterized by elevation of the serum ALP level. Second, we were not able to detect DILI that did not meet our diagnostic criteria, such as delayed DILI, even the hepatocellular-type DILI. This type of detection requires a different set of criteria.

Conclusions

The generation of a medication history database enabled us to detect ADEs with ALT elevation through the chronological relationship between the medication period and occurrence of liver injury. As our method used only standardized data from EMRs, it was possible to analyze real-world data accumulated by EMRs in multiple hospitals. Although our method could not identify the causative medicine among concomitant medicines, it was possible to compare the risk of ADEs for different medicines.

Acknowledgments

This work was supported by the Japan Society for the Promotion of Science KAKENHI (grant #16K09171).

Conflicts of Interest

None declared.

References

1. Vouk K, Benter U, Amonkar MM, Marocco A, Stapelkamp C, Pfersch S, et al. Cost and economic burden of adverse events associated with metastatic melanoma treatments in five countries. J Med Econ 2016 Sep;19(9):900-912. [doi: 10.1080/13696998.2016.1184155] [Medline: 27123564]
2. Stark RG, John J, Leidl R. Health care use and costs of adverse drug events emerging from outpatient treatment in Germany: a modelling approach. BMC Health Serv Res 2011 Jan 13;11:9 [FREE Full text] [doi: 10.1186/1472-6963-11-9] [Medline: 21232111]
3. Arondekar B, Curkendall S, Monberg M, Mirakhur B, Oglesby AK, Lenhart GM, et al. Economic burden associated with adverse events in patients with metastatic melanoma. J Manag Care Spec Pharm 2015 Feb;21(2):158-164. [doi: 10.18553/jmcp.2015.21.2.158] [Medline: 25615005]
4. Lasser KE, Allen PD, Woolhandler SJ, Himmelstein DU, Wolfe SM, Bor DH. Timing of new black box warnings and withdrawals for prescription medications. JAMA 2002 May 01;287(17):2215-2220. [Medline: 11980521]
5. Jefferys DB, Leakey D, Lewis JA, Payne S, Rawlins MD. New active substances authorized in the United Kingdom between 1972 and 1994. Br J Clin Pharmacol 1998 Feb;45(2):151-156 [FREE Full text] [doi: 10.1046/j.1365-2125.1998.00651.x] [Medline: 9491828]
6. Frank C, Himmelstein DU, Woolhandler S, Bor DH, Wolfe SM, Heymann O, et al. Era of faster FDA drug approval has also seen increased black-box warnings and market withdrawals. Health Aff (Millwood) 2014 Aug;33(8):1453-1459. [doi: 10.1377/hlthaff.2014.0122] [Medline: 25092848]
7. Wysowski DK, Swartz L. Adverse drug event surveillance and drug withdrawals in the United States, 1969-2002: the importance of reporting suspected reactions. Arch Intern Med 2005 Jun 27;165(12):1363-1369. [doi: 10.1001/archinte.165.12.1363] [Medline: 15983284]
8. Simone LK, Brumbaugh J, Ricketts C. Medical devices, the FDA, and the home healthcare clinician. Home Healthc Nurse 2014;32(7):402-408. [doi: 10.1097/NHH.0000000000000107] [Medline: 24978574]
9. Hojo T. Regulatory science in practice (Pharmaceuticals and Medical Devices Agency). Yakugaku Zasshi 2017;137(4):439-442 [FREE Full text] [doi: 10.1248/yakushi.16-00244-4] [Medline: 28381721]
10. Moore TJ, Furberg CD, Mattison DR, Cohen MR. Completeness of serious adverse drug event reports received by the US Food and Drug Administration in 2014. Pharmacoepidemiol Drug Saf 2016 Jun;25(6):713-718. [doi: 10.1002/pds.3979] [Medline: 26861066]
11. Hazell L, Shakir SAW. Under-reporting of adverse drug reactions : a systematic review. Drug Saf 2006;29(5):385-396. [Medline: 16689555]
12. Shin J, Hunt CM, Suzuki A, Papay JI, Beach KJ, Cheetham TC. Characterizing phenotypes and outcomes of drug-associated liver injury using electronic medical record data. Pharmacoepidemiol Drug Saf 2013 Feb;22(2):190-198. [doi: 10.1002/pds.3388] [Medline: 23258383]
13. Dandala B, Joopudi V, Tsou C, Liang JJ, Suryanarayanan P. Extraction of information related to drug safety surveillance from electronic health record notes: joint modeling of entities and relations using knowledge-aware neural attentive models. JMIR Med Inform 2020 Jul 10;8(7):e18417 [FREE Full text] [doi: 10.2196/18417] [Medline: 32459650]

14. 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] [Medline: 29695376]
15. Ujiie S, Yada S, Wakamiya S, Aramaki E. Identification of adverse drug event-related Japanese articles: natural language processing analysis. JMIR Med Inform 2020 Nov 27;8(11):e22661 [FREE Full text] [doi: 10.2196/22661] [Medline: 33245290]
16. Onakpoya IJ, Heneghan CJ, Aronson JK. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. BMC Med 2016 Feb 04;14:10 [FREE Full text] [doi: 10.1186/s12916-016-0553-2] [Medline: 26843061]
17. Hassan A, Fontana RJ. The diagnosis and management of idiosyncratic drug-induced liver injury. Liver Int 2019 Jan;39(1):31-41. [doi: 10.1111/liv.13931] [Medline: 30003672]
18. Leise MD, Poterucha JJ, Talwalkar JA. Drug-induced liver injury. Mayo Clin Proc 2014 Jan;89(1):95-106. [doi: 10.1016/j.mayocp.2013.09.016] [Medline: 24388027]
19. Nathwani RA, Pais S, Reynolds TB, Kaplowitz N. Serum alanine aminotransferase in skeletal muscle diseases. Hepatology 2005 Feb;41(2):380-382. [doi: 10.1002/hep.20548] [Medline: 15660433]
20. Green RM, Flamm S. AGA technical review on the evaluation of liver chemistry tests. Gastroenterology 2002 Oct;123(4):1367-1384. [doi: 10.1053/gast.2002.36061] [Medline: 12360498]
21. Rockey DC, Seeff LB, Rochon J, Freston J, Chalasani N, Bonacini M, US Drug-Induced Liver Injury Network. Causality assessment in drug-induced liver injury using a structured expert opinion process: comparison to the Roussel-Uclaf causality assessment method. Hepatology 2010 Jun;51(6):2117-2126 [FREE Full text] [doi: 10.1002/hep.23577] [Medline: 20512999]
22. Rochon J, Protiva P, Seeff LB, Fontana RJ, Liangpunsakul S, Watkins PB, Drug-Induced Liver Injury Network (DILIN). Reliability of the Roussel Uclaf Causality Assessment Method for assessing causality in drug-induced liver injury. Hepatology 2008 Oct;48(4):1175-1183 [FREE Full text] [doi: 10.1002/hep.22442] [Medline: 18798340]
23. Chalasani NP, Hayashi PH, Bonkovsky HL, Navarro VJ, Lee WM, Fontana RJ, Practice Parameters Committee of the American College of Gastroenterology. ACG Clinical Guideline: the diagnosis and management of idiosyncratic drug-induced liver injury. Am J Gastroenterol 2014 Jul;109(7):950-66; quiz 967. [doi: 10.1038/ajg.2014.131] [Medline: 24935270]
24. Benichou C, Danan G, Flahault A. Causality assessment of adverse reactions to drugs—II. An original model for validation of drug causality assessment methods: case reports with positive rechallenge. J Clin Epidemiol 1993 Nov;46(11):1331-1336. [doi: 10.1016/0895-4356(93)90102-7] [Medline: 8229111]
25. Cheetham TC, Lee J, Hunt CM, Niu F, Reisinger S, Murray R, et al. An automated causality assessment algorithm to detect drug-induced liver injury in electronic medical record data. Pharmacoepidemiol Drug Saf 2014 Jun;23(6):601-608. [doi: 10.1002/pds.3531] [Medline: 24920207]
26. Pisapia R, Abdeddaim A, Mariano A, Rianda A, Vincenzi L, Taibi C, et al. Acute hepatitis associated with clopidogrel: a case report and review of the literature. Am J Ther 2015;22(1):e8-e13. [doi: 10.1097/MJT.0b013e318293b0d6] [Medline: 23846525]
27. Motola D, Biagi C, Leone R, Venegoni M, Lapi F, Cutroneo P, et al. Ticlopidine safety profile: a case/non-case study on the basis of the spontaneous ADRs reporting in Italy. Curr Drug Saf 2012 Apr;7(2):99-105. [doi: 10.2174/157488612802715717] [Medline: 22873494]
28. Laster J, Satoskar R. Aspirin-induced acute liver injury. ACG Case Rep J 2014 Oct;2(1):48-49 [FREE Full text] [doi: 10.14309/crj.2014.81] [Medline: 26157904]
29. Shigematsu H, Komori K, Tanemoto K, Harada Y, Nakamura M. Clopidogrel for Atherothrombotic Event Management in Patients with Peripheral Arterial Disease (COOPER) study: safety and efficacy of clopidogrel versus ticlopidine in Japanese patients. Ann Vasc Dis 2012;5(3):364-375 [FREE Full text] [doi: 10.3400/avd.oa.12.00039] [Medline: 23555538]
30. Danan G, Benichou C. Causality assessment of adverse reactions to drugs—I. A novel method based on the conclusions of international consensus meetings: application to drug-induced liver injuries. J Clin Epidemiol 1993 Nov;46(11):1323-1330. [doi: 10.1016/0895-4356(93)90101-6] [Medline: 8229110]
31. K. Sanofi K. URL: https://e-mr.sanofi.co.jp/ja-JP/-/media/EMS/Conditions/eMR/di/interview/plavix.pdf [accessed 2021-07-16]
32. CAPRIE Steering Committee. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). CAPRIE Steering Committee. Lancet 1996 Nov 16;348(9038):1329-1339. [doi: 10.1016/s0140-6736(96)09457-3] [Medline: 8918275]

Abbreviations

ADEs: adverse drug events
ALP: alkaline phosphatase
ALT: alanine aminotransferase
CAPRIE: clopidogrel versus aspirin in patients at risk of ischemic events
CPOE: computer physician order entry

DILI: drug-induced liver injury
EMR: electronic medical record
MHG: medication history generation
NCVC: National Cerebral and Cardiovascular Center
OUH: Osaka University Hospital
RUCAM: Roussel Uclaf Causality Assessment Method
TUH: Tottori University Hospital
UMI: unused medicine index

Edited by G Eysenbach; submitted 13.03.21; peer-reviewed by R Rubin, T Ashihara; comments to author 06.04.21; revised version received 16.04.21; accepted 17.04.21; published 01.11.21.

Please cite as:
Teramoto K, Takeda T, Mihara N, Shimai Y, Manabe S, Kuwata S, Kondoh H, Matsumura Y
Detecting Adverse Drug Events Through the Chronological Relationship Between the Medication Period and the Presence of Adverse Reactions From Electronic Medical Record Systems: Observational Study
JMIR Med Inform 2021;9(11):e28763
URL: https://medinform.jmir.org/2021/11/e28763
doi:10.2196/28763
PMID:33993103

Original Paper

Toward Personalized Web-Based Cognitive Rehabilitation for Patients With Ischemic Stroke: Elo Rating Approach

Alejandro Garcia-Rudolph[1,2,3], PhD; Eloy Opisso[1,2,3], PhD; Jose M Tormos[1,2,3], PhD; Vince Istvan Madai[4,5,6], PhD; Dietmar Frey[4], PhD; Helard Becerra[7], PhD; John D Kelleher[8], PhD; Montserrat Bernabeu Guitart[1,2,3], MD; Jaume López[1,2,3], MSc

[1]Institut Guttmann Hospital de Neurorehabilitacio, Badalona, Spain
[2]Universitat Autònoma de Barcelona, Barcelona, Spain
[3]Fundació Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol, Badalona, Spain
[4]Charité Lab for AI in Medicine, Charité Universitätsmedizin, Berlin, Germany
[5]QUEST Center for Transforming Biomedical Research, Berlin Institute of Health (BIH), Berlin, Germany
[6]Faculty of Computing, Engineering and the Built Environment, School of Computing and Digital Technology, Birmingham City University, Birmingham, United Kingdom
[7]School of Computer Science, University College Dublin, Dublin, Ireland
[8]Information, Communication and Entertainment Research Institute, Technological University Dublin, Dublin, Ireland

Corresponding Author:
Alejandro Garcia-Rudolph, PhD
Institut Guttmann Hospital de Neurorehabilitacio
Cami de Can Ruti s/n
Badalona
Spain
Phone: 34 934 97 77 00
Email: alejandropablogarcia@gmail.com

Abstract

Background: Stroke is a worldwide cause of disability; 40% of stroke survivors sustain cognitive impairments, most of them following inpatient rehabilitation at specialized clinical centers. Web-based cognitive rehabilitation tasks are extensively used in clinical settings. The impact of task execution depends on the ratio between the skills of the treated patient and the challenges imposed by the task itself. Thus, treatment personalization requires a trade-off between patients' skills and task difficulties, which is still an open issue. In this study, we propose Elo ratings to support clinicians in tasks assignations and representing patients' skills to optimize rehabilitation outcomes.

Objective: This study aims to stratify patients with ischemic stroke at an early stage of rehabilitation into three levels according to their Elo rating; to show the relationships between the Elo rating levels, task difficulty levels, and rehabilitation outcomes; and to determine if the Elo rating obtained at early stages of rehabilitation is a significant predictor of rehabilitation outcomes.

Methods: The PlayerRatings R library was used to obtain the Elo rating for each patient. Working memory was assessed using the DIGITS subtest of the Barcelona test, and the Rey Auditory Verbal Memory Test (RAVLT) was used to assess verbal memory. Three subtests of RAVLT were used: RAVLT learning (RAVLT075), free-recall memory (RAVLT015), and recognition (RAVLT015R). Memory predictors were identified using forward stepwise selection to add covariates to the models, which were evaluated by assessing discrimination using the area under the receiver operating characteristic curve (AUC) for logistic regressions and adjusted R^2 for linear regressions.

Results: Three Elo levels (low, middle, and high) with the same number of patients (n=96) in each Elo group were obtained using the 50 initial task executions (from a total of 38,177) for N=288 adult patients consecutively admitted for inpatient rehabilitation in a clinical setting. The mid-Elo level showed the highest proportions of patients that improved in all four memory items: 56% (54/96) of them improved in DIGITS, 67% (64/96) in RAVLT075, 58% (56/96) in RAVLT015, and 53% (51/96) in RAVLT015R (*P*<.001). The proportions of patients from the mid-Elo level that performed tasks at difficulty levels 1, 2, and 3 were 32.1% (3997/12,449), 31.% (3997/12,449), and 36.9% (4595/12,449), respectively (*P*<.001), showing the highest match between skills (represented by Elo level) and task difficulties, considering the set of 38,177 task executions. Elo ratings were significant predictors in three of the four models and quasi-significant in the fourth. When predicting RAVLT075 and DIGITS at discharge, we obtained R^2=0.54 and 0.43, respectively; meanwhile, we obtained AUC=0.73 (95% CI 0.64-0.82) and AUC=0.81 (95% CI 0.72-0.89) in RAVLT075 and DIGITS improvement predictions, respectively.

Conclusions: Elo ratings can support clinicians in early rehabilitation stages in identifying cognitive profiles to be used for assigning task difficulty levels.

(JMIR Med Inform 2021;9(11):e28090) doi:10.2196/28090

KEYWORDS

cognitive rehabilitation; Elo rating; predictors; stroke rehabilitation; web-based tasks

Introduction

Background

Stroke is currently considered one of the top global causes of disability, with most survivors of stroke in need of inpatient rehabilitation at specialized clinical centers [1]. Recent studies have reported that almost 40% of survivors of stroke sustain cognitive impairment [2]. The World Health Organization definition of cognitive impairment has been recently referred [3] to as *problems experienced by an individual in remembering things, making decisions, learning abilities or concentrating on tasks that affect their everyday life*.

Cognitive rehabilitation (neuropsychological rehabilitation) relies on brain plasticity to induce neuroplastic changes to compensate for cognitive impairments [4]. Brain injury is one of the key causes of cognitive impairment; however, other factors contribute to the ever-increasing number of people in need of cognitive rehabilitation (one of them being the global trend in population aging).

One of the most frequent cognitive problems reported by poststroke patients in their daily lives is related to memory loss [5,6]. To date, associations between factors for ischemic stroke and clinical outcomes have been analyzed predominantly in older rather than younger patients [7]; however, the incidence rates of ischemic stroke have increased in young adults in the United States [8] and also in Europe [9].

New strategies for providing cognitive rehabilitation services are constantly required and are continuously being integrated into clinical practice [10]. One such strategy is the use of web-based systems, and several of these systems have already been used to optimize cognitive interventions [11,12]. However, because of the relatively recent development of these services, the best strategies to integrate them into everyday clinical practice are still unclear [13]. Nevertheless, strategies targeting the personalization of the proposed activities for patients according to their specific needs appear to be more effective [14].

A typical cognitive rehabilitation program mainly provides exercises that require repetitive use of the impaired cognitive system in a progressively more demanding [15] sequence of tasks. The impact of a task or exercise execution depends on the ratio between the skills of the treated patient and the challenges involved in the execution of the task itself. Thus, determining the correct training schedule requires a quite precise trade-off between sufficient stimulation and sufficiently achievable tasks, which is far from trivial and is still an open issue, both empirically and theoretically [16,17].

Furthermore, prediction of specific outcomes after stroke rehabilitation is used by clinicians to improve the accuracy of prognoses, set attainable goals, reach shared decisions, personalize rehabilitation plans, and inform patients and relatives [18].

In this study, we propose the application of Elo ratings to provide clinicians with a ranking of patients at an early stage of cognitive rehabilitation by using the results of web-based cognitive rehabilitation tasks. We hypothesize that (1) such ranking of patients will allow clinicians to match patient’s skills with task difficulties, thereby enabling better treatment personalization, and (2) such a rating will be a significant predictor of patients’ outcomes for memory cognitive function. The original proposal of the Elo rating system was designed to rate chess players, and the rating system was named after its creator Arpad Elo [19].

The Elo system works as follows: an initial rating is assigned to each player every time a player plays a match. This rating is updated for both players depending on the result of the match. If the winner is the player with the higher rating, the update is small, and it is larger depending on how unexpected the victory is, according to their previous ratings [20].

The basic Elo rating system is used in several types of contests beyond chess, for example, football [21]; however, different applications have been extensively reported elsewhere. It has been used for eliciting user preferences in community-based sites [22], assessing security and vulnerability risks [23], ranking posts in web-based forums [24], rating patterns in videogames [25], detecting fabric defects in the textile industry [26], providing students with individualized learning materials in educational settings [20], studying traffic congestion in urban transportation [27], studying dominance hierarchies in behavioral and evolutionary animal ecology [28], forecasting sales and optimizing prices of new product releases [29], allocating resources for criminal justice to support supervision officers [30], and identifying people using facial comparative descriptions [31].

Nevertheless, to the best of our knowledge, Elo ratings have not been applied in cognitive rehabilitation in general or in the specific use–case of a web-based application where patients perform web-based cognitive tasks during their rehabilitation period.

Objectives

In this study, we propose that instead of considering matches between, for example, chess players, we consider matches between patients and web-based cognitive rehabilitation tasks.

The aims of this study are (1) to demonstrate the feasibility of the approach by presenting a synthetic data set where we obtain an Elo rating for each patient by considering each execution of a cognitive rehabilitation task by the patient as a match between the patient and the task; (2) to obtain the Elo rating of each patient in a real rehabilitation setting where adult patients with ischemic stroke follow cognitive rehabilitation by executing web-based rehabilitation tasks and use these Elo ratings to perform a stratification of patients into 3 groups according to their Elo rating (low, middle, and high); (3) to analyze the relationship among the three Elo rating levels and the proportion of tasks executed at three increasing difficulty levels (1, 2, and 3) with the rehabilitation outcomes in the memory cognitive function; and (4) to develop and internally validate four predictive models for auditory verbal learning memory and working memory outcomes using Elo ratings obtained at early stages of rehabilitation as independent variables and state-of-the-art variables (eg, sex, age, and length of stay). The first two models are developed for predicting auditory verbal learning memory and working memory at discharge and the other two for predicting improvements in auditory verbal learning memory and working memory at discharge.

Methods

Participants and Clinical Setting

The setting was the inpatient acquired brain injury rehabilitation unit of the Institut Guttmann hospital, a specialized clinical center certified in quality of care and patient safety (Joint Commission International since 2005 and consecutively recertified in 2009, 2012, and 2018). The initial study population consisted of 344 patients with ischemic stroke who were consecutively admitted for inpatient rehabilitation from March 2009 to September 2019. Patients were included in the study if they had been admitted within 180 days of the onset of an ischemic stroke. Patients who were admitted >180 days after a stroke (31/344, 9%), who had no cognitive assessment within a week after stroke rehabilitation admission (18/344, 5.2%), or had missing data (7/344, 2%) were excluded. Therefore, 83.7% (288/344) of the patients were available for analysis. Patients with aphasia were not included in the n=344 initial sample as they follow a different rehabilitation protocol involving a different set of cognitive assessments and, therefore, need to be analyzed separately (in future work).

At admission, each patient was assigned a physician who coordinated the rehabilitation team (a nurse, a neuropsychologist, a physiotherapist, an occupational therapist, a social worker, and a clinical psychologist based on the characteristics of the case). Therefore, admission and discharge cognitive assessments (as well as all clinical and demographic data analyzed in this study) were systematically recorded in the electronic health records of the hospital. The authors confirm that this study is compliant with the Helsinki Declaration of 1975, as revised in 2008, and it was approved by the Ethics Committee of Clinical Research of Institut Guttmann.

The participants were anonymized and nonidentifiable. A specific written informed consent was not required for participants to be included in this study; nevertheless, at admission to Institut Guttmann, participants provided written informed consent to be included in research studies addressed by the Institut Guttmann hospital.

Web-Based Cognitive Rehabilitation System

The Guttmann, NeuroPersonalTrainer web-based cognitive rehabilitation platform used in this study comprises a set of 149 different web-based cognitive rehabilitation tasks. There is no established previous order in which patients should execute such tasks. Therefore, every patient executed (eventually) a different subset of them in a different order during their rehabilitation process, taking between 2 and 6 months, distributed over two to five sessions a week. During each session, the patient executed between 4 to 10 cognitive rehabilitation tasks, and the total duration of one session ranged between 45 minutes to 1 hour. Each task mainly addressed one of the following functions: memory, executive functioning, attention, gnosias, calculus, orientation, language, and social cognition. Immediately after each execution of a task, the patient received a feedback on performance (ranging from 0-100, as the percentage of compliance), with 0% being the lowest level of compliance and 100% being the highest.

Cognitive Assessments at Admission

Before starting web-based cognitive rehabilitation using the Guttmann, NeuroPersonalTrainer platform, every patient was assessed once using standardized tests specifically validated for the population under study. Specific linguistic abilities were assessed using three subtests of the Barcelona test [32,33]: (1) repetition (maximum score=10), (2) denomination (maximum score=14), and (3) comprehension (maximum score=16). For assessing verbal fluency, the phonetic verbal fluency test [34] was used. The Trail Making Test was used to assess executive functioning [35] and the Wechsler Adult Intelligence Test–III [36] to assess visuospatial construction and perception.

Cognitive Assessments at Admission and Discharge: Memory Variable

In this study (without loss of generality), we assessed improvements in the memory cognitive function using the Rey Auditory Verbal Memory Test (RAVLT) [37] and the DIGITS subtest of Barcelona test [32]. RAVLT comprises three subtests: RAVLT learning (RAVLT075), free-recall memory (RAVLT015), and recognition (RAVLT015R). In RAVLT075, the patient was asked to recall as many words as possible from a list of 15 words, repeated five times. After a latency of 20 minutes, the patient was asked to recall the words (RAVLT015), and then the patient heard a list of 50 words containing the 15 initial sets that had to be recognized by the patient (RAVLT015R).

The DIGITS subtest (direct version) of the Barcelona test addresses working memory, and the patient was asked to repeat a series of numbers of variable lengths (3-9) until they failed in two consecutive series, reporting the largest series before failure [32].

Elo Rating Formulation

The Elo rating system [19] is formally defined as [20]: given a rating estimate θ_i for each player *i*, the result of a match between players *i* and *j* is represented by $R_{ij} \in \{0,1\}$.

The actual ratings of each player are used to estimate the probability that player *i* wins:

which is used to update the ratings as follows, based on the Bradley-Terry model [38]:

where K is a constant parameter that controls how quickly θ_i changes, with large K values resulting in θ_i changing quickly and small K values resulting in θ_i changing slowly. In this study, we considered three extensions to the original Elo system: Glicko [39], Glicko-2 [40], and Stephenson [41]. Glicko models introduce a measure of reliability to assess the accuracy of the rating; that is, the rating deviation. Stephenson rating can be of interest in our context as it introduces a parameter that considers the strengths of the opponents, [41] being in our case, player *i*, the patient, and player *j*, the cognitive rehabilitation task.

Regression Models

Overview

Demographic and clinical state-of-the-art variables such as age, gender, marital status, and variables related to the rehabilitation program, such as the time in between the onset of stroke and initiation of the rehabilitation program or length of stay, were considered as candidate predictors. Categorical variables were dichotomized: female=0, male=1; low level of education=0, high level of education=1 (depending on the number of years of education); married=1, not married=0. Forward stepwise selection was used to add covariates to the models, which were evaluated by assessing discrimination using area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity for logistic regressions and to maximize R^2 and adjusted R^2 for linear regressions. The variance inflation factor and tolerance (1/variance inflation factor) were used to test the multicollinearity of independent variables (tolerance ≤0.40 indicates a multicollinearity problem) [42]. The Durbin-Watson (D-W) test was used to assess the assumption of independent errors (D-W should be close to 2 to meet the assumption of independence [42]). The Elo rating algorithm calculations (including Glicko, Glicko-2, and Stephenson) were applied using the PlayerRatings R package [41]. R v3.5.1 (R Foundation for Statistical Computing) was used for all statistical analyses. The level of significance was set at P=.05.

Dependent Variables

In linear regressions, the dependent variables were RAVLT075 and DIGITS at discharge. In logistic regressions, the aim was to predict improvement in RAVLT075 (if RAVLT075 at discharge–RAVLT075 at admission ≥5, then improvement=true; else, improvement=false) and improvement in DIGITS (if DIGITS at discharge–DIGITS at admission ≥1, then improvement=true; else, improvement=false).

Results

Demographic Characteristics and Cognitive Assessments

Table 1 shows the demographic characteristics and clinical assessments of the 288 included patients.

The mean age at the time of the lesion was 51 (SD 9) years. The proportion of participants aged <65 years was 93.8% (270/288) (as opposed to most studies addressing ischemic stroke, we analyzed working-age participants). In relation to sex, in our data set, the proportion was 67.7% (195/288) men and 32.3% (93/288) women, which seems to suggest a bias in favor of men. Nevertheless, it somehow reflects reality in the general population, where the proportion of men experiencing ischemic stroke is larger than that of women [43-45]; however, women experience more hemorrhagic strokes [46].

Table 1. Demographics and clinical assessments (N=288)[a].

Variables	Admission	Discharge
Age (years), mean (SD)	51 (9)	N/A[b]
Age <65 years, n (%)	270 (93.8)	N/A
Males, n (%)	195 (67.7)	N/A
Marital status (married), n (%)	180 (62.5)	N/A
Educational level, n (%)		
Read and write	9 (3.1)	N/A
Primary	114 (50)	N/A
Secondary	88 (30.5)	N/A
Higher	66 (22.9)	N/A
NIHSS[c], median (IQR)	11 (7-15)	N/A
TMT[d] A, mean (SD)	82 (64)	N/A
TMT B, mean (SD)	157 (90)	N/A
PMR[e], mean (SD)	27 (12)	N/A
VC[f]–CUBS, mean (SD)	23 (12)	N/A
VP[g]-IMAGES, mean (SD)	17 (3)	N/A
VP–WAIS III[h], mean (SD)	37 (15)	N/A
Barcelona test–repetition, mean (SD)	9 (1)	N/A
Barcelona test–denomination, mean (SD)	13 (1)	N/A
Barcelona test–comprehension, mean (SD)	15 (1)	N/A
Barcelona test–DIGITS, mean (SD)	3 (1)	4 (1)
Barcelona test–DIGITS, median (IQR)	4 (3-4)	4 (3-5)
RAVLT[i] 075, mean (SD)	37 (10)	43 (11)
RAVLT075, median (IQR)	37 (30-45)	44 (35-52)
RAVLT015, mean (SD)	6 (3)	8 (3)
RAVLT015, median (IQR)	7 (5-9)	9 (6-12)
RAVLT015R, mean (SD)	10 (4)	11 (3)
RAVLT015R, median (IQR)	12 (8-14)	13 (10-14)
Length of stay (days), mean (SD)	88 (36)	N/A
Length of stay (days), median (IQR)	84 (55-113)	N/A
Time since onset to rehab admission (days), mean (SD)	55 (35)	N/A
Time since onset to rehab admission (days), median (IQR)	43 (29-75)	N/A

[a] Results are presented as mean (SD), median (IQR), or percentage, when appropriate.

[b] N/A: not applicable.

[c] NIHSS: The National Institutes of Health Stroke Scale.

[d] TMT: Trail Making Test.

[e] PMR test assesses the capacity of word generation according to an initial letter (P, M, and R).

[f] VC: visual construction.

[g] VP: visual perception.

[h] WAIS-III: Wechsler Adult Intelligence Test–III.

[i] RAVLT: Rey Auditory Verbal Memory Test.

Elo Rating: Feasibility Case

We initially ran the four different Elo rating approaches (standard Elo, Glicko, Glicko-2, and Stephenson) in a reduced data set of 20 patients, each of whom executed the same task 20 times. Two screenshots of the selected task are presented in Figure 1. The task addresses executive functioning (planning), and the objective is to move a blue ball from an initial position in a maze to the final position, minimizing the number of moves. The bar at the right indicates the time left to perform the task. Figure 1 top shows the initial position of the ball, and Figure 1 bottom shows the status 20 seconds later when the objective was accomplished.

The included 20 patients were stratified into three categories according to their compliance in the maze task as follows:

- low compliance={id1, id2, id3, id4, id5, id6};
- mid compliance={id7, id8, id9, id10, id11, id12, id13};
- high compliance={id14, id15, id16, id17, id18, id19, id20}.

Figure 2 presents the boxplots of the obtained results in the maze task at each execution in the 3 groups, showing their different levels of compliance.

Figure 1. Two screenshots of the maze task, showing the initial position of the blue ball (top) and its position at the end of the task (bottom).

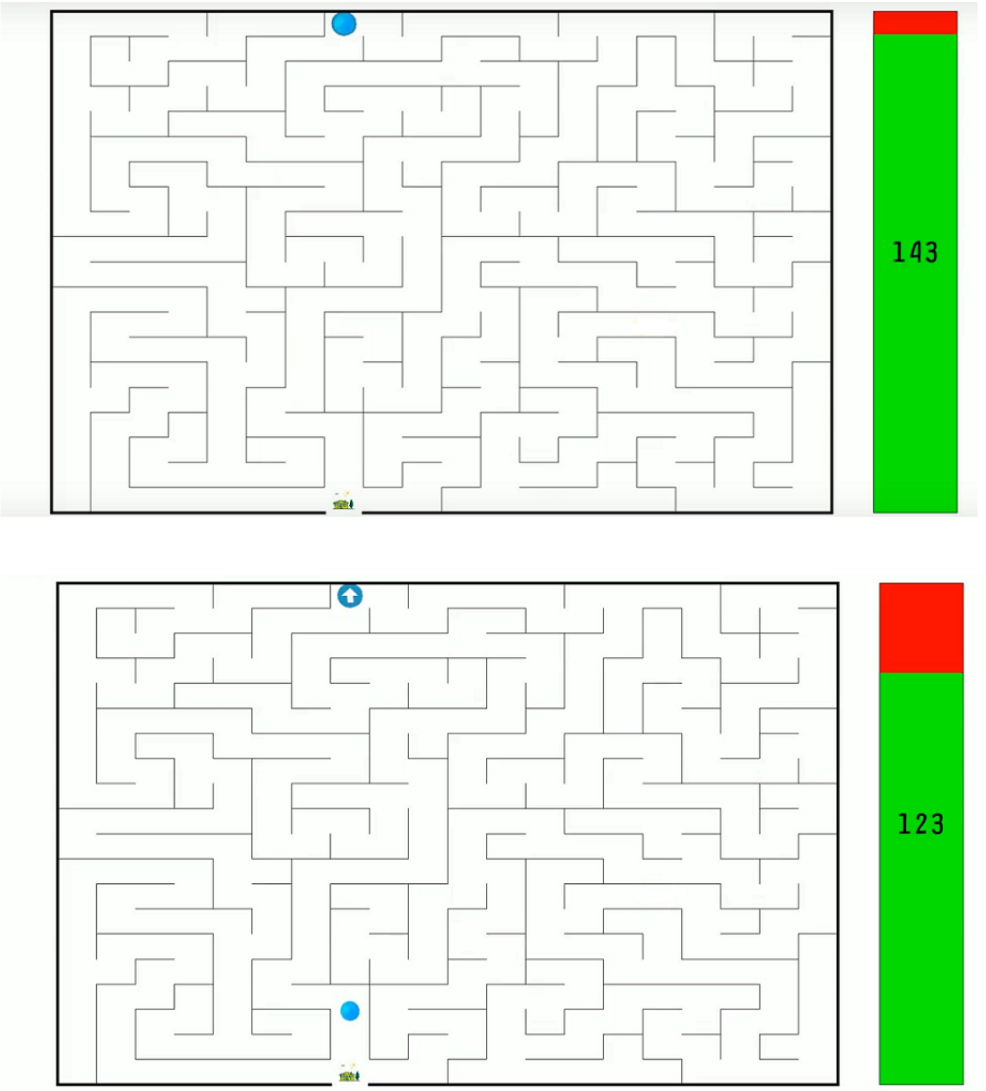
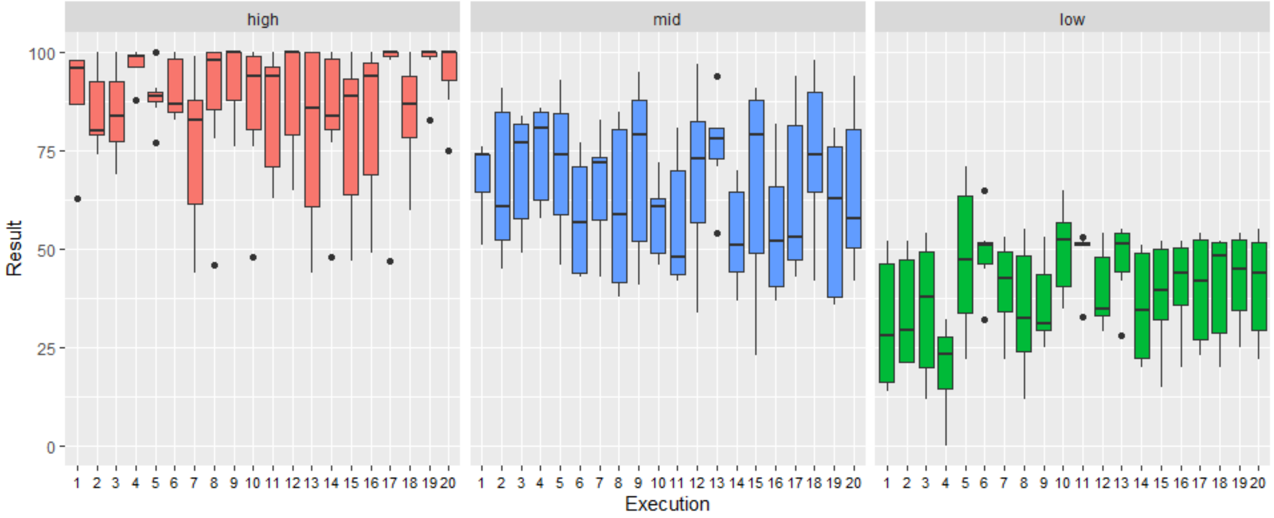


Figure 2. Boxplots of the obtained results in the maze task at each execution in the 3 groups, showing the high, middle, and low levels of compliance in the task.



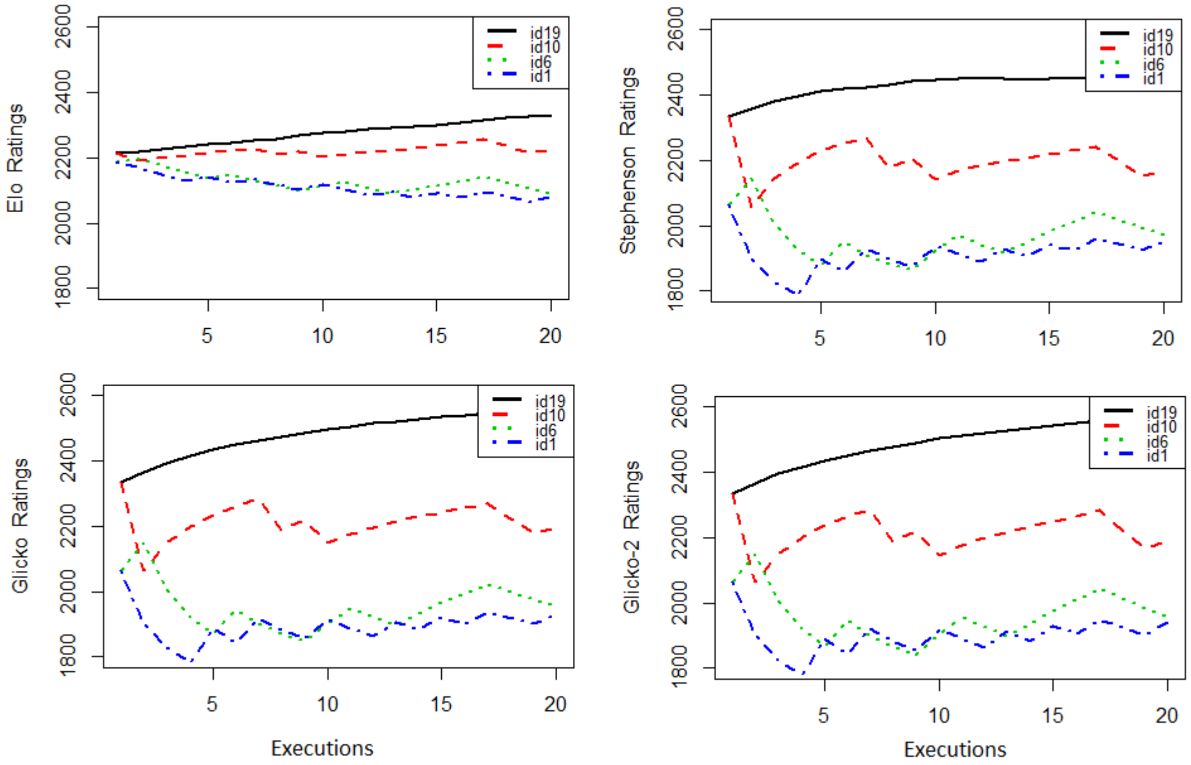
We then ran the four Elo rating systems with default values for the initial ratings and K. We considered that when a patient gets a result >50%, they win the match against the maze; however, if their result is <50%, the maze wins. The ratings obtained using the Glicko approach are presented in Table 2. Patients are ordered in Table 2 according to their obtained ratings. Table 2 shows patients id19, id20, id16, id14, id17, id15, and id18 at the first seven positions. Similarly, patients from the midcompliance group are in positions 8-14, and patients from the low compliance group are in the bottom positions. The maze task itself is also considered as a player; it played all 400 matches, winning 118 and losing 282.

Table 2. Glicko ratings after 20 executions of the maze task (n=20 synthetic patients).

Player	Glicko rating (deviation)	Games	Win	Loss
id19	2565 (146.26)	20	20	0
id20	2565 (146.26)	20	20	0
id16	2450 (124.22)	20	19	1
id14	2368 (118.61)	20	18	2
id17	2364 (124.24)	20	18	2
id15	2293 (113.49)	20	17	3
id18	2273 (119.90)	20	17	3
id12	2240 (104.33)	20	16	4
id10	2195 (97.57)	20	15	5
id13	2190 (99.44)	20	15	5
id9	2177 (103.78)	20	15	5
id11	2143 (95.22)	20	14	6
id7	2122 (101.43)	20	14	6
id8	2112 (91.53)	20	13	7
id3	2035 (86.81)	20	10	10
Maze	1999 (36.77)	400	118	282
id2	1981 (87.77)	20	9	11
id5	1965 (88.01)	20	9	11
id4	1960 (87.11)	20	8	12
id6	1957 (87.46)	20	8	12
id1	1930 (89.33)	20	7	13

Figure 3 shows the obtained ratings using all four approaches for patient representatives of each of the compliance groups; we plotted id1 and id6 patients from the low-level group, id10 from the midlevel group, and id19 from the high level of compliance group to visualize how the Elo ratings represented their compliance levels.

Figure 3. Elo ratings using all four approaches (traditional Elo, Stephenson, Glicko, and Glicko-2) for patient representatives of each of the compliance groups; id1 and id6 (low level), id10 (midlevel) and id19 (high level).



Cognitive Task Executions in Guttmann, NeuroPersonalTrainer Platform

Overview

Table 3 summarizes all task executions during the whole rehabilitation process for all 288 included patients. A total of 44,814 task executions were performed in 5088 sessions during the period under study. Each patient performed 155 task executions on average. When considering the different functions addressed by the tasks, the most frequently executed were those addressing memory (18,183 executions), comprising almost 40.57% (18,183/44,814) of the total executions.

Table 3. Cognitive rehab task executions (N=288 patients).

Description	Values
Total number of task executions	44,814
Executions per patient, mean (SD)	155 (113.2)
Total number of sessions	5008
Sessions executed per patient, mean (SD)	17 (11.5)
Tasks executed per session per patient, mean (SD)	9 (4.4)
Total number of memory tasks executed	18,183
Total number of executive functioning tasks executed	14,061
Total number of attention tasks executed	8062
Total number of gnosias tasks executed	1795
Total number of calculus tasks executed	1695
Total number of orientation tasks executed	741
Total number of language tasks executed	261
Total number of social cognition tasks executed	16
Memory task results, mean (SD)	53.1 (36.4)
Executive functioning tasks results, mean (SD)	49.6 (38.7)
Attention task results, mean (SD)	59.4 (36.7)
Gnosias task results, mean (SD)	74.4 (30.8)
Calculus task results, mean (SD)	72.9 (35.8)
Orientation task results, mean (SD)	75.6 (38.0)
Language task results, mean (SD)	55.5 (38.4)
Social cognition task results, mean (SD)	56.7 (37.1)

Preprocessing: Removing Less Executed Tasks

As introduced in the section *Web-Based Cognitive Rehabilitation System*, the Guttmann, NeuroPersonalTrainer cognitive platform includes 149 different web-based tasks. There is no established previous order or frequency in which patients should execute such tasks; therefore, in this section, we analyze task execution frequencies. As shown in Table S1 (Multimedia Appendix 1), several tasks were very infrequently executed. As detailed in Table S2 (Multimedia Appendix 1), 68 tasks accounted for 38,177 executions. Therefore, 45.6% (68/149) of all available tasks accounted for 85.18% (38,177/44,814) of all executions. In this section, we analyzed these 68 tasks (executed by all N=288 patients) and stratified them into three difficulty levels, considering their input parameter configurations during the 38,177 executions.

Ranking Patients Using the Initial 50 Task Executions: Elo Rating

We used the Stephenson rating with default parameters, considering the following criteria:

- If the result ≤39%, then the task wins.
- If 40% ≤ result ≤ 64%, then the result is a draw.
- If the result ≥65%, then the patient wins.

The Stephenson ratings were obtained by considering the first 50 task executions for every patient. We then stratified all 288 patients into 3 groups (each group comprised n=96 patients), according to their Elo ratings (low, middle, and high). Table 4 shows the memory assessments at admission and discharge, percentage of patients that improved, mean number of executed tasks, and obtained result comparisons for the three Elo levels (low, mid, and high) obtained using the 50 initial task executions for n=288 patients, with 96 patients in each Elo group that performed 38,177 task executions of the most frequent 68 tasks during rehabilitation.

Table 4. Memory assessments at admission and discharge, percentage of patients that improved, mean number of executed tasks, and obtained results comparisons for the three Elo levels (low, middle, and high) obtained using the 50 initial task executions (N=288 patients, 96 patients in each Elo group that executed 38,177 tasks during rehabilitation).

Variables	Low Elo (n=12,431)	Mid Elo (n=12,449)	High Elo (n=13,297)	*P* value
Sex (female), n (%)	4396 (35.36)	4607 (37.01)	3793 (28.52)	<.001
Age (years) when starting rehabilitation, mean (SD)	52 (8)	51 (8)	48 (10)	<.001
DIGITS at admission, mean (SD)	3.4 (0.9)	3.8 (0.9)	4.0 (1.0)	<.001
RAVLT[a] 075 at admission, mean (SD)	37 (9)	36 (10)	38 (10)	<.001
RAVLT015 at admission, mean (SD)	6 (3)	6 (3)	7 (3)	<.001
RAVLT015R at admission, mean (SD)	10 (4)	10 (4)	11 (4)	<.001
Length of stay (days), mean (SD)	104 (37)	105 (35)	106(40)	.05
Executed tasks, mean (SD)	245 (129)	222 (122)	241 (124)	<.001
Obtained results in tasks, mean (SD)	37 (36)	56 (36)	68 (33)	<.001
DIGITS at discharge, mean (SD)	3.5 (1.0)	4.5 (0.9)	4.4 (0.8)	<.001
RAVLT075 at discharge, mean (SD)	42 (11)	45 (11)	45 (12)	<.001
RAVLT015 at discharge, mean (SD)	8 (3)	9 (3)	9 (4)	<.001
RAVLT015R at discharge, mean (SD)	11 (3)	12 (3)	12 (3)	<.001
DIGITS IMP[b] (yes), n (%)	2859 (22.99)	7059 (56.7)	5353 (40.26)	<.001
RAVLT075 IMP (yes), n (%)	5308 (42.69)	8356 (67.12)	7484 (56.28)	<.001
RAVLT015 IMP (yes), n (%)	6136 (49.36)	7325 (58.84)	7482 (56.26)	<.001
RAVLT015R IMP (yes), n (%)	6802 (54.72)	6683 (53.68)	7132 (53.64)	<.001
Task difficulty level, n (%)				
Level 1	4812 (38.71)	3997 (32.11)	3536 (26.59)	<.001
Level 2	3999 (32.17)	3857 (30.98)	4346 (32.68)	<.001
Level 3	3620 (29.12)	4595 (36.91)	5415 (40.72)	<.001

[a] RAVLT: Rey Auditory Verbal Learning Test.

[b] IMP: improved.

Importance of Elo Rating in Predicting Outcomes: RAVLT075 and DIGITS

Table 5 presents the obtained predictors of RAVLT075 at discharge (model 1), 54% of the variance explained and the obtained predictors of DIGITS at discharge (model 2), 43% of the variance explained.

When the Elo rating feature is excluded from model 1, it explains 52% of the variance, and when it is excluded from model 2, the resulting model explains 42%.

Table 5. Multivariate linear regressions, nonstandard β (95% CI), standard β, Durbin-Watson (D-W) test, variance inflation factor, and R^2 and adjusted R^2 for RAVLT075 and DIGITS at discharge (N=288).

Variables	β (95% CI)	Standard β	1/VIF[a]	*P* value	R^2	Adjusted R^2
Model 1 predictors of RAVLT[b] 075 at discharge						
Elo rating	.01 (.01 to .02)	.09	0.95	.02	0.55	0.54
RAVLT075 at admission	.76 (.67 to .85)	.66	0.92	<.001	0.55	0.54
LOS[c]	.04 (.01 to .06)	.13	0.98	.002	0.55	0.54
Sex	−2.48 (−4.48 to −0.48)	–0.10	0.93	.01	0.55	0.54
Age (years)	−0.09 (−0.19 to .01)	–0.07	0.92	.06	0.55	0.54
D-W[d]=1.89	N/A[e]	N/A	N/A	.37	0.55	0.54
Model 2 predictors of DIGITS at discharge						
Elo rating	.00 (.00 to .00)	.10	0.88	.02	0.44	0.43
DIGITS at admission	.63 (.54 to .72)	.63	0.91	<.001	0.44	0.44
LOS	.00 (.00 to .00)	.05	0.98	.22	0.44	0.44
Sex	.04 (−0.15 to .23)	.01	0.95	.67	0.44	0.44
Age (years)	.00 (.00 to .01)	.02	0.94	.53	0.44	0.44
D-W=2.01	N/A	N/A	N/A	.95	0.44	0.44

[a] VIF: variance inflation factor.

[b] RAVLT: Rey Auditory Verbal Memory Test.

[c] LOS: length of stay.

[d] D-W: Durbin-Watson test.

[e] N/A: not applicable.

Importance of Elo Rating in Predicting Improvement: RAVLT075 and DIGITS

Table 6 presents the models used for predicting improvement in RAVLT075 and DIGITS. We used the criteria to decide whether a patient improved as described in the *Dependent Variables* section; 50.6% (146/288) of patients improved in RAVLT075, and 34% (98/288) of patients improved in DIGITS. We used the same Elo ratings as described in the *Ranking Patients Using the Initial 50 Task Executions: Elo Rating* section. Model 3 yielded an AUC of 0.73 (95% CI 0.64-0.82) for improvement in RAVLT075, with an accuracy=0.64 (95% CI 0.54-0.72), specificity=0.55, and sensitivity=0.73. Model 4 yielded an AUC of 0.81 (95% CI 0.72-0.89) for improvement in DIGITS, with an accuracy=0.73 (95% CI 0.64-0.81), specificity=0.22 and sensitivity=0.97. Models 3 and 4 are detailed in Table 6. When the Elo rating was excluded as an independent variable for model 3, the model yielded an AUC of 0.66 (95% CI 0.56-0.76) for improvement in RAVLT075, with an accuracy=0.62 (95% CI 0.52-0.71), specificity=0.62, and sensitivity=0.62. When the Elo rating was excluded as an independent variable for model 4, the model yielded an AUC of 0.73 (95% CI 0.62-0.83) for improvement in DIGITS, with an accuracy=0.72 (95% CI 0.62-0.80), specificity=0.34, and sensitivity=0.92. As shown in Table S3 (Multimedia Appendix 1), RAVLT075 was highly correlated with RAVLT015 and RAVLT015R at admission and at discharge.

Table 6. Multivariable logistic regressions, nonstandard β, odds ratio (95% CI), variance inflation factor for RAVLT075, and DIGITS improvement at discharge (N=288).

Variables	Odds ratio (95% CI)	β coefficients	1/VIF[a]	*P* value
Model 3[b] predictors of RAVLT[c] 075 improvement at discharge				
Rating	1.00 (1.00-1.00)	.61	0.93	.02
RAVLT075 at admission	0.95 (0.92-0.97)	−0.95	0.88	<.001
LOS[d]	1.00 (1.00-1.01)	.67	0.98	<.001
Sex	0.64 (0.37-1.11)	−0.40	0.92	.12
Age	0.97 (0.94-0.99)	−0.52	0.92	.04
Model 4[e] predictors of DIGITS improvement at discharge				
Rating	1.00 (0.99-1.00)	.57	0.86	.06
DIGITS at admission	0.38 (0.26-0.52)	−2.04	0.86	<.001
LOS	1.00 (1.00-1.01)	.71	0.97	.01
Sex	1.02 (0.57-1.85)	.02	0.96	.92
Age	1.00 (0.97-1.02)	.01	0.95	.95

[a] VIF: variance inflation factor.

[b] Area under the receiver operating characteristic curve=0.73 (95% CI 0.64-0.82), accuracy=0.64 (95% CI 0.5451-0.7281), specificity=0.55, and sensitivity=0.73.

[c] RAVLT: Rey Auditory Verbal Memory Test.

[d] LOS: length of stay.

[e] Area under the receiver operating characteristic curve=0.81 (95% CI 0.72-0.89), accuracy=0.73 (95% CI 0.64-0.81), specificity=0.22, and sensitivity=0.97.

Discussion

Principal Findings

To the best of our knowledge, in this study, Elo ratings were applied in the context of web-based cognitive rehabilitation tasks for the first time. We demonstrated the feasibility of using Elo ratings by using a publicly available R library (PlayerRatings) [41] on a synthetic use–case of 20 patients executing one task 20 times.

We then obtained the Elo ratings for each patient in a real rehabilitation setting where 288 adult patients with ischemic stroke followed cognitive rehabilitation, executing 68 different web-based rehabilitation tasks 38,177 times. We then performed a stratification of the patients into 3 groups (96 patients each) according to their Elo rating (low, middle, and high). We have shown the relationships among the three Elo rating levels and the proportion of tasks executed at three increasing difficulty levels (1, 2, and 3) with the rehabilitation outcomes in the memory cognitive function. We then developed four predictive models, where the Elo rating variables were significant in three of them (and quasi-significant in the fourth) for auditory verbal learning memory and working memory outcomes. We found that including Elo ratings as independent variables increased the model performance (for both linear and logistic regressions).

Clinical Implications

Several web-based cognitive rehabilitation platforms integrate some kind of stratification of patients as an initial step for treatment personalization. The web-based platform used in this study integrates an automatic therapy planning functionality—the intelligent therapy assistant (ITA) [47]. The ITA takes a set of patients' cognitive profiles as the starting point, obtained using cluster analysis on the baseline cognitive evaluation. When a new patient starts cognitive training in Guttmann, NeuroPersonalTrainer, the ITA dynamically assigns the patient to the appropriate cluster. The ITA then schedules different cognitive tasks during a user-defined rehabilitation period for the new patient. Therefore, an important clinical implication of our results in this study involves the ITA (or any other data-driven therapy assistant) starting point: using Elo rating as a starting point, alternative to cluster analysis.

Obtaining an initial Elo rating for each patient is a simple process (in terms of both implementation and interpretation of results). As remarked in previous research, for example, in the field of educational tutoring systems, Elo rating use is encouraged because of its simplicity [20]. As shown in Table 3, the mean number of tasks executed by a patient in a session is 9, so in about five sessions (usually 2 weeks), an Elo rating for each patient obtained using the first 50 task executions will be available.

Therapists can then use the Elo rating to assign the patient to a skill level. In this study, in Table 4, we present the results using three skill levels, each of them with the same number of patients (96; or one-third of the N=288 total participants). Table 4 shows that, for example, 67% (64/96) of patients in the mid-Elo group improved in the RAVLT075 item, and 58% (56/96) of patients in the mid-Elo group improved in the RAVLT015 item. Meanwhile, for example, only 23% (22/96) of patients in the low Elo group improved in the DIGITS item. The low Elo group performed 29.1% (3,617/12,431) of their tasks at difficulty level

3, whereas the mid-Elo group performed 31% (3,859/12,449) of their tasks at difficulty level 2. This seems to suggest that patients in the low Elo group could have performed a higher proportion of tasks at difficulty level 1, which is more appropriate to their skills. Patients in the mid-Elo group performed a higher proportion of tasks according to their skill levels, which seems to be related to a higher proportion of patients obtaining improvements in the four memory items presented in Table 4.

Another clinical implication was noted on in a recent systematic review on computerized cognitive training [48]. The review highlighted the need to develop interventions focused on specific cognitive functions by means of concrete training or rehabilitation activities (or tasks). Our results contribute in that sense; considering, for example, model 1 for predicting RAVLT075 at discharge, we obtained a standard β=.09 for the Elo rating variable. Therefore, for every 113 points obtained in the Elo rating, an extra point in RAVLT075 at discharge is obtained. If we consider, for example, in the maze task presented in Figure 1, patient id12 (Elo ranking=2240) and patient id8 (Elo ranking=2112) presented in Table 2, the difference between their Elo ratings is 128 points, with both patients belonging to the intermediate compliance group. Similar Elo rating scores were obtained for the final sample of N=288. Therefore, therapists can identify at the early stages of the rehab process–specific cognitive tasks where patients are close to obtaining a draw or a win (result ≥40%) and address different strategies [48] to improve performance in such specific tasks.

Limitations of This Study

Several limitations to the study need to be highlighted. First, we conducted a single-center study, an advantage of which is that data were obtained and included by clinicians trained in neurological rehabilitation, and all patients were managed under the same stroke rehabilitation protocols. The Guttmann, NeuroPersonalTrainer platform has already been integrated into the clinical practice of several acquired brain injury centers; nevertheless, their patients were not included in this analysis. A multicenter stroke study may include an initial preprocessing phase, wherein patients are grouped according to their initial National Institutes of Health Stroke Scale severity to avoid additional heterogeneity. Thereafter, Elo rating techniques, such as those proposed in this study, maybe applied within such groups. External validation assessments common to all participating centers are also an important aspect to be addressed in this future multicenter study. Second, the studied health area belongs mainly to the urban population, with a small rural population or populations from other regions. Third, our analysis lacked computerized tomography or magnetic resonance imaging examinations that describe the presence of contusion, hematoma, hemorrhage, ischemia, or other signs of parenchymal lesions in the frontal, temporal, parietal, occipital, and cerebellar lobes or diffuse axonal injury. Fourth, our sample did not include any patients with missing data. All data used as inputs were complete. Fifth, our analysis did not include indicators of mental health or other comorbidities. Persons who experience a stroke may have one or more preexisting medical comorbidities at the time of injury (eg, alcohol use and depression). Therefore, we plan to include comorbidity analyses in future research studies. Sixth, in all our Elo rating calculations, we used the default value for the K constant. Several approaches to K optimization have been reported, such as hill climbing, gradient descendent, or Bayesian [20], which can also be addressed in future work. Finally, the criteria for defining wins, draws, and losses in our Elo ratings were also constant for every task, and another possible improvement could be to fit such criteria according to the task difficulty level, considering the strength of the opponents (patients' skills and task difficulty levels) that can be addressed using the Stephenson extension [41].

Comparison with Prior Work

Cluster analysis has been extensively proposed in previous research to address heterogeneity in patients with acquired brain injury [49-51] and as an initial step for patient profiling. Most previous studies use commercial software products for cluster analysis, which are, in turn, not integrated into the web-based cognitive rehabilitation platform.

In a recent study, Faria et al [52] presented a framework for the creation of personalized cognitive rehabilitation tasks based on a participatory design strategy. They selected 11 paper-and-pencil tasks from standard clinical practice and parameterized them with multiple parameter configurations. A modeling approach was used to quantitatively determine how the task parameters affect each of the cognitive domains (memory, executive functions, attention, and language). For modeling this relationship, the parameters of each task were used as predictors of the demands in each cognitive domain. In our case, the parameters of each task were used by experts to assign a difficulty level to each task (difficulty level 1, 2, and 3, as presented in Table 4), where each task aims to address one main cognitive domain (memory, executive functions, attention, and language).

Conclusions

We have shown the feasibility of Elo ratings for identifying patients' profiles at the early stages of cognitive rehabilitation in a real clinical setting. Elo ratings can be used to match skills with task difficulties, aiming to maximize improvements in specific cognitive functions. Such Elo ratings are also significant in predicting cognitive outcomes. Elo ratings increased the models' performance (for both linear and logistic regressions). Generalization of the use of Elo ratings beyond patients with stroke to any other population with acquired brain injury requiring cognitive rehabilitation in any web-based platform is straightforward because of the simplicity of existing open-access Elo rating implementations.

Acknowledgments

This research was partially funded by PRECISE4Q Personalized Medicine by Predictive Modelling in Stroke for better Quality of Life–European Union's Horizon 2020 research and innovation program under grant agreement No. 777107.

Authors' Contributions

AGR conceived the study; AGR, HB, and EO collected, selected, and cleaned the data; JDK, AGR, and HB statistically analyzed the data. AGR drafted the manuscript, and JDK, VIM, DF, HB, MBG, EO, JL, and JMT revised the manuscript critically for important intellectual content and approved the final manuscript. AGR, HB, JDK, VIM, DF, EO, MBG, and JMT received funding for this study.

Conflicts of Interest

AGR, JMT, EO, JL, and MBG work at Institut Guttmann, Hospital de Neurorehabilitació, proprietary of the Guttmann, NeuroPersonalTrainer platform. VIM reported receiving personal fees from ai4medicine outside the submitted work. There is no connection, commercial exploitation, transfer, or association between the projects of ai4medicine and the results presented in this work. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1
All initial 44,814 web-based task executions, selection of the most frequently executed 68 tasks (38,177 executions), and correlation analysis of the Rey Auditory Verbal Memory Test and DIGITS assessments.
[DOCX File , 28 KB - medinform_v9i11e28090_app1.docx]

References

1. Harari Y, O'Brien MK, Lieber RL, Jayaraman A. Inpatient stroke rehabilitation: prediction of clinical outcomes using a machine-learning approach. J Neuroeng Rehabil 2020 Jun 10;17(1):71 [FREE Full text] [doi: 10.1186/s12984-020-00704-3] [Medline: 32522242]
2. Rohde D, Gaynor E, Large M, Mellon L, Bennett K, Williams DJ, et al. Cognitive impairment and medication adherence post-stroke: a five-year follow-up of the ASPIRE-S cohort. PLoS One 2019 Oct 17;14(10):e0223997 [FREE Full text] [doi: 10.1371/journal.pone.0223997] [Medline: 31622438]
3. Liew HP. Cognitive disparities between US- and foreign-born individuals. J Public Health (Berl) 2020 Feb 13:s10389 (forthcoming) [FREE Full text] [doi: 10.1007/s10389-020-01218-x]
4. Galetto V, Sacco K. Neuroplastic changes induced by cognitive rehabilitation in traumatic brain injury: a review. Neurorehabil Neural Repair 2017 Sep;31(9):800-813. [doi: 10.1177/1545968317723748] [Medline: 28786307]
5. van Rijsbergen MW, Mark RE, de Kort PL, Sitskoorn MM. Subjective cognitive complaints after stroke: a systematic review. J Stroke Cerebrovasc Dis 2014 Mar;23(3):408-420. [doi: 10.1016/j.jstrokecerebrovasdis.2013.05.003] [Medline: 23800498]
6. Lamb F, Anderson J, Saling M, Dewey H. Predictors of subjective cognitive complaint in postacute older adult stroke patients. Arch Phys Med Rehabil 2013 Sep;94(9):1747-1752. [doi: 10.1016/j.apmr.2013.02.026] [Medline: 23529143]
7. Lutski M, Zucker I, Shohat T, Tanne D. Characteristics and outcomes of young patients with first-ever ischemic stroke compared to older patients: the national acute stroke ISraeli registry. Front Neurol 2017 Aug 21;8:421 [FREE Full text] [doi: 10.3389/fneur.2017.00421] [Medline: 28871237]
8. Ramirez L, Kim-Tenser MA, Sanossian N, Cen S, Wen G, He S, et al. Trends in acute ischemic stroke hospitalizations in the United States. J Am Heart Assoc 2016 May 11;5(5):e003233 [FREE Full text] [doi: 10.1161/JAHA.116.003233] [Medline: 27169548]
9. Tibæk M, Dehlendorff C, Jørgensen HS, Forchhammer HB, Johnsen SP, Kammersgaard LP. Increasing incidence of hospitalization for stroke and transient ischemic attack in young adults: a registry-based study. J Am Heart Assoc 2016 May 11;5(5):e003158 [FREE Full text] [doi: 10.1161/JAHA.115.003158] [Medline: 27169547]
10. Gates NJ, Sachdev PS, Singh MA, Valenzuela M. Cognitive and memory training in adults at risk of dementia: a systematic review. BMC Geriatr 2011 Sep 25;11:55 [FREE Full text] [doi: 10.1186/1471-2318-11-55] [Medline: 21942932]
11. No authors listed. Effect of computer-based cognitive rehabilitation (CBCR) for people with stroke: a systematic review and meta-analysis. NeuroRehabilitation 2015;37(3):487. [doi: 10.3233/NRE-151288] [Medline: 26640137]
12. Kueider AM, Parisi JM, Gross AL, Rebok GW. Computerized cognitive training with older adults: a systematic review. PLoS One 2012;7(7):e40588 [FREE Full text] [doi: 10.1371/journal.pone.0040588] [Medline: 22792378]
13. Thompson G, Foth D. Cognitive-Training Programs for Older Adults: What are they and can they enhance mental fitness? Educ Gerontol 2006 Sep 1;31(8):603-626. [doi: 10.1080/03601270591003364]
14. Whitmer AJ, Gotlib IH. Switching and backward inhibition in major depressive disorder: the role of rumination. J Abnorm Psychol 2012 Aug;121(3):570-578. [doi: 10.1037/a0027474] [Medline: 22468767]

15. Sohlberg M. Cognitive Rehabilitation. An Interactive Neuropsychological Approach. New York: Guilford Press; 2001.
16. Digital Media: Transformations in Human Communication. Bern, Switzerland: Peter Lang; 2005.
17. Wilms IL. The computerized cognitive training alliance - a proposal for a therapeutic alliance model for home-based computerized cognitive training. Heliyon 2020 Jan 31;6(1):e03254 [FREE Full text] [doi: 10.1016/j.heliyon.2020.e03254] [Medline: 32042977]
18. Bates BE, Xie D, Kwong PL, Kurichi JE, Ripley DC, Davenport C, et al. Development and validation of prognostic indices for recovery of physical functioning following stroke: part 2. PM R 2015 Jul;7(7):699-710. [doi: 10.1016/j.pmrj.2015.01.012] [Medline: 25633635]
19. Elo A. The Rating of Chess Players, Past and Present. California: Ishi Press; May 2008.
20. Pelánek R. Applications of the Elo rating system in adaptive educational systems. Comp Educ 2016 Jul;98:169-179. [doi: 10.1016/j.compedu.2016.03.017]
21. Hvattum L, Arntzen H. Using ELO ratings for match result prediction in association football. Int J Forecast 2010;26(3):460-470 [FREE Full text] [doi: 10.1016/j.ijforecast.2009.10.002]
22. Hacker S, von Ahn L. Matchin: eliciting user preferences with an online game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2009 Presented at: CHI '09: SIGCHI Conference on Human Factors in Computing Systems; Apr 4-9, 2009; Boston MA USA. [doi: 10.1145/1518701.1518882]
23. Pieters W, van der Ven S, Probst C. A move in the security measurement stalemate: elo-style ratings to quantify vulnerability. In: Proceedings of the New Security Paradigms Workshop. 2012 Presented at: NSPW '12: New Security Paradigms Workshop; Sept 18-21, 2012; Bertinoro Italy. [doi: 10.1145/2413296.2413298]
24. Sarma A, Sarma A, Gollapudi S, Panigrahy R. Ranking mechanisms in twitter-like forums. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. 2010 Presented at: WSDM'10: Third ACM International Conference on Web Search and Data Mining; Feb 4-6, 2010; New York. [doi: 10.1145/1718487.1718491]
25. Coulom R. Computing “elo ratings” of move patterns in the game of go. ICGA J 2007 Dec 01;30(4):198-208. [doi: 10.3233/icg-2007-30403]
26. Kang X, Zhang E. A universal defect detection approach for various types of fabrics based on the Elo-rating algorithm of the integral image. Text Res J 2019 Apr 10;89(21-22):4766-4793. [doi: 10.1177/0040517519840636]
27. Wang YZ, Yao Z, Wang C, Ren J, Chen Q. The impact of intelligent transportation points system based on Elo rating on emergence of cooperation at Y intersection. Appl Math Comput 2020 Apr 1;370:124923 [FREE Full text] [doi: 10.1016/j.amc.2019.124923]
28. Sánchez-Tójar A, Schroeder J, Farine DR. A practical guide for inferring reliable dominance hierarchies and estimating their uncertainty. J Anim Ecol 2018 May;87(3):594-608. [doi: 10.1111/1365-2656.12776] [Medline: 29083030]
29. Yang L, Dimitrov S, Mantin B. Forecasting sales of new virtual goods with the Elo rating system. J Revenue Pricing Manag 2014 Oct 3;13(6):457-469 [FREE Full text] [doi: 10.1057/rpm.2014.26]
30. Tasharrofi S, Barnes JC. Developing an elo-rating system for criminal justice practitioners: a superior method for resource allocation? SocArXiv Papers 2019 Nov 11:A (forthcoming) [FREE Full text] [doi: 10.31235/osf.io/y3h2e]
31. Reid D, Nixon M. Human identification using facial comparative descriptions. In: Proceedings of the 2013 International Conference on Biometrics (ICB). 2013 Presented at: 2013 International Conference on Biometrics (ICB); Jun 4-7, 2013; Madrid, Spain. [doi: 10.1109/icb.2013.6612962]
32. Peña-Casanova J, Jarne Esparcia A, Guardia Olmos J. Programa integrado de exploración neuropsicológica — test barcelona: validez de contenidos. Revista de Logopedia, Foniatría y Audiología 1991 Jan;11(2):96-107 [FREE Full text] [doi: 10.1016/s0214-4603(91)75507-1]
33. Quintana M, Peña-Casanova J, Sánchez-Benavides G, Langohr K, Manero RM, Aguilar M, Neuronorma Study Team. Spanish multicenter normative studies (Neuronorma project): norms for the abbreviated Barcelona Test. Arch Clin Neuropsychol 2011 Mar 11;26(2):144-157. [doi: 10.1093/arclin/acq098] [Medline: 21149392]
34. Artiola IF, Hermosillo RD, Heaton R, Pardee RI. Manual de Normas y Procedimientos para la Batería Neuropsicológica en Español. Tucson: mPress; 1999.
35. Reitan R, Wolfson D. The Halstead-Reitan Neuropsychological Test Battery: Theory and Clinical Interpretation. Tucson, AZ: Neuropsychological Press; 1985.
36. Ryan J, Lopez S. Wechsler Adult Intelligence Scale-III. In: Understanding Psychological Assessment. Perspectives on Individual Differences. Boston, MA: Springer; 2001.
37. Schmid M. Rey Auditory and Verbal Learning Test: A Handbook. Los Angeles: Western Psychological Services; 1996.
38. Bradley R, Terry M. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 1952 Dec;39(3/4):324-345. [doi: 10.2307/2334029]
39. Glickman ME. Parameter estimation in large dynamic paired comparison experiments. J R Statist Soc C 1999 Aug;48(3):377-394. [doi: 10.1111/1467-9876.00159]
40. Glickman ME. Dynamic paired comparison models with stochastic variances. J Appl Stat 2001 Aug;28(6):673-689. [doi: 10.1080/02664760120059219]
41. Stephenson A, Sonas J. PlayerRatings: Dynamic updating methods for player ratings estimation. rdrr.io. 2016. URL: https://rdrr.io/cran/PlayerRatings/ [accessed 2021-01-01]

42. O'brien RM. A caution regarding rules of thumb for variance inflation factors. Qual Quant 2007 Mar 13;41(5):673-690. [doi: 10.1007/s11135-006-9018-6]
43. Di Carlo A, Lamassa M, Baldereschi M, Pracucci G, Basile AM, Wolfe CD, European BIOMED Study of Stroke Care Group. Sex differences in the clinical presentation, resource use, and 3-month outcome of acute stroke in Europe: data from a multicenter multinational hospital-based registry. Stroke 2003 May;34(5):1114-1119. [doi: 10.1161/01.STR.0000068410.07397.D7] [Medline: 12690218]
44. Kapral MK, Fang J, Hill MD, Silver F, Richards J, Jaigobin C, Investigators of the Registry of the Canadian Stroke Network. Sex differences in stroke care and outcomes: results from the Registry of the Canadian Stroke Network. Stroke 2005 Apr;36(4):809-814. [doi: 10.1161/01.STR.0000157662.09551.e5] [Medline: 15731476]
45. Simpson C, Wilson C, Hannaford P, Williams D. Evidence for age and sex differences in the secondary prevention of stroke in Scottish primary care. Stroke 2005 Aug;36(8):1771-1775. [doi: 10.1161/01.STR.0000173398.99163.9e] [Medline: 16040591]
46. Niewada M, Kobayashi A, Sandercock PA, Kamiński B, Członkowska A, International Stroke Trial Collaborative Group. Influence of gender on baseline features and clinical outcomes among 17,370 patients with confirmed ischaemic stroke in the international stroke trial. Neuroepidemiology 2005;24(3):123-128. [doi: 10.1159/000082999] [Medline: 15637449]
47. Solana J, Cáceres C, García-Molina A, Chausa P, Opisso E, Roig-Rovira T, et al. Intelligent Therapy Assistant (ITA) for cognitive rehabilitation in patients with acquired brain injury. BMC Med Inform Decis Mak 2014 Jul 19;14:58 [FREE Full text] [doi: 10.1186/1472-6947-14-58] [Medline: 25038823]
48. Sigmundsdottir L, Longley WA, Tate RL. Computerised cognitive training in acquired brain injury: a systematic review of outcomes using the International Classification of Functioning (ICF). Neuropsychol Rehabil 2016 Oct;26(5-6):673-741. [doi: 10.1080/09602011.2016.1140657] [Medline: 26965034]
49. Sherer M, Davis LC, Sander AM, Nick TG, Luo C, Pastorek N, et al. Factors associated with word memory test performance in persons with medically documented traumatic brain injury. Clin Neuropsychol 2015;29(4):522-541. [doi: 10.1080/13854046.2015.1052763] [Medline: 26063081]
50. Ringdahl EN, Becker ML, Hussey JE, Thaler NS, Vogel SJ, Cross C, et al. Executive function profiles in pediatric traumatic brain injury. Dev Neuropsychol 2019;44(2):172-188. [doi: 10.1080/87565641.2018.1557190] [Medline: 30590952]
51. Harman-Smith YE, Mathias JL, Bowden SC, Rosenfeld JV, Bigler ED. Wechsler Adult Intelligence Scale-Third Edition profiles and their relationship to self-reported outcome following traumatic brain injury. J Clin Exp Neuropsychol 2013;35(8):785-798. [doi: 10.1080/13803395.2013.824554] [Medline: 23947758]
52. Faria AL, Pinho MS, Badia SB. Capturing expert knowledge for the personalization of cognitive rehabilitation: study combining computational modeling and a participatory design strategy. JMIR Rehabil Assist Technol 2018 Dec 06;5(2):e10714 [FREE Full text] [doi: 10.2196/10714] [Medline: 30522994]

Abbreviations

AUC: area under the receiver operating characteristic curve
ITA: intelligent therapy assistant
RAVLT: Rey Auditory Verbal Memory Test

Edited by G Eysenbach; submitted 20.02.21; peer-reviewed by K Harrington; comments to author 12.03.21; revised version received 04.05.21; accepted 16.05.21; published 10.11.21.

Please cite as:
Garcia-Rudolph A, Opisso E, Tormos JM, Madai VI, Frey D, Becerra H, Kelleher JD, Bernabeu Guitart M, López J
Toward Personalized Web-Based Cognitive Rehabilitation for Patients With Ischemic Stroke: Elo Rating Approach
JMIR Med Inform 2021;9(11):e28090
URL: https://medinform.jmir.org/2021/11/e28090
doi:10.2196/28090
PMID:34757325

Original Paper

Implementation of an Anticoagulation Practice Guideline for COVID-19 via a Clinical Decision Support System in a Large Academic Health System and Its Evaluation: Observational Study

Surbhi Shah[1], MBBS; Sean Switzer[1], DO; Nathan D Shippee[1], PhD; Pamela Wogensen[2], BS; Kathryn Kosednar[2], BSN, RN; Emma Jones[3], MD; Deborah L Pestka[4], PharmD, PhD; Sameer Badlani[2], MD; Mary Butler[5], PhD; Brittin Wagner[5], PhD; Katie White[5], EdD, MBA; Joshua Rhein[6], MD; Bradley Benson[6], MD; Mark Reding[6], MD; Michael Usher[6], MD, PhD; Genevieve B Melton[3], MD, PhD; Christopher James Tignanelli[3], MS, MD

[1]University of Minnesota, Minneapolis, MN, United States
[2]Information Technology, Fairview Health Services, Minneapolis, MN, United States
[3]Department of Surgery, University of Minnesota, Minneapolis, MN, United States
[4]College of Pharmacy, University of Minnesota, Minneapolis, MN, United States
[5]School of Public Health, University of Minnesota, Minneapolis, MN, United States
[6]Department of Medicine, University of Minnesota, Minneapolis, MN, United States

Corresponding Author:
Christopher James Tignanelli, MS, MD
Department of Surgery
University of Minnesota
420 Delaware St SE, MMC 195
Minneapolis, MN, 55455
United States
Phone: 1 6126261968
Email: ctignane@umn.edu

Abstract

Background: Studies evaluating strategies for the rapid development, implementation, and evaluation of clinical decision support (CDS) systems supporting guidelines for diseases with a poor knowledge base, such as COVID-19, are limited.

Objective: We developed an anticoagulation clinical practice guideline (CPG) for COVID-19, which was delivered and scaled via CDS across a 12-hospital Midwest health care system. This study represents a preplanned 6-month postimplementation evaluation guided by the RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework.

Methods: The implementation outcomes evaluated were reach, adoption, implementation, and maintenance. To evaluate effectiveness, the association of CPG adherence on hospital admission with clinical outcomes was assessed via multivariable logistic regression and nearest neighbor propensity score matching. A time-to-event analysis was conducted. Sensitivity analyses were also conducted to evaluate the competing risk of death prior to intensive care unit (ICU) admission. The models were risk adjusted to account for age, gender, race/ethnicity, non-English speaking status, area deprivation index, month of admission, remdesivir treatment, tocilizumab treatment, steroid treatment, BMI, Elixhauser comorbidity index, oxygen saturation/fraction of inspired oxygen ratio, systolic blood pressure, respiratory rate, treating hospital, and source of admission. A preplanned subgroup analysis was also conducted in patients who had laboratory values (D-dimer, C-reactive protein, creatinine, and absolute neutrophil to absolute lymphocyte ratio) present. The primary effectiveness endpoint was the need for ICU admission within 48 hours of hospital admission.

Results: A total of 2503 patients were included in this study. CDS reach approached 95% during implementation. Adherence achieved a peak of 72% during implementation. Variation was noted in adoption across sites and nursing units. Adoption was the highest at hospitals that were specifically transformed to only provide care to patients with COVID-19 (COVID-19 cohorted hospitals; 74%-82%) and the lowest in academic settings (47%-55%). CPG delivery via the CDS system was associated with improved adherence (odds ratio [OR] 1.43, 95% CI 1.2-1.7; *P*<.001). Adherence with the anticoagulation CPG was associated with a significant reduction in the need for ICU admission within 48 hours (OR 0.39, 95% CI 0.30-0.51; *P*<.001) on multivariable logistic regression analysis. Similar findings were noted following 1:1 propensity score matching for patients who received adherent versus nonadherent care (21.5% vs 34.3% incidence of ICU admission within 48 hours; log-rank test *P*<.001).

Conclusions: Our institutional experience demonstrated that adherence with the institutional CPG delivered via the CDS system resulted in improved clinical outcomes for patients with COVID-19. CDS systems are an effective means to rapidly scale a CPG across a heterogeneous health care system. Further research is needed to investigate factors associated with adherence at low and high adopting sites and nursing units.

(JMIR Med Inform 2021;9(11):e30743) doi:10.2196/30743

KEYWORDS

COVID-19; anticoagulation; clinical practice guideline; evidence-based practice; clinical decision support; implementation science; RE-AIM

Introduction

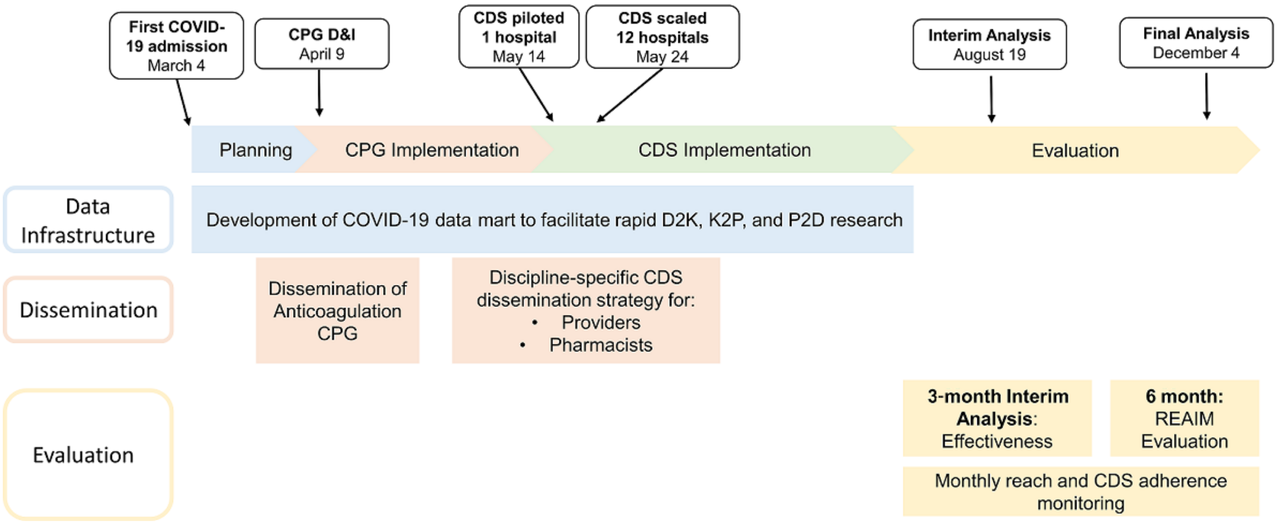
COVID-19, caused by SARS-CoV-2, has infected millions of people worldwide. This disease has shown many unique attributes including a hypercoagulable profile [1-4]. COVID-19–associated coagulopathy results in widespread macrovascular and microvascular thrombosis that contributes to multisystem organ failure and thus contributes to significant mortality and morbidity [5]. Observational and recent randomized controlled studies involving COVID-19 and other viral pneumonias have suggested that routine anticoagulation is associated with improved clinical outcomes for hospitalized patients [6-10]. Considering this, our health care system developed a clinical practice guideline (CPG) delivered as a clinical decision support (CDS) system to facilitate guideline-driven anticoagulation for COVID-19 patients.

CDS technology solutions offer a mechanism that in support of the learning health system (LHS) facilitates long-term process and quality measure improvements [11,12]. CDS systems leverage electronic health records (EHRs) to deliver process-specific information to health care teams, aiding clinical decision-making. When designed well and implemented effectively, CDS systems have been shown to improve adherence with evidence-based practice and, in some cases, improve clinical outcomes [11-13]. Unfortunately, the best practices for successful implementation and scaling of CDS are still unknown [14]. Furthermore, very little is known about developing, implementing, and scaling CDS interventions during a pandemic with a rapidly changing evidence base and strained clinical resources across diverse sites.

On April 9, 2020, our institution developed and disseminated a 3-tiered CPG for anticoagulation in COVID-19 in collaboration with national experts (Multimedia Appendix 1). Given the rapid evolution of evidence, this CPG has evolved over time to reflect the best practice based on the available evidence at the time [4,7]. To maximize the dissemination and reach of the CPG, a CDS solution was developed to deliver the CPG, including both passive and interruptive alerts, piloted at a single site on May 14, 2020, and was successfully scaled across a 12-hospital Midwest health care system on May 24, 2020. An interim reach, adoption, and effectiveness evaluation occurred 3 months following implementation on August 19, 2020 (Figure 1).

This study represents a preplanned 6-month implementation evaluation guided by the RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework [15] of the anticoagulation CPG CDS system for patients admitted with COVID-19 between March 4, 2020, and December 4, 2020, across a large 12-hospital Midwest health care system.

Figure 1. Overall development, dissemination, implementation, and evaluation strategy. CDS: clinical decision support; D&I: Dissemination and Implementation; CPG: clinical practice guideline; D2K: data to knowledge; K2P: knowledge to practice; P2D: practice to data; RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance.


First COVID-19 admission March 4
CPG D&I April 9
CDS piloted 1 hospital May 14
CDS scaled 12 hospitals May 24
Interim Analysis August 19
Final Analysis December 4
Planning
CPG Implementation
CDS Implementation
Evaluation
Data Infrastructure
Development of COVID-19 data mart to facilitate rapid D2K, K2P, and P2D research
Dissemination
Dissemination of Anticoagulation CPG
Discipline-specific CDS dissemination strategy for:
• Providers
• Pharmacists
Evaluation
3-month Interim Analysis: Effectiveness
6 month: REAIM Evaluation
Monthly reach and CDS adherence monitoring

Methods

Context and Evidence Synthesis

A COVID-19 evidence-based medicine (EBM) team was created in March 2020 to rapidly review, catalogue, and publicly disseminate evidence related to proposed COVID-19 therapeutics including anticoagulation management [16]. Due to a lack of high-quality evidence from randomized controlled trials and a lack of expert guidelines, the EBM team developed a novel rubric to grade COVID-19 evidence [17]. Using this rubric, a multidisciplinary team, including COVID-19 physicians, LHS researchers, pharmacists, public health epidemiologists, and medical librarians, reviewed and graded the evidence to date for anticoagulation in COVID-19.

CPG Development

Guided by the EBM team's recommendations, system hematology service line leads (S Shah and MR) oversaw the development of a CPG in collaboration with national hematology experts (Multimedia Appendix 1). This CPG was initially disseminated beginning April 9, 2020. Significant controversy existed at the time regarding the appropriate anticoagulation strategy, with some studies suggesting no anticoagulation (*conservative treatment approach*) due to concerns for bleeding and disseminated intravascular coagulopathy, and some studies suggesting tissue plasminogen activator infusions (*aggressive treatment approach*) for patients with severe respiratory failure [8,18]. Ultimately, the CPG adhered to a "middle of the road" approach by instituting universal prophylactic weight-based anticoagulation for all patients. Similar to anticoagulation CPGs for other disease processes [19], we incorporated a risk stratification model, whereby the intensity of anticoagulation was increased to moderate intensity (0.5 mg/kg BID enoxaparin or low intensity heparin infusion in case of renal failure) for patients with a history of thrombosis, cancer, admission to the intensive care unit (ICU), or D-dimer >10 times the upper limit of normal. The CPG is a "living" framework, and has since undergone several iterations of modifications (upstratification of patients with a prior history of deep venous thrombosis or cancer, and ICU patients, and exclusion of pregnant patients) based on evolution of the evidence.

CDS Development, Dissemination, and Implementation

The COVID-19 anticoagulation CPG CDS system was developed by the M Health Fairview clinical informatics development team (KK and PW) in collaboration with the Associate Chief Medical Informatics Office for CDS (S Switzer) in May 2020. Multimedia Appendix 2 displays a process map for the CDS system. In brief, the CDS solution includes the following (representative screenshots displayed in Multimedia Appendix 3 and Multimedia Appendix 4): a tiered anticoagulation order set, a passive "reminder" of the anticoagulation CPG for COVID-19 patients without anticoagulation orders displayed in the EHR admission and transfer navigator, 3 "triggers" that activate interruptive alerts, and various interactive and interruptive alerts.

The interruptive best practice advisories were essentially "safety checks" to surveil if patients were on venous thromboembolism (VTE) chemoprophylaxis on admission or if the criteria for VTE risk changed (eg, increase in D-dimer above the threshold or transfer to the ICU) and were only triggered for providers with ordering privileges. To ensure the best practice advisory would not trigger for patients who have recovered from COVID-19, an infection status of *Recovered COVID-19* was built into the EHR.

Development, dissemination, and implementation followed our system protocol SCALED (scaling acceptable CDS) for CDS user-centered design, pilot testing, scaling, and evaluation. Prior to pilot testing, the CDS underwent iterative user interface/user experience improvement during May 2020. The CDS was piloted on May 14, 2020, and scaled on May 24, 2020. To support adoption and usability, a discipline-specific CDS dissemination strategy was carried out in May 2020 (Figure 1). To ensure embeddedness of the CDS in provider and pharmacist workflow, dissemination overlapped with implementation for 1 month after the intervention went live or was "turned on." The specific dissemination strategies are presented in Textbox 1.

Textbox 1. Dissemination strategies.

Providers

- The clinical decision support (CDS) system was presented to intensivist, hospital medicine, and primary care practice groups via formal didactic methods.
- Our system utilized a daily workflow document for intensivists and hospital medicine providers caring for COVID-19 patients representing best practices, recent publications, and ongoing trials. The anticoagulation CDS was integrated into this workflow document and remained as a constant on this document throughout the implementation period.
- In the university setting, CDS was presented at grand rounds on divisional, departmental, and medical school platforms.

Pharmacy

- The CDS system was presented routinely at the System-wide Anticoagulation Committee.

All

- Our system utilized a COVID-19 intranet COVID-19 resource page. This CDS was placed within the system guidelines for management of COVID-19 patients.
- The CDS was also posted on the University of Minnesota evidence-based medicine COVID-19 website, on a public facing webpage for COVID-19 evidence-based practice geared toward clinicians.

Data Extraction and Evaluation

Members of the study team (CJT, GBM, and MU) developed a COVID-19 data mart to facilitate near real-time evaluation of the CDS. Structured query language was used to automate daily export of COVID-19 EHR data into the data repository. A preprocessing pipeline was developed and implemented using Python 3.7.3. (CreateSpace) and Stata 16 (StataCorp) to generate a flat file for each patient, including patient anticoagulation risk stratification, tier of anticoagulation received, reach, adherence, clinical outcomes (in-hospital and out-of-hospital mortality, complications, ICU admission, and mechanical ventilation), comorbidities, home medications, inpatient medications received (eg, remdesivir, tocilizumab, and steroids), daily laboratory and vital data, and demographics. LogicStream Health (LogicStream Health Inc), an analysis platform for EHR content, was utilized to evaluate order set, passive, and interruptive alert utilization.

CDS RE-AIM Evaluation

A preplanned 6-month implementation evaluation was conducted with guidance from the RE-AIM framework.

Reach was defined as the number of patients admitted each month, who received CDS (CDS reach) or appropriate anticoagulation (CPG reach) (numerator) over the number of patients admitted each month with COVID-19 (denominator). These 2 definitions of reach were used to facilitate internal performance monitoring. For example, the state where CPG reach is high but CDS reach is low represents integration of the CPG into normal workflow without the need for CDS.

Adoption was defined at the implementation site and nursing unit level as the number of patients admitted each month, who received guideline-concordant anticoagulation therapy (numerator) over the number of patients admitted each month with COVID-19 (denominator). It was not possible to define adoption accurately at the provider level as we were unable to assign a single provider responsible for a patient's initial care. Patient's receive orders from a variety of provider types, including house staff, advanced practice providers, or attending physicians either within the emergency department or the inpatient team, and thus we are unable to assign a single provider responsible for CPG adherence.

An *implementation* evaluation was conducted to investigate the effect of various CDS alert methods (passive and/or interruptive alerts) on anticoagulation CPG adherence. *Adherence* was defined at the patient level as receiving guideline-concordant care within 24 hours of admission. Additionally, CPG fidelity was evaluated for each VTE risk stratification (Multimedia Appendix 1).

To evaluate *maintenance*, following a wash-out period without any continued dissemination, we evaluated adherence during months 5 and 6 after implementation.

Statistical Approach

To evaluate *effectiveness*, the association of CPG adherence on admission (at the patient level) with clinical outcomes was assessed via multiple methods.

First, multivariable logistic regression was used for binary dependent variables and negative binomial regression was used for continuous variables with a skewed distribution (hospital length of stay) using all 2406 patients who either received adherent care (n=1650) or did not receive adherent care (n=853). All models were risk adjusted using the confounding variables included below.

Second, 1:1 nearest neighbor propensity score matching was used to create cohorts of patients who received CPG-adherent care (exposure or treatment of interest). Univariate logistic regression was then used to compare the need for ICU admission within 48 hours (primary outcome) for patients who received (vs did not receive) CPG-adherent care on admission (exposure). Kaplan-Meier curves were also estimated via a time-to-event analysis (censored at 48 hours following hospital admission) and compared using the log-rank test. Propensity scores were estimated with logistic regression using the confounding variables listed below. Two evenly matched groups were formed with the common caliper set at 0.01. Following matching, there were 1342 patients included (671 patients in each propensity-matched cohort). Standardized difference was evaluated prior to and after propensity matching to ensure that the standardized difference was <0.1 in the propensity-matched cohort for each confounding variable (Multimedia Appendix 5). The distribution of propensity scores was well balanced between propensity-matched cohorts (Multimedia Appendix 6).

Third, to account for the competing risk of death prior to ICU admission by 48 hours, which occurred in 5 patients, a competing risk regression model (censored at 48 hours following hospital admission) was used in the propensity-matched cohort. Cumulative incidence curves were also estimated. Due to the importance of age as a confounding variable, age-stratified cumulative incidence curves were also generated.

The primary clinical outcome for the above models was the need for ICU admission within 48 hours of hospital admission. This endpoint was chosen clinically as the primary outcome because adherence with anticoagulation best practices is hypothesized to reduce microthrombosis and macrothrombosis events, and minimize progression of the disease and critical illness.

Secondary outcomes of interest for the above models were also evaluated, including all-cause in-hospital mortality, the need for ICU admission at any time during hospitalization, the need for mechanical ventilation, hospital length of stay, and the development of VTE or bleeding complications. Additionally, a binary composite outcome metric was developed and coded as positive if a patient had all-cause in-hospital mortality, required ICU admission, required mechanical ventilation, or required a hospital length of stay greater than 7 days.

The exposure or treatment of interest for the above models was defined as a binary variable if patients received guideline-adherent care on hospital admission.

With regard to confounding variables for the above models, variables known to be associated with the outcome of more severe COVID-19 infection (defined as requiring ICU admission or mechanical ventilation) were included as confounding variables for all analyses. This list of variables was developed by our team of subject matter experts with clinical and research expertise managing patients with COVID-19. All models were risk adjusted to account for patient-level baseline demographics (age, gender, race/ethnicity, English vs non-English speaking, and area deprivation index [a marker of neighborhood socioeconomic status] [20]), the month of admission, in-hospital treatments for COVID-19 (remdesivir, tocilizumab, and steroids), BMI, Elixhauser comorbidity index, the most aberrant vital signs within the first 24 hours of hospital admission (minimum saturation/FiO2 ratio, minimum systolic blood pressure, and maximum respiratory rate), the initial hospital of treatment, and the source of admission (home, emergency department, skilled nursing facility, intrahospital transfer, prescheduled admission for surgery, and admission from a clinic/office appointment).

With regard to subgroup analysis, of the 2503 patients, initial D-dimer, C-reactive protein, creatinine, and absolute neutrophil to absolute lymphocyte ratio data were present for 1181 patients. As these laboratory values have been shown on admission to be predictive of worse clinical outcomes [21-23], a secondary analysis was conducted in these 1181 patients.

With regard to data missingness, overall missingness was low (<2.04% for any individual variable, with 3.9% of patients missing at least one covariate). Given the low rate of missingness, imputation was deemed unnecessary [24].

Statistical analyses were performed using Stata MP, version 16 (StataCorp). Statistical significance was defined as a *P* value <.05.

Data Availability

The data underlying this article were provided by M Health Fairview (Minneapolis, MN) with permission from M Health Fairview Research and IT. Data will be shared on request to the corresponding author with the permission of M Health Fairview.

Results

Patient Characteristics

A total of 2503 patients required in-hospital admission during the study period, with polymerase chain reaction–confirmed COVID-19 (Multimedia Appendix 7). The median patient age was 64.9 years (IQR 48.4-77.7 years). Of the patients, 1180 (47.1%) were male and 262 (10.5%) had in-hospital death. The baseline characteristics of patients who received CPG-adherent (vs nonadherent) care are shown in Table 1. Similarly, the baseline unadjusted clinical outcomes by CPG adherence are shown in Multimedia Appendix 8.

Table 1. Patient characteristics.

Characteristic	Did not receive adherent anticoagulation (n=853)	Received adherent anticoagulation (n=1650)	*P* value[a]
Age (years), median (IQR)	60.1 (35.2-75.7)	66.2 (52.7-78.4)	<.001
Race, n (%)			.60
White	476 (55.8)	945 (57.3)	
Black	117 (13.7)	187 (11.3)	
Asian	105 (12.3)	211 (12.8)	
Hispanic	62 (7.3)	114 (6.9)	
Declined	74 (8.7)	161 (9.8)	
Other	19 (2.2)	32 (1.9)	
Male, n (%)	350 (41.0)	830 (50.3)	<.001
Area deprivation index quintile, n (%)			.58
0%-19%	168 (19.7)	312 (18.9)	
20%-39%	256 (30.0)	499 (30.2)	
40%-59%	231 (27.1)	478 (29.0)	
60%-79%	114 (13.4)	227 (13.8)	
80%-100%	84 (9.8)	134 (8.1)	
Non-English Speaking, n (%)	233 (27.3)	477 (28.9)	.40
Elixhauser comorbidity index, median (IQR)	4.0 (1.0-8.0)	5.0 (2.0-8.0)	<.001
BMI, median (IQR)	28.6 (24.6-33.6)	29.8 (25.7-35.4)	<.001
Lowest systolic blood pressure in the first 24 hours (mmHg), median (IQR)	111.0 (98.0-124.0)	113.0 (100.0-127.0)	.01
Highest respiratory rate in the first 24 hours (bpm), median (IQR)	22.0 (18.0-29.0)	24.0 (20.0-32.0)	<.001
Lowest S/F[b] ratio in the first 24 hours, median (IQR)	438.1 (320.0-459.5)	355.6 (286.4-447.6)	<.001
Initial D-dimer, median (IQR)	1.2 (0.7-2.3)	1.1 (0.6-2.0)	.02
Initial CRP[c], median (IQR)	64.3 (24.0-125.0)	72.0 (30.8-132.0)	.18
Initial creatinine, median (IQR)	1.0 (0.8-1.4)	1.0 (0.8-1.3)	.39
Initial NLR[d], median (IQR)	5.1 (3.1-8.4)	4.9 (3.0-8.6)	.97
Received remdesivir, n (%)	203 (24.2)	843 (51.1)	<.001
Received tocilizumab, n (%)	26 (3.0)	90 (5.5)	.007
Received steroids, n (%)	153 (17.9)	575 (34.8)	<.001
Admission month of 2020, n (%)			.06
March	18 (2.1)	20 (1.2)	
April	64 (7.5)	103 (6.2)	
May	90 (10.6)	238 (14.4)	
June	51 (6.0)	103 (6.2)	
July	65 (7.6)	121 (7.3)	
August	100 (11.7)	159 (9.6)	
September	70 (8.2)	112 (6.8)	
October	143 (16.8)	291 (17.6)	
November	252 (29.5)	503 (30.5)	
Implementation site, n (%)			<.001

Characteristic	Did not receive adherent anticoagulation (n=853)	Received adherent anticoagulation (n=1650)	*P* value[a]
Hospital 0	13 (1.5)	59 (3.6)	
Hospital 1	14 (1.6)	26 (1.6)	
Hospital 2	182 (21.3)	363 (22.0)	
Hospital 3	12 (1.4)	21 (1.3)	
Hospital 4	18 (2.1)	51 (3.1)	
Hospital 5	134 (15.7)	268 (16.2)	
Hospital 6	135 (15.8)	264 (16.0)	
Hospital 7	56 (6.6)	163 (9.9)	
Hospital 8	58 (6.8)	50 (3.0)	
Hospital 9	110 (12.9)	148 (9.0)	
Hospital 10	72 (8.4)	152 (9.2)	
Hospital 11	49 (5.7)	85 (5.2)	
Source of admission, n (%)			<.001
Home	391 (46.0)	672 (40.8)	
Emergency department	374 (44.0)	815 (49.5)	
Skilled nursing facility	36 (4.2)	62 (3.8)	
External hospital transfer	33 (3.9)	87 (5.3)	
Admission for surgery	8 (0.9)	1 (0.1)	
Clinic	8 (0.9)	11 (0.7)	

[a] The Pearson chi-square test was used to compare categorical and binary variables, and the Wilcoxon rank-sum test was used to compare continuous variables with a skewed distribution.

[b] S/F: oxygen saturation to fraction of inspired oxygen.

[c] CRP: C-reactive protein.

[d] NLR: neutrophil-to-lymphocyte ratio.

Reach

Figure 2 displays the reach of the CPG by month. Reach was purposefully measured in 2 ways (CPG reach or CDS reach). CPG reach was measured by the percentage of patients who received appropriate anticoagulation. CDS reach was measured by the percentage of patients whose providers received a CDS "reminder" to adhere to the CPG (Figure 2). Figure 2 displays the combined CPG and CDS reach (blue line) compared with CDS reach alone (red line). In an ideal setting, reach (blue line) would approach 1.0 and CDS reach (red line) would approach 0. This would represent the state where all patients are receiving adherence without the need for interruptive CDS and would reflect complete uptake of the CPG by providers. Baseline reach of the anticoagulation CPG in April was approximately 61%. System-wide implementation of a CDS strategy in May resulted in 97% reach. The reduced triggering of CDS after August represents increased ordering of any anticoagulation for COVID-19 patients.

CPG reach improved following implementation of the CDS system. CPG reach peaked during piloting and scaling of the CDS system with an adherence rate of 74.4%. In the 6 months since scaling, adherence averaged 67% (Figure 3).

Figure 2. Average implementation reach by month. The blue line represents the combined CPG (patient received adherent anticoagulation) and CDS reach (patient's ordering providers received the CDS system suggesting adherent anticoagulation) by month. The red line represents only CDS reach. CDS: clinical decision support; CPG, clinical practice guideline.

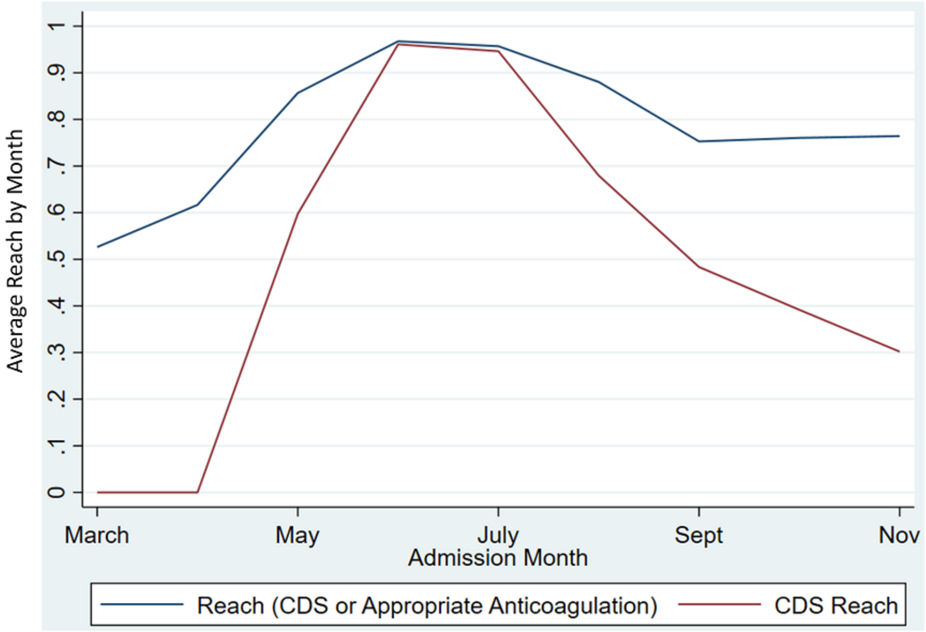
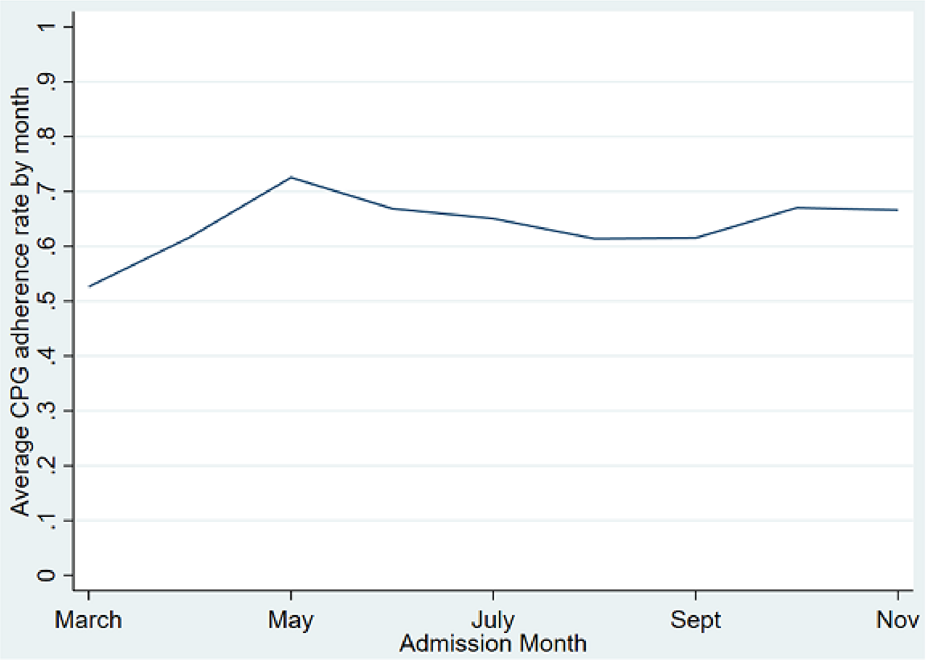


Figure 3. Average implementation reach by month. (A) Average CPG reach by health care system by month. CDS: clinical decision support; CPG: clinical practice guideline.



Effectiveness

The *primary hypothesis* tested was if adherence with the anticoagulation CPG on hospital admission was associated with a reduced need for ICU management by 48 hours. Adherence with the anticoagulation CPG was independently associated with reduced need for ICU admission within 48 hours of hospital admission on multivariable logistic regression analysis (odds ratio [OR] 0.39, 95% CI 0.30-0.51; *P*<.001) (Table 2 and Multimedia Appendix 9). In the propensity-matched cohorts, patients who received CPG-adherent care on admission had a 21.46% incidence of ICU admission within 48 hours compared with 34.28% for patients who did not receive adherent care on admission (chi-square *P*<.001; logistic regression OR 0.52; *P*<.001). A time-to-event analysis was also conducted. Patients who received adherent care on admission (vs patients who did not) were more likely to not require ICU admission by 48 hours (log-rank test *P*<.001; Multimedia Appendix 10). Five patients died prior to 48 hours, and thus, to account for this competing risk, a competing risk-regression analysis was conducted. Patients who received adherent care according to the CPG on admission had significantly reduced hazards for ICU admission by 48 hours when accounting for the competing risk of death (subhazard ratio 0.58, *P*<.001). Cumulative incidence functions are provided in Multimedia Appendix 11. As older patients may elect for comfort measures and die in the hospital ward in lieu of aggressive ICU care, an age-stratified cumulative incidence function is provided in Multimedia Appendix 12.

Secondary outcome analysis identified that adherence with the anticoagulation CPG was associated with reduced need for ICU admission at any point during hospitalization (OR 0.53, 95% CI 0.42-0.69; *P*<.001) and reduced all-cause in-hospital mortality (OR 0.67, 95% CI 0.48-0.94; *P*=.02; Table 2). Adherence with the anticoagulation CPG significantly reduced the odds of death, ICU admission, requirement for mechanical ventilation, and hospital length of stay greater than 7 days (OR 0.75, 95% CI 0.60-0.94; *P*=.01). Adherence with the anticoagulation CPG was independently associated with reduced bleeding complications (OR 0.39, 95% CI 0.21-0.72; *P*=.003), but not VTE complications (OR 0.87, 95% CI 0.65-1.17; *P*=.40). Adherence with the anticoagulation CPG was independently associated with an increased hospital length of stay (incident rate ratio [IRR] 1.15, 95% CI 1.08-1.22; *P*<.001). This effect persisted on excluding patients who had in-hospital death (IRR 1.13, 95% CI 1.06-1.2; *P*<.001). None of the other secondary analyses reached statistical significance (Table 2).

Table 2. Likelihood of adherence with the clinical practice guideline on multivariable logistic regression.

Variable	Odds ratio for CPG[a] adherence (vs nonadherence)	95% CI	*P* value	C-statistic[b]
Model 1: Risk adjustment without initial labs (n=2406)				
ICU[c] admission within 48 hours	0.39	0.30-0.51	<.001	0.87
ICU admission	0.53	0.42-0.69	<.001	0.87
Required mechanical ventilation	1.18	0.79-1.77	.40	0.93
All-cause in-hospital mortality	0.67	0.48-0.94	.02	0.88
Composite outcome[d]	0.75	0.60-0.94	.01	0.82
VTE[e] complication	0.87	0.65-1.17	.40	0.79
Bleeding complication	0.39	0.21-0.73	.003	0.83
Model 2: Risk adjustment including initial labs (n=1181)				
ICU admission within 48 hours	0.28	0.19-0.43	<.001	0.90
ICU admission	0.44	0.29-0.64	<.001	0.89
Required mechanical ventilation	1.20	0.67-2.20	.50	0.94
All-cause in-hospital mortality	0.92	0.56-1.52	.70	0.88
Composite outcome[d]	0.61	0.42-0.87	.006	0.82
VTE complication	1.05	0.64-1.71	.90	0.81
Bleeding complication	0.47	0.17-1.26	.10	0.87

[a] CPG: clinical practice guideline.

[b] C-statistic or concordance statistic was calculated for each model.

[c] ICU: intensive care unit.

[d] Composite outcome is defined as need for ICU admission, mechanical ventilation, all-cause in-hospital mortality, or hospital length of stay greater than 7 days.

[e] VTE: venous thromboembolism.

In the model that included initial D-dimer, C-reactive protein, creatinine, and the neutrophil-to-lymphocyte ratio, adherence with the anticoagulation CPG was independently associated with reduced need for ICU admission within 48 hours of hospital admission (OR 0.28, 95% CI 0.19-0.43; *P*<.001) or the need for ICU admission at any time during hospitalization (OR 0.44, 95% CI 0.29-0.64; *P*<.001) (Table 2). Adherence with the anticoagulation CPG significantly reduced the odds of death, ICU admission, requirement for mechanical ventilation, or hospital length of stay greater than 7 days (OR 0.61, 95% CI 0.42-0.87; *P*=.006). Adherence with the anticoagulation CPG was independently associated with an increased hospital length of stay (IRR 1.12, 95% CI 1.03-1.22; *P*=.008). This effect persisted on excluding patients who had in-hospital death (IRR 1.1, 95% CI 1.004-1.2; *P*=.04). None of the other secondary analyses reached statistical significance (Table 2).

Adoption

To investigate adoption rates across the system, we evaluated adoption by hospital. Our system includes 12 hospitals, 2 university settings that include resident and fellow trainees, 2 COVID-19 cohorted hospitals [25] staffed by attending physicians and advanced practice providers, and 8 community hospitals staffed by attending physicians and advanced practice providers. Adoption was the highest at the COVID-19 cohorted hospitals and lowest at the university hospitals (Multimedia Appendix 13). Variability was similarly noted across nursing units (Multimedia Appendix 14). No discernable difference was noted in adoption analyses performed by patient race/ethnicity, encounter type, or investigation for COVID-19 status on admission (versus known COVID-19).

Implementation and Maintenance

Adherence was evaluated in the context of CDS. Adherence when CDS was delivered was 70% as compared to 62% without CDS (OR 1.43, 95% CI 1.2-1.7; *P*<.001).

The CDS 5 Rights Framework recommends delivery of CDS at the "right" time in workflow. Four passive CDS elements were included to facilitate CPG adherence within various areas of the EHR and clinical workflow. For example, passive CDS was delivered within EHR navigators used during the admission, discharge, or transfer (ADT) workflow, during the rounding navigator, and within the general EHR order environment. Adherence with anticoagulation was the highest when these passive elements were integrated within the admission (75%), rounding (75%), or transfer (80%) navigators (Multimedia Appendix 15) than when outside of an EHR care navigator (57%).

We then sought to investigate the relationship between adherence and passive versus interruptive CDS intervention formats. Overall, 1423 (56.9%) patients had no CDS elements that were passive or interruptive. On the other hand, 699 (27.9%) patients had passive-only CDS delivered to providers, 111 (4.4%) patients had interruptive-only CDS delivered to providers, and 270 (10.8%) patients had a combination of passive and interruptive CDS delivered to providers. The combination of passive CDS and interruptive alerts was associated with the highest adherence with the anticoagulation CPG (Multimedia Appendix 16).

Variation in adherence was noted across baseline risk groups. Patients in the moderate-risk group were less likely to receive adherent care (606/1016, 59.6%) compared with patients in the high- (145/223, 65%) and low-risk groups (899/1264, 71%). Following implementation, wash-out maintenance stabilized at 67% in months 5 to 6 (October to November 2020).

Discussion

Real-World Application of the LHS

This study represents the completed iteration of a continuous LHS cycle [26]. Adherence with the anticoagulation CPG was associated with significantly improved clinical outcomes. Adoption improved following the delivery of the CPG within a CDS system. Despite these improvements, variation was found in adoption across hospitals and units. Adoption was the highest at hospitals specializing in treating patients with COVID-19 and was the lowest in tertiary academic hospitals. An evaluation of CDS delivery methods identified that the combination of passive and interruptive alerts was associated with the highest adherence rate.

This study provides an important and early example of the real-world application of the LHS during COVID-19, a period with surged clinical resources and uncertain evidence base. Critical to our success was the early development of a COVID-19 data mart that included highly granular structured and unstructured patient-level data. Integration with an EHR analysis solution (LogicStream Health) facilitated near real-time evaluation of CDS alert activities by providers.

Our health care system has a rigorous and validated protocol for the development, implementation, scaling, and evaluation of user-centered CDS systems with over 20 use cases implemented each year overseen by various enterprise CDS committees. Typically, the process of development, implementation, and scaling requires months, and in this case, it occurred in a matter of weeks. COVID-19 provided a heightened sense of urgency and purpose in health care research and quality improvement that resulted in rapid progress in CDS development. The dedicated EBM team facilitated prompt CPG updates in response to rapidly changing evidence. Augmented stakeholder engagement and buy-in from the informatics development team were also critical elements for success. The combination of expedited access to fully preprocessed and analyzable EHR data updated daily along with extraordinary team engagement and stakeholder support were critical for rapid implementation.

Despite these successes, the room for optimization was identified from this analysis. First, in an attempt to minimize alert fatigue, the CDS system was initially designed to only trigger for patients with COVID-19 but not on anticoagulation. While it was successful for the months of June and July in achieving near 95% reach, based on the data presented in this study, we hypothesize that providers became comfortable attempting to order anticoagulation independent of the order set, resulting in patients being on incorrect anticoagulation and preventing corrective triggering of the CDS system. Others have shared this experience, where an attempt to develop a user-centered CDS system minimizing alerts resulted in a system that was overly passive and could not change behavior [27]. Despite all the negative press for interruptive alerts, we were surprised that interruptive alerts and the combination of interruptive alerts and passive CDS were associated with improved adherence compared to passive CDS alone. It is possible due to the COVID-19 pandemic and the augmented sense of unity and purpose between clinicians and quality improvement researchers that interruptive alerts were received more favorably.

Second, we identified a large variation in adoption across hospitals and nursing units. Our academic health system is unique in the sense that we created specialty cohorted hospitals for COVID-19 patients [25] needing care across our academic health medical center, which was staffed by attending clinicians well versed with institutional guidelines. Despite the availability of the same resources at all sites, adoption of the CPG via the use of the CDS system was much lower at noncohorted sites. Specifically, we identified that adoption was very poor at university sites where the majority of orders are placed by house staff compared with advanced practice providers at other sites.

Third, following each LHS evaluation cycle (practice to data), it is imperative that positive findings are disseminated widely. We were surprised that adherence did not improve following our interim effectiveness analysis in August 2020, which identified a significant and independent improvement in clinical outcomes with anticoagulation. The unified theory of acceptance and use of technology posits that a key construct affecting technology use intention (in this case, using the CDS system) is performance expectancy [28]. Essentially, if the provider believes that the CDS system will improve patient outcomes, they will have higher intentions to use it. In response to this RE-AIM evaluation, we will pilot an education intervention, provide continuing medical education (CME) credit as an added incentive for participation at sites with lower adoption, and assess the impact at our next Plan-Do-Study-Act (PDSA) cycle.

In the early phase of the pandemic, there was widespread confusion about the underlying pathogenic mechanism and its implications for patient outcomes, leading to highly variable practice in the medical community. To date, limited data exist on the management approach for COVID-19–associated coagulopathy, particularly in high-risk critically ill populations and for patients who are either managed as outpatients or posthospital discharge patients [29,30]. At the time of writing this manuscript, there were 147 (32 from the United States) randomized trials ongoing or recently completed to assess different anticoagulation approaches in COVID-19 [31]. Since the completion of this evaluation in December 2020, results of multiple COVID-19 anticoagulation randomized trials have since been published. While a formal review of the literature was outside the scope of this study, controversy persists as 2 recently published open-label randomized trials offered conflicting evidence, with one suggesting a lack of benefit from therapeutic-dose anticoagulation in critically ill patients [10] and the other showing significant improvement in survival and increased organ support-free days in noncritically ill patients [9]. However, both these studies had multiple limitations [32] and both evaluated therapeutic anticoagulation doses, whereas our consensus guideline includes a tiered approach including intermediate anticoagulation for specific high-risk subsets. During the study period, our tiered approach recommended an intermediate dose for critically ill patients. In March 2021, the INSPIRATION trial reported its findings and did not identify an advantage of intermediate-dose versus prophylactic-dose anticoagulation for critically ill patients with COVID-19 [33]. Although this was an observational study, it provides additional support that adherence with a tiered approach for anticoagulation in patients with COVID-19 is associated with improved clinical outcomes and, in our health care setting, reduced bleeding complications. Interestingly, we noted that adherence was associated with reduced bleeding but not VTE complications. This may suggest that this approach does not impact large vessel VTE, but improved outcomes overall suggest that patients may be developing fewer microvessel thrombi, causing less systemic complications that typically lead to ICU admissions and adverse outcomes.

COVID-19 is a global emergency; given the lack of robust/consistent guidelines from leading societies, institutions had to develop a local approach to create a uniform plan of care and an approach for its implementation. This is specifically problematic in larger systems with multiple hospitals. Our health system was particularly vulnerable to this issue owing to significant heterogeneity resulting from a recent merger (different instances of the same EHR, heterogeneous administrative policies, and site-specific management protocols). As described above, there were concerns for an increased risk of bleeding, and many individual practitioners in our system were apprehensive to order anticoagulation in patients with COVID-19, leading to variable VTE prevention strategies and adverse patient outcomes.

Our study has several limitations. First, our institutional preference for implementation evaluation is typically a mixed-methods approach. However, due to contact precautions surrounding COVID-19 and significantly increased provider workload, it was not feasible to perform a qualitative analysis of staff. Thus, this represents a quantitative-only approach, which may not fully discern specific trends. To expand on hypotheses that arose from this research, a future direction includes a voluntary survey of health care providers (initiated on December 14, 2020) surrounding their familiarity with COVID-19 institutional guidelines and their experiences interacting with the CDS system. Survey question development was guided by unified theory of acceptance and use of technology constructs for technology acceptance. Additionally, while we identified an association with adherence and the effectiveness of anticoagulation, it is important to not misconstrue this analysis. We did not evaluate anticoagulation versus no anticoagulation and the association with outcomes, but rather we evaluated adherence with the guideline versus nonadherence with the guideline. Thus, nonadherent patients could have been receiving either more or less aggressive anticoagulation than the comparison group. In our institution, adherence was associated with improved clinical outcomes; however, this may not be generalizable to other institutions with different baseline cultural practices for anticoagulation management. VTE and bleeding complications were extracted using structured EHR data. These events may be underreported. Our relatively small cohort size (n=2503) and single health care system represent additional limitations of the study.

Conclusion

This study provides an early example of the real-world application of the LHS during COVID-19. With or without a pandemic, there is a need for the implementation of evidence-based practice that is most up-to-date. Traditionally, the largest barrier to this effort has been the need for making major changes in the workflow. With the widespread use of EHRs and increasing consolidation of health care systems, the application of a CPG through the use of a CDS system can offer an easy tool for implementation without adding confusion related to workflow changes, thus bringing uniformity in care at every level in the system and influencing the quality of care and patient outcomes.

Acknowledgments

The authors gratefully acknowledge Fairview Health Services IT staff Iva Ninkovic, Karl Koch, Phill Stokkeland, Sandy Soukup, and Eric Murray. This research was supported by the Agency for Healthcare Research and Quality (AHRQ) and Patient-Centered Outcomes Research Institute (PCORI) (grant K12HS026379) (S Shah and CJT) and the National Institutes of Health's National Center for Advancing Translational Sciences (grant KL2TR002492). Additional support for Minnesota Learning Health System Mentored Career Development Program (MN-LHS) scholars was offered by the University of Minnesota Office of Academic Clinical Affairs and the Division of Health Policy and Management, University of Minnesota School of Public Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the AHRQ, PCORI, or MN-LHS. This study was also supported by the University of Minnesota Clinical and Translational Science Award (grant UL1TR000114).

Authors' Contributions

S Shah contributed to study design, data collection, data analysis, data interpretation, writing, and critical revision. S Switzer contributed to study design, data collection, data analysis, data interpretation, writing, and critical revision. NDS contributed to study design, data interpretation, writing, and critical revision. KK contributed to study design, data interpretation, writing, and critical revision. PW contributed to study design, data interpretation, writing, and critical revision. EJ contributed to study design, data interpretation, writing, and critical revision. DLP contributed to study design, data interpretation, writing, and critical revision. SB contributed to study design, data interpretation, writing, and critical revision. MB contributed to study design, data interpretation, writing, and critical revision. BW contributed to study design, data interpretation, writing, and critical revision. KW contributed to study design, data interpretation, writing, and critical revision. JR contributed to study design, data collection, data interpretation, writing, and critical revision. BB contributed to study design, data collection, data interpretation, writing, and critical revision. MR contributed to study design, data interpretation, writing, and critical revision. MU contributed to study design, data collection, data analysis, data interpretation, writing, and critical revision. GBM contributed to study design, data collection, data analysis, data interpretation, writing, and critical revision. CJT contributed to study design, data collection, data analysis, data interpretation, writing, and critical revision.

Conflicts of Interest

None declared.

Multimedia Appendix 1

M Health Fairview COVID-19 anticoagulation clinical practice guideline.

[PNG File , 567 KB - medinform_v9i11e30743_app1.png]

Multimedia Appendix 2

Process map of the M Health Fairview clinical decision support system.

[PNG File , 4755 KB - medinform_v9i11e30743_app2.png]

Multimedia Appendix 3

Screenshots of the COVID-19 anticoagulation clinical decision support system's passive and interruptive elements.

[PNG File , 296 KB - medinform_v9i11e30743_app3.png]

Multimedia Appendix 4

Screenshots of the COVID-19 anticoagulation clinical decision support system's surveillance of appropriate anticoagulation prophylaxis.

[PNG File , 316 KB - medinform_v9i11e30743_app4.png]

Multimedia Appendix 5

Propensity score standardized differences prior to and after matching.

[DOCX File , 14 KB - medinform_v9i11e30743_app5.docx]

Multimedia Appendix 6

Propensity score distribution in propensity-matched cohorts. Blue/untreated refers to the propensity-matched patient cohort that did not receive CPG adherent care on hospital admission. Red/treated refers to the propensity-matched patient cohort that did receive CPG adherent care on hospital admission. CPG: clinical practice guideline.

[PNG File , 1544 KB - medinform_v9i11e30743_app6.png]

Multimedia Appendix 7

Study diagram for the selection of patients from the COVID-19 inpatient database.
[PNG File , 56 KB - medinform_v9i11e30743_app7.png]

Multimedia Appendix 8
Clinical outcomes.
[DOCX File , 14 KB - medinform_v9i11e30743_app8.docx]

Multimedia Appendix 9
Full model output for multivariable logistic regression evaluating the association of adherence with the clinical practice guideline on hospital admission with the need for intensive care within 48 hours.
[DOCX File , 17 KB - medinform_v9i11e30743_app9.docx]

Multimedia Appendix 10
Kaplan-Meier failure estimates (for intensive care unit admission by 48 hours). Kaplan-Meier failure estimates for patients who received CPG adherent care (red line) versus those who did not receive CPG adherent care (blue line). CPG: clinical practice guideline.
[PNG File , 2833 KB - medinform_v9i11e30743_app10.png]

Multimedia Appendix 11
Cumulative incidence function of intensive care unit (ICU) admission by 48 hours. Cumulative incidence function of ICU admission by 48 hours for patients who received CPG adherent care (red line) versus those who did not receive CPG adherent care (blue line). CPG: clinical practice guideline.
[PNG File , 3224 KB - medinform_v9i11e30743_app11.png]

Multimedia Appendix 12
Age-stratified cumulative incidence function of intensive care unit (ICU) admission by 48 hours. Age-stratified cumulative incidence function of ICU admission by 48 hours for patients who received CPG adherent care versus those who did not receive CPG adherent care. CPG: clinical practice guideline.
[PNG File , 4642 KB - medinform_v9i11e30743_app12.png]

Multimedia Appendix 13
Mean adoption by implementation site.
[PNG File , 86 KB - medinform_v9i11e30743_app13.png]

Multimedia Appendix 14
Mean adoption by nursing unit.
[PNG File , 80 KB - medinform_v9i11e30743_app14.png]

Multimedia Appendix 15
Adherence with the anticoagulation CPG by passive CDS elements. CDS: clinical decision support; CPG: clinical practice guideline.
[PNG File , 2132 KB - medinform_v9i11e30743_app15.png]

Multimedia Appendix 16
Adherence with the anticoagulation CPG by CDS type. CDS: clinical decision support; CPG: clinical practice guideline.
[PNG File , 68 KB - medinform_v9i11e30743_app16.png]

References

1. Mei H, Luo L, Hu Y. Thrombocytopenia and thrombosis in hospitalized patients with COVID-19. J Hematol Oncol 2020 Dec 01;13(1):161-161 [FREE Full text] [doi: 10.1186/s13045-020-01003-z] [Medline: 33261634]
2. Kyriakoulis KG, Kokkinidis DG, Kyprianou IA, Papanastasiou CA, Archontakis-Barakakis P, Doundoulakis I, et al. Venous thromboembolism in the era of COVID-19. Phlebology 2021 Mar 10;36(2):91-99. [doi: 10.1177/0268355520955083] [Medline: 33249999]
3. Atallah B, Mallah S, AlMahmeed W. Anticoagulation in COVID-19. Eur Heart J Cardiovasc Pharmacother 2020 Jul 01;6(4):260-261 [FREE Full text] [doi: 10.1093/ehjcvp/pvaa036] [Medline: 32352517]

4. Klok F, Kruip M, van der Meer N, Arbous M, Gommers D, Kant K, et al. Incidence of thrombotic complications in critically ill ICU patients with COVID-19. Thromb Res 2020 Jul;191:145-147 [FREE Full text] [doi: 10.1016/j.thromres.2020.04.013] [Medline: 32291094]
5. Bikdeli B, Madhavan MV, Jimenez D, Chuich T, Dreyfus I, Driggin E, Global COVID-19 Thrombosis Collaborative Group, Endorsed by the ISTH, NATF, ESVM,the IUA, Supported by the ESC Working Group on Pulmonary CirculationRight Ventricular Function. COVID-19 and Thrombotic or Thromboembolic Disease: Implications for Prevention, Antithrombotic Therapy, and Follow-Up: JACC State-of-the-Art Review. J Am Coll Cardiol 2020 Jun 16;75(23):2950-2973 [FREE Full text] [doi: 10.1016/j.jacc.2020.04.031] [Medline: 32311448]
6. Tang N, Bai H, Chen X, Gong J, Li D, Sun Z. Anticoagulant treatment is associated with decreased mortality in severe coronavirus disease 2019 patients with coagulopathy. J Thromb Haemost 2020 May 27;18(5):1094-1099. [doi: 10.1111/jth.14817] [Medline: 32220112]
7. Paranjpe I, Fuster V, Lala A, Russak AJ, Glicksberg BS, Levin MA, et al. Association of Treatment Dose Anticoagulation With In-Hospital Survival Among Hospitalized Patients With COVID-19. J Am Coll Cardiol 2020 Jul 07;76(1):122-124 [FREE Full text] [doi: 10.1016/j.jacc.2020.05.001] [Medline: 32387623]
8. Obi AT, Tignanelli CJ, Jacobs BN, Arya S, Park PK, Wakefield TW, et al. Empirical systemic anticoagulation is associated with decreased venous thromboembolism in critically ill influenza A H1N1 acute respiratory distress syndrome patients. J Vasc Surg Venous Lymphat Disord 2019 May;7(3):317-324. [doi: 10.1016/j.jvsv.2018.08.010] [Medline: 30477976]
9. The ATTACC, ACTIV-4a, and REMAP-CAP Investigators. Therapeutic Anticoagulation with Heparin in Noncritically Ill Patients with Covid-19. N Engl J Med 2021 Aug 26;385(9):790-802. [doi: 10.1056/nejmoa2105911]
10. The REMAP-CAP, ACTIV-4a, and ATTACC Investigators. Therapeutic Anticoagulation with Heparin in Critically Ill Patients with Covid-19. N Engl J Med 2021 Aug 26;385(9):777-789. [doi: 10.1056/nejmoa2103417]
11. Macheel C, Reicks P, Sybrant C, Evans C, Farhat J, West MA, et al. Clinical Decision Support Intervention for Rib Fracture Treatment. J Am Coll Surg 2020 Aug;231(2):249-256.e2. [doi: 10.1016/j.jamcollsurg.2020.04.023] [Medline: 32360959]
12. Nguyen AS, Yang S, Thielen BV, Techar K, Lorenzo RM, Berg C, et al. Clinical Decision Support Intervention and Time to Imaging in Older Patients with Traumatic Brain Injury. J Am Coll Surg 2020 Sep;231(3):361-367.e2. [doi: 10.1016/j.jamcollsurg.2020.05.023] [Medline: 32561447]
13. Teoh D, Vogel RI, Langer A, Bharucha J, Geller MA, Harwood E, et al. Effect of an Electronic Health Record Decision Support Alert to Decrease Excess Cervical Cancer Screening. J Low Genit Tract Dis 2019;23(4):253-258. [doi: 10.1097/lgt.0000000000000484]
14. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis. JMIR Med Inform 2018 Apr 18;6(2):e24 [FREE Full text] [doi: 10.2196/medinform.8912] [Medline: 29669706]
15. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
16. Ingraham N, Tignanelli C. Fact Versus Science Fiction: Fighting Coronavirus Disease 2019 Requires the Wisdom to Know the Difference. Crit Care Explor 2020 Apr;2(4):e0108 [FREE Full text] [doi: 10.1097/CCE.0000000000000108] [Medline: 32426750]
17. University of Minnesota Evidence Based Medicine Homepage for COVID-19. URL: https://covidebm.umn.edu/ [accessed 2020-12-08]
18. Wang J, Hajizadeh N, Moore EE, McIntyre RC, Moore PK, Veress LA, et al. Tissue plasminogen activator (tPA) treatment for COVID-19 associated acute respiratory distress syndrome (ARDS): A case series. J Thromb Haemost 2020 Jul 11;18(7):1752-1755 [FREE Full text] [doi: 10.1111/jth.14828] [Medline: 32267998]
19. Tignanelli CJ, Gipson J, Nguyen A, Martinez R, Yang S, Reicks PL, et al. Implementation of a Prophylactic Anticoagulation Guideline for Patients with Traumatic Brain Injury. Jt Comm J Qual Patient Saf 2020 Apr;46(4):185-191. [doi: 10.1016/j.jcjq.2019.11.007] [Medline: 31899154]
20. Ingraham N, Purcell L, Karam B, Dudley RA, Usher MG, Warlick CA, et al. Racial and Ethnic Disparities in Hospital Admissions from COVID-19: Determining the Impact of Neighborhood Deprivation and Primary Language. J Gen Intern Med 2021 May 18:1-9 [FREE Full text] [doi: 10.1007/s11606-021-06790-w] [Medline: 34003427]
21. Ponti G, Maccaferri M, Ruini C, Tomasi A, Ozben T. Biomarkers associated with COVID-19 disease progression. Crit Rev Clin Lab Sci 2020 Sep 05;57(6):389-399 [FREE Full text] [doi: 10.1080/10408363.2020.1770685] [Medline: 32503382]
22. Lusczek E, Ingraham N, Karam B, Proper J, Siegel L, Helgeson ES, et al. Characterizing COVID-19 clinical phenotypes and associated comorbidities and complication profiles. PLoS One 2021 Mar 31;16(3):e0248956 [FREE Full text] [doi: 10.1371/journal.pone.0248956] [Medline: 33788884]
23. Ingraham NE, Lotfi-Emran S, Thielen BK, Techar K, Morris RS, Holtan SG, et al. Immunomodulation in COVID-19. The Lancet Respiratory Medicine 2020 Jun;8(6):544-546. [doi: 10.1016/s2213-2600(20)30226-5] [Medline: 32380023]
24. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. BMC Med Res Methodol 2017 Dec 06;17(1):162 [FREE Full text] [doi: 10.1186/s12874-017-0442-1] [Medline: 29207961]

25. Robbins A, Beilman GJ, Amdahl B, Welton M, Tignanelli C, Olson AP, et al. Transforming a Long-Term Acute Care Hospital into a COVID-19-Designated Hospital. Surg Infect (Larchmt) 2020 Nov;21(9):729-731. [doi: 10.1089/sur.2020.155] [Medline: 32697625]
26. Friedman CP, Rubin JC, Sullivan KJ. Toward an Information Infrastructure for Global Health Improvement. Yearb Med Inform 2017 Sep 11;26(01):16-23. [doi: 10.15265/iy-2017-004]
27. Mann D, Hess R, McGinn T, Mishuris R, Chokshi S, McCullagh L, et al. Adaptive design of a clinical decision support tool: What the impact on utilization rates means for future CDS research. Digit Health 2019 Feb 06;5:2055207619827716 [FREE Full text] [doi: 10.1177/2055207619827716] [Medline: 30792877]
28. Venkatesh V, Morris MG, Davis GB, Davis FD. User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
29. Patell R, Midha S, Kimani S, Martin R, Neparidze N, Jaglal M, et al. Variability in Institutional Guidance for COVID-19-Associated Coagulopathy in the United States. Thromb Haemost 2020 Dec 22;120(12):1725-1732 [FREE Full text] [doi: 10.1055/s-0040-1715837] [Medline: 32828072]
30. Dicks AB, Weinberg I. Further Evidence Supporting the Use of Prophylactic Anticoagulation in Hospitalized Patients With COVID-19. JAMA Netw Open 2021 Jun 01;4(6):e2112403 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.12403] [Medline: 34115133]
31. Search results. ClinicalTrials.gov. URL: https://clinicaltrials.gov/ct2/results?cond=Covid19&term=thrombosis+&cntry=US&state=&city=&dist= [accessed 2020-12-15]
32. ten Cate H. Surviving Covid-19 with Heparin? N Engl J Med 2021 Aug 26;385(9):845-846. [doi: 10.1056/nejme2111151]
33. INSPIRATION Investigators, Sadeghipour P, Talasaz A, Rashidi F, Sharif-Kashani B, Beigmohammadi MT, et al. Effect of Intermediate-Dose vs Standard-Dose Prophylactic Anticoagulation on Thrombotic Events, Extracorporeal Membrane Oxygenation Treatment, or Mortality Among Patients With COVID-19 Admitted to the Intensive Care Unit: The INSPIRATION Randomized Clinical Trial. JAMA 2021 Apr 27;325(16):1620-1630 [FREE Full text] [doi: 10.1001/jama.2021.4152] [Medline: 33734299]

Abbreviations

CDS: clinical decision support
CPG: clinical practice guideline
EBM: evidence-based medicine
EHR: electronic health record
ICU: intensive care unit
IRR: incident rate ratio
LHS: learning health system
OR: odds ratio
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
VTE: venous thromboembolism

Edited by G Eysenbach; submitted 27.05.21; peer-reviewed by Y Chu; comments to author 17.06.21; revised version received 17.08.21; accepted 17.09.21; published 18.11.21.

Please cite as:
Shah S, Switzer S, Shippee ND, Wogensen P, Kosednar K, Jones E, Pestka DL, Badlani S, Butler M, Wagner B, White K, Rhein J, Benson B, Reding M, Usher M, Melton GB, Tignanelli CJ
Implementation of an Anticoagulation Practice Guideline for COVID-19 via a Clinical Decision Support System in a Large Academic Health System and Its Evaluation: Observational Study
JMIR Med Inform 2021;9(11):e30743
URL: https://medinform.jmir.org/2021/11/e30743
doi:10.2196/30743
PMID:34550900

Review

The Role of Electronic Medical Records in Reducing Unwarranted Clinical Variation in Acute Health Care: Systematic Review

Tobias Hodgson[1*], BSc, MBA, PhD; Andrew Burton-Jones[1*], BCom, MIS, PhD; Raelene Donovan[2*], BSc, MBBS; Clair Sullivan[3*], MBBS, MD

[1]The University of Queensland Business School, The University of Queensland, St Lucia, Australia

[2]Princess Alexandra Hospital, Metro South Health, Woolloongabba, Australia

[3]The University of Queensland Centre for Health Services Research, The University of Queensland, Herston, Australia

[*]all authors contributed equally

Corresponding Author:
Tobias Hodgson, BSc, MBA, PhD
The University of Queensland Business School
The University of Queensland
39 Blair Drive
St Lucia, 4067
Australia
Phone: 61 733468100
Email: t.hodgson@business.uq.edu.au

Abstract

Background: The use of electronic medical records (EMRs)/electronic health records (EHRs) provides potential to reduce unwarranted clinical variation and thereby improve patient health care outcomes. Minimization of unwarranted clinical variation may raise and refine the standard of patient care provided and satisfy the quadruple aim of health care.

Objective: A systematic review of the impact of EMRs and specific subcomponents (PowerPlans/SmartSets) on variation in clinical care processes in hospital settings was undertaken to summarize the existing literature on the effects of EMRs on clinical variation and patient outcomes.

Methods: Articles from January 2000 to November 2020 were identified through a comprehensive search that examined EMRs/EHRs and clinical variation or PowerPlans/SmartSets. Thirty-six articles met the inclusion criteria. Articles were examined for evidence for EMR-induced changes in variation and effects on health care outcomes and mapped to the quadruple aim of health care.

Results: Most of the studies reported positive effects of EMR-related interventions (30/36, 83%). All of the 36 included studies discussed clinical variation, but only half measured it (18/36, 50%). Those studies that measured variation generally examined how changes to variation affected individual patient care (11/36, 31%) or costs (9/36, 25%), while other outcomes (population health and clinician experience) were seldom studied. High-quality study designs were rare.

Conclusions: The literature provides some evidence that EMRs can help reduce unwarranted clinical variation and thereby improve health care outcomes. However, the evidence is surprisingly thin because of insufficient attention to the measurement of clinical variation, and to the chain of evidence from EMRs to variation in clinical practices to health care outcomes.

(JMIR Med Inform 2021;9(11):e30432) doi:10.2196/30432

KEYWORDS

clinical variation; unwarranted clinical variation; electronic health record; EHR; electronic medical record; EMR; PowerPlan; SmartSet; acute care; eHealth; digital health; health care; health care outcomes; outcome; review; standard of care; hospital; research; literature; variation; intervention

Introduction

Variation in Health Care

Any health care service seeks to raise and refine the standard of care it provides to patients and to satisfy the quadruple aim of health care, that is, to improve patient care, population health, cost of care, and clinician experience [1,2]. It is commonly accepted that achieving this aim involves minimizing unwarranted clinical variation, that is, unjustified differences between health care processes or outcomes compared with peers, or with a gold standard [3].

Health care clinical practice variation has been observed, studied, and documented for many decades [4,5]. There are a plethora of potential causes of variation, such as the individuals involved (clinician and patient), their level of agency or motivation, organizational or system factors (eg, capacity) and the nature of the evidence available (clinical and scientific) [6,7]. The method of diffusion of best practice clinical knowledge and clinician adoption of these guidelines and standards has been long been identified as a potential cause of variation [8,9].

Many countries mandate efforts to reduce unwarranted clinical variation in health care provided [10]. While some level of variation is required for innovation and learning, low levels of variation are generally thought to be best [11]. As stated in [12], “The idea is to hold on to variation across patients (to meet the needs of individual patients) and to limit variation across clinicians (which is driven by individual clinician preferences or differences in knowledge and experience)”.

Variation is unwarranted if it is not justified by clinical imperatives, patient needs or preferences, or innovation. In its most basic form, clinical variation that leads to positive outcomes may be warranted, whereas variation that leads to negative outcomes is deemed unwarranted. Many health care services have looked to electronic medical record (EMR) systems to reduce unwarranted variation and thereby improve outcomes [10].

EMRs

EMR use has become virtually ubiquitous in health services in developed countries [13]. EMRs can offer many benefits, including improvements in billing and cost management, reporting and analytics, real-time access to data by clinicians, information sharing, treatment management, patient safety, and clinical decision making [14-21].

EMRs provide the means to both monitor and address clinical variation through the provision of best practice guidelines and clinical decision support (CDS) to improve care and reduce waste [22]. At the same time, EMRs can also create variation by offering users multiple ways to perform a task. Work-as-done by clinicians also often varies from the work-as-imagined expectation of EMR designers [23]; as a result, it is an empirical question as to whether EMRs actually reduce unwarranted clinical variation.

Theoretical Framework

Studying how EMRs may affect unwarranted clinical variation requires understanding 3 elements: why clinical variation occurs, why and how EMRs may reduce clinical variation, and how measuring and altering variation are operationalized in practice (Textbox 1).

Textbox 1. Clinical variation factors. CDS: clinical decision support, EMR: electronic medical record.

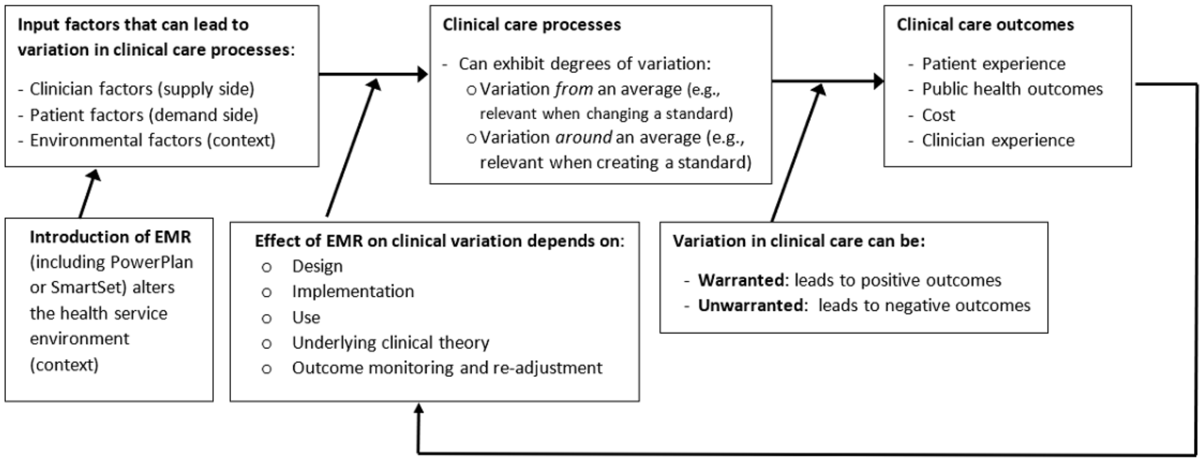
- Clinical variation can occur due to supply-side, demand-side, or contextual factors [24]:
  - Clinician factors (supply side): expertise, training and experience, preference, practice style;
  - Consumer factors (demand side): case complexity, consumer preference, social determinants of health; and
  - Environmental factors (context): local guidelines, available resources, hospital case mix.
- EMRs may reduce clinical variation through their ability to control process delivery and outcomes. It is common for health services to tackle clinical variation through EMR-related process control efforts (eg, clinical guidelines and pathways), and process design and development efforts [25,26]. EMRs hold promise for reducing unwarranted clinical variation because they can help tackle each of the 3 aforementioned factors:
  - Clinician factors: EMRs can constrain clinicians to perform similarly via restrictions to particular behaviors or range of behaviors.
  - Consumer factors: EMRs can inform and guide patients in a consistent manner via patient portals, and they can help standardize the reporting of patient outcomes.
  - Environmental factors: EMRs can provide standardized decision support and data that health services can use to monitor and improve operations and achieve greater consistency.
- Understanding precisely how an EMR can reduce unwarranted variation requires opening the EMR “black box” and assessing its components. One set of EMR components designed to help reduce unwarranted clinical variation is CDS. There are numerous CDS tools and features in the marketplace, with EMR vendors naming and implementing components in proprietary ways. This review focuses on 2 CDS components from the 2 most prevalent EMR vendors globally (>50% of acute care market), with products that have similar aims to help reduce unwarranted clinical variation: PowerPlan (Cerner Corporation) and SmartSet (Epic Systems Corporation) [27-29]:
  - PowerPlan: “A Power Plan is a group of orders under a single title designed to support a procedure or a process.” [30]
  - SmartSet: “A documentation template. A group of orders and other elements, such as notes, chief complaints, SmartGroup Panels, and levels of service, that are commonly used together to document a specific type of visit.” [31].

EMRs can implement tools to guide and constrain practice; however, clinicians do not always use these interventions as intended. For example, they may focus on using a PowerPlan to make ordering easier rather than using it to reduce variation. For this reason, it is important to empirically test whether in practice they reduce clinical variation as intended.

To understand how EMR interventions might or might not reduce unwarranted clinical variation as intended, a variation in clinical care framework was devised (Figure 1). The framework highlights the expected factors that must be accounted for if EMRs are to reduce unwarranted clinical variation. That is, the expectation is that EMRs—through their components—should help reduce unwarranted clinical variation if the following factors are considered:

- *Design:* if the EMR and its PowerPlan or SmartSet components are configured to reduce unwarranted variation.
- *Implementation:* if the goal of reducing unwarranted variation is kept in focus during implementation.
- *Use:* if clinicians use the EMR as intended.
- *Clinical theory:* if the clinical logic or theory underlying the design of the intervention and the clinical practice is mature (rather than lacking evidence and having ambiguity, allowing variation among clinicians).
- *Monitoring and intervention:* if the health service monitors outcomes that flow from changes in clinical variation and iteratively improve the design and use of the EMR based on this learning feedback loop.

Figure 1. Variation in clinical care - theoretical framework. EMR: electronic medical record.


Input factors that can lead to variation in clinical care processes:
- Clinician factors (supply side)
- Patient factors (demand side)
- Environmental factors (context)
Clinical care processes
- Can exhibit degrees of variation:
o Variation from an average (e.g., relevant when changing a standard)
o Variation around an average (e.g., relevant when creating a standard)
Clinical care outcomes
- Patient experience
- Public health outcomes
- Cost
- Clinician experience
Introduction of EMR (including PowerPlan or SmartSet) alters the health service environment (context)
Effect of EMR on clinical variation depends on:
o Design
o Implementation
o Use
o Underlying clinical theory
o Outcome monitoring and re-adjustment
Variation in clinical care can be:
- Warranted: leads to positive outcomes
- Unwarranted: leads to negative outcomes

If the use of an EMR can lead to changes in unwarranted clinical variation, how can this variation be measured? The framework (Figure 1) suggests that there are 2 archetypal changes in variation (Textbox 2), conveying different meanings of "variance."

Combinations of these 2 archetypes may also occur. For instance, a health service may implement an EMR to both change a standard and encourage clinicians to achieve greater consistency around that standard.

Textbox 2. Changes in variation.

- *Variation from a level or a standard:* For example, assume that a health service has a guideline for a clinical practice. If clinicians follow the guideline, with appropriate variation in adherence and excellent outcomes, this will be reflected in an average level on that practice with variation around the average. If the health service shifts the guideline, unwarranted variation can be viewed as the degree to which the distribution of behavior fails to shift to the new standard and improve outcomes. Statistically, this can be tested by comparing the average practices (accounting for the variation around each average) before and after the intervention, (eg, via a *t* test).
- *Variation around a level or a standard:* For example, assume that a health service has no guideline for a practice, and clinicians just follow their own practices. Assume also that the average behavior is close to the desired level, but the variation around this average is concerning. If the health service then implements a guideline to reduce this variation, unwarranted clinical variation and monitoring of outcomes can be operationalized as the degree to which the level of variance in practices fails to be reduced. Statistically, this can be tested by a change in the level of variance (eg, range or SD).

The implication of these different meanings of clinical variation is that researchers need to be precise as to which type of variation and associated outcomes they are studying and how. In short, studying changes in variation requires careful attention to measurement.

Finally, variation is only unwarranted if it impairs outcomes, such as any of the quadruple aims of health care (Figure 1). That is, variance itself is not the outcome, nor it is necessarily negative. Rather, the aim is to learn how to design, implement, use, and monitor the EMR and find the "right" level of variation to achieve the best outcomes.

Objective of This Review

We aim to summarize the existing literature on the effects of EMRs on variation in clinical care processes and patient outcomes as mapped to the quadruple aim of health care. To account for the specifics of EMR systems, and for the specific ways that variation can occur, searches were conducted not only for the effects of EMRs in general, but also for the components of EMRs (PowerPlan and SmartSet). Studies were coded for changes in clinical variation and for how changes in variance affected both process and patient outcomes.

Because of differences in tools and methods used to achieve clinical standardization between the primary and acute care settings (eg, case complexity, technology utilized), this study focuses purely on the acute sector and hospital-based EMRs.

Methods

Eligibility Criteria

A Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)–compliant systematic review of studies examining clinical variation and EMRs was undertaken [32].

Inclusion Criteria

To be included in the review, studies were required to meet the following criteria:

- The article is published in English.
- The topic of the study is relevant to clinical variation and EMRs.
- Articles published after 2000, due to the prominence of EMR/electronic health record (EHR) articles published since then (Multimedia Appendix 1).
- Participants to include clinicians performing medical care duties or patients receiving medical treatment.
- Measured outcomes were reported, whether immediate (eg, test results) or longer term (eg, length of stay, economic), and whether measured objectively or by self-report.
- Peer-reviewed studies only.
- Empirical studies are either qualitative or quantitative (or mixed): quantitative studies may have included experimental and observational study designs such as randomized controlled trials, quasi-experimental studies, before-and-after studies, case–control studies, cohort studies, and cross-sectional studies.
- Quality improvement initiatives.
- Articles focused on acute care settings (including ambulatory specialist care).

Exclusion Criteria

Studies were excluded from the review if they met any of the following criteria:

- Abstracts in which full study data were unavailable.
- Nonempirical studies.
- Outcome measures of expected variation (not actual).
- Articles with a care focus of primary care.

Information Sources

Searches were made on ACM Digital Library, CINAHL, EMBASE/MEDLINE, Google Scholar, IEEE Xplore, PubMed, Scopus, and Web of Science for articles from the year 2000 to November 2020.

The search query used was: “EHR” OR “EMR” AND “practice variation” OR “clinical variation” OR “unwanted variation” OR “unwarranted variation” OR “reduction in waste” OR “PowerPlan” OR “SmartSet”.

As noted earlier, understanding how EMRs have their effects requires opening the “black box” of the EMR to study its components, in this case PowerPlan and SmartSet. However, a given study may use these proprietary terms or instead use more generic terms. Including these specific vendor EMR components in the search string with an OR term increased the extent to which articles that examined clinical variation, even if an article did not specifically use those words (ie, to increase the level of recall), would be found.

Additional applicable search terms were assessed but excluded as they added no additional search results (eg, medical-order entry systems). The term “order sets” was excluded from the search as they are not necessarily electronic (often paper based) and many studies focus on discrete point-in-time events (eg, prescribing anithrombotics) rather the patient’s entire care process (as implemented in PowerPlan and SmartSet for specific conditions). As noted earlier, the focus of this study was on clinical care processes and outcomes.

The ultimate searches were undertaken in February 2021. Both backward and forward citation searching were undertaken for all included articles with a quality score over 50% (35/36 studies; Multimedia Appendix 2) [33-67]. Forward searches were undertaken with the assistance of Anne O’Tate, PubMed, Google Scholar, and Scopus [68].

Once duplicates were removed, these searches resulted in 4622 potential articles. Titles and abstracts were then identified and screened, with 3935 initial further exclusions, with 40 cases having only partial text available or requiring further information to make an assessment, resulting in 687 full texts that were retrieved and evaluated.

Interrater agreement during the screening phase was assessed based on 30 randomly sampled papers screened by 2 reviewers (TH [first author] and TL [research assistant]). The observed agreement was 90% (27/30), with an acceptable κ (Cohen κ) of 0.67 [69]. Given the reliable coding, the remainder of the screening phase was undertaken by TH.

Each of the included articles was assessed independently by 2 reviewers (TH and TL) against the inclusion criteria. After assessment, 36 studies remained [33-67,70] (Figure 2) (Table 1). In instances of doubtful eligibility, a consensus assignment was made after deliberation (5 articles were excluded). The 2 reviewers also measured the disposition of these studies as positive, mixed, or negative based on how the authors of the study discussed the outcomes.

Figure 2. Systematic review flow diagram (after Preferred Reporting Items for Systematic Reviews and Meta-Analyses [PRISMA] [32]).

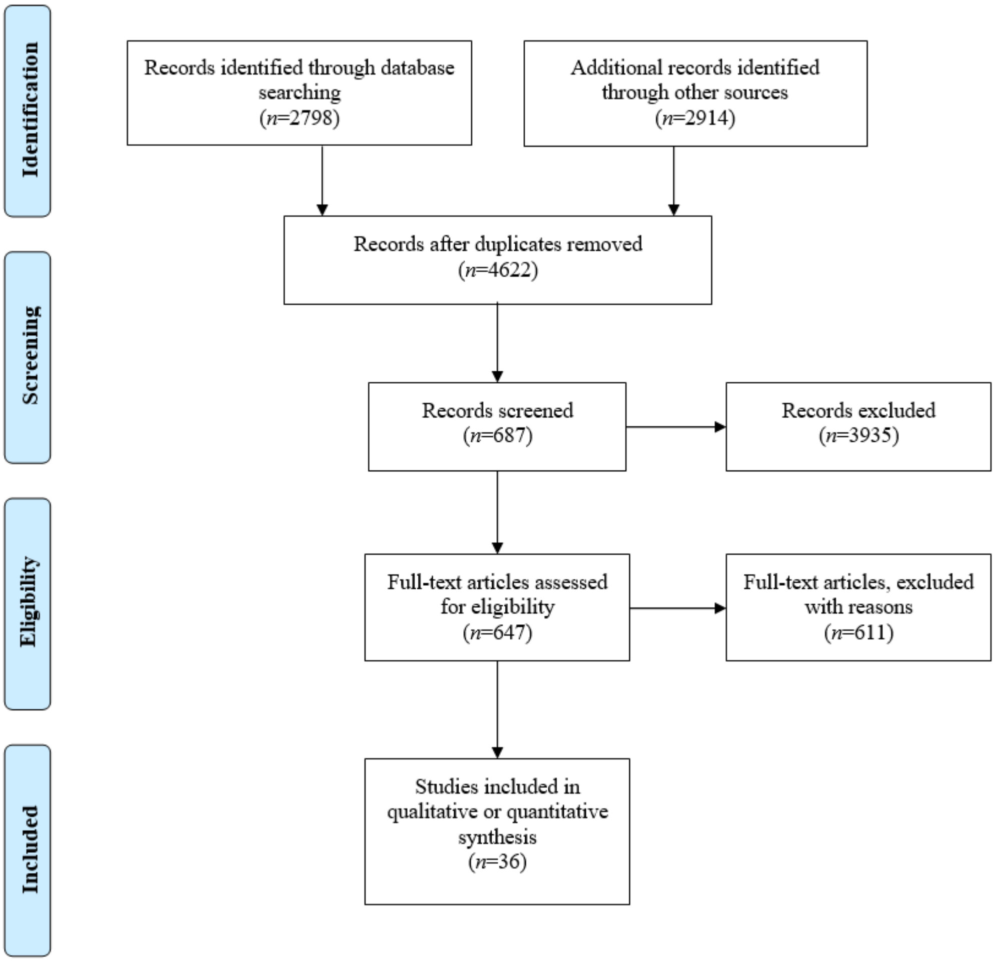

Identification
Screening
Eligibility
Included
Records identified through database searching (n=2798)
Additional records identified through other sources (n=2914)
Records after duplicates removed (n=4622)
Records screened (n=687)
Records excluded (n=3935)
Full-text articles assessed for eligibility (n=647)
Full-text articles, excluded with reasons (n=611)
Studies included in qualitative or quantitative synthesis (n=36)

Table 1. Summary of included studies (N=36).

Study author	EMR[a]/EHR[b] vendor	Study disposition	Quality assessment (QATS-DD[c]), %	Variation in clinical care processes	Variance type[d]	Patient care[e]	Population health[e]	Costs/ efficiency[e]	Clinician experience[e]
Adelson et al [33]	EPIC (SmartSet)	Positive	54.17	Orders/prescription	2	Clinical events		Length of events	Quality (clinician)
Akenroye et al [34]	Vendor not stated	Positive	64.58	Orders/prescription	2			Costs	Quality (clinician)
Amland et al [35]	Cerner (PowerPlan)	Positive	78.57	Patient assessment	2	Clinical events			
Asan et al [36]	EPIC (SmartSet)	Negative	66.67	Care provision	5				Clinical burden
Attaar et al [37]	Other: Allscripts Sunrise Clinical Manager	Positive	66.67	Orders/prescription	4	Quality (patient)		Length of stay	
Ballesca et al [38]	EPIC (SmartSet)	Positive	66.67	Orders/prescription	2	Clinical events Test measures		Length of stay	
Borok et al [39]	EPIC (SmartSet)	Positive	61.90	Orders/prescription Referrals	4				Clinical burden
Bradywood et al [40]	Vendor not stated	Positive	80.95	Clinical care pathway	4	Quality (Patient) Clinical events		Length of stay Length of events	
Chisolm et al [41]	Vendor not stated	Positive	77.08	Orders/prescription	5			Costs Length of stay	Quality (clinician)
Dort et al [42]	Vendor not stated	Positive	73.81	Clinical care pathway	5	Clinical events		Length of stay Length of events	
Ebinger et al [43]	Vendor not stated	Positive	66.67	Care provision	4	Clinical events		Costs Length of stay	
Geltman et al [44]	Vendor not stated	Mixed	71.43	Patient assessment	2	Test measures			
Goga et al [45]	Vendor not stated	Positive	54.76	Orders/prescription	4				
Gulati et al [46]	Cerner (PowerPlan)	Positive	76.19	Orders/prescription	2	Clinical events		Length of stay Length of events	
Hendrickson et al [47]	Vendor not stated	Positive	78.57	Orders/prescription	2	Clinical events		Number of tests	
Hooper et al [48]	Vendor not stated	Positive	66.67	Patient assessment	2	Test measures			
Horton et al [49]	EPIC (SmartSet)	Positive	59.52	Orders/prescription	2	Quality (patient) Clinical events		Test measures	
Jacobs et al [50]	Other: ICIS, a web-based EHR	Positive	71.43	Ordering	2				

Study author	EMR[a]/EHR[b] vendor	Study disposition	Quality assessment (QATSDD[c]), %	Variation in clinical care processes	Variance type[d]	Patient care[e]	Population health[e]	Costs/efficiency[e]	Clinician experience[e]
Karajgikar et al [51]	Cerner (PowerPlan)	Positive	54.76	Orders/prescription	5	Clinical events		Length of events Length of stay	
Kicker et al [67]	Vendor not stated	Positive	57.14	Ordering Preparation	2			Costs Length of events	
Lewin et al [52]	Vendor not stated	Positive	59.52	Orders/prescription Use of intervention	4			Costs Length of stay	
Lindberg et al [53]	EPIC (SmartSet)	Positive	76.19	Patient assessment	2	Test levels			
Lindberg et al [54]	EPIC (SmartSet)	Positive	73.81	Patient assessment	5	Test levels			
Morrisette et al [55]	Cerner (PowerPlan)	Mixed	69.05	Ordering	4			Costs Length of events	
Prevedello et al [56]	Other: Percipio; Medicalis Corp	Mixed	73.81	Patient assessment	2			Test measures	
Reynolds et al [57]	EPIC (SmartSet)	Negative	61.90	Orders/prescription	4				
Rooholamini et al [58]	Cerner (PowerPlan)	Positive	59.52	Orders/prescription Patient assessment	2	Clinical events		Costs Length of events	
Rosovsky et al [70]	EPIC (SmartSet)	Positive	45.24	Ordering	4				
Sim et al [59]	Other: AllScripts	Positive	69.05	Ordering	2				
Sonstein et al [60]	EPIC (SmartSet)	Positive	69.05	Ordering	4	Clinical events		Length of stay	
Soo et al [61]	Cerner (PowerPlan)	Negative	68.75	Ordering	4			Length of events	Clinical burden
Studer et al [65]	Vendor not stated	Positive	61.90	Orders/prescription	2	Clinical events			
Teich et al [66]	Vendor not stated	Positive	42.86	Ordering	2				
Terasaki et al [62]	EPIC (SmartSet)	Positive	64.29	Patient assessment	2				
Wang et al [63]	EPIC (SmartSet)	Positive	52.38	Orders/prescription	4	Quality (patient)		Volume of drugs	
Webber et al [64]	Cerner (PowerPlan)	Positive	57.14	Ordering	4			Costs	

[a] EMR: electronic medical record.

[b] EHR: electronic health record.

[c] QATSDD: Quality Assessment Tool for Studies with Diverse Designs.

[d] 1=Mean constant; variance change, 2=mean change; variance change, 3=mean change; variance constant, 4=mean change; variance unknown, 5=mean unknown; variance unknown (or N/A, assumed only).

[e] Where the outcomes were not observed within the study table cells remain empty.

Study data including the intervention, population, study design, and effects were extracted by both reviewers using a standardized template within Covidence systematic review software (Multimedia Appendix 3) [71]. Data quality was assessed via a bespoke Covidence template employing the Quality Assessment Tool for Studies with Diverse Designs (QATSDD), a 16-item mixed methods quality assessment tool (Multimedia Appendix 2) [72].

Risk of Bias

The studies were examined to determine the risk of drawing biased inferences [73]. Five risks were identified (Textbox 3).

Textbox 3. Risk of bias. EMR: electronic medical record.

1. Publication bias: most papers (30/36, 83%) reported positive results [33-35,37-43,45-54,58-60,62-66,70], with a minority reporting mixed [44,55,56] or negative results [36,57,61] (3/36, 8% for both). The completeness of results including nonsignificant effects was not always assured.
2. Selection: participation in the trials varied from compulsory to voluntary. Where the study was voluntary, it was more likely that those with interest in, and with a positive opinion toward, EMRs participated [36,41,57,58,67].
3. Randomization of intervention: this only occurred in 1 study which randomized the use of the SmartSet intervention using block randomization, stratified by provider subspecialty [57].
4. Performance: the studies were all composed of unblinded trials, and in many cases the participants of the study knew if they were utilizing the intervention or not.
5. Time lag bias: some papers were reporting on data collected much earlier than publication date (eg, Teich et al [66] was based on 1993 data) [41,66].

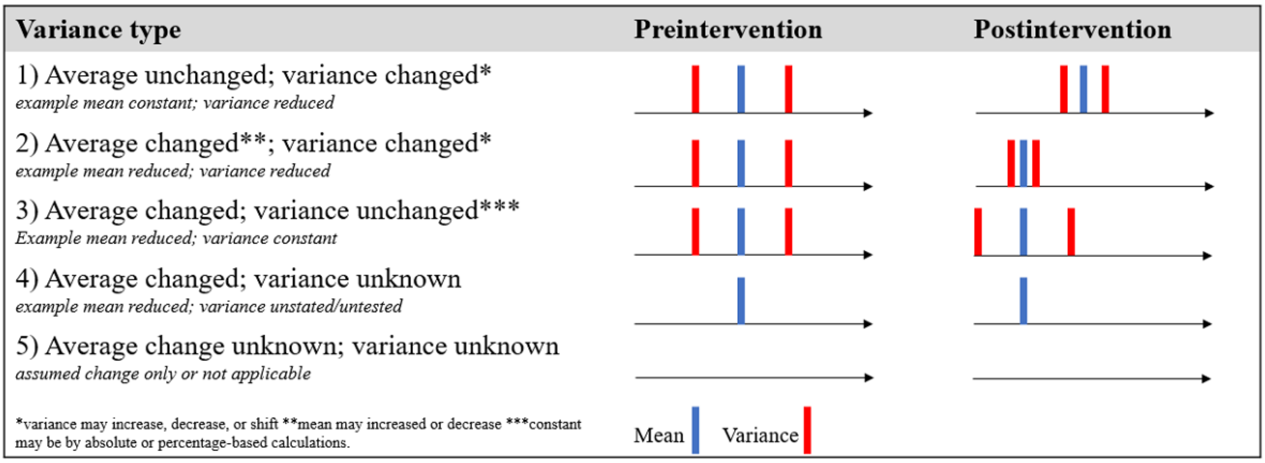
Recruitment

The recruitment of participants for clinicians utilizing the interventions was voluntary in all but 2 studies, and existing clinic/hospital EMR data were utilized for patient data [33,55].

Coding

Following the earlier description of how variation in clinical practices can be observed, studies were coded for 5 types of variation, each reflecting different patterns in the change of a distribution (Figure 3 and Multimedia Appendix 3). Types 1 and 3 refer to the 2 archetypes noted earlier (“variance from” and “variance around”), whereas Type 2 reflects their combination. Type 4 reflects the possibility that a study refers to changes in average behavior without reporting changes in variance. Type 5 is where change is assumed but not measured.

Figure 3. How changes in variance can be operationalized in clinical practice.


Variance type
Preintervention
Postintervention
1) Average unchanged; variance changed*
example mean constant; variance reduced
2) Average changed**; variance changed*
example mean reduced; variance reduced
3) Average changed; variance unchanged***
Example mean reduced; variance constant
4) Average changed; variance unknown
example mean reduced; variance unstated/untested
5) Average change unknown; variance unknown
assumed change only or not applicable
*variance may increase, decrease, or shift **mean may increased or decrease ***constant may be by absolute or percentage-based calculations.
Mean
Variance

As this study’s aim is to learn the effects of the EMR on changes in clinical variation, the focus is on variance types 1, 2, and 3, which reflect different ways in which clinical variation can be expressed. By contrast, Types 4 and 5 do not provide clear measures of variation.

To code the study’s disposition, 2 reviewers (TH and TL) coded the overall disposition of a study as either positive, mixed, or negative, based on the following criteria:

- Positive: a majority of studies stated expected outcomes were met.
- Mixed: some elements of expected outcomes were met, some not (with an approximate 50/50 split).
- Negative: intervention not used, majority of expected outcomes not met, or reverse outcomes seen.

Disposition reflects the authors’ overall conclusions in that study in favor of or against the EMR or the intervention. It is not a measure of whether a study measured clinical variation or outcomes. Interrater agreement on study disposition was calculated using Cohen κ, and showed high levels of agreement (33/36, 92%, κ=0.71) [69].

Clinical outcomes were coded according to the quadruple aim of health care: quality of patient care, population health, cost/efficiency, and clinician experience [1,2].

Results

Almost all the studies were based on the implementation of an intervention (new or refined) into a clinical setting (35/36, 97%) with 1 qualitative analysis of EMRs by clinicians [36]. Most studies were quality or process improvement based (28/36, 78%) [33-35,37,39-45,47-49,51,52,54,55,58,59,61-67] or best practice/evidence-based intervention related (27/36, 75% for both) [33-35,37,38,40-42,45,47,48,50,52-60,62-66,70]. Over half of the studies examined EMR elements such as order sets (23/36, 64%) [33,34,36-38,40-42,46,47,49-51,52,54,55,58,60,61,64-66,70] and care pathways/treatment plans (22/36, 61%) [33-36,39-43,46-48,50,52,54,58,60,62,63,65,66,70]. Many papers addressed the minimization or elimination of a particular drug prescription/use (17/36, 47%) [39,40,45,46,49,51-54,57,58,60,63,65-67,70].

Of the papers where the specific EMR used by the health facility was identified (24/36, 67%), half were Epic (12/24, 50%) [33,36,38,39,49,53,54,57,59,60,62,63], some Cerner (7/24, 29%) [35,46,51,55,58,61,64], and few with other vendors (5/24, 21%) [37,50,52,56,59].

Regarding overall disposition, most studies reported positive results (30/36, 83%) [33-35,37-43,45-54,58-60,62-67,70], while a minority reported mixed [44,55,56] or negative results [36,57,61] (3/36, 8% each). That is, the authors concluded in most studies that the EMR was used successfully as part of an initiative to address clinical variation.

However, most studies did not measure or report variation. Of the 5 codes for coding variance (Figure 3), no studies reported Type 1 or 3, half reported Type 2 (18/36, 50%) [33-35,38,44,46-50,53,56,58,59,62,65-67], some reported Type 4 (13/36, 36%) [37,39,40,43,45,52,55,57,60,61,63,64,70], and a few reported Type 5 (5/36, 14%) [36,41,42,51,54]. The studies that reported results for variation coded as Types 2 and 4 generally examined how an intervention led to changes in the average of a clinical behavior. Such studies reflected Type 2 variation if they explicitly referred to measures of variance in addition to average practices or if the distribution of the variable examined was such that a change in the average clearly implied a change in variance (the dependency between the average and variance of a distribution is dependent on the type of distribution).

For example, if clinician behaviors were coded in a study as adhering or not adhering to a guideline, the rate of adherence would reflect a binomial distribution and so an increase in adherence (eg, from 60% to 80%), implying both an increase in the average behavior and a reduction in variance. Where this connection between a change in a behavior and the change in variance was not explicitly reported or could not be inferred clearly from the distribution, this reflected a Type 4 change. That is, the 13 studies coded as Type 4 found that the EMR affected clinical practices but not necessarily clinical variation.

Regarding the quadruple aims of health care outcomes, over half of the studies addressed individual care outcomes (19/36, 53%) [33,35,37,38,40,42-44,46-49,51,53,54,58,60,63,65], many examined efficiency (21/36, 58%) [33,34,37,38, 40-43,46,47,49,51,52,55,56,58,60,61,63,64,67], a handful examined clinician experience (6/36, 17%) [33,34,36,39,41,61], and none examined population health outcomes (Table 1). Some studies examined just 1 quadruple aim of health care outcome (13/36, 36%) [35,36,39,44,48,52-56,64,65,67], most studies examined 2 outcomes (15/36, 42%) [34,37,38,40-43, 46,47,49,51,58,60,61,63], 1 study examined 3 outcomes [33], and none of the studies examined all 4 outcomes associated with the widely accepted quadruple aims of health care.

Of the studies that measured changes in variation (18/36, 50%), many (11/18, 61%) [33,35,38,44,46-49,53,58,65] examined follow-on changes in clinical-care outcomes, half assessed cost outcomes (9/18, 50%) [33,34,38,46,47,49,56,58,67], few examined clinician experience outcomes (2/18, 11%) [33,34], and none addressed public health outcomes. In other words, even though studies generally reported positive findings (in terms of overall study disposition), this positive conclusion was based on a partial (rather than comprehensive) assessment of outcomes.

There was heterogeneity in study data quality, with QATSDD scores ranging from a low of 43% through to a high of 81% and a mean of 65% across all included studies (Multimedia Appendix 2).

Discussion

Principal Findings

This review finds some evidence to justify that EMRs can help reduce unwanted clinical variation and thereby improve health care outcomes. The evidence, however, is not strong. This reflects that (1) study quality was not high, (2) not many studies examined the effect, and (3) clinical variation and outcomes were not examined consistently (different outcome measures across studies) or comprehensively (rarely studying more than 1 outcome).

Surprisingly, while all the studies retrieved by our search discussed clinical variation, few studies measured it, and even fewer tied these changes in clinical variation to a broad set of health care outcome measures.

The theoretical framework proposed earlier can be used to understand the results of the review and identify directions for research. Specifically, 5 factors can enhance the EMR's effects on unwarranted clinical variation and follow-on health care outcomes: design, implementation, use, clinical theory, and outcome monitoring and re-adjustment (Figure 1). These factors were examined only sporadically across studies with an average of 3 addressed per paper, and only 4 of the 36 retrieved studies examined all 5 factors [34,41,43,49].

Design

Intervention design was discussed in most studies (27/36, 75%) [33,34,36,37,40-50,53-55,58,59,61-65,67,70] but not in depth. While not a core focus of the studies, design-related issues that may affect clinical variation were identified, such as in the insights that “design characteristics that are intended to make documentation more efficient can have unintended consequences” and that “some of the suboptimal design characteristics of the EHR may be exacerbated by user-related practices.” [36].

Implementation

Almost all the studies (35/36, 97%) [33-35,37-67,70] examined the implementation of a new or refined intervention into a clinical setting, but specific implementation details were found in fewer studies (23/36, 64%) [33-35,37,40-46, 48-50,52,55,56,59-62,64,70]. The introduction of EMRs and their components are in large part a change management process, with both situational and psychological aspects to consider [74]. The successful implementation of change requires the participation, commitment, and support of key organizational stakeholders throughout the life span of the process to provide the highest chance of success [75,76].

Use

One way to improve outcomes is to educate and train users to employ the EMR more effectively. The role of education and training was addressed in the majority of studies (25/36, 69%) [34,37,39-41,43,44,46-49,52-58,60-64,66,70] and frequently mentioned as critical for the intervention’s success or failure. Education/training was also identified as requiring primary focus in those studies deemed as having a negative or mixed disposition [36,44,55-57,59]. A multifaceted approach with local super-user support, high-quality training materials, and education and feedback sessions is likely to help. For instance, a 2018 study by Robinson [77] of Kaiser Permanente saw a significant increase in the use of many order sets after the implementation of a 3-day intensive EMR education intervention specifically tailored for the physicians with interactive teaching methods.

Underlying Clinical Theory

The interventions in the retrieved studies were all developed on underlying clinical theory that explicitly or implicitly directs clinical practice via pathway, program, or guidelines. These varied from locally developed standards established from journal articles and consensus guidelines, or more commonly the implementation of established national or peak body guidelines. Given that clinical care should be tailored to the needs of patients in the local setting, how best to identify and customize the appropriate underlying theory for a guideline and how stringently to implement it in the EMR are open questions that require further research.

Outcome Monitoring and Re-adjustment

Only 10 studies addressed monitoring of clinical outcomes and re-adjustment of the interventions. Even when addressed, they were typically confined to the implementation phase, rather than long-term and ongoing monitoring and revisions. Using EMRs to implement feedback loops and quality management life cycles can help health care organizations improve safety and quality and become learning organizations [78]. Intermountain Healthcare has shown how this can be achieved via repeated cycles of *create*, *distribute*, *use*, *monitor*, and *feedback* [79,80].

Limitations

Despite steps taken to perform high-quality searching, sample bias may still exist. Because this is an understudied topic, the required search terms and meta-tags on the topic are not yet mature and validated. As a result, different search terms could potentially have retrieved additional relevant publications. Gray literature (such as internal health service reports) may also exist on the topics that were not retrieved. The time span of included studies was broad, covering over 20 years, but a longer time span may have identified additional papers.

Differences in the design and scope of the retrieved papers prevented direct comparisons among studies and meta-analytic tests. Judgment also needed to be exercised when coding articles. While interrater reliability tests suggested that the coding was reliable, some subjectivity inevitably remained. Finally, the context faced by a health service (eg, its resources and patient mix) influences how an EMR can help. Given the small number of studies in this area and their heterogeneity, it was not possible to pinpoint the most salient elements of context.

Expanding the study to include nonacute health care settings, articles in languages other than English, and specifying additional EMR vendors may provide valuable insight into additional means and methods available to address EMR-based clinical variation beyond those identified within this review.

Comparison With Prior Work

Existing studies and reviews on comparable topics were examined and while there is much existing work addressing the effects of EMRs on health care quality and outcomes, and measuring various criteria (efficiency, guideline adherence, errors, clinical outcomes), none adequately or directly address these aspects through the lens of clinical variation and outcomes [81,82]. No previous studies related variation and clinical outcomes back to the quadruple aims of health care. The ability to map variation in EMR-related clinical care processes and outcomes to all 4 of the quadruple aims (patient experience, public health, cost, and clinician experience) sets this review apart from any prior work in the field (Figure 1).

Conclusions

EMRs and their components such as PowerPlans/SmartSets are not a panacea, but rather tools to assist health care provision. It is widely thought that evidence-based clinical guidelines play an essential role in promoting quality of care and minimizing unwanted variation [83]. Ideally, EMRs should be able to improve both the average clinical practices and reduce unwarranted variation. However, the effects of unwarranted variation on clinical outcomes are unclear and understudied.

This review finds some evidence to suggest that unwarranted variation can be reduced, but the evidence is not strong. Many studies focused on technical outcomes (eg, adoption, reduction in variation), rather than on the clinical health care outcomes themselves. More research is needed to learn how EMRs can be implemented and used to reduce unwarranted variation; however, it is important to remember that reduction in clinical variation itself is not the desired outcome. Rather, improved health care outcomes are the ultimate goal.

It is critical that these health care outcomes are clearly defined and monitored, in concert with the ongoing reduction in variation driven by EMRs as a mechanism, to create a continuous learning health care system with appropriate governance to keep iteratively improving health care outcomes over time.

Recommendations

Additional empirical research on EMRs and how their elements such as PowerPlans/SmartSets affect clinical variation and patient outcomes is needed. More attention needs to be given on how to: (1) measure clinical variation and unwarranted variation; (2) improve the effects of an EMR on reducing unwarranted clinical variation; (3) measure multiple elements of the quadruple aim of health care in a single study; and (4) articulate and test the chain of evidence from the EMR to changes in clinical variation to outcomes.

Acknowledgments

This research was funded via an Australian Research Council Linkage Projects grant (LP170101154). Tri Lam (TL) was the second reviewer who assisted in the phase of article screening and data extraction.

Authors' Contributions

AB-J and CS conceived the study and its design. TH conducted the research, the primary analysis, and the initial drafting of the paper. AB-J, CS, and RD contributed to the analysis and drafting of the paper and all authors approved the final manuscript. TH is the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Electronic medical record/electronic health record prevalence in published studies figure.
[PNG File , 52 KB - medinform_v9i11e30432_app1.png]

Multimedia Appendix 2
Quality assessment summary table using the Quality Assessment Tool for Studies with Diverse Designs (QATSDD).
[DOCX File , 14 KB - medinform_v9i11e30432_app2.docx]

Multimedia Appendix 3
Data extraction coding table.
[DOCX File , 9 KB - medinform_v9i11e30432_app3.docx]

References

1. Sikka R, Morath JM, Leape L. The Quadruple Aim: care, health, cost and meaning in work. BMJ Qual Saf 2015 Oct;24(10):608-610. [doi: 10.1136/bmjqs-2015-004160] [Medline: 26038586]
2. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff (Millwood) 2008;27(3):759-769. [doi: 10.1377/hlthaff.27.3.759] [Medline: 18474969]
3. Australian Commission on Safety and Quality in Health Care. Healthcare Variation. 2019. URL: https://www.safetyandquality.gov.au/our-work/healthcare-variation [accessed 2021-09-29]
4. Glover JA. The incidence of tonsillectomy in school children: (section of epidemiology and state medicine). Proc R Soc Med 1938 Aug;31(10):1219-1236 [FREE Full text] [Medline: 19991659]
5. Wennberg J, Gittelsohn. Small area variations in health care delivery. Science 1973 Dec 14;182(4117):1102-1108. [doi: 10.1126/science.182.4117.1102] [Medline: 4750608]
6. Kennedy PJ, Leathley CM, Hughes CF. Clinical practice variation. Med J Aust 2010 Oct 18;193(S8):S97-S99. [doi: 10.5694/j.1326-5377.2010.tb04021.x] [Medline: 20955142]
7. Sutherland K, Levesque J. Unwarranted clinical variation in health care: Definitions and proposal of an analytic framework. J Eval Clin Pract 2020 Jun;26(3):687-696 [FREE Full text] [doi: 10.1111/jep.13181] [Medline: 31136047]
8. Phelps CE. Diffusion of information in medical care. J Econ Perspect 1992;6(3):23-42. [doi: 10.1257/jep.6.3.23] [Medline: 10128077]
9. Timmermans S. From autonomy to accountability: the role of clinical practice guidelines in professional power. Perspect Biol Med 2005;48(4):490-501. [doi: 10.1353/pbm.2005.0096] [Medline: 16227662]

10. Australian Commission on Safety and Quality in Health Care. National Safety and Quality Health Service Standards: User Guide for the Review of Clinical Variation in Health Care. 2020. URL: https://tinyurl.com/72yj64ke [accessed 2021-09-29]
11. Harrison R, Manias E, Mears S, Heslop D, Hinchcliff R, Hay L. Addressing unwarranted clinical variation: A rapid review of current evidence. J Eval Clin Pract 2019 Feb;25(1):53-65. [doi: 10.1111/jep.12930] [Medline: 29766616]
12. Imison C, Castle-Clarke S, Watson R, Edwards N. Delivering the Benefits of Digital Health Care. London, UK: Nuffield Trust; 2016.
13. WHO Global Observatory for eHealth. Global Diffusion of EHealth: Making Universal Health Coverage Achievable: Report of the Third Global Survey on EHealth. Geneva, Switzerland: World Health Organization; 2016.
14. Hoover R. Benefits of using an electronic health record. Nursing2018 2016;46(7):21-22. [doi: 10.1097/01.Nurse.0000484036.85939.06]
15. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011;4:47-55 [FREE Full text] [doi: 10.2147/RMHP.S12985] [Medline: 22312227]
16. Menachemi N, Brooks RG. Reviewing the benefits and costs of electronic health records and associated patient safety technologies. J Med Syst 2006 Jun;30(3):159-168. [doi: 10.1007/s10916-005-7988-x] [Medline: 16848129]
17. Jang J, Yu SH, Kim C, Moon Y, Kim S. The effects of an electronic medical record on the completeness of documentation in the anesthesia record. Int J Med Inform 2013 Aug;82(8):702-707. [doi: 10.1016/j.ijmedinf.2013.04.004] [Medline: 23731825]
18. Tang PC, LaRosa MP, Gorden SM. Use of computer-based records, completeness of documentation, and appropriateness of documented clinical decisions. J Am Med Inform Assoc 1999;6(3):245-251 [FREE Full text] [doi: 10.1136/jamia.1999.0060245] [Medline: 10332657]
19. Gunter TD, Terry NP. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. J Med Internet Res 2005 Mar 14;7(1):e3 [FREE Full text] [doi: 10.2196/jmir.7.1.e3] [Medline: 15829475]
20. Musen MA, Middleton B, Greenes RA. Clinical decision-support systems. In: Biomedical Informatics. Berlin, Germany: Springer; 2014:643-674.
21. Raposo VL. Electronic health records: Is it a risk worth taking in healthcare delivery? GMS Health Technol Assess 2015;11:Doc02 [FREE Full text] [doi: 10.3205/hta000123] [Medline: 26693253]
22. Pelletier LR. Information-Enabled Decision-Making in Health Care: EHR-Enabled Standardization, Physician Profiling and Medical Home. 2010. URL: https://web.wpi.edu/Pubs/ETD/Available/etd-042510-120618/unrestricted/Dissertation_Lori_Pelletier_FINAL.pdf [accessed 2021-10-05]
23. Thomas J, Dahm MR, Li J, Smith P, Irvine J, Westbrook JI, et al. Variation in electronic test results management and its implications for patient safety: A multisite investigation. J Am Med Inform Assoc 2020 Aug 01;27(8):1214-1224 [FREE Full text] [doi: 10.1093/jamia/ocaa093] [Medline: 32719839]
24. Detsky AS. Regional variation in medical care. N Engl J Med 1995 Aug 31;333(9):589-590. [doi: 10.1056/NEJM199508313330911] [Medline: 7623911]
25. McLaughlin CP. Why variation reduction is not everything: a new paradigm for service operations. Int J of Service Industry Mgmt 1996 Aug;7(3):17-30. [doi: 10.1108/09564239610122938]
26. McLaughlin C, Johnson S. Inherent variability in service operations: identification, measurement and implications. In: Services Management: New Directions and Perspectives. London, UK: Cassell; 1995:226-229.
27. KLAS Research. Global (Non-US) EMR Market Share. 2019. URL: https://klasresearch.com/report/global-non-us-emr-market-share-2019/1460 [accessed 2021-10-05]
28. KLAS Research. US Hospital EMR Market Share. 2020. URL: https://klasresearch.com/report/us-hospital-emr-market-share-2020/1616 [accessed 2021-10-05]
29. Fierce Healthcare. Epic, Meditech Gain U.S. Hospital Market Share as Other EHR Vendors Lose Ground.: Fierce Health care; 2020. URL: https://www.fiercehealthcare.com/tech/epic-meditech-gain-u-s-hospital-market-share-as-other-ehr-vendors-lose-ground [accessed 2021-10-05]
30. Heckel K. Power Plans. 2014. URL: https://docport.columbia-stmarys.org/gradepoint/internet/CPOEProviderGuide/PowerPlans.pdf
31. BJC HealthCare. EPIC- Resources - How To Speak Epic. 2020. URL: https://www.epic1.org/Resources/How-to-Speak-Epic [accessed 2021-10-05]
32. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009 Jul 21;6(7):e1000100 [FREE Full text] [doi: 10.1371/journal.pmed.1000100] [Medline: 19621070]
33. Adelson KB, Qiu YC, Evangelista M, Spencer-Cisek P, Whipple C, Holcombe RF. Implementation of electronic chemotherapy ordering: an opportunity to improve evidence-based oncology care. J Oncol Pract 2014 Mar;10(2):e113-e119. [doi: 10.1200/JOP.2013.001184] [Medline: 24371301]
34. Akenroye AT, Stack AM. The development and evaluation of an evidence-based guideline programme to improve care in a paediatric emergency department. Emerg Med J 2016 Feb;33(2):109-117. [doi: 10.1136/emermed-2014-204363] [Medline: 26150121]

35. Amland RC, Dean BB, Yu H, Ryan H, Orsund T, Hackman JL, et al. Computerized clinical decision support to prevent venous thromboembolism among hospitalized patients: proximal outcomes from a multiyear quality improvement project. J Healthc Qual 2015;37(4):221-231. [doi: 10.1111/jhq.12069] [Medline: 26151096]
36. Asan O, Nattinger AB, Gurses AP, Tyszka JT, Yen TWF. Oncologists' views regarding the role of electronic health records in care coordination. JCO Clin Cancer Inform 2018 Dec;2:1-12 [FREE Full text] [doi: 10.1200/CCI.17.00118] [Medline: 30652555]
37. Attaar A, Wei J, Brunetti L. Evaluating adherence to guideline-directed infection management pre- and postimplementation of an electronic order set. J Pharm Pract 2021 Oct;34(5):721-726. [doi: 10.1177/0897190020903854] [Medline: 32054402]
38. Ballesca MA, LaGuardia JC, Lee PC, Hwang AM, Park DK, Gardner MN, et al. An electronic order set for acute myocardial infarction is associated with improved patient outcomes through better adherence to clinical practice guidelines. J Hosp Med 2014 Mar;9(3):155-161. [doi: 10.1002/jhm.2149] [Medline: 24493376]
39. Borok J, Udkoff J, Vaida F, Murphy J, Torriani F, Waldman A, et al. Transforming acne care by pediatricians: An interventional cohort study. J Am Acad Dermatol 2018 Nov;79(5):966-968 [FREE Full text] [doi: 10.1016/j.jaad.2018.04.055] [Medline: 29753064]
40. Bradywood A, Farrokhi F, Williams B, Kowalczyk M, Blackmore CC. Reduction of inpatient hospital length of stay in lumbar fusion patients with implementation of an evidence-based clinical care pathway. Spine (Phila Pa 1976) 2017 Feb;42(3):169-176. [doi: 10.1097/BRS.0000000000001703] [Medline: 27213939]
41. Chisolm DJ, McAlearney AS, Veneris S, Fisher D, Holtzlander M, McCoy KS. The role of computerized order sets in pediatric inpatient asthma treatment. Pediatr Allergy Immunol 2006 May;17(3):199-206. [doi: 10.1111/j.1399-3038.2005.00362.x] [Medline: 16672007]
42. Dort JC, Sauro KM, Chandarana S, Schrag C, Matthews J, Nakoneshny S, et al. The impact of a quality management program for patients undergoing head and neck resection with free-flap reconstruction: longitudinal study examining sustainability. J Otolaryngol Head Neck Surg 2020 Jun 23;49(1):42 [FREE Full text] [doi: 10.1186/s40463-020-00437-2] [Medline: 32571424]
43. Ebinger JE, Porten BR, Strauss CE, Garberich RF, Han C, Wahl SK, et al. Design, challenges, and implications of quality improvement projects using the electronic medical record: case study: a protocol to reduce the burden of postoperative atrial fibrillation. Circ Cardiovasc Qual Outcomes 2016 Sep;9(5):593-599 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.116.003122] [Medline: 27553597]
44. Geltman PL, Fried LE, Arsenault LN, Knowles AM, Link DA, Goldstein JN, et al. A planned care approach and patient registry to improve adherence to clinical guidelines for the diagnosis and management of attention-deficit/hyperactivity disorder. Acad Pediatr 2015;15(3):289-296. [doi: 10.1016/j.acap.2014.12.002] [Medline: 25906699]
45. Goga JK, Depaolo A, Khushalani S, Walters JK, Roca R, Zisselman M, et al. Lean methodology reduces inappropriate use of antipsychotics for agitation at a psychiatric hospital. Consult Pharm 2017 Jan 01;32(1):54-62. [doi: 10.4140/TCP.n.2017.54] [Medline: 29221501]
46. Gulati S, Zouk AN, Kalehoff JP, Wren CS, Davison PN, Kirkpatrick DP, et al. The use of a standardized order set reduces systemic corticosteroid dose and length of stay for individuals hospitalized with acute exacerbations of COPD: a cohort study. Int J Chron Obstruct Pulmon Dis 2018;13:2271-2278 [FREE Full text] [doi: 10.2147/COPD.S165665] [Medline: 30100717]
47. Hendrickson MA, Wey AR, Gaillard PR, Kharbanda AB. Implementation of an electronic clinical decision support tool for pediatric appendicitis within a hospital network. Pediatr Emerg Care 2018 Jan;34(1):10-16 [FREE Full text] [doi: 10.1097/PEC.0000000000001069] [Medline: 28277414]
48. Hooper DK, Kirby CL, Margolis PA, Goebel J. Reliable individualized monitoring improves cholesterol control in kidney transplant recipients. Pediatrics 2013 Apr;131(4):e1271-e1279 [FREE Full text] [doi: 10.1542/peds.2012-2374] [Medline: 23478865]
49. Horton JD, Corrigan C, Patel T, Schaffer C, Cina RA, White DR. Effect of a standardized electronic medical record order set on opioid prescribing after tonsillectomy. Otolaryngol Head Neck Surg 2020 Aug;163(2):216-220. [doi: 10.1177/0194599820911721] [Medline: 32178580]
50. Jacobs BR, Hart KW, Rucker DW. Reduction in clinical variance using targeted design changes in Computerized Provider Order Entry (CPOE) order sets: impact on hospitalized children with acute asthma exacerbation. Appl Clin Inform 2012;3(1):52-63 [FREE Full text] [doi: 10.4338/ACI-2011-01-RA-0002] [Medline: 23616900]
51. Karajgikar ND, Manroa P, Acharya R, Codario RA, Reider JA, Donihi AC, et al. Addressing pitfalls in management of diabetic ketoacidosis with a standardized protocol. Endocr Pract 2019 May;25(5):407-412. [doi: 10.4158/EP-2018-0398] [Medline: 30657360]
52. Lewin SM, McConnell RA, Patel R, Sharpton SR, Velayos F, Mahadevan U. Improving the quality of inpatient ulcerative colitis management: promoting evidence-based practice and reducing care variation with an inpatient protocol. Inflamm Bowel Dis 2019 Oct 18;25(11):1822-1827. [doi: 10.1093/ibd/izz066] [Medline: 30980712]
53. Lindberg SM, Anderson CK. Improving gestational weight gain counseling through meaningful use of an electronic medical record. Matern Child Health J 2014 Nov;18(9):2188-2194 [FREE Full text] [doi: 10.1007/s10995-014-1467-2] [Medline: 24627233]

54. Lindberg SM, DeBoth A, Anderson CK. Effect of a best practice alert on gestational weight gain, health services, and pregnancy outcomes. Matern Child Health J 2016 Oct;20(10):2169-2178 [FREE Full text] [doi: 10.1007/s10995-016-2052-7] [Medline: 27395382]
55. Morrisette M, Hammer J, Anderson W, Norton H, Green M, Gesin G. Impact of a multifaceted intervention on prescribing of proton pump inhibitors for stress ulcer prophylaxis in the critically ill. Arch Crit Care Med 2015 May 30;1(2):e1. [doi: 10.17795/accm-3969]
56. Prevedello LM, Raja AS, Ip IK, Sodickson A, Khorasani R. Does clinical decision support reduce unwarranted variation in yield of CT pulmonary angiogram? Am J Med 2013 Nov;126(11):975-981 [FREE Full text] [doi: 10.1016/j.amjmed.2013.04.018] [Medline: 24157288]
57. Reynolds EL, Burke JF, Banerjee M, Callaghan BC. Randomized controlled trial of a clinical decision support system for painful polyneuropathy. Muscle Nerve 2020 May;61(5):640-644. [doi: 10.1002/mus.26774] [Medline: 31811650]
58. Rooholamini SN, Clifton H, Haaland W, McGrath C, Vora SB, Crowell CS, et al. Outcomes of a clinical pathway to standardize use of maintenance intravenous fluids. Hosp Pediatr 2017 Dec;7(12):703-709. [doi: 10.1542/hpeds.2017-0099] [Medline: 29162640]
59. Sim EY, Tan DJA, Abdullah HR. The use of computerized physician order entry with clinical decision support reduces practice variance in ordering preoperative investigations: A retrospective cohort study. Int J Med Inform 2017 Dec;108:29-35. [doi: 10.1016/j.ijmedinf.2017.09.015] [Medline: 29132628]
60. Sonstein L, Clark C, Seidensticker S, Zeng L, Sharma G. Improving adherence for management of acute exacerbation of chronic obstructive pulmonary disease. Am J Med 2014 Nov;127(11):1097-1104 [FREE Full text] [doi: 10.1016/j.amjmed.2014.05.033] [Medline: 24927911]
61. Soo G, Wong Doo N, Burrows J, Ritchie A, Zhang J, Burke R. Improving the adoption of an electronic clinical decision support tool and evaluating its effect on venous thromboembolism prophylaxis prescribing at a Sydney tertiary teaching hospital. J Pharm Pract Res 2019 Jul 09;49(6):508-516. [doi: 10.1002/jppr.1562]
62. Terasaki J, Singh G, Zhang W, Wagner P, Sharma G. Using EMR to improve compliance with clinical practice guidelines for management of stable COPD. Respir Med 2015 Nov;109(11):1423-1429 [FREE Full text] [doi: 10.1016/j.rmed.2015.10.003] [Medline: 26475055]
63. Wang EJ, Helgesen R, Johr CR, Lacko HS, Ashburn MA, Merkel PA. Targeted program in an academic rheumatology practice to improve compliance with opioid prescribing guidelines for the treatment of chronic pain. Arthritis Care Res (Hoboken) 2021 Oct;73(10):1425-1429. [doi: 10.1002/acr.24354] [Medline: 32558375]
64. Webber EC, Warhurst HM, Smith SS, Cox EG, Crumby AS, Nichols KR. Conversion of a single-facility pediatric antimicrobial stewardship program to multi-facility application with computerized provider order entry and clinical decision support. Appl Clin Inform 2013;4(4):556-568 [FREE Full text] [doi: 10.4338/ACI-2013-07-RA-0054] [Medline: 24454582]
65. Studer A, Billings K, Thompson D, Ida J, Rastatter J, Patel M, et al. Standardized order set exhibits surgeon adherence to pain protocol in pediatric adenotonsillectomy. Laryngoscope 2021 Jul;131(7):E2337-E2343. [doi: 10.1002/lary.29314] [Medline: 33314128]
66. Teich JM, Merchia PR, Schmiz JL, Kuperman GJ, Spurr CD, Bates DW. Effects of computerized physician order entry on prescribing practices. Arch Intern Med 2000 Oct 09;160(18):2741-2747 [FREE Full text] [doi: 10.1001/archinte.160.18.2741] [Medline: 11025783]
67. Kicker JS, Hill HS, Matheson CK. Better pairing propofol volume with procedural needs: a propofol waste reduction quality improvement project. Hosp Pediatr 2018 Oct;8(10):604-610. [doi: 10.1542/hpeds.2018-0010] [Medline: 30206112]
68. Smalheiser NR, Zhou W, Torvik VI. Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. J Biomed Discov Collab 2008 Feb 15;3:2 [FREE Full text] [doi: 10.1186/1747-5333-3-2] [Medline: 18279519]
69. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968 Oct;70(4):213-220. [doi: 10.1037/h0026256] [Medline: 19673146]
70. Rosovsky RP, Barra ME, Roberts RJ, Parmar A, Andonian J, Suh L, et al. When pigs fly: a multidisciplinary approach to navigating a critical heparin shortage. Oncologist 2020 Apr;25(4):334-347 [FREE Full text] [doi: 10.1634/theoncologist.2019-0910] [Medline: 32154634]
71. Covidence Systematic Review Software. Melbourne, VIC, Australia: Veritas Health Innovation; 2020. URL: https://www.covidence.org/ [accessed 2021-10-05]
72. Sirriyeh R, Lawton R, Gardner P, Armitage G. Reviewing studies with diverse designs: the development and evaluation of a new tool. J Eval Clin Pract 2012 Aug;18(4):746-752. [doi: 10.1111/j.1365-2753.2011.01662.x] [Medline: 21410846]
73. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ. Cochrane Handbook for Systematic Reviews of Interventions. Hoboken, NJ: John Wiley & Sons; 2019.
74. Bencomo M. Implementing a Clinical Protocol: Early Enteral Nutrition Therapy in Critically Ill Patients. proquest.com. Peoria, IL: Saint Francis Medical Center College of Nursing; 2019. URL: https://tinyurl.com/t6bz9jyw [accessed 2020-06-22]
75. Kotter JP. Leading Change: Why Transformation Efforts Fail. Harvard Business Review. 1995. URL: https://hbr.org/1995/05/leading-change-why-transformation-efforts-fail-2 [accessed 2021-10-05]

76. Campbell RJ. Change management in health care. Health Care Manag (Frederick) 2008;27(1):23-39. [doi: 10.1097/01.hcm.0000285028.79762.a1] [Medline: 18510142]
77. Robinson KE, Kersey JA. Novel electronic health record (EHR) education intervention in large healthcare organization improves quality, efficiency, time, and impact on burnout. Medicine (Baltimore) 2018 Sep;97(38):e12319 [FREE Full text] [doi: 10.1097/MD.0000000000012319] [Medline: 30235684]
78. Al-Abri RK, Al-Hashmi IS. The learning organisation and health care education. Sultan Qaboos Univ Med J 2007 Dec;7(3):207-214 [FREE Full text] [Medline: 21748105]
79. Hulse NC, Lee J, Borgeson T. Visualization of order set creation and usage patterns in early implementation phases of an electronic health record. AMIA Annu Symp Proc 2016;2016:657-666 [FREE Full text] [Medline: 28269862]
80. Senge PM. The Fifth Discipline: The Art and Practice of the Learning Organization. Redfern, NSW: Currency Press; 2006.
81. Campanella P, Lovato E, Marone C, Fallacara L, Mancuso A, Ricciardi W, et al. The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. Eur J Public Health 2016 Feb;26(1):60-64. [doi: 10.1093/eurpub/ckv122] [Medline: 26136462]
82. Riza R, Nurwahyuni A. The implementation and outcome of clinical pathway: a systematic review. 2019 Presented at: The 5th International Conference on Public Health; February 13-14, 2019; Solo, Indonesia p. 677-686. [doi: 10.26911/theicph.2019.05.33]
83. Patel BN. Impact of implementing a computerised quality improvement intervention in primary healthcare. 2018. URL: https://ses.library.usyd.edu.au/handle/2123/18988 [accessed 2021-10-05]

Abbreviations

CDS: clinical decision support
EHR: electronic health record
EMR: electronic medical record
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QATSDD: Quality Assessment Tool for Studies with Diverse Designs

Edited by G Eysenbach; submitted 16.05.21; peer-reviewed by A Kazley; comments to author 17.06.21; revised version received 22.06.21; accepted 19.09.21; published 17.11.21.

Please cite as:
Hodgson T, Burton-Jones A, Donovan R, Sullivan C
The Role of Electronic Medical Records in Reducing Unwarranted Clinical Variation in Acute Health Care: Systematic Review
JMIR Med Inform 2021;9(11):e30432
URL: https://medinform.jmir.org/2021/11/e30432
doi:10.2196/30432
PMID:34787585

Original Paper

Introduction of Systematized Nomenclature of Medicine–Clinical Terms Coding Into an Electronic Health Record and Evaluation of its Impact: Qualitative and Quantitative Study

Tanya Pankhurst[1], MBBS, MSci, PhD; Felicity Evison[1], MSc; Jolene Atia[1], PhD; Suzy Gallier[1], BSc; Jamie Coleman[1], MBChB, MA, MD, FRCP; Simon Ball[1,2], MA, PhD, FRCP; Deborah McKee[1]; Steven Ryan[1], BSc (Hons), PGCE, MSc; Ruth Black[3], JD, EdD

[1]NHS Foundation Trust, University Hospitals Birmingham, Birmingham, United Kingdom

[2]Health Data Research UK (HDR-UK), University of Birmingham, Birmingham, United Kingdom

[3]Institute for Global Health Innovation (IGHI), Imperial College London, London, United Kingdom

Corresponding Author:
Tanya Pankhurst, MBBS, MSci, PhD
NHS Foundation Trust
University Hospitals Birmingham
Mindelsohn Way
Birmingham, B15 2TG
United Kingdom
Phone: 44 7811357984
Email: pankhurst.tanya@gmail.com

Abstract

Background: This study describes the conversion within an existing electronic health record (EHR) from the *International Classification of Diseases, Tenth Revision* coding system to the SNOMED-CT (Systematized Nomenclature of Medicine–Clinical Terms) for the collection of patient histories and diagnoses. The setting is a large acute hospital that is designing and building its own EHR. Well-designed EHRs create opportunities for continuous data collection, which can be used in clinical decision support rules to drive patient safety. Collected data can be exchanged across health care systems to support patients in all health care settings. Data can be used for research to prevent diseases and protect future populations.

Objective: The aim of this study was to migrate a current EHR, with all relevant patient data, to the SNOMED-CT coding system to optimize clinical use and clinical decision support, facilitate data sharing across organizational boundaries for national programs, and enable remodeling of medical pathways.

Methods: The study used qualitative and quantitative data to understand the successes and gaps in the project, clinician attitudes toward the new tool, and the future use of the tool.

Results: The new coding system (*tool*) was well received and immediately widely used in all specialties. This resulted in increased, accurate, and clinically relevant data collection. Clinicians appreciated the increased depth and detail of the new coding, welcomed the potential for both data sharing and research, and provided extensive feedback for further development.

Conclusions: Successful implementation of the new system aligned the University Hospitals Birmingham NHS Foundation Trust with national strategy and can be used as a blueprint for similar projects in other health care settings.

(JMIR Med Inform 2021;9(11):e29532) doi:10.2196/29532

KEYWORDS

coding standards; clinical decision support; Clinician led design; clinician reported experience; clinical usability; data sharing; diagnoses; electronic health records; electronic health record standards; health data exchange; health data research; International Classification of Diseases version 10 (ICD-10); National Health Service Blueprint; patient diagnoses; population health; problem list; research; Systematized Nomenclature Of Medicine–Clinical Terms (SNOMED-CT); use of electronic health data; user-led design

Introduction

Background

The Digital Age has brought fast and convenient technology to almost all industries, and latterly to health care. The current system in Britain is divided between primary health care, where general practitioners increasingly manage chronic disease, and secondary and tertiary care, which largely remain modeled on hospitals. General practitioner records have been electronic for 20 years [1], but hospital records are only being ubiquitously considered for digital conversion now. These electronic health records (EHRs) are variable and often either overly simplistic [2] or require so much data input that they overburden clinicians [3]. EHRs have grown up separately and most still do not allow easy sharing of data, a situation further compounded by variability of the data themselves and lack of standardized coding [4]. Paucity of data exchange is not only a barrier to treating patients [5] but also prevents remodeling of medical pathways [6] locking the National Health Service (NHS) into artificial divisions between primary and secondary care. NHS Digital has issued guidance on the use of standards, including the universal use of the SNOMED-CT (Systematized Nomenclature of Medicine–Clinical Terms) for diagnoses [7] and other standards for medications, data messaging, and patient identification [8].

University Hospitals Birmingham (UHB) is a large secondary and tertiary referral NHS Foundation Trust comprising 4 hospitals and community services across Birmingham, the second largest city in the United Kingdom. UHB has an annual turnover of £1.4 (US $1.88) billion and treats 2.2 million people per year [9]. Across its 4 sites it has 3000 inpatient beds and employs more than 20,000 people. UHB has been at the forefront of digital innovation and is unique in the United Kingdom in employing programmers to build its own EHR. The organization was recognized as a global digital exemplar in 2017 [10]. It has strong links to primary care and is engaged in building regional data-sharing platforms [11]. Therefore, it is highly influential in digital health care development regionally and nationally in the United Kingdom.

The University Hospitals Birmingham NHS Foundation Trust's (the Trust, hereafter) central EHR, named *Prescribing Information and Communication System* (*PICS*), [12] comprises contemporaneous clinical noting; electronic prescribing and medicines administration; electronic observations; electronic ordering; patient barcoding; and electronic results display and complex clinical decision support (CDS). UHB builds other software, including clinical and patient portals.

Electronic health care systems are ubiquitously used, and their design has been controlled by clinicians over many years. Diagnostic coding within PICS was based on the *International Classification of Diseases, Tenth Revision* (*ICD-10*) [13], which is the current system used for payment of hospitals. The *ICD-10* was originally designed for population studies and is, therefore, not ideally suited to hospital settings. Clinicians often cannot find diagnoses they need because detailed or rare conditions are listed under *any other...*. Consequently, diagnoses are often not recorded and are therefore lost to use within any CDS rules and lost to data exchange beyond the organizational boundaries.

Implementation of SNOMED-CT coding into PICS was predicted to allow clinicians to more accurately enter detailed diagnoses, enable more complex CDS, and facilitate more accurate data collection, provided the tool was intuitive. In addition, there was a strong interest from the central government and the public to share health data. Primary care has already encoded diagnoses using SNOMED-CT, and similar data collection in hospitals would enable data exchange across previously incompatible data entry applications. This approach enables the remodeling of medical pathways and begins to address the siloed nature of medical care.

Objectives

The organization's mission is to improve the health of patients and communities, and UHB is committed to transformation through consistent innovation across digital systems. The goals of UHB align with national expectations around standardized electronic data collection and the easy and seamless sharing of data [14]. In addition, clinicians require a more enriched and relevant system to record medical diagnoses. To address these goals, this study sought to establish coding within all components of the EHR based on the SNOMED-CT by April 2020. Therefore, the question was whether the current EHR, with all relevant patient data, could be migrated to the SNOMED-CT coding system, to optimize clinical use and CDS and facilitate data sharing across organizational boundaries for national programs and remodeling of medical pathways.

Methods

Overview

The formal methodology was designed to assess the success of the tool (utility or clinical relevance), success of mapping and safeguarding CDS rules in the software, and successful data collection for direct clinical care, research, and data exchange. This used pre-existing data sets, interviews, focus groups, and surveys of the stakeholders, and formal assessment of the use of the tool with quantitative data, using a mixed methods approach.

Existing Data Sets: Fixed Points

The design and build of the SNOMED-CT tool were created by clinicians and developers using an iterative model. Design meetings included clinicians, developers, business analysts, and a project manager, and a specification document was initially developed; development followed, and demonstrations of the software build and modifications were reviewed as they progressed over time. Many of the questions that the study explores in the postlaunch analysis (eg, "Is the tool useful?") can be analyzed in this predata set, where the clinicians and developers worked together (eg, to create a tool that they believed *would* have clinical utility). The project specification document, specifically the user stories, provides a fixed point for analysis (Multimedia Appendix 1).

Surveys

A survey targeted at a large sample of clinicians after the launch of the SNOMED-CT tool assessed clinician opinion on utility and improvements in clinical care associated with the tool. The survey was sent by email across the Trust via Typeform [15] to all junior doctors and all consultants on the master staff index. Senior nurses (the only nursing group using the tool) were sent the same survey via the divisional leads; Allied Health Professionals (AHPs) were sampled using group leaders.

The SNOMED-CT tool was designed to collect consistent and standardized data, independent of which health care workers from the multidisciplinary team were collecting it. Thus, the tool had to work for doctors, nurses, and AHPs and be intuitive and accessible to junior and senior staff. To qualitatively assess whether this was successful, the study used a wide sample of users.

The sampling strategy was typical purposeful sampling, aiming to represent the average clinical situation and sample across all types of staff using the new tool. The sampling aimed to highlight the experience of this tool and understand the typical, normal, and average experience of a clinician using the tool [16].

Survey questions were developed with a focus on understanding staff experiences and future expectations regarding the use of a tool in constructing a problem list within a patient record. The tool's utility in gathering data for direct clinical care, onward sharing, and the creation of research sets was explored. There were 7 elements to the survey (Multimedia Appendix 2) designed on a Likert scale [17] with one open-ended question, allowing for more detailed data collection.

From previous surveys of this type within the organization, the expected response rate was about 20%. The survey was designed for easy and rapid completion. Clinicians are busy, and traditionally ignore feedback requests. On the basis of previous experience and supported by market research [18], it was known that surveys exceeding 5 minutes would not be completed. The survey was therefore much shorter than a classical qualitative study survey, but accrued as much qualitative information as possible.

The validity and reliability of these items were tested by adequate engagement to saturation (the survey was repeatedly sent out to the staff lists until there were no new elements to answers) [19] and by triangulation or crystallization [20] of pre-existing data sets, focus groups, and data collection.

Interviews

Senior programmer interviews created an understanding of build success in relation to the original specification. Senior developers worked closely with clinicians and identified flaws in the design and implementation. Interviews with senior developers revealed difficulties and workarounds in the course of the build and illuminated reasons for subsequent clinician behavior. Interviews were designed as one-time sessions with each of the 2 principal developers in a meeting room in the building in which they worked. The interviews were informal and semistructured, asked open questions and allowed for follow-up and probing. The interviews were recorded and transcribed later, and notes were taken during the interviews. The questions (Multimedia Appendix 3) were designed to investigate, from the developers' perspective, the success of the project, and the aspects that they viewed as strengths and weaknesses. Interviews also queried how developers viewed users and requested opinions about the implications of the project beyond the software build. The questions were deliberately open [16].

In similar previous projects, it was noted that developers who worked closely with clinicians often had insight into the software and possible enhancements that would improve user experience and data exchange. Developers are often reserved in their judgment and often do not proffer an opinion unless specifically asked.

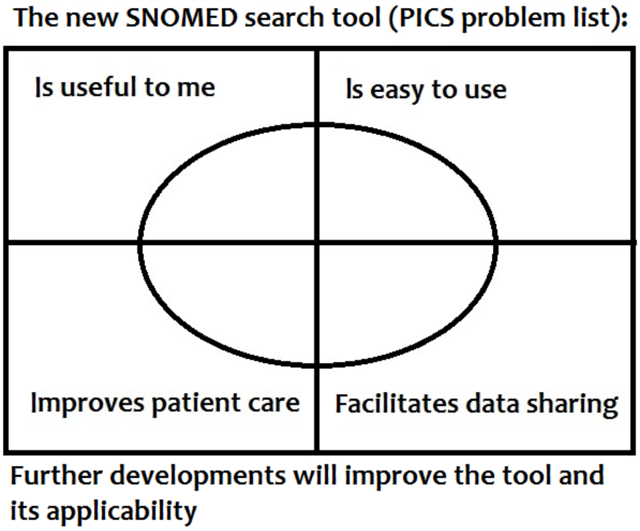
Focus Groups and Email Feedback

Focus groups aimed to understand the validity and utility of the final product from clinicians involved in the iterative design and build of the software. In addition, unsolicited feedback from clinicians is a frequent part of the EHR build, and email feedback was therefore included in the thematic analysis. There were up to 50 clinicians involved in the design of the EHR, but the focus groups included between 5 and 10 participants, as this was an optimal number to support vibrant discussion while allowing all voices to be heard [21], thus using nonprobability unique sampling [19]. Clinicians who were invited were the ones most actively involved in the design and build, and to whom the tool was most relevant (because of direct clinical care, a research interest, or an interest in onward data sharing). Additional email feedback was not requested, but extensive feedback from this source was provided and was therefore included in the analysis. Clinicians in the focus groups and those who provided email feedback were likely to be a subset of those responding to the survey.

Two focus groups were conducted with 6 and 8 clinicians. The focus groups were informal and semistructured with prompts (Figure 1), which not only allowed for expression of opinion on the entire conceptual framework but also allowed for discussion, so that aspects of the project that had not been previously understood by the investigators were revealed.

The questions were designed to understand the utility of the tool in terms of clinical use and acceptability. This requires careful assessment, as clinicians may welcome even clunky digital tools if they think that patient care is improved. In relation to the conceptual framework, questions on the applicability of the tool to direct clinical care, research, and data sharing were also explored.

Figure 1. Importance of the themes to the users. PICS: Prescribing Information and Communication System; SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms.


The new SNOMED search tool (PICS problem list):
Is useful to me
Is easy to use
Improves patient care
Facilitates data sharing
Further developments will improve the tool and its applicability

Data Analysis

Pre-existing Data Sets or Fixed Points

These predated the study and were analyzed by extracting data against the key themes of improving care, improving access to data, and ensuring usability of the tool. The data set comprised user stories from a specification document developed iteratively with developers and clinicians. Suggestions and actions incorporated during the data build were analyzed and used to understand whether the same points were discussed in the survey responses and focus groups (eg, were they addressed in the build? If they were not, why not?). The documents were coded in a similar manner to that of the focus groups, which allowed theming and issue spotting.

Surveys

Descriptive statistical analyses were conducted on the results after all the surveys were submitted. For all questions answered on a Likert scale, responses to each question were analyzed as whole numbers in each group. Binary questions were presented as percentages. The analysis of open questions gathered themes with the quoted examples.

Interviews, Focus Groups, and Email Feedback

Interviews and focus groups were recorded and transcribed, and thoughts, concerns, and conclusions documented by the investigator during or immediately after the interviews or focus groups. The transcription was coded and scored on a matrix (Figure 1) for frequency, concern, and time spent discussing each issue and where they overlapped using a multifactorial balancing test-like methodology [22]. Themes or issues were identified and cross-referenced with survey results and pre-existing data sets.

Informal email feedback was collected by the chief investigator and incorporated into the thematic analysis.

Use Data

Data were collected digitally on the use of the SNOMED-CT tool across the entire organization by the Informatics Department. The following data were collected: total number of users, by area and specialty; number of terms used, by user and by specialty; and application over time (for the first 4 months of use) to understand whether the use was increasing or waning. These data were presented graphically as absolute numbers and as the percentage use of the total user base.

Ethics

Human subjects were involved in this research, and participation was voluntary and confidential. Informed consent was obtained from all participants, including for recording and storing of information, and consent could be withdrawn at any time until results were collated in an anonymized manner. Any quotations from the interviews and focus groups were only used with consent. It is unlikely in a study of this type that interviews or focus groups would cause ethical issues to be raised as private behavior was not the subject of this study [16]. The data were stored securely and confidentially.

Bias arising from the dual role of the researcher as a designer or user and evaluator of the tool [16] was mitigated by processes already in place. All software build within the organization was controlled by a project group, with a project manager, business analyst, standard testing, and rollout procedures. Although researchers have invested in the project, many software developments to date have required extensive iterations to ensure utility, and this was expected to be necessary in this project also.

Results

Overview

Data collection took place during September 2019-December 2019 and was conducted at UHB, United Kingdom. Statistics for the problem list in PICS were gathered by the Informatics Department, from the date of the tool release into PICS software, September 2019, to the end of December 2019.

The 5 key themes with subthemes that emerged from this study are detailed in Table 1.

Table 1. Themes emerging from user feedback, with positive and negative aspects.

Theme and subtheme	Positive feedback	Negative feedback
Design commitment and project success		
In general	• All project objectives met • Positive user feedback	—[a]
Usability	• High use • Easy to find • Search tool affective	• Changes to user interface needed • Some terms not found • Too many choices
Engagement	• Things clear and in the same place	• Time constraints • Incentive low
Data quality	• Much more detailed data • Enriched data sets	• Incomplete data • Too much data
Improving patient care	• Use in safety rules • Improves communication	• Lists too long • Lists not sorted in my order
Research	• Data rich source	• Too many codes
Data sharing		
In general	• Standardized data collection • Sharing with general practitioners • Sharing with other hospitals	—
Standards and payment	• Provides standard	—

[a] No data obtained.

Questionnaires were sent to 585 junior doctors, 315 consultants, 40 senior nurses, and 20 AHPs. In total, 15.3% (147/960) responded to the survey. The survey was anonymous; thus, the role of the respondents was not recorded.

Design, Commitment, and Project Success

The project delivered the planned tool. The requirement for programmer time was high, up to 40 hours per week in the last month of the project, with extensive preceding groundwork for the project team. A total of 300,000 new codes were added to the EHR, and 15,000 old codes were mapped to new codes. Because of the considerably increased detail in SNOMED-CT compared with the *ICD-10*, this mapping was complicated and multiple SNOMED-CT codes had to be mapped back to each *ICD-10* code. In addition to coding, the existing EHR had over 1000 existing CDS rules, all of which were successfully mapped to new codes. All these changes required a new user interface (UI), which was iteratively built with clinicians.

The programmers who built the new tool felt that the overall project was well organized and well managed. However, programmers were frustrated at the time required for mapping and checking the codes. The programmers felt that coders should have been involved earlier and more extensively. Before the project, the programmers were concerned that loading thousands of new codes would slow down the application, reducing both the speed of the search and the EHR application itself. This concern did not materialize because search speeds and EHR stability were not impacted by the loading of a large number of new codes.

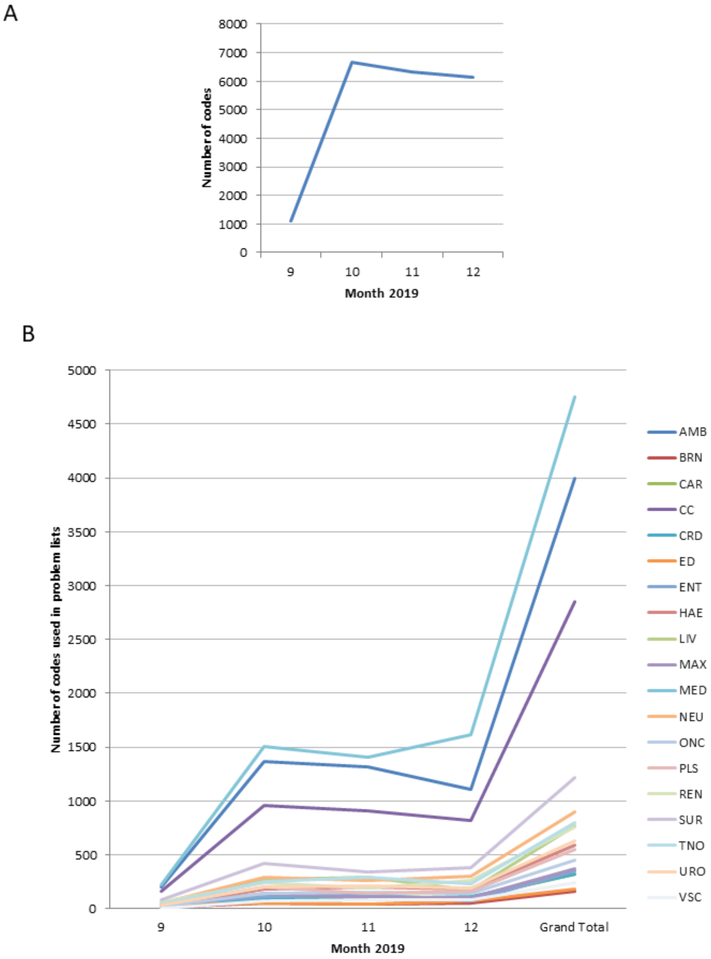
Programmers further commented that predicting problems in organizing and anticipating questions was difficult in innovation projects of this scope. Fixed data sources (Multimedia Appendix 1) included specification documents and these outlined *user stories* identifying what users articulated as their need. All phase 1 requirements were delivered. Some modifications were requested in the feedback, which are outlined in the themes below. Similar to the programmers, users fed back that the project was successful. Users had assumed that it would be successful, as the build of EHR in our organization has a good track record for the delivery of successful tools.

Usability

This tool has been widely used. After the first month, codes entered into the EHR rapidly rose to between 6000 and 7000 a month and continued to be high over the 4 months of the study (Figure 2). This was higher than expected, as no specific communications or *advertising* was put out to the organization. The expectation is that the number of codes entered each month will fall as more returning patients have most of their medical history added to the EHR. It is anticipated that a steady state will be reached in approximately 12 months with only new patients or new problems being added to the system.

Codes were used in all specialties. As expected, the highest use was in general medicine, short-stay ambulatory care, and critical care, but use was seen across all hospital specialties (Figure 2). Most users were doctors and a significant minority nurses. Small numbers of pharmacists and AHPs were starting to use the tool during the study period, and this is expected to rise as these groups of professionals generally follow doctors’ behavior.

Figure 2. Use of the new tool. (A) Monthly problem list use. (B) Use of problem lists by clinician type. AMB: ambulatory care; BRN: Burns Medicine; CAR: cardiac surgery; CC: critical care; CRD: cardiology; ED: emergency medicine; ENT: ear nose and throat; HAE: haematology; LIV: liver medicine; MAX: maxillo-facial surgery; MED: general medicine; NEU: neurology; ONC: oncology; PLS: plastic surgery; REN: renal medicine; SUR: general surgery; TNO: trauma and orthopaedics; URO: urology; VSC: vascular surgery.

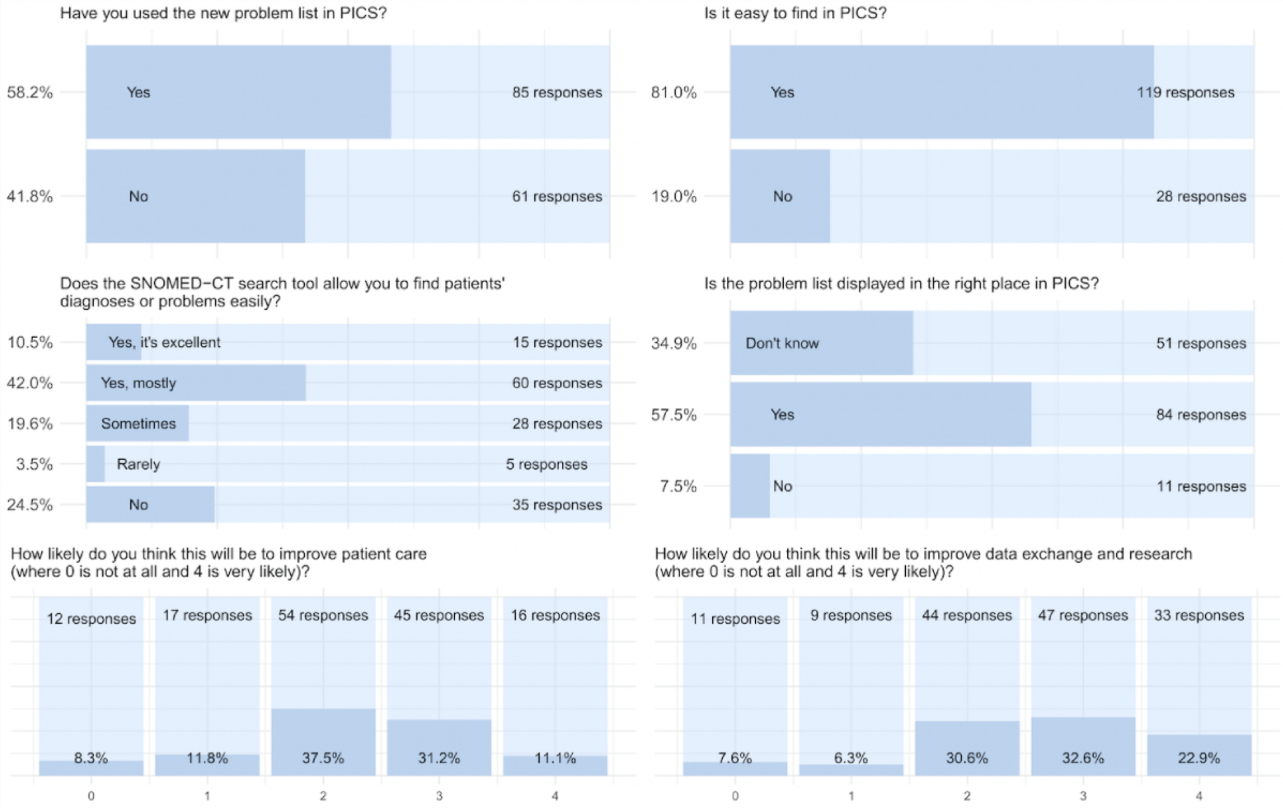


Most study participants thought the tool was easy to find (Figure 3) and that diagnoses could be searched for effectively. Most participants also thought that the new problem list was displayed in the right places in the EHR. There were several examples of user feedback requesting changes to the UI; dates did not appear until final commitment to the record that confused some users, and users suggested the addition of abbreviations in the search box.

Most of the required diagnoses seemed to be available, although ophthalmology was reported to be sparse, and some specific diagnoses were not found. Some doctors asked how to add diagnoses to the SNOMED-CT itself. Overall, however, users more frequently commented that there were too many diagnoses rather than too few, and they did not understand that synonyms mapped back to the same codes (Textbox 1).

More fundamentally, surgeons commented that *end* dates did not make sense for operations. They felt that even though the *problem* had ended, surgical history should persist on diagnostic lists. Specific dates for some diagnoses were also thought to be irrelevant. Textbox 1 below is a compilation of free text data from the combined methodology, detailing respondents' perspectives around the core areas of utility, diagnosis search, and dating diagnoses.

Figure 3. Questionnaire summary for user feedback. PICS: Prescribing Information and Communication System; SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms.



Textbox 1. Usability of new tool.

The tool was easy to use

- “More abbreviations would make search engine more efficient” [user]
- “Agree it’s unlikely that we will find anything better for our speciality” [consultant in cardiac surgery]

Users could find the right diagnoses but some additions were needed

- “Even a small number of extra terms for the VAD [ventricular assist device] complication would be very useful and of course there are new devices coming out that aren't covered” [doctor]
- “I’ve tried quite a few orthopaedic trauma diagnoses and it seems comprehensive” [orthopedic surgeon]
- “The search option for some of the diagnoses may need improvement. For example, when I type TB, several tuberculosis- related diagnoses are available. Miliary TB however only appears if I type in Miliary. Similarly some diagnoses seem to have been only tagged to human immunodeficiency virus rather than HIV (eg, Lymphoma with HIV)” [HIV consultant]
- “Only one condition (glaucoma) in ophthalmology section” [ophthalmology doctor]
- “There are still some diagnoses missing from the list, for example Non-specific interstitial pneumonia. Is there any facility for us to add diagnoses to the list?” [HIV consultant]

Large numbers of diagnoses to choose from

- “So many to choose from when updating problem list” [doctor]
- “Multiple names for the same thing an issue” [anesthetic consultant]
- “It is quite time consuming to enter data and sometimes it is hard to find the exact diagnosis needed” [user]
- “Many of the diagnoses are repeated so it is not clear which to put down. This may make auditing tricky” [user]
- “Agree. Better than ICD10. There are often multiple options for the same disease -are they automatically clustered together for analysis or would you need to do this manually? CT concept seems to be mainly irrelevant stuff” [hematology consultant]

Modifications of dates needed

- “One of the slight frustrations is that when you enter a procedure, for example repair of extensor tendon of hand, the start date has to be different from the end date” [hand surgeon]
- “Some diagnosis patients tend to remember the date like MI [myocardial infarction]. Others such as hypertension are hard to date exactly” [doctor]
- “It seems a little clunky; perhaps I am using it wrong. When I try to enter comorbidities, the SNOMED-CT diagnosis disappears I try to put in the start date” [user]
- “I commonly find a correct problem with search and when I click it just disappears rather than gets added. I am not sure what I am doing wrong but the interface may need attention” [user]

Engagement

The issue of clinicians not entering all relevant diagnoses was raised many times, but as these are persistent data, users thought these may accumulate over time. Doctors felt motivated to enter data but were restricted by time constraints and demotivated by the perception that other doctors and other staff members were not doing a share of the work. Nurses, patients, and coders were all expected to be potentially powerful contributors to the problem list. Finally, education was raised as important in user engagement.

Data Quality

Users expressed that the data needed to be correct and updated. Ensuring that all users entered correct and complete lists was thought to be potentially problematic, and deletion of irrelevant or incorrect items from the lists was raised as a concern. Typical feedback about the quality of data entry and barriers to the affective gathering of data are illustrated in Textbox 2.

Textbox 2. Data entry and data quality.

Data entry improved by the new tool

- "Once all the data was in there updating it wouldn't take too long" [hand surgeon]
- "In clinic now I feel inclined to do things properly" [oncologist]

Barriers to data entry

- "Ideally I should do this with all my patients but there is simply not enough time (or perhaps incentive) to do this for all the OPD [outpatient department] patients I see. I do this for my day case patients currently" [user]
- "Too much effort needs to go into entering data, a huge amount of time went into this in a previous Trust but responsibility and effort decrease over time and I'm worried that this may happen here" [nephrologist]
- "Used it today and the PPM [permenant pacemaker] hadn't been entered by the team...still relies on the entry! I hope that my entering the PPM will stay on prescribing information and communication system for all future episodes" [user]
- "The challenge will be making sure that people use it so that it becomes habit" [user]
- "I think the main challenge will be to get enough doctors, of all grades and specialties, to use it to populate the list accurately. The second challenge is how we will know."
- "I tend to put some stuff in. Not everyone is doing this. In radiology they are not adding" [vascular surgeon]

Quality of data entry

- "If I ask a junior doctor to add problems in the problem list they will miss lots of stuff. How can we ensure quality" [ICU consultant]
- "I think we need to take the history properly" [vascular surgeon]
- "We need to make sure things are deleted as well as added" [neurologist]
- "It will only improve things if the user is taken to it and encouraged to review and increase content" [user]
- "Data accuracy is an issue – if I'm coding other people's diagnoses I'd probably get this wrong" [doctor]

Involvement of patients and the multidisciplinary team

- "If engage with it will be useful, but often find there is only half the patients' past medical history is there. Is there scope that this could be update by pharmacists so when they are taking meds rec they ensure that it is updated on the system as well" [user]
- "The main problem I see with this is like before whether doctors are inputting 'problem list'. I know that when patients come to [the] angio[graphy suite] it is unlikely that any one will add anything. Can the coders do something with this??" [user]
- "Nurses are good at taking a history – they should put this in there" [ICU consultant]
- "Why don't we give the patients a copy and ask them 'is this accurate'?" [upper gastrointestinal surgeon]

Improving Patient Care

In the questionnaire, users were ambivalent about the tool positively impacting patient care (Figure 3). However, in focus groups, clinicians thought that the problem list positively contributed to patient care and could have wide implications for safety rules. It was also recognized as improving communication.

There were concerns about lists becoming too long and that loss of human sorting of lists could lead to a loss of nuance. Many doctors use clinical letters to present a list of patient problems, which are subtly changed in each specialty with a different ordering or grouping of problems. Doctors may also exclude problems that are less relevant to their specialty or add details that are not understood (or transcribed) by other specialists. Building relevant patient lists that cater for all is probably the most challenging aspect of building relevant electronic lists. The current situation in which some information is missed or excluded by doctors may be dangerous, but an alternative that overloads clinicians with information is also not ideal. Pragmatic solutions, where "one size fits all," will require extensive clinician collaborations and consensus. Detailed feedback from the various methodologies that exemplify these points is presented in Textbox 3.

Textbox 3. Impact of the tool on patient care.

Positive impact on patient care

- Having a problem list improves patient care and safety
  - "In general surgery we need to know about a lot of other specialties. If these were summarised in a problem list this would really help" [upper gastrointestinal surgeon]
  - "From a clinical point of view, it ought to be a good thing certainly because just looking at what is in ICD 10 which is what we were using, there is an awful lot in there that really is not clinical classification suitable for individual patients whereas in SNOMED that should be, they should all be suitable for individual patients for at least the disease and the implications that we are using should be. So yes, I think that it is probably worthwhile" [programmer]
  - "Yes useful if its filled out for example, steroids – if these are missed but are on the problem list this would be very useful" [upper gastrointestinal surgeon]
  - "If you get wrong things in the record, like 'heart attack' this may prevent chemotherapy in the future" [consultant oncologist]
- Correct coding has implications for safety rules
  - "Once this replaces paper notes as a single instance of truth, it will be very valuable and can be used to drive decision support and indication based prescribing or decision support" [doctor]
- Improvement in communication
  - "I think having the problem list is very helpful, but I think a clerking page on prescribing information and communication system for the admission would be very helpful especially with patients who have been in for a while" [user]
  - "For people in highly specialised specialities they don't cross over into other peoples' fields. In general surgery we need to know about a lot of other specialities. If these were summarised in a problem list this would really help" [upper gastrointestinal surgeon]
  - "For ED/pre-assessment - this is where it's so important to document this" [neurologist]
  - "If you saw a new patient with 30-40 letters you would be missing very important information that you would really want to know" [upper gastrointestinal surgeon]
  - "I think it's important because when people go away and come back lots of important information is there" [HIV consultant]
  - "The problem is sometimes not too much information but too little" [hand surgeon]

Negative impact on patient care

- Lists may get too long
  - "Terms are so detailed that a patient could end up with several codes for the same diagnosis, and this will make the problem list overly complicated and the letters cluttered" [consultant geriatrician]
  - "It's lovely to have this list in the medical record but I'm not sure a 30 year list is all relevant and can lead to confusion, failure to triage and fatigue" [consultant nephrologist]
- Human nuance may be lost
  - "For me in cardiology I do detailed problem lists. There is not sufficient detail in this tool" [cardiologist]
  - "It's difficult you can't replicate the problem list [that a human created] at the top of the clinical letter without there being too much information" [consultant nephrologist]
  - "Yes we need more detail on the lists we need the notes 'this patient requires steroids' – is this part of SNOMED?" [upper gastrointestinal surgeon]
  - "I don't use this at all I only use my letters. I only summarise for my specialty" [liver medicine]

Use in Research or Audit

The principal concern regarding the use of codes for research was the overlapping of codes and lack of granularity. However, users also noted that the tool was already being used in audits. Many clinicians have recognized the potential for research and expected that digital development could contribute extensively.

Use in Data Sharing

In the questionnaire, users generally thought that the tool would improve data exchange (Figure 3). It was generally agreed that more standard and detailed coding made data sharing easier and that this was desirable for patient care. Users were interested in sharing the data with general practitioners and other hospitals. Clinicians were also interested in the idea that data could be pulled into the hospital system for acute and outpatient care. User feedback is presented in Textbox 4. Of note is that users remained concerned about the number of codes available, although there is some recognition that these can be mapped to each other or grouped for the purposes of research and clinical care.

Standards and Payment

Data sharing was also discussed in relation to payments and uploads to national registries. Programmers were the only group that discussed the NHS standards (Textbox 4).

Textbox 4. Use in research, audit, data sharing and payment.

The tool is useful in research

- “The only difficultly I find with the patient group I see is that there tend to be multiple overlapping codes which could be used and unless easy consensus is reached across the organisation you may get lots of codes of one condition. We used to be able to see the 'code' behind the 'description'. This would still be useful, particularly for future informatics searches based on code” [doctor]
- “Assuming it is used correctly it will be helpful (more for research than for clinical care) but it can be tricky to know what to code and how to code - there are often a lot of similar diagnoses on SNOMED which can make it confusing” [user]
- “I used to use this list in order to code patients with a view to doing audit on out patients - however when we have requested the list of patients with a certain diagnosis for audit purposes we are told that we can only search for in patients this way. As a result people have stopped coding outpatient episodes as they can no longer see the relevance. It would be good to be able to generate a list of certain outpatient diagnoses on request so that we can improve patient care” [user]
- “More granularity for research is likely to be needed but with less overlap” [user]
- “Used by NIV [non-invasive ventilation] physio team to record patient as receiving acute NIV data. This should allow data to be collated for the COPD [chronic obstructive pulmonary disease] national audit. I would like to know if it can be used locally to pull data for NIV service audit at QEHB [Queen Elizabeth Hospital Birmingham]” [user]
- “If the information was structured this would be a valuable source of data for any future admissions, audit and research” [acute medicine consultant]
- “The more coded information we have the more useful it will be for the future – information in the system means computers will be able to pull this and write queries, your own brain can’t do this when there is too much information” [consultant nephrologist]
- “In the old days there was no information, computing has revolutionised this but there is now irrelevant data overload” [HIV consultant]

Data sharing

- In general
  - “Well certainly in terms of sharing, the codes that we are using should have, it is difficult to say agreed semantics but there should be less disagreement over what the semantics of them are. Certainly within the use of local codes that were being used maybe in terms of ICD10 codes, I can think of some ICD10 codes that, depending on whether you interpret the description of the code as being about a single patient or about a patient population which are completely different things because of the use of the word and where it kind of gets turned around when it is a bucket to really mean for a single patient” [programmer]
  - “It is something which seems to be highly detailed so presumably is going to make it easier for sharing data between systems which is, I guess it has to be an aim in the long run for most organisations” [programmer]
  - “So we must facilitate data sharing and this improves patient care doesn’t it?” [hands surgeon]
- With general practitioners
  - “Discharge codes should go into letters and be sent to the GP” [fixed point data]
  - “Also by definition it would be nice if that were put on the discharge letter (not that I usually see what is actually on them), but I guess that by “finishing” the code, it isn’t put anywhere?” [consultant in hand and plastic surgery]
  - “GP Systems are really clunky, this needs to be transcribed, if this was intelligent it would help GP colleagues a lot” [HIV consultant]
- With other health care providers
  - “So I think I understand how this feeds into data sharing. Improving clinical care – when I go to the Women’s Hospital I am completely blind, if this could join things together that would be great” [consultant nephrologist]
  - “So the Ambulance Service have very little understanding about what has gone on with the patient, and end of life. We don’t know the post mortem data” [hands surgeon]
- To pull data into the organization
  - “Prescribing information and communication system needs to be able to pull things in from other systems” [plastics and hands surgeon]
  - “We need the clerking in emergency department” [vascular surgeon]
  - “On the post take ward round we are checking letters, if it was there it would be much quicker” [upper gastrointestinal surgeon]
  - “Confused elderly patient, this is important when you are trying to get GP information – it will be fantastic” [ICU consultant]
  - “There are lots of patients who have had other things in other hospitals, but no proper data sharing exists” [consultant cardiologist]

Standards and payment

- “Plus it is mandated by the National Health Service” [programmer]
- “Well, yes it is something that has been mandated” [programmer]

- "If you do things incorrectly the clinical commissioning group will take money off you" [upper gastrointestinal surgeon]
- "We have been caught out by poor coding" [ICU consultant]
- "In Transcatheter Aortic Valve Implantation the list is strange for payment, on prescribing information and communication system we list these in the right order so that the coders get it right – this is very difficult and specific to do – it's all manual. It's now a game – there are points" [consultant cardiologist]
- "One of the issues is that codes may be hidden in notes and you might not get your money" [upper gastrointestinal surgeon]

The Importance of Various Themes to Users

To understand which aspects of the tool were most important to users, feedback from the focus groups, interviews, emails, and free comment text in the questionnaires were marked to understand where comments or conversation had been based on particular themes, and analyzed for time spent (Figure 1). In the focus groups and interviews, this was measured as the actual time spent in discussing an item; in the questionnaire, and email feedback, this was the length of free text comments written by respondents.

Overall, users were most concerned about the ease of use of the tool and its direct use. They were then equally concerned with patient care, data sharing, and further development. All users talked about all the themes except for programmers who, as expected, did not talk about usefulness, as they were not users. Questionnaire comments focused on how easy it was to use the tool, with some focus on usefulness and patient care.

Interestingly, the 2 focus groups spent time on different themes, despite having been shown the same facilitation tools. One group discussed data sharing extensively and the other future developments. Neither focus group discussed patient care in depth, which may be because of the seniority of the doctors in these groups, for whom the importance of good patient care is already assumed to be high.

Validity and Reliability of Data

Data were triangulated from fixed-point data sets, surveys, interviews, and focus groups with the data collected on actual use statistics. Survey data were collected until saturation or redundancy was reached. To reduce nonresponder bias, the survey was constructed carefully using proven techniques to ensure experience, opinion, and emotional response without deploying leading questions or resulting in dead-end answers [16]. Reminders were deployed to increase response rates, and the survey was deliberately short so that it was easy to complete. The survey was digital to ensure the ease of returning completed responses. The principal investigator has a relationship with many clinicians across the organization, and this was exploited to increase response rates. Member checks ensured internal validity by soliciting feedback on preliminary findings from participants to check the correct interpretation [23].

Discussion

Principal Findings

This study demonstrates the successful conversion of a hospital electronic record to SNOMED-CT with high clinician acceptability, which forms the basis of data sharing, innovative use of data, and availability of coded clinical information for research.

SNOMED-CT as an ontology has been developed over several decades [24] and represents a detailed and multifaceted database that can be used to accurately describe patients' conditions, medical and surgical history, and current problems. The ontology is divided into several levels, allowing both specific detail and broader general categories, and these are linked together in a web structure [7].

There are several coding structures currently in use in the English NHS, including the *ICD-10* and Office of Population Censuses and Surveys Version 4, which are the current methods of payment for hospitals [13,25]. In general, these methods do not provide the depth or detail required to describe patient conditions. This lack of detail leads clinicians to look for better ways to encode clinical information [26-28].

Data Exchange

Clinicians and patients increasingly demand safe exchange of clinical data to optimize clinical care [29]. This can be leveraged in secondary data collection for research to safeguard future population health [30]. The interoperability of health systems, and in particular, moving data between systems, has been discussed extensively but has not been realized to date [4]. One of the major hurdles is nonstandard data collection [31]. It is difficult to transfer data where mapping from one coding system to another is needed, especially when the mapping is from one generalized diagnosis to many more detailed ones. In an ideal situation, the primary data sources would collect data in a single standard coding structure, facilitating easy data exchange [14]. Clinicians in our study recognized the need for data exchange and welcomed a tool that actively facilitates sharing.

The national strategy in England is clearly set out by NHS Digital with aspirations for all aspects of the English NHS to use SNOMED-CT from 2020. Conversion is under way in primary care and secondary care, acute care, mental health, community systems, dentistry and other systems used in direct patient care must use SNOMED-CT as the clinical terminology, before April 1, 2020 [7]. The conversion of commercial systems has been slow [32], but continues to increase.

Previous Transitions

There are no published examples of SNOMED-CT conversion in a complex hospital EHR that supports CDS, but there are several case studies of smaller successful and unsuccessful implementations of SNOMED-CT in EHRs. Lee et al [33] studied 13 implementations, 5 of which were hospital-wide, and 4 of these involved diagnostic lists. Clinicians found that SNOMED-CT did not always include all diagnoses that they required. In a small Portuguese study [34], user acceptance was high, and the adoption of SNOMED-CT enabled system interoperability. More recently, in Wales, the successful development of a neurology database using SNOMED-CT has influenced direct patient care, allowing accurate web-based reporting and informing clinical research [35].

Currently, there is very little published literature and there are no peer-reviewed papers describing the impact of primary or whole system secondary care transition to SNOMED-CT in England. Many of these implementations also reported benefits not directly related to SNOMED-CT, but rather to transition from paper to digital recording [33]. The principal trends from previous implementations were reflected in our study—the hierarchy was challenging to use, some concepts were ambiguous, and there was some syntactic inconsistency. The combining of multiple terms used in SNOMED-CT was also challenging and consequently created difficulties in data retrieval [33-35]. Success trends could also be found; implementations were successful as the UI was simple and immediate value was demonstrated to clinicians. Synonyms allowed clinicians to easily find terms and reuse the data.

Interface

Clinicians not only require a coding system that encompasses all the diagnoses and procedures that affect patients; they also want a seamless UI for data entry [33]. By providing a more complex and detailed coding structure, an unintended consequence may be that the data entry application is so complicated that clinicians disengage, resulting in a paradoxical decrease in acquired data [36,37]. The SNOMED-CT database is complex and needs to be presented in a way that makes sense to clinicians [38,39]. The interface must be intuitive and not overwhelming, and the search engine must return specific terms in a helpful order [40]. Data exchange must be considered at inception [4], and mapping back to other coding structures must be possible, for example, to national data sets for reporting [41] and to the *ICD-10* for payment [42]. Of note, where enriched data are collected by clinicians, mapping back to the *ICD-10* for payment can be effective in increasing hospital income [43]. Reflecting these points, we found in our study that clinicians accepted the use of SNOMED-CT, but care was needed in the presentation and utility of this vast database in clinical EHR. In general, clinicians report success in using the tool, finding the depth of the coding useful and applicable [33]. Clinicians, particularly doctors, are in general interested primarily in providing excellent patient care, and drivers for successful projects are therefore efficient patient safety [26,44] and research for population health [45].

The organization has a long history of user-led design, and the strength of the EHR is that it is "designed for clinicians by clinicians." Increasingly successful digital health systems are reported to have a human-based design at their core [4,46-48]. Bringing developers and users together is reported to increase usability and value [49,50], forming the basis of user-based co-design [51].

Embedded in Complex EHR With CDS

In addition to the literature on SNOMED-CT and its implementation, the design of the EHR is crucial to the success of the project. Independent of the coding structure behind it, the EHR must be highly usable to engage clinicians, and must be intuitive and helpful or risks disengagement [3]. Data sharing for direct health care, or for population research to safeguard future generations, is only possible if data are entered into the record in the first place. Therefore, interface design is crucial [40,46], more so in implementations involving extremely complex coding systems such as SNOMED-CT. In addition to design, standards that are universally adopted (or enforced) are required to allow data sharing and interoperability of systems to work [5,8,52]. Sophisticated CDS or artificial intelligence algorithms cannot be leveraged if clinical records are devoid of data.

Beyond the transition of EHR to SNOMED-CT, there is limited literature on its subsequent use in CDS and artificial intelligence. Previous studies in this area have briefly discussed the use of SNOMED-CT in CDS [53,54], but have not explored the complex mapping that is needed to link existing rules to multiple codes as a result of the one-to-many mapping. SNOMED-CT has the potential to greatly enrich medical records [34] and therefore enhance the safety and quality of care via CDS [53]. The detailed diagnostic codes in SNOMED-CT may also allow the development of more precise CDS in the future. The extensive work needed to map all the codes in an existing system was completed, although it took time.

In addition, the transition of complex EHR to SNOMED-CT must consider some unique problems. SNOMED-CT is compositional, negating the need for a large number of specific terms. Therefore "left pneumothorax as a complication of a chest drain" does not exist in SNOMED-CT; terms must be combined to reach concepts that are applicable to patients. Compositional systems allow greater reuse of data without the need for human intervention for the interpretation of the categories [55]. The challenge is to allow clinicians to choose from the richness of the database, including combining of concepts, while safeguarding the CDS rules so that any concept related to the conditions in the rule is fired by all of the relevant SNOMED-CT codes that clinicians choose.

Health Ecosystem

This project took place in a large secondary and tertiary care hospital ecosystem consisting of 4 hospitals and community services. This ecosystem has nascent links to primary care records via a project that does not yet exchange data but uses a "look up and leave" data sharing platform [56]. Beyond the institution itself, the EHR has been acquired by 2 other hospitals in the West Midlands [57,58]. Similar ecosystems in the United Kingdom are now increasingly implementing EHR [59,60], and in North America, many have already done so [61]. There are limited examples of acute care ecosystems that implement a change in coding, but a few exist [34,62]. None of these studies, however, achieved recoding in a complex EHR that supported complex CDS rules, and none of these examples reported on data exchange after the coding change.

Of note, although this study is restricted to the acute hospital setting, the whole of the NHS is influenced by projects of this kind. By converting coding central to one of the largest acute trusts in the country whose influence through its software extends beyond the institution itself, this project contributes to the facilitation of national data exchange.

Limitations and Delimitations

The limitations of this study are largely related to the time-limited nature of the study; thus, a pragmatic approach had to be taken for the number of respondents for the survey and the number of focus groups that could be undertaken. The study protocols aimed to continue the survey until saturation was reached and to reduce bias from nonresponders. However, this was not entirely possible, and the study therefore needed to focus on deep analysis of smaller data sets rather than unrealistic expectations of analyzing extensive data sets.

Limitations beyond the control of the investigator were the truthfulness of the responders, the attendance of participants to invited groups, and consent for research inclusion.

The research was only able to study the response to the new tool for the first 4 months after its launch. This may not have been representative, as novel software may be used extensively initially with a reduction in use over time, or conversely, may increase as more clinicians become committed to data sharing. Pre-existing data sets were brief notes, capturing requirements for software build and modification. Although imperfect, these were valuable resources for comparing the attitudes and requirements of both stakeholders before and after the launch of the tool. Finally, the study could not study the repetitive loops of development, and it is likely that iterative changes to the tool are needed.

Lessons Learned

The following lessons were learned during the development and implementation process:

1. Clinicians will accept the task of adding SNOMED-CT–coded diagnoses into an electronic record provided the UI is intuitive, but they will disengage if the process is too complex.
2. The underlying EHR must be highly usable to engage clinicians.
3. Bringing developers and users together increases usability and value by engaging in user-based co-design.
4. Clinicians are driven by patient safety, data exchange across the health system, and research for population health.
5. Sophisticated and CDS or artificial intelligence algorithms cannot be leveraged without this data collection.

Conclusions

This study demonstrates the successful conversion of a hospital electronic record to the updated standardized coding of diagnoses, with high clinician acceptability. The project began with clinician demand, followed by the building of a clinically led tool and extensive examination of its success and requirement for iteration. Starting in one institution, this has implications for the entire NHS. This tool is important because it supports national aspirations for enhancing patient care through CDS and allows for data sharing, innovation, and research. The project is a blueprint for capturing the data that form the bedrock for driving artificial intelligence and full digitalization of health care, and it is therefore a benchmark for transformation and innovation, influencing the aspiration to safeguard health for future generations.

Acknowledgments

Many thanks to all the University Hospitals Birmingham staff in the digital teams and clinicians who contributed to the success of the project. The National Health Service Digital Academy delivered by Imperial College London, University of Edinburgh, and Harvard Medical School supported the analysis of the project. The work was funded by University Hospitals Birmingham.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Fixed documents and excerpts from specification documents (user stories).
[DOCX File , 20 KB - medinform_v9i11e29532_app1.docx]

Multimedia Appendix 2
Survey of clinicians regarding acceptance of the problem list tool in the prescribing information and communication system.
[DOCX File , 14 KB - medinform_v9i11e29532_app2.docx]

Multimedia Appendix 3
Interview questions for developers working on the project.
[DOCX File , 14 KB - medinform_v9i11e29532_app3.docx]

References

1. Benson T. Why general practitioners use computers and hospital doctors do not--part 1: incentives. Br Med J 2002 Nov 09;325(7372):1086-1089 [FREE Full text] [doi: 10.1136/bmj.325.7372.1086] [Medline: 12424171]
2. Torres YR, Huang J, Mihlstin M, Juzych MS, Kromrei H, Hwang FS. The effect of electronic health record software design on resident documentation and compliance with evidence-based medicine. PLoS One 2017 Sep 21;12(9):e0185052 [FREE Full text] [doi: 10.1371/journal.pone.0185052] [Medline: 28934326]
3. Gawande A. Why doctors hate their computers. The New Yorker. 2018. URL: https://www.newyorker.com/magazine/2018/11/12/why-doctors-hate-their-computers [accessed 2021-09-07]
4. Miller H, Johns L. Interoperability of electronic health records: a physician-driven redesign. Manag Care 2018 Jan;27(1):37-40. [Medline: 29369771]
5. Marcheschi P. Relevance of ehealth standards for big data interoperability in radiology and beyond. Radiol Med 2017 Jun;122(6):437-443. [doi: 10.1007/s11547-016-0691-9] [Medline: 27815798]
6. 5 new value pathways fueling the big data revolution in healthcare. McKinsey & Company. URL: https://getreferralmd.com/2013/04/5-new-value-pathways-fueling-the-big-data-revolution-in-healthcare/, [accessed 2021-09-07]
7. SNOMED CT. NHS Digital. URL: https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct [accessed 2021-09-07]
8. Standards and collections. NHS Digital. URL: https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections [accessed 2021-09-07]
9. University Hospitals Birmingham (UHB). 2021. URL: https://www.uhb.nhs.uk/about-us.htm [accessed 2021-09-07]
10. Global digital exemplars. NHS England. URL: https://www.england.nhs.uk/digitaltechnology/info-revolution/exemplars/ [accessed 2021-09-07]
11. Birmingham and Solihull. NHS England. URL: https://www.england.nhs.uk/integratedcare/stps/view-stps/birmingham-and-solihull/, [accessed 2021-09-07]
12. Birmingham systems prescribing information and communications system. University Hospitals Birmingham NHS Foundation Trust. URL: https://www.uhb.nhs.uk/birmingham-systems-pics.htm [accessed 2021-09-07]
13. International statistical classification of diseases and related health problems 10th revision. World Health Organisation. 2016. URL: https://icd.who.int/browse10/2016/en [accessed 2021-09-07]
14. Benson T, Grieve G. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR. Cham: Springer; Jun 23, 2016.
15. Typeform. 2020. URL: https://www.typeform.com/ [accessed 2021-09-07]
16. Patton MQ. Qualitative Research and Evaluation Methods. Thousand Oaks, California, United States: SAGE Publications; 2014.
17. Likert R. A technique for the measurement of attitudes. Archives of Psychology. 1932. URL: https://legacy.voteview.com/pdf/Likert_1932.pdf [accessed 2021-09-07]
18. Determine survey length and number of questions. Zarca Interactive. 2019. URL: https://www.zarca.com/Online-Survey-Resource/Survey-Best-Practices/before-the-survey-process-begins/determine-survey-length.html [accessed 2021-09-24]
19. Merriam S, Tisdell EJ. Qualitative Research: A Guide to Design and Implementation, 4th Edition. San Francisco: Jossey-Bass; 2015.
20. Denzin NK, Lincoln YS, editors. The Sage Handbook of Qualitative Research, 3rd ed. Thousand Oaks: Sage Publications; 2005.
21. Hennink MM. Focus Group Discussions. New York, NY: Oxford University Press; 2014.
22. MacFadden P. The Balancing Test. Boston College Law Review. Volume. 29. 1988. URL: https://lawdigitalcommons.bc.edu/cgi/viewcontent.cgi?article=1868&context=bclr [accessed 2021-09-07]
23. Maxwell J. Qualitative Research Design: An Interactive Approach, 3rd Edition. Thousand Oaks: SAGE Publications; 2013.
24. Cornet R, de Keizer N. Forty years of SNOMED: a literature review. BMC Med Inform Decis Mak 2008 Oct 27;8 Suppl 1:S2 [FREE Full text] [doi: 10.1186/1472-6947-8-S1-S2] [Medline: 19007439]
25. NHS Classifications OPCS-4. NHS Digital. URL: https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/10 [accessed 2021-09-07]
26. Elevitch FR. SNOMED CT: electronic health record enhances anesthesia patient safety. AANA J 2005 Oct;73(5):361-366. [Medline: 16261852]
27. Nyström M, Vikström A, Nilsson GH, Ahlfeldt H, Orman H. Enriching a primary health care version of ICD-10 using SNOMED CT mapping. J Biomed Semantics 2010 Jun 17;1(1):7 [FREE Full text] [doi: 10.1186/2041-1480-1-7] [Medline: 20618919]
28. Spencer A, Horridge K, Downs D. Empowering clinical data collection at the point of care. Arch Dis Child 2015 Sep;100(9):815-817. [doi: 10.1136/archdischild-2014-307972] [Medline: 26153504]
29. Cohen JK. Who do patients want to share health data with? Beckers Hospital Review. URL: https://www.beckershospitalreview.com/data-analytics-precision-medicine/who-do-patients-want-to-share-health-data-with.html [accessed 2021-09-07]

30. Balio CP, Apathy NC, Danek RL. Health information technology and accountable care organizations: a systematic review and future directions. EGEMS (Wash DC) 2019 Jul 08;7(1):24 [FREE Full text] [doi: 10.5334/egems.261] [Medline: 31328131]
31. Standardising clinical information. Public Health Scotland. URL: https://www.isdscotland.org/Products-and-Services/Terminology-Services/Information-for-Clinicians/Standardising-Information/index.asp [accessed 2021-09-07]
32. Giannangelo K, Fenton SH. SNOMED CT survey: an assessment of implementation in EMR/EHR applications. Perspect Health Inf Manag 2008 May 20;5:7 [FREE Full text] [Medline: 18509501]
33. Lee D, Cornet R, Lau F, de KN. A survey of SNOMED CT implementations. J Biomed Inform 2013 Feb;46(1):87-96 [FREE Full text] [doi: 10.1016/j.jbi.2012.09.006] [Medline: 23041717]
34. Duarte J, Castro S, Santos M, Abelha A, Machado J. Improving quality of electronic health records with SNOMED. Procedia Technol 2014;16:1342-1350. [doi: 10.1016/j.protcy.2014.10.151]
35. Wardle M, Spencer A. Implementation of SNOMED CT in an online clinical database. Future Healthc J 2017 Jun;4(2):126-130 [FREE Full text] [doi: 10.7861/futurehosp.4-2-126] [Medline: 31098449]
36. Zheng K, Padman R, Johnson MP, Diamond HS. An interface-driven analysis of user interactions with an electronic health records system. J Am Med Inform Assoc 2009;16(2):228-237 [FREE Full text] [doi: 10.1197/jamia.M2852] [Medline: 19074301]
37. Ratwani RM, Hettinger AZ, Fairbanks RJ. Barriers to comparing the usability of electronic health records. J Am Med Inform Assoc 2017 Apr 01;24(e1):191-193 [FREE Full text] [doi: 10.1093/jamia/ocw117] [Medline: 27572813]
38. López-García P, Schulz S. Can SNOMED CT be squeezed without losing its shape? J Biomed Semantics 2016 Sep 21;7(1):56 [FREE Full text] [doi: 10.1186/s13326-016-0101-1] [Medline: 27655655]
39. Bakhshi-Raiez F, de Keizer NF, Cornet R, Dorrepaal M, Dongelmans D, Jaspers MW. A usability evaluation of a SNOMED CT based compositional interface terminology for intensive care. Int J Med Inform 2012 May;81(5):351-362. [doi: 10.1016/j.ijmedinf.2011.09.010] [Medline: 22030036]
40. 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] [Medline: 29126196]
41. Intensive Care National Research and Audit Centre (ICNARC). 2019. URL: https://www.icnarc.org/ [accessed 2021-09-07]
42. Clinical classifications. NHS Digital. URL: https://digital.nhs.uk/services/terminology-and-classifications/clinical-classifications [accessed 2021-09-07]
43. Fung KW, Xu J, Rosenbloom ST, Campbell JR. Using SNOMED CT-encoded problems to improve ICD-10-CM coding-a randomized controlled experiment. Int J Med Inform 2019 Jun;126:19-25 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.03.002] [Medline: 31029260]
44. Sefton G, Lane S, Killen R, Black S, Lyon M, Ampah P, et al. Accuracy and efficiency of recording pediatric early warning scores using an electronic physiological surveillance system compared with traditional paper-based documentation. Comput Inform Nurs 2017 May;35(5):228-236 [FREE Full text] [doi: 10.1097/CIN.0000000000000305] [Medline: 27832032]
45. Unlocking the power of digital health for the population. Orcha. URL: https://www.orcha.co.uk/Review/114347/ [accessed 2021-09-07]
46. Beaudry J. Bringing human-centered design to healthcare. Design UVMMC. 2017 Jun 15. URL: https://medium.com/design-uvmmc/bringing-human-centered-design-to-healthcare-5d8ede2aee3b [accessed 2021-09-07]
47. Designing with people. Helen Hamlyn Centre for Design. URL: https://oecd-opsi.org/toolkits/designing-with-people/ [accessed 2021-09-07]
48. Lee J. Patient centred participatory design. Prescribe Design. URL: http://prescribedesign.com/portfolio/participatory-design-in-healthcare/ [accessed 2021-09-07]
49. Pagliari C. Design and evaluation in eHealth: challenges and implications for an interdisciplinary field. J Med Internet Res 2007;9(2):e15 [FREE Full text] [doi: 10.2196/jmir.9.2.e15] [Medline: 17537718]
50. Designing for usability. The University of Texas Health Science Center at Houston. URL: https://sbmi.uth.edu/nccd/ehrusability/design/ [accessed 2021-09-07]
51. Experience-based co-design toolkit. Point Of Care Foundation. URL: https://www.pointofcarefoundation.org.uk/resource/experience-based-co-design-ebcd-toolkit/step-by-step-guide/1-experience-based-co-design/ [accessed 2021-09-07]
52. 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 Feb 17;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
53. Al-Hablani B. The use of automated SNOMED ct clinical coding in clinical decision support systems for preventive care. Perspect Health Inf Manag 2017;14(Winter):1f [FREE Full text] [Medline: 28566995]
54. Martínez-Salvador B, Marcos M, Mañas A, Maldonado JA, Robles M. Using SNOMED CT expression constraints to bridge the gap between clinical decision-support systems and electronic health records. Stud Health Technol Inform 2016;228:504-508. [doi: 10.3233/978-1-61499-678-1-504] [Medline: 27577434]
55. Nyström M, Vikström A, Nilsson G, Orman H, Ahlfeldt H. Visualization of disease distribution with SNOMED CT and ICD-10. Stud Health Technol Inform 2010;160(Pt 2):1100-1103. [Medline: 20841854]

56. Your care connected. National Health Service. URL: https://midlandsyourcareconnected.nhs.uk/ [accessed 2021-09-07]
57. Stevens L. Birmingham children's goes live with paediatric e-prescribing. Digital Health. 2017. URL: https://www.digitalhealth.net/2017/05/birmingham-childrens-goes-live-with-pics-e-prescribing/ [accessed 2021-09-07]
58. Whitfield L. Royal orthopaedic first to pick PICS. Digital Health. 2015. URL: https://www.digitalhealth.net/2015/08/royal-orthopaedic-first-to-pick-pics/ [accessed 2021-09-07]
59. Mongain B. UK's countess of chester hospital signs 15-year EHR deal with Cerner. Healthcare IT News. 2018. URL: https://www.healthcareitnews.com/news/uks-countess-chester-hospital-signs-15-year-ehr-deal-cerner [accessed 2021-09-07]
60. Stevens L. Three years on from Cambridge's epic big bang go-live. Digital Health. 2017. URL: https://www.digitalhealth.net/2017/08/three-years-on-cambridge-epic/ [accessed 2021-09-07]
61. Monica K. Top 50 most popular hospital inpatient EHR systems in US. EHR Intelligence. 2017. URL: https://ehrintelligence.com/news/top-50-most-popular-hospital-inpatient-ehr-systems-in-us [accessed 2021-09-07]
62. Sedano FJ, Cuadrado M, Rebolledo EM, Clemente Y, Balazote P, Delgado A. Implementation of SNOMED CT to the medicines database of a general hospital. Stud Health Technol Inform 2009;148:123-130. [Medline: 19745242]

Abbreviations

AHP: Allied Health Professionals
CDS: clinical decision support
EHR: electronic health record
ICD-10: International Classification of Diseases, Tenth Revision
NHS: National Health Service
PICS: Prescribing Information and Communication System
SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms
UHB: University Hospitals Birmingham
UI: user interface

Edited by C Lovis; submitted 11.04.21; peer-reviewed by S Sabarguna; comments to author 05.06.21; revised version received 15.06.21; accepted 17.06.21; published 23.11.21.

Please cite as:
Pankhurst T, Evison F, Atia J, Gallier S, Coleman J, Ball S, McKee D, Ryan S, Black R
Introduction of Systematized Nomenclature of Medicine–Clinical Terms Coding Into an Electronic Health Record and Evaluation of its Impact: Qualitative and Quantitative Study
JMIR Med Inform 2021;9(11):e29532
URL: https://medinform.jmir.org/2021/11/e29532
doi:10.2196/29532
PMID:34817387

Original Paper

Health Professionals’ Perspectives on Electronic Medical Record Infusion and Individual Performance: Model Development and Questionnaire Survey Study

Rai-Fu Chen[1*], PhD; Ju-Ling Hsiao[2*], PhD

[1]Department of Information Management, Chia-Nan University of Pharmacy and Science, Tainan City, Taiwan

[2]Department of Pharmacy, Chia-Nan University of Pharmacy and Science, Tainan City, Taiwan

[*]all authors contributed equally

Corresponding Author:
Ju-Ling Hsiao, PhD
Department of Pharmacy
Chia-Nan University of Pharmacy and Science
Number 60, Sec 1, Erren Road
Rende District
Tainan City, 71710
Taiwan
Phone: 886 6 2664911 ext 5106
Email: mayo5012@gmail.com

Abstract

Background: Electronic medical records (EMRs) are integrated information sources generated by health care professionals (HCPs) from various health care information systems. EMRs play crucial roles in improving the quality of care and medical decision-making and in facilitating cross-hospital health information exchange. Although many hospitals have invested considerable resources and efforts to develop EMRs for several years, the factors affecting the long-term success of EMRs, particularly in the EMR infusion stage, remain unclear.

Objective: The aim of this study was to investigate the effects of technology, user, and task characteristics on EMR infusion to determine the factors that largely affect EMR infusion. In addition, we examined the effect of EMR infusion on individual HCP performance.

Methods: A questionnaire survey was used to collect data from HCPs with >6 months experience in using EMRs in a Taiwanese teaching hospital. A total of 316 questionnaires were distributed and 211 complete copies were returned, yielding a valid response rate of 66.8%. The collected data were further analyzed using WarpPLS 5.0.

Results: EMR infusion (R^2=0.771) was mainly affected by user habits (β=.411), portability (β=.217), personal innovativeness (β=.198), technostress (β=.169), and time criticality (β=.168), and individual performance (R^2=0.541) was affected by EMR infusion (β=.735). This finding indicated that user (habit, personal innovativeness, and technostress), technology (portability), and task (mobility and time criticality) characteristics have major effects on EMR infusion. Furthermore, the results indicated that EMR infusion positively affects individual performance.

Conclusions: The factors identified in this study can extend information systems infusion theory and provide useful insights for the further improvement of EMR development in hospitals and by the government, specifically in its infusion stage. In addition, the developed instrument can be used as an assessment tool to identify the key factors for EMR infusion, and to evaluate the extent of EMR infusion and the individual performance of hospitals that have implemented EMR systems. Moreover, the results can help governments to understand the urgent needs of hospitals in implementing EMR systems, provide sufficient resources and support to improve the incentives of EMR development, and develop adequate EMR policies for the meaningful use of electronic health records among hospitals and clinics.

(JMIR Med Inform 2021;9(11):e32180) doi:10.2196/32180

KEYWORDS

health care professional; electronic medical records; IS infusion; individual performance; EHR; electronic health record; performance; perspective; information system; integration; decision-making; health information exchange; questionnaire

Introduction

Background

Electronic medical records (EMRs), as an important health information technology (HIT), have been developed to solve problems arising from the use of paper medical records, including the difficulty in searching for information, incompleteness of information, illegibility of handwriting, difficulty in management and storage, and inaccessibility [1]. EMRs are computerized medical information systems that collect, store, and display patient information generated by health care professionals (HCPs) from various health care information systems, including hospital information systems (HISs), picture archiving and communication systems, laboratory information systems, radiology information systems, and others [2]. EMR systems can be regarded as both electronic health record (EHR) systems and clinical information systems to provide clinical (including patient history and clinical notes, prescription management and patient demographics, and patient care management), communicational (visualization of results, communication with other institutions, and electronic transfers), and administrative (appointments scheduling and distance access, and billing and data security) functionalities [3]. Therefore, EMRs play a crucial role in improving the care quality, continuity, safety, efficiency, and decision-making in health care, and facilitate the cross-hospital exchange of health information [1,2,4-9].

Prior studies found positive correlations between information technology (IT)/information system (IS) utilization and individual and organizational performance [10-12]. With the wide-ranging use of HITs in the health care industry, understanding the usage behavior of HCPs is an important research topic for further HIT development. Underutilization of a system is often a major problem contributing to lack of complete infusion or integration of the implemented IS into employees' daily work or organizational processes after it is introduced [13,14]. Underutilization of a system has been considered as one of the key causes for a system not meeting the initial expectations in increasing productivity and yielding reasonable returns [15-17]. IS infusion, the final stage of the IS development process in Cooper and Zmud's [18] IS implementation model, is defined as using the system to its full potential in an extended, integrative, and emergent way. Organizations can fully leverage their investments in IS infusion because users voluntarily go beyond standardized system usage and exploit the system's full potential to improve their task performance [19,20]. Despite being recognized as critical to the long-term success of an IS, particularly for full realization of its potential [19], relatively little attention has focused on how infusion occurs [3,14,21-23]. In addition, some studies argued that most IS infusion studies have mainly focused on technological aspects at an organizational level rather than an individual level [22-24]. This is a problem as the results obtained from such studies cannot offer useful suggestions and improvements to major system users for further explorative, integrative, and future use, and may further cause negative effects in individual and organizational performance derived from use of the IT/IS.

Although EMRs can provide clinical and operational benefits, EMR adoption is lagging because of user resistance and other barriers [2]. Most EMR-related studies have mainly focused on issues that arise in the early stage of EMR development, particularly for investigating the factors or barriers affecting EMR adoption or acceptance [2,9,25-32]. Trudel et al [33] found that an increase in EMR adoption does not lead to physicians' progress in using EMR systems during EMR assimilation, which are the routinization and infusion stages mentioned by Cooper and Zmud [18]. Raymond et al [34] called for a deeper understanding of the factors leading to greater performance outcomes from EMR systems after extended EMR use. Moreover, Bhattacherjee [35] confirmed that the long-term success of an IT/IS depends on its continued use rather than its first use, and the influencing factors toward the use of an IT/IS may vary in various IT/IS implementation stages. Ng and Kim [36] argued that IS infusion requires the authentic motivation of users, which is not the case for IS adoption and continuance, and they indicated a lack of understanding about the authentic motivations leading to infusion in the existing literature. Due to the paucity of HIT infusion studies [21,22], the factors (motivations) concerning EMR infusion in the context in which EMR systems are being integrated in clinical settings and incorporated in routine practice should be evaluated. Identification and understanding of the key determinants and consequents of EMR infusion will be helpful to minimize the gaps between IT practitioners and HCPs in EMR design and implementation, and enable further improvements to meet organizational expectations.

Numbeo [37] reported that the health care system of Taiwan ranks first among 93 countries. The successful implementation of HIS and EMRs with full integration of various ITs/ISs in health care institutes is considered the key to the early success of cross-hospital exchange of EMRs in Taiwan. For example, during the COVID-19 pandemic, authorized HCPs in Taiwanese clinics or hospitals can request patient medical records within 3 months and check patients' travel history, occupation, contact history, and cluster history using the patient's health insurance card information to help reduce the infection risk to HCPs and to enable electronic transfer, if necessary, for providing better patient care. Although many Taiwanese hospitals have invested substantial resources and efforts to develop EMRs, with long-term government support for eHealth and EMR development since 2009, factors that affect the long-term success of EMR infusion and performance remain unclear. This situation calls for further research from a practical perspective to provide significant insights for EMR development, particularly in the EMR infusion stage, and the results may further provide contributions to HIT development.

Moreover, the initial success of EMR adoption and acceptance does not ensure its long-term success in terms of its incorporation into the daily operating procedures of hospitals in the technology infusion stage. Therefore, the purpose of this study was to understand the determinants and consequences (performance impacts) of EMR infusion at the individual level. The research questions addressed in this study were: (1) What are the salient factors (motivations) of technology, task, and user influencing EMR infusion by HCPs? (2) How do these factors (motivations) influence EMR infusion and performance?

Prior Related Studies

IT Infusion and Performance

ITs and ISs have the potential to substantially improve the operational effectiveness and efficiency of an organization if used appropriately. IT implementation is a dynamic IT adaption process in organizations with various stages, including initiation, adoption, adaptation, acceptance, routinization, and infusion, as identified by Cooper and Zmud [18] based on a technological diffusion approach. This general IT implementation model has been widely accepted and used in the IS discipline by various users and contexts for exploring the key factors influencing different implementation stages of specific systems. The IT postadoption stages, including IT acceptance, routinization, and infusion as defined by Cooper and Zmud [18], are often the main focus of such research because they are highly relevant to actual IT usage and organizations can observe the realized benefits obtained through IT usage [15,36].

Saga and Zmud [20] investigated the nature and determinants of IT acceptance, routinization, and infusion, and they identified key variables and determinants related to various IT postadoption stages. They found that standardized use, use perceived as “normal,” and administration infrastructure development are key characteristics of IT/IS routinization, whereas extended use, integrative use, and emergent use of IT/IS can be represented as the measurement variables for IT/IS infusion. For the IT implementation stages, IT assimilation and IT infusion are two relevant and similar concepts but with slight differences in nature and the applied theories. Armstrong and Sambamurthy [38] defined IT/IS assimilation as “the success achieved by firms in utilizing the capabilities of IT/IS to enhance their business performance.” IT/IS assimilation can bring significant business value if IT/IS becomes a routinized element of value-chain activities and business strategies in a firm [18,38]. They further argued that IT assimilation “focuses more on the extent of which IT has been infused into specific business activities” and “how effectively IT is enabling the conduct of those activities relative to rivals.” Thus, IT/IS assimilation can be viewed as a broader and integrative stage across IT/IS routinization and infusion stages [18].

Some well-known theoretical models have been proposed and validated for IT/IS postadoption stages, including technology adoption [17,39,40], IS continuance [35], and IS infusion [41,42]. Technology acceptance models (TAMs) are widely used for IT/IS evaluation in the technology adoption stage [17,39,40], whereas IS continuance models are used for evaluation in the IS routinization stage [35]. Ng and Kim [36] argued that most IS research has placed substantial attention on initial system adoption and continuance; however, only few IS infusion studies have been performed to date, which have produced mixed and inconclusive results due to the lack of consideration of the authentic motivation of users in IS infusion. Tennant et al [23] argued that existing implemented systems are often underutilized or not used effectively. They suggested focusing on a deeper level of usage (ie, the infusion stage) that can enhance work tasks and performance. Tennant et al [23] summarized the definitions of infusion used by researchers into two main types: (1) use of IT in a more comprehensive and integrated manner to support the higher-level aspects of organizational work, resulting in the use of IT at its full potential [18]; and (2) the extent to which the full potential of innovation has been embedded in an organization’s (or individual’s) work system [43].

Ng and Kim [36] investigated the relationships between user empowerment, which is regarded as the authentic motivation derived from psychological empowerment theory, and IS infusion. They also examined the moderating effect of habit between user empowerment dimensions and IS infusion types. The results showed that user empowerment dimensions have significant effects on the IS infusion types. In addition, the moderating role of habit between motivations of user empowerment and IS infusion (extended use and integrative use) was confirmed. O’Connor et al [22] and Tennant et al [23] argued that the majority of infusion studies have largely focused on technological aspects at an organizational level rather than an individual level. This may lead to difficulty in the extent to which the full potential of IT/IS can be embedded in an organization’s or individual’s work system. Tennant et al [23] proposed a conceptual research model based on concepts derived from the system usage perspective reported by Burton-Jones and Straub [44] and the task-technology fit (TTF) perspective described by Goodhue and Thompson [10] to examine IS infusion and performance. According to the research model, factors related to the characteristics of the system, individual or user, and task are the antecedents of IS infusion, which may subsequently affect an individual’s performance. Although the IS infusion model proposed by Tennant et al [23] provided a starting point to understand the nature of IS infusion and its effects on consequences, the model should be further validated in various contexts.

Performance evaluation is a key surrogate measure for IS success within the IS discipline in the postimplementation stage [10-12]. Goodhue and Thompson [10] proposed a TTF model by integrating the perspectives of fit and utilization focus to explore the antecedents (task, technology, and individual factors) of TTF, and the relationship between TTF, utilization, and individual performance. They reported that IT/IS can positively affect individual performance if the technology is used (utilization) and is a good fit for the supported task. Some studies have adopted the TTF model to investigate task, technology, and individual fit, and to explore their effects on individual performance [45,46]. By contrast, other studies have examined IT/IS models by incorporating the TTF construct across various contexts (user groups, technologies, and tasks) [47-49]. For example, Hsaio and Chen [45] examined the TTF of mobile nursing ISs for nursing performance on the basis of the TTF model. Dishaw and Strong [48] integrated two IT utilization models (TTF and TAM, as an extension of the TAM with the TTF construct) and found that the integrated model had more explanatory power than each individual model (TTF or TAM) and facilitated a more favorable understanding of IT utilization.

The TTF model has been considered as the core of investigating IS infusion and its effect on individual performance [21-23]. In a study based on the TTF perspective, Tennant et al [23] indicated that future studies related to IS infusion should incorporate the elements of user, task, and system, because IS infusion involves the use of technologies that assist individuals in performing their tasks. Hsaio and Chen [50] and O'Connor et al [21,22] proposed research models by incorporating user, task, and system characteristics as the determinants to investigate mobile health (mHealth) continuance and infusion based on the TTF perspective. In addition, Goodhue and Thompson [10] indicated that the feedback mechanism of performance may affect subsequent TTF, utilization, and performance. For the long-term evaluation of an IS, the fit among the task, technology, and individual should be evaluated, and the IS should be continuously used for supporting the tasks. Thus, this process of IS infusion may affect individual performance.

EMR-Related Studies

HIMSS Analytics defined an EMR as "an application environment composed of the clinical data repository, clinical decision support, controlled medical vocabulary, order entry, computerized provider order entry, pharmacy, and clinical documentation applications." Moreover, the environment contains patients' EMRs across inpatient and outpatient services, and is used by HCPs to document, monitor, and manage health care delivery within a health care organization [2]. Boonstra and Broekhuis [2] mentioned that EMRs are computerized medical information systems that collect, store, and display patient information. Yamamoto and Khan [9] indicated that the perceived advantages of EMRs include:

> *optimizing the documentation of patient encounters, improving the communication of information among physicians, improving access to patient medical information, reducing errors, optimizing billing, improving reimbursement for services, forming a data repository for research and quality improvement, and reducing the use of papers.*

Furthermore, EMRs designed to document patient care information and enable data sharing among clinicians [29,51] can disrupt work practices, thus causing a significant decrease in productivity with their initial use [52,53].

EMR systems are viewed as both EHR systems and clinical information systems in hospital settings to provide clinical, communicational, and administrative functionalities [3]. Among the investigated three functionalities, Raymond et al [3] found that clinical functionalities can explain why certain primary care physicians make more extended use of EHRs than others because clinical functionalities can better support or fit their main medical tasks. The core components of an EHR are administrative functions, computerized physician order entry, lab systems, radiology systems, pharmacy systems, and clinical documentation [54]. Raymond et al [3] indicated that EHR systems have been slowly adopted by health care providers in the United States due to the challenges of costly software packages, system security, patient confidentiality, and unknown future government regulations.

Although the appropriate use of EMRs can improve quality, continuity, safety, and efficiency in health care, they have not been widely adopted in health care institutions as expected because of resistance among HCPs [2]. To understand the key barriers to the use of EMRs, Boonstra and Broekhuis [2] performed a comprehensive systematic review of studies related to the use of EMRs among physicians in the early stage of EMR development in health care institutions. They identified the following 8 key barriers affecting physicians' acceptance of EMR implementation: organizational, process change, financial, technical, time, psychological, social, and legal barriers. O'Donnell et al [55] investigated primary care physicians' attitudes toward EMR adoption and proposed a clinical adoption framework to understand the disparate performance among health care institutions in EMR implementation. Some studies have explored EMR adoption based on technological evaluation theories, namely the theory of planned behavior and the unified theory of acceptance and use of technology (UTAUT) [56-58]. Ahmed et al [56] investigated predicators of intention to use EMR based on the expanded UTAUT model (UTAUT2). They identified performance expectancy, effort expectancy, social influence, facilitating conditions, and computer literacy as key predicators of intention to use EMRs by health care providers; however, hedonic motivation and habit had no significant effect on intention to use EMRs. In addition, Sayyah Gilani et al [59] performed a study on EMR continuance intention of HCPs based on technology continuance theory. They found that EMR continuance was influenced by attitude and satisfaction. Attitude is mainly influenced by perceived ease of use, perceived usefulness, and satisfaction, whereas satisfaction is influenced by perceived usefulness and confirmation. In turn, perceived usefulness is affected by perceived ease of use and confirmation.

However, previous studies have argued that the scope of these theories is limited because they do not address the causes related to the adoption process, specifically for the postimplementation stage [28,57,60]. In addition, Trudel et al [33] reported the presence of the ceiling effect on EMR assimilation based on their observation that the growth of EMR adoption does not lead to physicians' progress in using EMR systems. Raymond et al [34] called for a deeper understanding of the factors leading to greater performance outcomes from EMR systems after extended EMR use. Goh et al [61] proposed a dynamic process model based on adaptive structuration theory for improving understanding of the interplay between HIT and patterns of clinical work embodied in daily routines. O'Connor et al [21] investigated the determinants (factors adapted from technology, user, and task) of individual infusion of mHealth technologies and their subsequent outcomes. In their conceptual model, technology, user, and task were considered key enablers of successful mHealth infusion, and mHealth infusion led to improvements in preventative care, greater decision-making, and reduced medical errors. O'Connor et al [22] further proposed an mHealth infusion model to investigate the effects of the characteristics of technology (availability, maturity, and portability), individuals (habits, self-efficacy, and technology trust), and tasks (time criticality, interdependence, and mobility) on the integrative and exploratory use of mHealth infusion by HCPs and the relationship between mHealth infusion and performance.

Bhattacherjee [35] indicated that the long-term success of an IS depends on its continued use in the postimplementation stage, and proposed the expectation confirmation model (ECM) to understand IS continuance by examining the relationships among confirmation, perceived usefulness, satisfaction, and IS continuance. In addition, the author indicated that the key factors determined for IS acceptance are not necessarily crucial in the IS postimplementation or infusion stage. This finding implied that factors affecting ISs may exert different effects on different IS implementation stages. Hsaio and Chen [50] investigated the determinants of HCPs' perspectives on mHealth continuance and performance based on the ECM proposed by Bhattacherjee [35] and the mHealth infusion model proposed by O'Connor et al [22]. They found that mHealth continuance is mainly affected by perceived usefulness, technology maturity, individual habits, task mobility, and user satisfaction, whereas individual performance is substantially affected by mHealth continuance.

The aforementioned findings provide the theoretical basis for this study in terms of understanding the effects of the antecedents of EMR infusion (task, technology, and individual characteristic) on the exploratory, integrative, and future (emergent) use of EMRs, and the subsequent effects on individual performance.

Methods

Research Model

On the basis of the results reported by Hsaio and Chen [50] and O'Connor et al [22] regarding mHealth infusion, and those by Goodhue and Thompson [10] regarding the TTF model, we constructed an EMR infusion model to investigate the effects of technology, user, and task characteristics on EMR infusion by HCPs and their subsequent effects on individual performance. In addition, some variables mentioned by Hsaio and Chen [50] and O'Connor et al [22] were included to accommodate the EMR utilization and health care context in Taiwan. In this study, we excluded self-efficacy and technology trust from the user characteristics reported by O'Connor et al [22] because self-efficacy was reported to be a nonsignificant factor affecting HCPs' IT acceptance [62,63]. EMRs have been developed, used (not mandatorily), and incorporated into daily procedures in Taiwanese hospitals for several years; thus, technology trust was not considered in this model. In addition, personal innovativeness of IT and technostress were added as variables in the research model. Rai et al [64] indicated that personal innovativeness positively and significantly affected mHealth usage intention and assimilation. Technostress is an emergent problem derived from the pervasive use of technologies, which may significantly affect individual productivity and daily life [65]. We considered that EMR use in its infusion stage may be affected by technostress exerting an effect on EMR infusion and individual performance. However, the effects of personal innovativeness and technostress on EMR infusion and performance should be empirically validated. Thus, the research model (Figure 1) involved nine EMR infusion antecedents: accessibility, maturity, portability (technology), time criticality, interdependence, mobility (task), EMR infusion, and individual performance (the outcomes of EMR infusion). The measurement, operational definition, and number of items for the variables investigated in this study are summarized in Table 1.

The technology characteristics examined in this study included accessibility (availability), maturity, and portability; these factors determine the functionality or usability of an EMR. The task characteristics examined in this study included time criticality, interdependence, and mobility. The user characteristics examined in this study included personal innovativeness in IT, technostress, and habit. These characteristics represent individual traits and perceptions toward the use of IT.

Figure 1. Research model. EMR: electronic medical record. H: hypothesis.

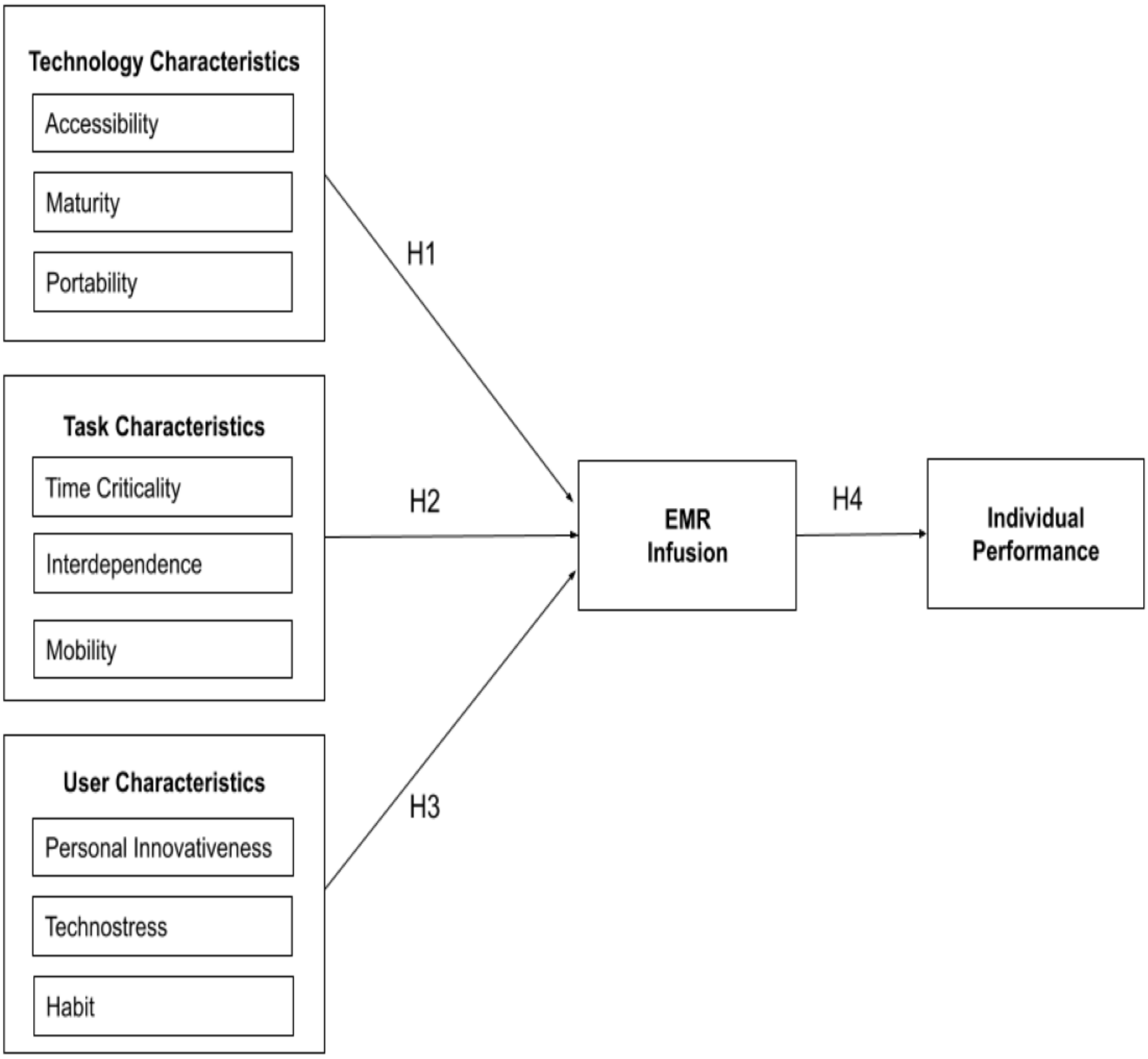


Table 1. Measurement and operational definitions of variables.

Construct	Operational definition	References	Number of items
Technology characteristics			
Accessibility	The ability of accessing EMR[a] information when required	Gebauer et al [66], Liang et al [67]	3
Maturity	The existence of a level of EMR quality that is perceived as satisfactory and the perceived need for system improvement by the user	O'Connor et al [22], Liang et al [67]	5
Portability	The degree of ease associated with transporting the EMR	O'Connor et al [22], Gebauer et al [66]	3
Task characteristics			
Time criticality	The urgency when accessing information through the EMR	O'Connor et al [22], Liang et al [67]	3
Interdependence	The degree to which completing tasks using the EMR requires interaction with other people	Hsiao and Chen [45], Gebauer et al [66]	3
Mobility	The extent to which a task is being performed in different locations using the EMR	Gebauer et al [66], Liang et al [67]	3
User characteristics			
Personal innovativeness	Willingness to try out any new technology	Agarwal and Prasad [68], Thatcher and Perrewe [69]	4
Technostress	A problem of adaptation resulting from the health care professional's inability to cope with EMR use in a healthy manner	Ragu-Nathan et al [70], Tu et al [71]	5
Habit	The extent to which an individual tends to use the EMR automatically	Limayem et al [72], Venkatesh et al [73]	3
EMR infusion	The extent of EMR infusion related to the exploratory, integrative, and future use of EMR	O'Connor et al [22], Tennant et al [23]	9
Individual performance	The extent to which EMR use (continuance) can improve the health care professional's efficiency, effectiveness, and quality of medical activities	Junglas et al [74]	8

[a] EMR: electronic medical record.

EMR infusion is the extent of the exploratory, integrative, and future use of the EMR. Individual performance is defined as the improvement in HCPs' efficiency, effectiveness, and quality of medical activities through EMR continuance (use). O'Connor et al [22] and Tennant et al [23] reported that IT assimilation and infusion are two types of uses that are beyond routine use and refer to a deeper level of usage that enhances work tasks and performance. Technology, task, and user characteristics are considered key antecedents that affect IS infusion or continuance [22,23,50] and the extent of IS infusion can affect individual performance. However, not all technology, task, and user characteristics examined by Hsaio and Chen [50] were reported to be significant antecedents, implying that the factors affecting IS infusion may vary among different technology, task, and user groups. Thus, the antecedents of EMR infusion should be empirically examined and tested for long-term EMR system evaluation. We proposed four research hypotheses with nine subhypotheses, which are summarized in Textbox 1.

Textbox 1. Research hypotheses and subhypotheses.

H1: The technology characteristics of health care professionals (HCPs) significantly affect electronic medical record (EMR) infusion

- H1a: The accessibility of an EMR significantly affects EMR infusion
- H1b: The maturity of an EMR significantly affects EMR infusion
- H1c: The portability of an EMR significantly affects EMR infusion

H2: The task characteristics of HCPs significantly affect EMR infusion

- H2a: Task time criticality significantly affects EMR infusion
- H2b: Task interdependence significantly affects EMR infusion
- H2c: Task mobility significantly affects EMR infusion

H3: The user characteristics of HCPs significantly affect EMR infusion

- H3a: Personal innovativeness in information technology significantly affects EMR infusion
- H3b: Technostress significantly affects EMR infusion
- H3c: Individual habit significantly affects EMR infusion

H4: EMR infusion significantly affects individual performance

Instrument and Respondents

The respondents were HCPs (doctors and nurses) who worked at different departments in the case hospital, and we elaborated rigorous instrument design processes to minimize the potential risks of common method bias derived from a single-respondent questionnaire-based survey as reported by Podsakoff et al [75]. In the first stage, we established 50 preliminary measurement items by referencing the literature and validated the instrument to ensure its appropriateness for the purpose of this study. Individual performance was measured using 8 items adapted from the study of Junglas et al [74]. EMR infusion was assessed with 9 items adapted from the studies of O'Connor et al [22] and Tennant et al [23]. Technology characteristics, namely accessibility, maturity, and portability, were measured using 11 items adapted from studies conducted by Gebauer et al [66], Liang et al [67], and O'Connor et al [22]. Task characteristics, namely time criticality, interdependence, and mobility, were measured using 9 items adapted from Gebauer et al [66], Hsiao and Chen [45], Liang et al [67], and O'Connor et al [22]. User characteristics, namely personal innovativeness, technostress, and habit, were measured using 12 items adapted from Agarwal and Prasad [68], Limayem et al [72], Ragu-Nathan et al [70], Thatcher and Perrewe [69], Tu et al [71], and Venkatesh et al [73].

The second stage of the questionnaire design focused on the evaluation and selection of the measurement items. To evaluate the content validity of the questionnaire, the content validity index (CVI) was calculated, and a threshold value of >0.80 was recommended [76]. Two experts on EMRs and one health informatics professional were invited to examine the content validity of the instrument. One item in habit was excluded because its value was <0.80, and the average CVI of the questionnaire was 0.98. Ultimately, a final research questionnaire consisting of 49 items was developed, and each item of the questionnaire was scored on a 5-point Likert scale (1 for strongly disagree and 5 for strongly agree). The detailed description of the research instrument can be found in Multimedia Appendix 1. The questionnaire consisted of two major parts. The first part focused on the demographic data of respondents, namely age, sex, educational level, and experience using the EMR in the hospital. The second part included measurement items related to the antecedents of EMR, EMR infusion, and individual performance.

To ensure that the data collected were in line with the research objectives, only HCPs with >6 months experience using the EMR in the case hospital were included as participants in this study. A total of 120 doctors and 500 nurses from hospitals in southern Taiwan were recruited. The case hospital has been using EMRs since 2009 to improve their quality of care, continuity, safety, efficiency, and medical decision-making. In addition, the case hospital obtained the EMR accreditation issued by the Ministry of Health and Welfare of Taiwan. This certification indicates that the core EMR functionality can provide the services of cross-hospital health information exchange and EHRs with the support of the existing IT and IS infrastructure.

Model Evaluation

Measurement Model

The collected data were further analyzed using WarpPLS 5.0 [77], with the partial-least squares (PLS) approach to perform extensive, scalable, and flexible casual modeling [78]. Chin [79] suggested using the PLS technique to analyze measurement and structural models. Several common model data fit and quality indices are recommended for WarpPLS because of their advantages compared with other variance-based structural equation modeling (SEM) methods, including the average path coefficient (APC), average R^2 (ARS), average adjusted R^2 (AARS), average block variance inflation factor (AVIF), average full collinearity variance inflation factor (AFVIF), Tenenhaus goodness of fit (GoF), and R^2 contribution ratio (RSCR) [77]. Kock [77] indicated that the addition of latent variables into a model can increase the ARS value but reduce the APC value; however, both ARS and APC can increase concurrently only when the addition of variance can improve the predictive and explanatory qualities of the overall model. The AARS is used to correct inappropriate increases in R^2 coefficients when predictors cannot adequately improve the explanatory value of each latent variable [77]. Both AVIF and AFVIF are often used to evaluate the collinearity of a model if new latent variables are added. Kock [80] mentioned that if all variance inflation factors resulting from a full collinearity test are equal to or lower than 3.3, the model can be considered free of common method bias for PLS-SEM. The GoF is an index used to evaluate the explanatory power of a model, and RSCR is used to examine the degree to which a model is free from negative R^2 effects [77].

Structural Model

WarpPLS 5.0 with the bootstrap resampling method was used to analyze the structural model, which was mainly evaluated using the path coefficient (β) and R^2 value. Path coefficients represent the strength and direction of the relationship between variables, and they are meaningful to research if they achieve a statistically significant level. R^2 values indicate the total variance of dependent variables explained by influencing variables, demonstrating the predictive power of the investigated model.

Ethical Considerations

To address potential ethical concerns, our study protocol and informed consent forms were reviewed and approved by the institutional review board (IRB) of St. Martin De Porres Hospital in Taiwan before the distribution and collection of surveys. After receiving approval from the IRB of the target hospital (IRB-15B-021), we obtained voluntary and verbal consent from the study participants. The requirement to document consent was waived. The responses of HCPs were anonymous and unidentified.

Results

Demographic Data and Descriptive Statistics

The questionnaire was administered to collect data from HCPs who worked at the case hospital at the time of the survey and had at least 6 months of experience in using EMRs. A total of 316 questionnaires were distributed and 211 complete copies were returned, resulting in a valid response rate of 66.8%. Voluntary participation might explain the relatively high response rate. According to the collected demographic data (Table 2), most of the respondents were aged <40 years, were women, and had a bachelor’s or higher degree. Furthermore, over 70% of the respondents worked in the nursing department and the remaining respondents worked in the medical department. In addition, 93.4% (203/211) of the respondents had >1 year of experience in using EMRs, indicating the representativeness of the participants. Among the investigated demographic variables, experience in using the EMR and respondents’ department were found to be significantly related to EMR infusion and performance responses based on analysis of variance. This is reasonable because doctors are heavy EMR users, and users with abundant experience in using EMRs tend to have more positive attitudes toward exploratory, integrative, and future EMR use, and a better performance evaluation.

Table 2. Demographic characteristics of the respondents (N=211).

Characteristic	Respondents, n (%)
Age (years)	
<30	73 (34.6)
31-40	91 (43.1)
41-50	31 (14.7)
51-60	11 (5.2)
> 60	5 (2.4)
Gender	
Male	52 (24.7)
Female	159 (75.3)
Education level	
Junior college	29 (13.8)
Bachelor	169 (80.1)
Master (or higher)	13 (6.1)
Experience in hospital (years)	
<1	8 (3.8)
1-3	42 (19.9)
3-6	56 (26.6)
6-9	36 (17.0)
>9	69 (32.7)
Experience in using EMRs[a] (years)	
<1	14 (6.6)
1-3	102 (48.4)
3-5	60 (28.4)
>5	35 (16.6)
Department	
Nursing	150 (71.1)
Medical	61 (28.9)

[a] EMR: electronic medical record.

Table 3 shows the descriptive statistics of responses to the original questionnaire, including 49 items of the 11 investigated constructs used in this study. Among the constructs, the mean scores were the highest for time criticality, followed by performance, interdependence, portability, habit, technostress, mobility, accessibility, EMR infusion, maturity, and personal innovativeness in IT. Time criticality, performance, and interdependence were the top three variables and had mean scores greater than 4.0, whereas personal innovativeness in IT had the lowest mean score of the investigated variables. This implied that participating HCPs have a more positive evaluation toward EMR use for supporting their task characteristics (time criticality and interdependence); however, they do not demonstrate excellent personal innovativeness in IT. We further evaluated the item appropriateness and consistency of the measured constructs among domain experts. In total, 42 items were used for further PLS-SEM analysis and 7 items were excluded.

Table 3. Descriptive statistics of constructs and their respective items.

Construct	Score, mean (SD)	Range
ACC[a]	3.85 (0.65)	1-5
ACC1	3.94 (0.55)	2-5
ACC2	3.83 (0.66)	2-5
ACC3	3.77 (0.73)	1-5
POR[b]	3.96(.64)	2-5
POR1[c]	4.10 (0.68)	2-5
POR2	3.88 (0.64)	2-5
POR3	3.91 (0.61)	2-5
MAT[d]	3.78 (0.67)	1-5
MAT1	3.68 (0.71)	2-5
MAT2	3.64 (0.76)	1-5
MAT3	3.85 (0.60)	2-5
MAT4[c]	3.88 (0.61)	2-5
MAT5	3.85 (0.60)	2-5
TC[e]	4.05 (0.62)	1-5
TC1[c]	3.89 (0.66)	1-5
TC2	4.08 (0.60)	2-5
TC3	4.18 (0.61)	2-5
INT[f]	4.02 (0.65)	2-5
INT1	4.09 (0.65)	2-5
INT2	4.02 (0.64)	2-5
INT3	3.96 (0.67)	2-5
MOB[g]	3.89 (0.66)	1-5
MOB1	3.91 (0.71)	1-5
MOB2	3.93 (0.60)	2-5
MOB3	3.83 (0.68)	2-5
PI[h]	3.60 (0.72)	1-5
PI1	3.48 (0.70)	1-5
PI1	3.43 (0.75)	1-5
PI3[c]	3.72 (0.73)	2-5
PI4[c]	3.76 (0.70)	2-5
TS[i]	3.90 (0.67)	2-5
TS1	3.90 (0.71)	2-5
TS2	3.90 (0.61)	2-5
TS3	4.10 (0.64)	2-5
TS4	4.15 (0.62)	2-5
TS5[c]	3.47 (0.78)	2-5
HAB[j]	3.96 (0.57)	3-5
HAB1	4.02 (0.56)	3-5

Construct	Score, mean (SD)	Range
HAB2	3.93 (0.57)	3-5
HAB3	3.93 (0.59)	3-5
INF[k]	3.82 (0.64)	1-5
INF1 (EU[l]1)	3.71 (0.69)	2-5
INF2 (EU2	3.62 (0.71)	2-5
INF3 (EU3)	3.64 (0.66)	2-5
INF4 (IU[m]1)	3.97 (0.60)	2-5
INF5 (IU2)	3.91 (0.60)	3-5
INF6 (IU3)	3.96 (0.58)	3-5
INF7 (FU[n]1)	3.97 (0.65)	2-5
INF8 (FU2)	4.00 (0.61)	2-5
INF9 (FU3)[c]	3.59 (0.67)	1-5
PER[o]	4.05 (0.58)	2-5
PER1	4.20 (0.59)	3-5
PER2	4.15 (0.55)	3-5
PER3	4.02 (0.56)	3-5
PER4	4.00 (0.57)	3-5
PER5	4.00 (0.53)	3-5
PER6	4.06 (0.56)	3-5
PER7	4.03 (0.61)	3-5
PER8[c]	3.94 (0.63)	2-5

[a] ACC: accessibility.

[b] POR: portability.

[c] Excluded from further analysis.

[d] MAT: maturity.

[e] TC: time criticality.

[f] INT: interdependence.

[g] MOB: mobility.

[h] PI: personal innovativeness in information technology.

[i] TS: technostress.

[j] HAB: habit.

[k] INF: electronic medical record infusion.

[l] EU: exploratory use.

[m] IU: integrative use.

[n] FU: future use.

[o] PER: performance.

Measurement Model

As shown in Table 4, the results demonstrated that all of the model data fit and quality indices met the criteria suggested by Kock [77]. All of the APC, ARS, and AARS values exceeded the recommended values, thereby indicating a highly satisfactory fit. The AVIF and AFVIF values indicated the absence of the collinearly problem in the model, demonstrating that the model can be considered free of common method bias as suggested by Kock [80]. The GoF value also exceeded the suggested value, indicating a satisfactory fit, and the RSCR value indicated an excellent fit. These findings validated the data fit and quality indices of the EMR infusion model.

Table 4. Model fit and quality indices.

Fit indices	Value	Criteria	Result
Average path coefficient	0.193 (P<.001)	P<.05	Fit
Average R^2	0.529 (P<.001)	P<.05	Fit
Average adjusted R^2	0.521 (P<.001)	P<.05	Fit
Average block VIF[a]	1.954	Acceptable if ≤5, ideally ≤ 3.3	Fit
Average full collinearity VIF	2.324	Acceptable if ≤5, ideally ≤3.3	Fit
Tenenhaus GoF[b]	0.618	Small, ≥0.1; medium, ≥0.25; large, ≥0.36	Fit
R^2 contribution ratio	1.000	Acceptable if ≥0.9, ideally=1	Fit

[a] VIF: variance inflation factor.

[b] GoF: goodness of fit.

We further evaluated the reliability and validity (convergent and discriminant validity) of the instrument as suggested by previous studies [79,81,82]. Hair et al [82] suggested evaluating the reliability and internal consistency of each construct as reflective indicators by performing principal component analysis. They recommended a cut-off value of >.70 for Cronbach α and composite reliability (CR). Furthermore, Fornell and Larcker [81] indicated that the value of average variance extracted (AVE) should exceed 0.5, and each square correlation should have adequate convergent and discriminant validity. As shown in Table 5, the Cronbach α (>.798), CR (>0.734), and AVE (>0.604) values of all constructs exceeded the recommended cut-off values, indicating satisfactory reliability and convergent validity. One criterion for adequate discriminant validity is that the square root of the AVE for each construct should exceed the correlation coefficient between the construct and other constructs in the research model [79]. All square roots of AVE values in this study (diagonals in Table 5) were higher than the correlation coefficients (off-diagonals in Table 5), indicating satisfactory discriminant validity. These findings demonstrated the adequate reliability, convergent validity, and discriminant validity of the model.

Table 5. Correlations among variables, and the reliability and validity of the research model.

Variables	Correlations[a]											AVE[b] (>0.5)	CR[c] (>0.7)	Cronbach α (>.7)
	ACC[d]	POR[e]	MAT[f]	TC[g]	INT[h]	MB[i]	TS[j]	PI[k]	HB[l]	INF[m]	PER[n]			
ACC	0.792	0.686	0.643	0.582	0.581	0.501	0.382	0.598	0.412	0.550	0.456	0.628	0.761	.830
POR	0.686	0.830	0.715	0.538	0.414	0.374	0.439	0.481	0.619	0.663	0.576	0.689	0.753	.822
MAT	0.643	0.715	0.787	0.546	0.480	0.321	0.370	0.375	0.449	0.538	0.460	0.619	0.801	.870
TC	0.582	0.538	0.546	0.899	0.653	0.576	0.482	0.294	0.430	0.518	0.498	0.808	0.871	.855
INT	0.581	0.414	0.480	0.653	0.856	0.658	0.429	0.336	0.309	0.517	0.447	0.732	0.857	.881
MB	0.501	0.374	0.321	0.576	0.658	0.831	0.537	0.443	0.295	0.570	0.321	0.691	0.823	.868
TS	0.382	0.439	0.370	0.482	0.429	0.537	0.805	0.353	0.531	0.635	0.453	0.812	0.875	.866
PI	0.598	0.481	0.375	0.294	0.336	0.443	0.353	0.901	0.318	0.499	0.314	0.648	0.827	.798
HB	0.412	0.619	0.449	0.430	0.309	0.295	0.531	0.318	0.865	0.675	0.704	0.749	0.899	.925
INF	0.550	0.663	0.538	0.518	0.517	0.570	0.635	0.499	0.675	0.777	0.803	0.604	0.734	.801
PER	0.456	0.576	0.460	0.498	0.447	0.321	0.453	0.314	0.704	0.662	0.803	0.645	0.892	.931

[a] The values in the diagonal are square roots of the AVE and the off-diagonal elements are the correlation coefficients (*r*) among constructs.

[b] AVE: average variance extracted.

[c] CR: composite reliability.

[d] ACC: accessibility.

[e] POR: portability.

[f] MAT: maturity.

[g] TC: time criticality.

[h] INT: interdependence.

[i] MOB: mobility.

[j] PI: personal innovativeness in information technology.

[k] TS: technostress.

[l] HAB: habit.

[m] INF: electronic medical record infusion.

[n] PER: performance.

Structural Model

As shown in Figure 2, among the four major hypotheses (including nine subhypotheses) of this study, those related to technology characteristics (only H1c was positively supported) and task characteristics (only H2a and H2c were positively supported) were partially supported. However, all of the subhypotheses (H3a, H3b, and H3c) related to user characteristics were significantly supported. The results revealed that EMR infusion was mainly affected by portability among technology characteristics; time criticality and mobility among task characteristics; and personal innovation, technostress, and habit among user characteristics. Habit (H3a), portability (H1c), personal innovativeness (H3a), technostress (H3b), time criticality (H2a), and mobility (H2c) were identified as key factors affecting EMR infusion according to their relative effects on EMR infusion by ranking. In addition, individual performance was significantly affected by EMR infusion (H4). We further examined the direct effects, indirect effects, and total effects of the research variables, specifically for the mediation analysis of technology, task, and user characteristics (through infusion) on performance. As shown in Table 6, technostress, habit, personal innovativeness in IT, and mobility had significant mediating (indirect) effects through infusion on EMR performance.

Figure 2. Results of model validity.

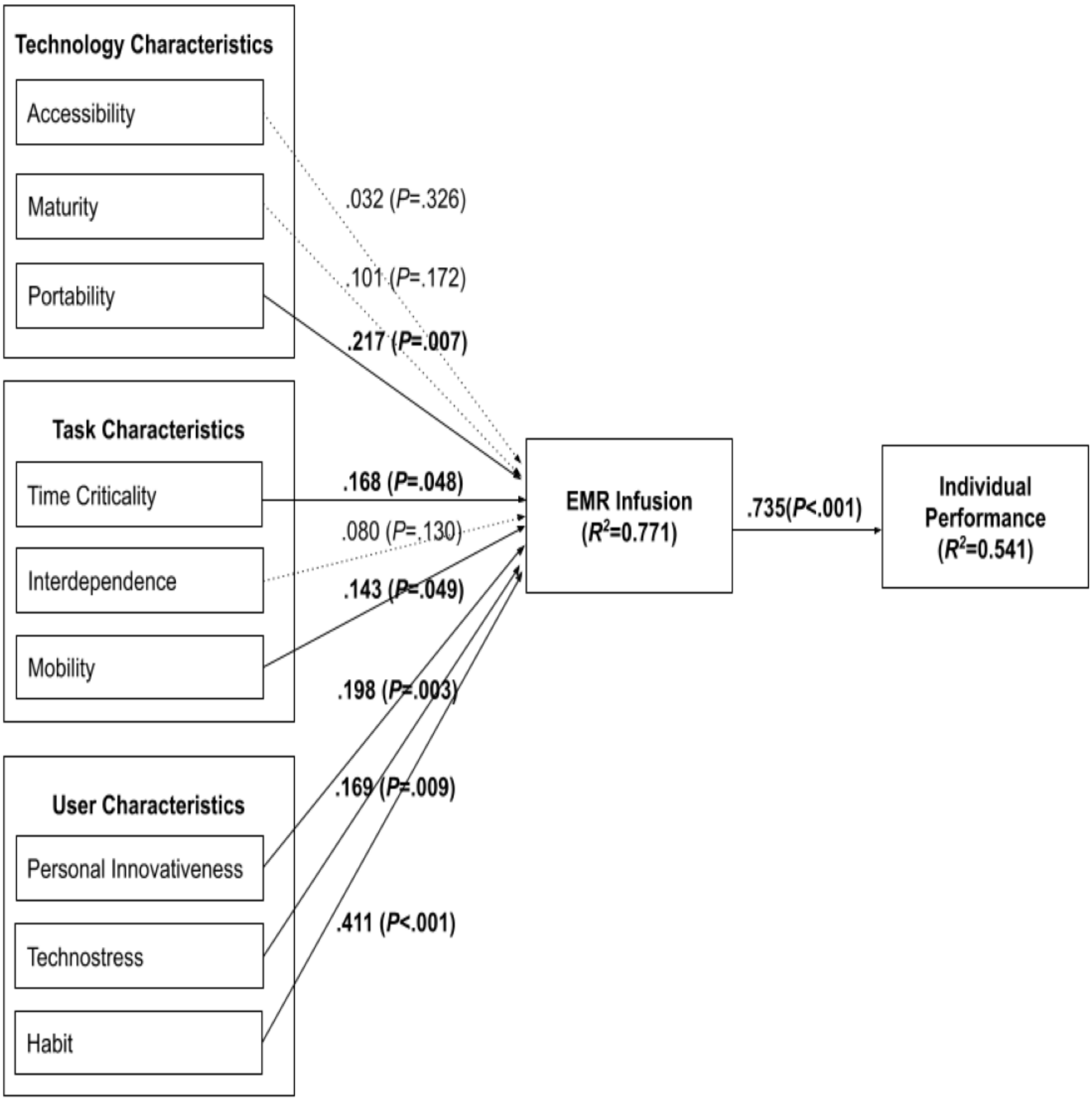


Table 6. Direct, indirect, and total effects (β values) of research variables.

Variable	TS[a]	HAB[b]	INT[c]	TC[d]	MOB[e]	MAT[f]	POR[g]	ACC[h]	PI[i]	INF[j]
INF (direct effect)	.169 (P=.005)	.411 (P<.001)	.141	.031	.145	.059	.217	.047	.110	N/A[k]
PER[l]										
Direct effect	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	.735 (P<.001)
Indirect effect	.124 (P=.03)	.302 (P=.001)	.104 (P=.04)	–.023	.106 (P=.04)	.043	.159	–.034	.081	N/A
Total effect	.124 (P=.005)	.302 (P=.001)	.104	.023	.106	.043	.159	.034	.081	.735 (P<.001)

[a] TS: time criticality.
[b] HAB: habit.
[c] INT: interdependence.
[d] TC: time criticality.
[e] MOB: mobility.
[f] MAT: maturity.
[g] POR: portability.
[h] ACC: accessibility.
[i] PI: personal innovativeness in information technology.
[j] INF: electronic medical record infusion.
[k] N/A: not applicable.
[l] PER: performance.

Discussion

Principal Findings

The results revealed that EMR infusion (R^2=0.771) was mainly affected by user habits (β=.411), portability (β=.217), personal innovativeness (β=.198), technostress (β=.169), and time criticality (β=.168), whereas individual performance (R^2=0.541) was affected by EMR infusion (β=.735). This finding indicated that user (habit, personal innovativeness, and technostress), technology (portability), and task (mobility and time criticality) characteristics have major influences on EMR infusion. Furthermore, the results indicated that EMR infusion positively affected individual performance. Consistent with the findings of previous IS infusion and TTF studies [10,21-23,50], EMR infusion was found to be affected by technology, task, and user (individual) characteristics. However, not all investigated technology, task, and user characteristics were found to be significant for EMR infusion. The results of this study revealed that among technology characteristics, only portability (β=.217) affected EMR infusion; however, the effects of accessibility and maturity on EMR infusion were not as expected. Portability is the degree of ease associated with transporting EMRs [22,66]. This finding is in accordance with the result reported by O’Connor et al [22] but is not consistent with that reported by Hsaio and Chen [50] for mHealth infusion. EMRs are computerized medical information systems that collect, store, and display patient information [2]; thus, they can be easily transported to HCPs to fulfill their needs. This finding may explain why portability was found to be a significant factor affecting EMR infusion.

O’Connor et al [22] reported that task characteristics, namely time criticality, interdependence, and mobility, were considered to be salient factors affecting mHealth infusion; however, Hsaio and Chen [50] found that mobility was the only significant factor affecting mHealth continuance. We found that among task characteristics, time criticality (β=.168) and mobility (β=.143) affected EMR infusion. The tasks of HCPs are complex and time-critical, and require mobility, specifically for providing services in inpatient and emergency departments. Real-time and accurate information obtained from an EMR is critical to increase efficiency and effectiveness in patient care duties [45]. These findings explain why time criticality and mobility were significant factors affecting EMR infusion. Consistent with the results of Hsaio and Chen [45,50], interdependence was found to be insignificant for EMR infusion.

This study found that user characteristics, namely personal innovativeness (β=.198), technostress (β=.217), and habit (β=.411), significantly affect EMR infusion. All three user characteristics exerted significantly stronger effects compared with those of technology and task characteristics on EMR infusion. This finding implied that user characteristics are the key antecedents of EMR infusion. The individual habit of EMR use showed a consistent result with the findings of O’Connor et al [22] and Hsaio and Chen [50]; however, personal innovativeness was not observed to be a significant factor in the context of mHealth [50]. Previous studies have reported that habit can affect future behavior if technology use becomes a habit as routine behavior [83,84], which is consistent with the findings of this study. The result of personal innovativeness in this study is also in accordance with that reported by Rai et al [64], who confirmed that personal innovativeness positively and significantly affected IS usage intention and assimilation. Consistent with the results reported by La Torre et al [65], we found that technostress significantly affected EMR infusion and individual productivity.

In this study, EMR infusion referred to the extent of EMR infusion related to the exploratory, integrative, and future use of EMRs, whereas individual performance was defined as the improvement in HCPs' efficiency, effectiveness, and quality of medical activities through EMR continuance (use). Previous studies have indicated that IT assimilation and infusion are two types of use that are beyond routine use, and refer to a deeper level of usage that enhances work tasks and performance [22,23]. Consistent with findings of previous studies [22,23,50], this study indicated that EMR infusion significantly affected its outcomes (individual performance). We also found that technostress, habit, personal innovativeness in IT, and mobility have significant mediating (indirect) effects through EMR infusion on EMR performance. This implied that technostress, habit, and personal innovativeness in IT have both positive and significant direct effects on EMR infusion and indirect effects on EMR performance. Moreover, mobility was found to only have positive and significant indirect effects on EMR performance. Therefore, we should pay more attention to these significant factors of EMR infusion and performance.

We performed an additional analysis to determine HCPs' performance based on EMR use. As shown in Table 3 (PER1-PER8), the top three items were as follows in descending order: PER1 ("EMR use accelerates information exchange with other members of the health care team"; score=4.20), PER2 ("EMR use reduces information retrieval time in clinical care practice"; score=4.15), and PER6 ("EMR use can make me more efficient at patient care"; score=4.06). Of all the items investigated, most respondents only provided a high and positive evaluation toward EMR performance, and only the mean value of the item "EMR use facilitates estimating and managing the costs of patient care" did not exceed 4 (mean 3.94, SD 0.63), indicating that this item had a slightly lower value than the other items. The results confirmed that EMR use can improve HCPs' performance (mean 4.05, SD 0.58) related to efficiency, effectiveness, and quality of medical activities.

Contributions to Medical Informatics Theory

Previous studies have suggested investigating the IS infusion chain, including the antecedents of IS infusion and the outcomes of IS infusion at the individual level with broader considerations, to examine the extent to which the full potential of ITs and ISs has been embedded in an organization's or individual's work system [22,23]. This study attempted to determine the reasons underlying the ceiling effect in EMR assimilation observed by Trudel et al [33]. From the theoretical perspective of TTF, the results of this study are in accordance with those reported by Goodhue and Thompson [10], who indicated that IT/IS can positively affect individual performance if the technology has been used continuously (utilization) and is a good fit for the supported task (TTF) in the EMR context. For the long-term evaluation of an IS, the fit among task, technology, and individual should be evaluated for the IS, and the IS should be continuously used for supporting the tasks. Thus, the IS infusion process can substantially affect individual performance. The appropriate, integrative, and exploratory use of EMRs in the infusion stage can significantly improve quality of care, continuity, safety, efficiency, and medical decision-making, and facilitate the exchange of cross-hospital health information and EHRs [1,2,4-9].

This is one of the few studies to specifically focus on EMR infusion by considering technology, task, and user aspects, and to examine EMR infusion effects on individual performance. The results of this study can be helpful for extending IS infusion research and identifying critical factors affecting EMR assimilation and infusion where EMRs have been deeply incorporated into the daily operating procedures of hospitals. In addition, EMR design and implementation should meet HCP task needs, particularly for time critically and mobility, and technology needs, specifically for EMR portability. Moreover, future studies can extend the results of this study by incorporating different behavioral theories and factors as the antecedents of IS infusion, and investigating their effects on IS infusion and individual performance. We found that accessibility (availability) and maturity among technology characteristics and interdependence among task characteristics were insignificant factors for EMR infusion. Therefore, additional studies should be performed to validate the factors investigated in the research model because their effects on IS infusion may vary depending on technology, task, and user groups. In addition, inspired by Kim et al [85] on what clinical information is valuable to doctors toward using a mobile EMR, we suggest that further studies pay attention to investigating the critical clinical information and functionalities of EMRs used by different HCPs in EMR infusion and to what extent they can effectively support their tasks with long-term technology use.

Contributions to Medical Informatics Practice

The key factors identified in this study provide useful insights for the further improvement of EMR development in hospitals and the government, specifically for the infusion stage. In addition, the developed instrument can be used as an assessment tool for identifying the key considerations of EMR infusion, and for evaluating the extent of the EMR infusion and individual performance of hospitals that have implemented EMR systems. The results can help the government to understand the urgent needs of hospitals in implementing EMRs, provide sufficient resources and support for improving the incentives of EMR development, and develop adequate EMR policies for the widespread use of health information exchanges and EHRs. Future studies should focus attention on these characteristics, specifically user characteristics (personal innovativeness, technostress, and habit) and task characteristics (time critical and mobility), to further facilitate EMR infusion. Our findings indicate that the routine use of EMRs by HCPs in their daily workflow processes can reduce their technostress related to the use of EMRs and increase their perceived personal innovativeness, thus promoting EMR infusion.

Limitations

This study has several limitations. First, we collected samples from a regional teaching hospital in Taiwan, restricting the generalization of the findings to other medical institutions. Second, the data were derived from questionnaires provided to participants with more than 6 months experience in using EMRs. Respondents answered questions based on their perceptions, experiences, and understanding. Thus, the data collected may not be adequately objective. However, due to the nature of this study (exploratory research), the quality of the collected data is acceptable. Furthermore, this study was based on a sample of voluntary participants. This type of recruitment is not considered to have negatively affected the results because this approach is commonly used in the field.

Acknowledgments

The authors sincerely extend their gratitude and recognition to the Ministry of Science and Technology of Taiwan (MOST109-2410-H-041-001) and Chia-Nan University of Pharmacy and Science of Taiwan (CN10417) for funding this study.

Authors' Contributions

All authors participated sufficiently in the work to take public responsibility for appropriate portions of the content. RFC and JLH were responsible for study conception and design. JLH performed data acquisition. RFC and JLH performed analysis and interpretation of data. RFC and JLH drafted the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire for EMR infusion and performance.
[DOCX File , 17 KB - medinform_v9i11e32180_app1.docx]

References

1. Ayatollahi H, Bath PA, Goodacre S. Paper-based versus computer-based records in the emergency department: staff preferences, expectations, and concerns. Health Informatics J 2009 Sep;15(3):199-211 [FREE Full text] [doi: 10.1177/1460458209337433] [Medline: 19713395]
2. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Health Serv Res 2010 Aug 06;10:231 [FREE Full text] [doi: 10.1186/1472-6963-10-231] [Medline: 20691097]
3. Raymond L, Paré G, Marchand M. Extended use of electronic health records by primary care physicians: Does the electronic health record artefact matter? Health Informatics J 2017 Apr 01;25(1):146045821770424482 [FREE Full text] [doi: 10.1177/1460458217704244] [Medline: 28434279]
4. Angst CM, Devaraj S, D'Arcy J. Dual role of IT-assisted communication in patient care: a validated structure-process-outcome framework. J Manag Inf Syst 2014 Dec 08;29(2):257-292. [doi: 10.2753/mis0742-1222290209]
5. Garets D, Davis M. Electronic Medical Records vs Electronic Health Records: Yes, There Is a Difference. A HIMSS Analytics White paper. Teh and Associates. URL: https://www.tehandassociates.com/wp-content/uploads/2017/03/WP_EMR_EHR.pdf [accessed 2021-07-17]
6. Klar R. Selected impressions on the beginning of the electronic medical record and patient information. Methods Inf Med 2004;43(5):537-542. [Medline: 15702216]
7. The introduction of EMR promotion in Taiwan. Ministry of Health and Welfare. 2015. URL: https://emr.mohw.gov.tw/emr/introduction.aspx [accessed 2021-07-17]
8. Wen H, Chang W, Hsu M, Ho C, Chu C. An assessment of the interoperability of electronic health record exchanges among hospitals and clinics in Taiwan. JMIR Med Inform 2019 Mar 28;7(1):e12630 [FREE Full text] [doi: 10.2196/12630] [Medline: 30920376]
9. Yamamoto LG, Khan ANGA. Challenges of electronic medical record implementation in the emergency department. Pediatr Emerg Care 2006 Mar;22(3):184-91; quiz 192. [doi: 10.1097/01.pec.0000203821.02045.69] [Medline: 16628105]
10. Goodhue D, Thompson R. Task-technology fit and individual performance. MIS Quart 1995 Jun;19(2):213-236. [doi: 10.2307/249689]
11. DeLone WH, McLean ER. Information systems success: the quest for the dependent variable. Inf Syst Res 1992 Mar;3(1):60-95. [doi: 10.1287/isre.3.1.60]
12. DeLone WH, McLean ER. The DeLone and McLean Model of information systems success: a ten-year update. J Manag Inf Syst 2014 Dec 23;19(4):9-30. [doi: 10.1080/07421222.2003.11045748]
13. Aral S, Brynjolfsson E, Van AM. Information, technology,information worker productivity: Task-level evidence. 2006 Presented at: ICIS2006 Proceedings; 2006; Milwaukee, WI p. 285-306. [doi: 10.3386/w13172]
14. Kim H, Gupta S. A user empowerment approach to information systems infusion. IEEE Trans Eng Manage 2014 Nov;61(4):656-668. [doi: 10.1109/tem.2014.2354693]
15. Jasperson J, Carter PE, Zmud RW. A comprehensive conceptualization of post-adoptive behaviors associated with information technology enabled work systems. MIS Quart 2005;29(3):525. [doi: 10.2307/25148694]

16. Marcolin BL, Compeau DR, Munro MC, Huff SL. Assessing user competence: conceptualization and measurement. Inf Syst Res 2000 Mar;11(1):37-60. [doi: 10.1287/isre.11.1.37.11782]
17. Venkatesh V, Morris M, Davis G, Davis F. User acceptance of information technology: toward a unified view. MIS Quart 2003;27(3):425-478. [doi: 10.2307/30036540]
18. Cooper RB, Zmud RW. Information technology implementation research: a technological diffusion approach. Manag Sci 1990 Feb;36(2):123-139. [doi: 10.1287/mnsc.36.2.123]
19. Hsieh P, Rai A, Xu S. Extracting business value from IT: a sensemaking perspective of post-adoptive use. Manag Sci 2011 Nov;57(11):2018-2039. [doi: 10.1287/mnsc.1110.1398]
20. Saga V, Zmud R. The nature determinants of IT acceptance, routinization and infusion. In: Levine L, editor. Diffusion, transfer and implementation of information technology. Amsterdam: Elsevier; 1994:67-86.
21. O'Connor Y, O’Rahailligh PJ, O’Donoghue J. Individual infusion of m-health technologies: Determinants and outcomes. 2012 Presented at: 20th European Conference on Information Systems; 2012; Barcelona, Spain p. 64.
22. O'Connor Y, O'Reilly P, O'Donoghue J. M-health infusion by healthcare practitioners in the national health services (NHS). Health Policy Technol 2013 Mar;2(1):26-35. [doi: 10.1016/j.hlpt.2012.12.002]
23. Tennant V, Mills A, Chin W. Investigating information system infusion at the individual level: re-conceptualisation and operationalization. 2011 Jul 7 Presented at: 15th Pacific Asia Conference on Information Systems; 2011; Brisbane, Australia URL: https://aisel.aisnet.org/pacis2011/189
24. Hassandoust F, Techatassanasoontorn AA, Tan FB. Factors influencing the infusion of information systems: a literature review. Pacif Asia J Assoc Inf Syst 2016;8(1):1-32. [doi: 10.17705/1pais.08101]
25. Davidson E, Heslinga D. Bridging the IT adoption gap for small physician practices: an action research study on electronic health records. Inf Syst Manag 2006 Dec 22;24(1):15-28. [doi: 10.1080/10580530601036786]
26. DesRoches CM, Campbell EG, Rao SR, Donelan K, Ferris TG, Jha A, et al. Electronic health records in ambulatory care--a national survey of physicians. N Engl J Med 2008 Jul 03;359(1):50-60. [doi: 10.1056/NEJMsa0802005] [Medline: 18565855]
27. Jung SY, Hwang H, Lee K, Lee H, Kim E, Kim M, et al. Barriers and facilitators to implementation of medication decision support systems in electronic medical records: mixed methods approach based on structural equation modeling and qualitative analysis. JMIR Med Inform 2020 Jul 22;8(7):e18758 [FREE Full text] [doi: 10.2196/18758] [Medline: 32706717]
28. 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] [Medline: 26831123]
29. B. Meinert D. Resistance to Electronic Medical Records (EMRs): A Barrier to Improved Quality of Care. Issues Informing Sci Inf Technol 2005;2:494-504. [doi: 10.28945/846]
30. Miller RH, Sim I. Physicians' use of electronic medical records: barriers and solutions. Health Aff (Millwood) 2004;23(2):116-126. [doi: 10.1377/hlthaff.23.2.116] [Medline: 15046136]
31. Mohd H, Mohamad S. Acceptance model of electronic medical record. J Adv Inf Manag Stud 2005 Jun;2(1):75-92.
32. Randeree E. Exploring physician adoption of EMRs: a multi-case analysis. J Med Syst 2007 Dec;31(6):489-496. [doi: 10.1007/s10916-007-9089-5] [Medline: 18041282]
33. Trudel M, Marsan J, Paré G, Raymond L, Ortiz de Guinea A, Maillet É, et al. Ceiling effect in EMR system assimilation: a multiple case study in primary care family practices. BMC Med Inform Decis Mak 2017 Apr 20;17(1):46 [FREE Full text] [doi: 10.1186/s12911-017-0445-1] [Medline: 28427405]
34. Raymond L, Paré G, Ortiz de Guinea A, Poba-Nzaou P, Trudel M, Marsan J, et al. Improving performance in medical practices through the extended use of electronic medical record systems: a survey of Canadian family physicians. BMC Med Inform Decis Mak 2015 Apr 14;15:27 [FREE Full text] [doi: 10.1186/s12911-015-0152-8] [Medline: 25888991]
35. Bhattacherjee A. Understanding information systems continuance: an expectation-confirmation model. MIS Quart 2001 Sep;25(3):351-370. [doi: 10.2307/3250921]
36. Ng E, Kim H. Investigating information systems infusion and the moderating role of habit: A user empowerment perspective. 2009 Presented at: ICIS 2009 Proceedings; 2009; Phoenix, AZ URL: https://aisel.aisnet.org/icis2009/137
37. Health Care Index by Country 2020. Numbeo. 2020. URL: https://www.numbeo.com/health-care/rankings_by_country.jsp?title=2020&displayColumn=0 [accessed 2021-07-17]
38. Armstrong CP, Sambamurthy V. Information technology assimilation in firms: the influence of senior leadership and IT infrastructures. Inf Syst Res 1999 Dec;10(4):304-327. [doi: 10.1287/isre.10.4.304]
39. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quart 1989 Sep;13(3):319-340. [doi: 10.2307/249008]
40. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag Sci 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
41. Jones E, Sundaram S, Chin W. Factors leading to sales force automation use: a longitudinal analysis. J Pers Sell Sales Manag 2002;XXII(3):145-156. [doi: 10.1080/08853134.2002.10754303]
42. Wang W, Hsieh P. Beyond routine: Symbolic adoption, extended use, and emergent use of complex information systems in the mandatory organizational context. 2006 Presented at: ICIS 2006; 2006; Milwaukee, WI URL: http://aisel.aisnet.org/icis2006/48

43. Zmud RW, Apple LE. Measuring technology incorporation/infusion. J Product Innov Manag 1992 Jun;9(2):148-155. [doi: 10.1111/1540-5885.920148]
44. Burton-Jones A, Straub DW. Reconceptualizing system usage: an approach and empirical test. Inf Syst Res 2006 Sep;17(3):228-246. [doi: 10.1287/isre.1060.0096]
45. Hsiao J, Chen R. An investigation on task-technology fit of mobile nursing information systems for nursing performance. Comput Inform Nurs 2012 May;30(5):265-273. [doi: 10.1097/NCN.0b013e31823eb82c] [Medline: 22156768]
46. Lin TC. Mobile nursing information system utilization: the task-technology fit perspective. Comput Inform Nurs 2014 Mar;32(3):129-137. [doi: 10.1097/CIN.0000000000000039] [Medline: 24419090]
47. Abbas SK, Hassan HA, Asif J, Ahmed B, Hassan F, Haider SS. Integration of TTF, UTAUT, and ITM for mobile banking adoption. Int J Manag Sci Eng Manag 2018;4(5):375-379. [doi: 10.22161/ijaems.4.5.6]
48. Dishaw MT, Strong DM. Extending the technology acceptance model with task–technology fit constructs. Inf Manag 1999 Jul;36(1):9-21. [doi: 10.1016/S0378-7206(98)00101-3]
49. Shih Y, Chen C. The study of behavioral intention for mobile commerce: via integrated model of TAM and TTF. Qual Quant 2011 Aug 30;47(2):1009-1020. [doi: 10.1007/s11135-011-9579-x]
50. Hsiao JL, Chen RF. Understanding determinants of health care professionals' perspectives on mobile health continuance and performance. JMIR Med Inform 2019 Mar 18;7(1):e12350 [FREE Full text] [doi: 10.2196/12350] [Medline: 30882353]
51. Lim SY, Jarvenpaa SL, Lanham HJ. Barriers to interorganizational knowledge transfer in post-hospital care transitions: review and directions for information systems research. J Manag Inf Syst 2015 Dec 16;32(3):48-74. [doi: 10.1080/07421222.2015.1095013]
52. Kellermann AL, Jones SS. What it will take to achieve the as-yet-unfulfilled promises of health information technology. Health Aff (Millwood) 2013 Jan;32(1):63-68. [doi: 10.1377/hlthaff.2012.0693] [Medline: 23297272]
53. Mennemeyer ST, Menachemi N, Rahurkar S, Ford EW. Impact of the HITECH Act on physicians' adoption of electronic health records. J Am Med Inform Assoc 2016 Mar;23(2):375-379 [FREE Full text] [doi: 10.1093/jamia/ocv103] [Medline: 26228764]
54. Seymour T, Frantsvog D, Graeber T. Electronic health records (EHR). Am J Health Sci 2012 Jul 13;3(3):201-210. [doi: 10.19030/ajhs.v3i3.7139]
55. O'Donnell A, Kaner E, Shaw C, Haighton C. Primary care physicians' attitudes to the adoption of electronic medical records: a systematic review and evidence synthesis using the clinical adoption framework. BMC Med Inform Decis Mak 2018 Nov 13;18(1):101 [FREE Full text] [doi: 10.1186/s12911-018-0703-x] [Medline: 30424758]
56. Ahmed MH, Bogale AD, Tilahun B, Kalayou MH, Klein J, Mengiste SA, et al. Intention to use electronic medical record and its predictors among health care providers at referral hospitals, north-West Ethiopia, 2019: using unified theory of acceptance and use technology 2(UTAUT2) model. BMC Med Inform Decis Mak 2020 Sep 03;20(1):207 [FREE Full text] [doi: 10.1186/s12911-020-01222-x] [Medline: 32883267]
57. Duncan T. An examination of physician resistance related to electronic medical records adoption. Thesis Doctor of Business Administration. Walden Dissertations and Doctoral Studies. 2015. URL: http://scholarworks.waldenu.edu/dissertations/1257/ [accessed 2021-07-17]
58. Seeman E, Gibson, S. S. Predicting acceptance of electronic medical records: Is the technology acceptance model enough? SAM Adv Manag J 2009;74(4):21.
59. Sayyah Gilani M, Iranmanesh M, Nikbin D, Zailani S. EMR continuance usage intention of healthcare professionals. Inform Health Soc Care 2017 Mar;42(2):153-165. [doi: 10.3109/17538157.2016.1160245] [Medline: 27100821]
60. Duncan T, Rahim E. Challenges in healthcare post-EMR adoption. 2018 May Presented at: 13th Midwest United States Association for Information Systems; 2018; Missouri, USA URL: https://aisel.aisnet.org/mwais2018/9
61. Goh JM, Gao G, Agarwal R. Evolving work routines: adaptive routinization of information technology in healthcare. Inf Syst Res 2011 Sep;22(3):565-585. [doi: 10.1287/isre.1110.0365]
62. Chen RF, Hsiao JL. An investigation on physicians' acceptance of hospital information systems: a case study. Int J Med Inform 2012 Dec;81(12):810-820. [doi: 10.1016/j.ijmedinf.2012.05.003] [Medline: 22652011]
63. Hsiao JL, Chang HC, Chen RF. A study of factors affecting acceptance of hospital information systems: a nursing perspective. J Nurs Res 2011 Jun;19(2):150-160. [doi: 10.1097/JNR.0b013e31821cbb25] [Medline: 21586992]
64. Rai A, Chen L, Pye J, Baird A. Understanding determinants of consumer mobile health usage intentions, assimilation, and channel preferences. J Med Internet Res 2013 Aug 02;15(8):e149 [FREE Full text] [doi: 10.2196/jmir.2635] [Medline: 23912839]
65. La Torre G, De Leonardis V, Chiappetta M. Technostress: how does it affect the productivity and life of an individual? Results of an observational study. Public Health 2020 Dec;189:60-65. [doi: 10.1016/j.puhe.2020.09.013] [Medline: 33166856]
66. Gebauer J, Shaw MJ, Gribbins ML. Task-technology fit for mobile information systems. J Inf Technol 2010 Sep 01;25(3):259-272. [doi: 10.1057/jit.2010.10]
67. Liang TP, Huang CW, Yeh YH, Lin B. Adoption of mobile technology in business: a fit ‐ viability model. Industr Mngmnt Data Systems 2007 Oct 02;107(8):1154-1169. [doi: 10.1108/02635570710822796]
68. Agarwal R, Prasad J. A conceptual and operational definition of personal innovativeness in the domain of information technology. Inf Syst Res 1998 Jun;9(2):204-215. [doi: 10.1287/isre.9.2.204]

69. Thatcher JB, Perrewe PL. An empirical examination of individual traits as antecedents to computer anxiety and computer self-efficacy. MIS Quart 2002 Dec;26(4):381-396. [doi: 10.2307/4132314]
70. Ragu-Nathan TS, Tarafdar M, Ragu-Nathan BS, Tu Q. The consequences of technostress for end users in organizations: conceptual development and empirical validation. Inf Syst Res 2008 Dec;19(4):417-433. [doi: 10.1287/isre.1070.0165]
71. Tu Q, Wang K, Shu Q. Computer-related technostress in China. Commun ACM 2005 Apr;48(4):77-81. [doi: 10.1145/1053291.1053323]
72. Limayem M, Hirt SG, Cheung CM. How habit limits the predictive power of intention: the case of information systems continuance. MIS Quart 2007;31(4):705-737. [doi: 10.2307/25148817]
73. Venkatesh V, Thong JY, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Quart 2012;36(1):157-178. [doi: 10.2307/41410412]
74. Junglas I, Abraham C, Ives B. Mobile technology at the frontlines of patient care: Understanding fit and human drives in utilization decisions and performance. Decis Support Syst 2009 Feb;46(3):634-647. [doi: 10.1016/j.dss.2008.11.012]
75. Podsakoff PM, MacKenzie SB, Lee J, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003 Oct;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
76. Petrick JF. Development of a multi-dimensional scale for measuring the perceived value of a service. J Leisure Res 2017 Dec 13;34(2):119-134. [doi: 10.1080/00222216.2002.11949965]
77. Kock N. WarpPLS 5.0 user manual. ScriptWarp Systems. 2015. URL: http://cits.tamiu.edu/WarpPLS/UserManual_v_5_0.pdf [accessed 2021-07-17]
78. Lowry PB, Gaskin J. Partial least squares (PLS) structural equation modeling (SEM) for building and testing behavioral causal theory: when to choose it and how to use it. IEEE Trans Profess Commun 2014 Jun;57(2):123-146. [doi: 10.1109/tpc.2014.2312452]
79. Chin WW. Bootstrap cross-validation indices for PLS path model assessment. In: Esposito V, Chin WW, Henseler J, Wang H, editors. Handbook of partial least squares concepts, methods and applications. Berlin, Germany: Springer Nature; 2010:83-97.
80. Kock N. Common method bias in PLS-SEM: a full collinearity assessment approach. Int J e-Collab 2015;11(4):1-10. [doi: 10.4018/ijec.2015100101]
81. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Market Res 1981 Feb;18(1):39-50. [doi: 10.2307/3151312]
82. Hair Jr J, Sarstedt M, Hopkins L, Kuppelwieser VG. Partial least squares structural equation modeling (PLS-SEM): an emerging tool in business research. Eur Bus Rev 2014 Mar 04;26(2):106-121. [doi: 10.1108/EBR-10-2013-0128]
83. Gefen D. TAM or just plain habit: a look at experienced online shoppers. J Organ End User Comp 2003;15(3):1-13. [doi: 10.4018/joeuc.2003070101]
84. Ouellette JA, Wood W. Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. Psychol Bull 1998 Jul;124(1):54-74. [doi: 10.1037/0033-2909.124.1.54]
85. Kim J, Lee Y, Lim S, Kim JH, Lee B, Lee J. What clinical information is valuable to doctors using mobile electronic medical records and when? J Med Internet Res 2017 Oct 18;19(10):e340 [FREE Full text] [doi: 10.2196/jmir.8128] [Medline: 29046269]

Abbreviations

AARS: average adjusted R2
AFVIF: average full collinearity variance inflation factor
APC: average path coefficient
ARS: average R2
AVE: average variance extracted
AVIF: average block variance inflation factor
CR: composite reliability
CVI: content validity index
ECM: expectation confirmation model
EHR: electronic health record
EMR: electronic medical record
GoF: Tenenhaus Goodness of Fit
HCP: health care professional
HIS: health information system
HIT: health information technology
IRB: institutional review board
IS: information system
IT: information technology

mHealth: mobile health
PLS: partial least squares
RSCR: R2 contribution ratio
SEM: structural equation model
TAM: technology acceptance model
TTF: task-technology fit
UTAUT: unified theory of acceptance and use of technology

Edited by G Eysenbach; submitted 18.07.21; peer-reviewed by L Raymond; comments to author 10.08.21; revised version received 07.10.21; accepted 17.10.21; published 30.11.21.

Please cite as:
Chen RF, Hsiao JL
Health Professionals' Perspectives on Electronic Medical Record Infusion and Individual Performance: Model Development and Questionnaire Survey Study
JMIR Med Inform 2021;9(11):e32180
URL: https://medinform.jmir.org/2021/11/e32180
doi:10.2196/32180
PMID:34851297

Original Paper

The Collaborative Metadata Repository (CoMetaR) Web App: Quantitative and Qualitative Usability Evaluation

Mark R Stöhr[1]; Andreas Günther[1], Prof Dr; Raphael W Majeed[1]

Justus-Liebig-University Giessen, Universities of Giessen and Marburg Lung Center (UGMLC), German Center for Lung Research (DZL), Gießen, Germany

Corresponding Author:
Mark R Stöhr
Justus-Liebig-University Giessen
Universities of Giessen and Marburg Lung Center (UGMLC)
German Center for Lung Research (DZL)
Klinikstraße 36
Gießen, 35392
Germany
Phone: 49 641 985 42117
Email: mark.stoehr@innere.med.uni-giessen.de

Abstract

Background: In the field of medicine and medical informatics, the importance of comprehensive metadata has long been recognized, and the composition of metadata has become its own field of profession and research. To ensure sustainable and meaningful metadata are maintained, standards and guidelines such as the FAIR (Findability, Accessibility, Interoperability, Reusability) principles have been published. The compilation and maintenance of metadata is performed by field experts supported by metadata management apps. The usability of these apps, for example, in terms of ease of use, efficiency, and error tolerance, crucially determines their benefit to those interested in the data.

Objective: This study aims to provide a metadata management app with high usability that assists scientists in compiling and using rich metadata. We aim to evaluate our recently developed interactive web app for our collaborative metadata repository (CoMetaR). This study reflects how real users perceive the app by assessing usability scores and explicit usability issues.

Methods: We evaluated the CoMetaR web app by measuring the usability of 3 modules: *core module*, *provenance module*, and *data integration module*. We defined 10 tasks in which users must acquire information specific to their user role. The participants were asked to complete the tasks in a live web meeting. We used the System Usability Scale questionnaire to measure the usability of the app. For qualitative analysis, we applied a modified think aloud method with the following thematic analysis and categorization into the ISO 9241-110 usability categories.

Results: A total of 12 individuals participated in the study. We found that over 97% (85/88) of all the tasks were completed successfully. We measured usability scores of 81, 81, and 72 for the 3 evaluated modules. The qualitative analysis resulted in 24 issues with the app.

Conclusions: A usability score of 81 implies very good usability for the 2 modules, whereas a usability score of 72 still indicates acceptable usability for the third module. We identified 24 issues that serve as starting points for further development. Our method proved to be effective and efficient in terms of effort and outcome. It can be adapted to evaluate apps within the medical informatics field and potentially beyond.

(JMIR Med Inform 2021;9(11):e30308) doi:10.2196/30308

KEYWORDS

usability; metadata; data visualization; semantic web; data management; data warehousing; communication barriers; quality improvement; biological ontologies; data curation

Introduction

The Importance of Metadata

Raw data are useless without metadata that characterizes and contextualizes its content. A number is meaningless without the information on which parameter it describes (eg, blood pressure) and a finding is of no use without its context (eg, sepsis as a comorbidity vs sepsis as cause of death). Metadata itself always needs context (eg, the concept it describes). In many cases, metadata are merely implied by column headers of tabular databases and the implicit knowledge of the few people working with the database. Many information scientists have researched the field of metadata, for example, Wilkinson et al [1], who published the FAIR (Findability, Accessibility, Interoperability, Reusability) principles, which is a guideline for well-designed metadata. Whenever data are reused (for analysis, validity checks, etc), the corresponding metadata must be attached to the actual data. Thus, explicitly formulated, rich, and comprehensive metadata are indispensable for any sustainable research project [2]. At present, most data processing is done automatically by computers, which necessitates all metadata to be available in machine-readable form [3]. In addition to data processing, metadata are used to describe data sets to a broader audience, such as the national or international research community. BioPortal [4], for example, is a comprehensive repository of biomedical ontologies interconnecting researchers globally. In addition, there are approaches for recording the variety of existing data and metadata repositories in public registers [5,6].

Metadata in the Field of Data Integration

Overview

Particularly in the context of data integration within large research networks, comprehensive metadata are essential. “Data integration is the problem of combining data residing at different sources, and providing the user with a unified view of these data” [7]. Although the process of exporting, transforming, and loading data is a huge task, this *unified view* is an achievement by itself. In medical informatics, the purpose of data integration is to promote translational research and to have access to a larger data pool for retrospective data analysis and prospective patient recruitment. The amount of integrated data and the way they are presented to users determine their acceptance and accessibility. If too few concepts are covered by a repository or if too few instances of data are integrated, researchers have no sufficient basis for analysis. If metadata are not presented accessibly, users will presume app shortcomings rather than investing in exploration time. This applies especially to entry-level users and, in most cases, yields in rejection of the software.

Data Integration: Main Components and Roles

Software-driven data integration involves multiple technical components: various *heterogeneous source databases* are harmonized and integrated into a *collective data repository*. All affected parameters, more precisely the canonical concepts behind these parameters, are annotated in a separate *metadata repository*. Both repositories are linked through identifiers [8-10]. *Configuration files* define the harmonization process of different source database schemata into a target schema. These configuration files vary in format and syntax, but all of them are written in a formal computer-readable language [11-14].

From the user perspective, these components are managed and elaborated by the following roles: *data providers* know the meaning of their data and its acquisition processes. In medical informatics, this knowledge is essential for data harmonization, because labels such as column names or form labels are not always sufficiently specific. According to Nadkarni and Marenco [15], “[...] column names may be quasi-gibberish, heavily abbreviated, and their names may follow arbitrary conventions that are idiosyncratic to the system designer or organization.” Rahm and Bernstein [16] showed that even automatic schema matching can only provide mapping candidates. The formulation of mapping rules is performed by the *local and central data managers* (responsible for the source databases and collective database repository, respectively) as they have the required technical background to maintain the formally written configuration files. *Data coordinators* elaborate the metadata repository content, incorporating multiple studies and registers with varying scopes and the focus of research. This process includes rating for relevance, harmonization, annotation, curation, and clustering. The clustering and hierarchical organization of metadata have a direct impact on the presentation of user interfaces. It determines how intuitively information can be found and used.

Information Access Barriers

To provide a data warehouse with comprehensive and accurate data, different roles need access to different classes of information residing in the described data integration system. We identified 3 cases in which access barriers prevent users from contributing their expertise [17,18]:

1. All users need access to the listing of all data elements represented in the data warehouse. These annotations and context information can be derived from the metadata repository and must be visualized.
2. Data managers and, in particular, data providers need full access to the mapping rules for data harmonization. They are only available in the formal language, which requires the respective information technology background. Data providers usually do not have that knowledge.
3. Data coordinators need access to the provenance information of the metadata to be able to curate it. “Especially in collaborative metadata development, a comprehensive annotation about ‘who contributed what, when and why’ is essential” [17].

In most cases, barrier (1) is resolved through metadata browsers [4,19,20]. For metadata repositories in the context of data integration, barriers (2) and (3) often form a huge gap between users and the required information.

The Implementation of Collaborative Metadata Repository

The German Center for Lung Research (German: Deutsches Zentrum für Lungenforschung [DZL]) implemented the collaborative metadata repository (CoMetaR), applying principles of collaborative metadata development and FAIR metadata warehousing [1,17,18,21]. It is based on open and commonly used standards. The DZL metadata constitutes a highly specified thesaurus specifically developed for lung research, and till July 2021, it contains 3.474 distinct concepts. CoMetaR supports storing a single thesaurus in the Resource Description Framework (RDF) format based on the Simple Knowledge Organization System (SKOS) and Dublin Core (DC) knowledge organization systems [22-24]. The ISO/IEC 21526 [25] standard explicitly "mandates the use of SKOS to provide user-interface surfaced content classification." Versioning occurs via Git, which also provides information about the changes among different versions [26,27]. The latest thesaurus version is loaded in a triple store and accessible through the SPARQL Protocol and RDF Query Language (SPARQL) interface [28]. This interface can be used to extract metadata information and, as in our case, to set up tree-like metadata in a data warehouse [29]. The extracted metadata information can also be used to generate a visual metadata representation similar to our user front end, the CoMetaR web app. This front end was developed to dissolve access barriers for all user roles and thereby support them in contributing to their expertise. However, it has yet to be proven scientifically that the CoMetaR web app meets the requirements for metadata management and data integration support.

This study evaluates the usability of 3 modules built for common tasks in the field of data integration and metadata maintenance.

Methods

Study Design

Overview

The usability evaluation performed was a combination of (1) the think aloud method and (2) usability questionnaires. By combining both methods, we wanted to measure both observable and perceived usability. The execution consisted of two phases: (1) a screen sharing–supported training specific to the respective user's roles and (2) solving of the given tasks by the participant with subsequent retrospection, including the completion of a usability questionnaire. All evaluations were performed by the same experimenter.

The Think Aloud Method

This method is commonly applied to the usability evaluations of web interfaces [30,31]. The idea behind the think aloud method is that participants verbalize their thoughts while performing given tasks. Their expressions were recorded and later transcribed and analyzed according to an interpretation model.

We decided not to record the participants but to make notes on their expressions as well as their app use behavior. These notes focused on usability, functional, and methodological issues. The advantage of this approach is a more comfortable setting for the user on the one hand and less effort for the experimenter on the other hand. The downside is the potential information loss because the experimenter already filters information.

As our interpretation model, we used the 7 categories described in ISO 9241-110 [32]: suitability for the task, conformity with user expectations, suitability for learning, suitability for individualization, self-descriptiveness, controllability, and error tolerance.

System Usability Scale

We used the System Usability Scale invented by Brooke in 1996 as a measurement tool for the usability of the app. This scale was introduced as a *quick and dirty* but a meaningful measurement tool for user experience [33,34]. It consists of 10 questions answered on a scale from 0 to 4. All questions are available in multiple languages, including German, which we used for our evaluation.

Materials

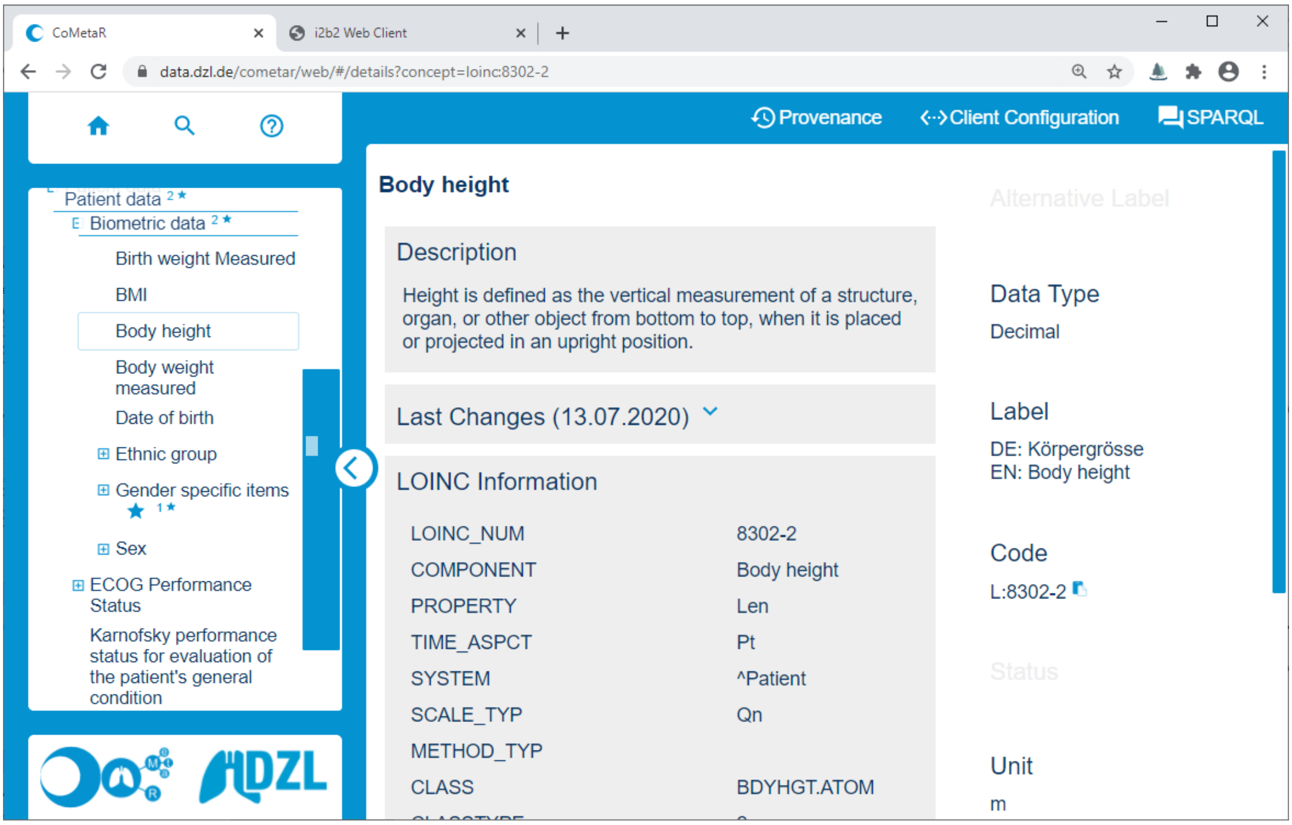
CoMetaR Modules

The CoMetaR web app is divided into a concept tree navigation area and a module area. Modules can be selected in the module menu in the top-right corner, as shown in Figure 1. In the following paragraphs, we will briefly describe the functionality of the 3 evaluated modules: the *core module*, *provenance module*, and *data integration module*. In the *Introduction* section, we described 3 user roles involved in the data integration process: data managers, data providers, and data coordinators. A user may perform more than one role. Each role makes use of the core module, whereas the data integration module and provenance module are more role-specific (see the Tasks section).

The core module functionality of the CoMetaR web app (Figure 1) involves browsing through all metadata concepts and showing the corresponding detailed information. Users can navigate the concept tree by expanding the nodes and retrieving details by clicking them. They can also use the search function to check if and where a concept is located in the thesaurus. Concept details are shown in the module area. They include core information like labels, alternative labels, data type, code, status (*on draft* yes or no), and unit. In addition, we present the author, description, and concept specifications. A dedicated panel shows the history of all changes that have been made to the selected concept. A button allows the export of the concept and all of its subconcepts with basic information in the CSV format.

As our metadata are growing and developing over time with many participants involved, we decided to provide the provenance module, which enables users to track all changes. These changes may be the additions, moves, or removal of concepts in the concept tree, but also modifications of their annotations. When selecting the provenance module (Figure 2), the affected concept tree elements receive icons that symbolize their changes for a given timespan. The default timespan is 1 month from the current date and can be adjusted in the module. The module itself shows all dates concerned with metadata changes in vertical order. Horizontal bars attached to such a date represent single uploads, and their width indicates the amount of change. Clicking a date or a single upload bar loads the respective changes and shows them in the concept tree underneath the corresponding concepts.

Figure 1. Screenshot of the collaborative metadata repository (CoMetaR) web app core module. Left side: concept tree. Right side: module content (concept details). Top-right corner: module navigation. Top-left corner: home button, search panel, and help panel.



Body height

Alternative Label

Description

Height is defined as the vertical measurement of a structure, organ, or other object from bottom to top, when it is placed or projected in an upright position.

Last Changes (13.07.2020)

LOINC Information

LOINC_NUM	8302-2
COMPONENT	Body height
PROPERTY	Len
TIME_ASPCT	Pt
SYSTEM	^Patient
SCALE_TYP	Qn
METHOD_TYP	
CLASS	BDYHGT.ATOM

Data Type

Decimal

Label

DE: Körpergrösse
EN: Body height

Code

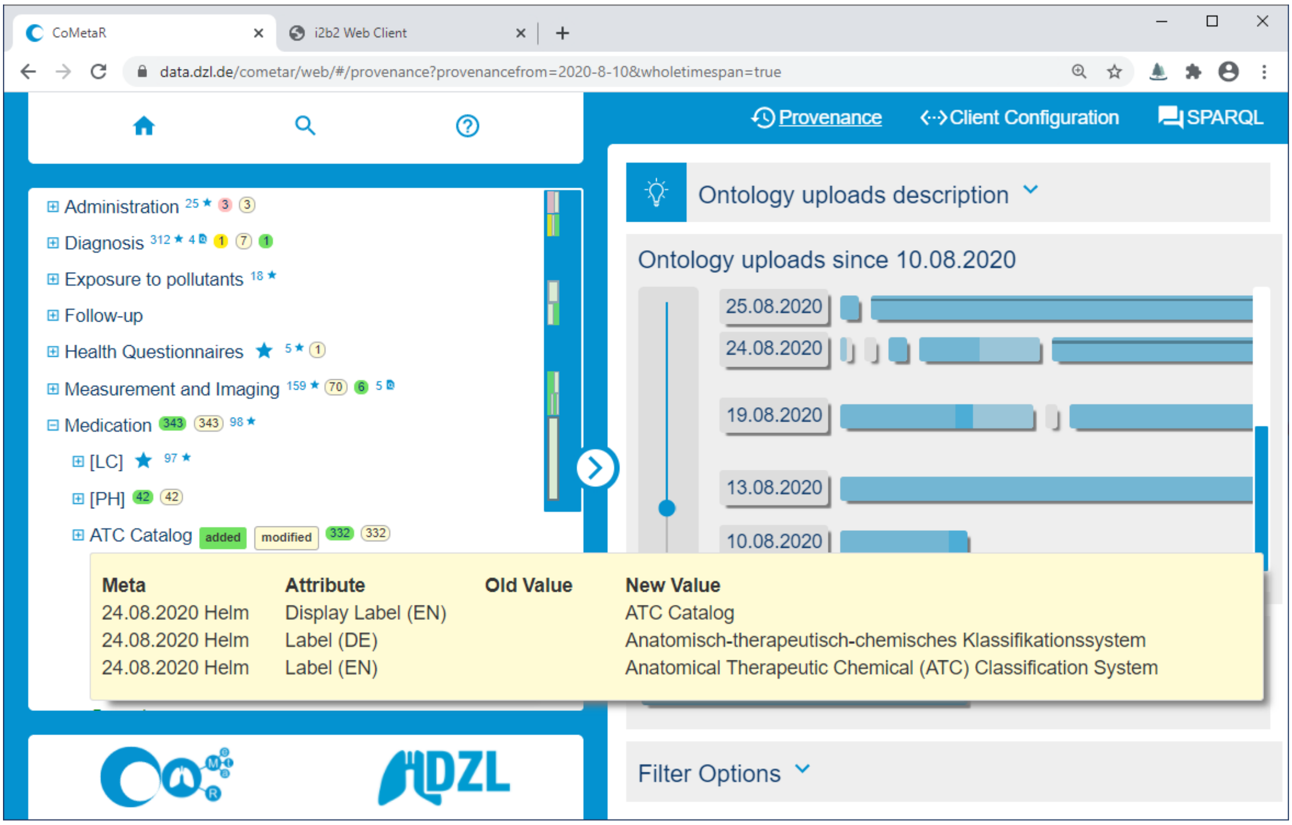
L:8302-2

Status

Unit

m

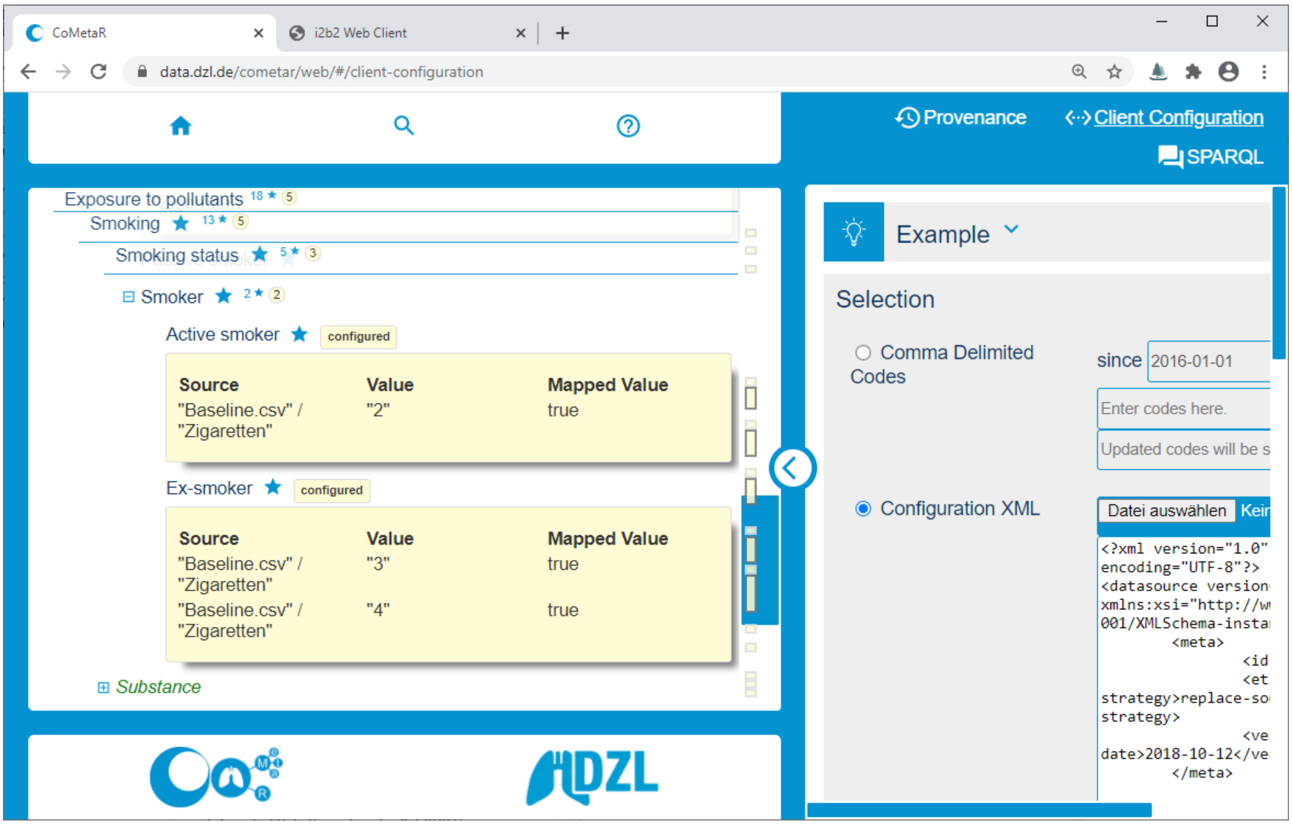
Figure 2. Screenshot of the collaborative metadata repository (CoMetaR) web app provenance module. Left side: concept tree with colorized annotations for added, moved, or removed and modified items. Light yellow box: information box for the item ATC Catalog on mouse-over. Right side: module content (upload history visualization). ATC: Anatomical Therapeutic Chemical.


CoMetaR
i2b2 Web Client
data.dzl.de/cometar/web/#/provenance?provenancefrom=2020-8-10&wholetimespan=true
Provenance
Client Configuration
SPARQL
Administration 25 ★ 3 3
Diagnosis 312 ★ 4 1 7 1
Exposure to pollutants 18 ★
Follow-up
Health Questionnaires ★ 5 ★ 1
Measurement and Imaging 159 ★ 70 6 5
Medication 343 343 98 ★
[LC] ★ 97 ★
[PH] 42 42
ATC Catalog added modified 332 332
Ontology uploads description
Ontology uploads since 10.08.2020
25.08.2020
24.08.2020
19.08.2020
13.08.2020
10.08.2020
Filter Options
DZL

Meta	Attribute	Old Value	New Value
24.08.2020 Helm	Display Label (EN)		ATC Catalog
24.08.2020 Helm	Label (DE)		Anatomisch-therapeutisch-chemisches Klassifikationssystem
24.08.2020 Helm	Label (EN)		Anatomical Therapeutic Chemical (ATC) Classification System

Our data integration process is supported by the data integration module. The integration process for a single data source is divided into 4 parts. (1) The export of data from the source system, (2) the preparation of data for the integration software, (3) configuration of the integration software, and (4) its execution. As the configuration file is written in formal language to be interpreted by software, it is not accessible for humans who lack the required technical background. To verify the configurations, the respective data providers must be able to access the formulated rules. For this task, they can upload the configuration file to the data integration module (Figure 3). All rules are then shown below the corresponding concept in the concept tree. In addition, we print notifications in the module area if any rule refers to a concept that does not exist (anymore) or that has been reintroduced. In such a case, the correct reference can be determined automatically, depending on the metadata's formal documentation. Subsequently, an updated version of the configuration file is offered for download. Note that this process does not invoke any kind of data upload; it is solely used to verify the configuration itself.

Figure 3. Screenshot of the collaborative metadata repository (CoMetaR) web app data integration module. Left side: concept tree. Light yellow boxes: corresponding mapping rules. Right side: module content (configuration file upload).



Source	Value	Mapped Value
"Baseline.csv" / "Zigaretten"	"2"	true

Source	Value	Mapped Value
"Baseline.csv" / "Zigaretten"	"3"	true
"Baseline.csv" / "Zigaretten"	"4"	true

since 2016-01-01

Enter codes here.

Updated codes will be s

```
<?xml version="1.0"
encoding="UTF-8"?>
<datasource version
xmlns:xsi="http://w
001/XMLSchema-insta
        <meta>
                <id
                <et
strategy>replace-so
strategy>
                <ve
date>2018-10-12</ve
        </meta>
```

Tasks

CoMetaR was designed to support data integration tasks. In the German Center for Lung Research, we have been practicing data integration since 2016 and identified information that is of high interest for data integration experts. For example, to match and map elements of the source data to the integrated data, the person formulating the rules needs to know which elements are part of the integrated metadata, what are their exact characteristics (method of measurement, scale, classification, etc), and how they are uniquely identified. If these characteristics change, the mapping rules must be adjusted. For various processes, people often want the metadata to be available in Microsoft Excel format, yielding the need for respective export capabilities. For these and further scenarios, we defined 10 tasks that verified CoMetaR's suitability in the field of lung research. The following tasks were composed by 2 experts, who have been internationally active in the field of data integration for >5 years. The composition process included brainstorming, discussion, and finally consensus. To assign modules to each participant, we considered their user roles as well as their everyday tasks. All users must solve core module tasks, all data coordinators must solve provenance module tasks, and all data managers who upload data must solve the data integration module tasks.

The first 4 tasks aim at the use of the core module. They test the ability to search for and find specific thesaurus elements and their annotations as well as the capability to export data:

1. Indicate which of the parameters *Never smoker* and *Opportunity smoker* are part of the DZL metadata.

2. Indicate code, datatype, and unit of the spirometry parameter Forced Expiratory Volume in 1 Second (*FEV1*) according to the metadata.

3. Regarding the last change of the concept *Comorbidities*, indicate its date and the modifications applied.

4. Describe in detail which individual steps you would take to print the subtree of *Biometric Data* in tabular form.

The following 2 tasks aim at the use of the provenance module. They test the ability to track changes within the thesaurus:

5. Indicate which concepts have been added, moved, or removed in the last month.

6. Pick one concept for which annotations have been changed in the last upload. Indicate who performed this change on which date.

The last 4 tasks aim at the use of the data integration module. They test the ability to verify individual upload client configurations:

7. Examine the configuration for falsely mapped concepts.

8. Examine the configuration for properly mapped concepts.

9. Examine the metadata for concepts that are not mapped in the configuration but you could provide.

10. Update your local configuration to meet changed concept references. Describe your approach.

Tasks 7, 8, and 9 must be seen as one task with 3 subtasks. The participants were asked to use their own configuration files designed for uploading the data they administered. Some configuration files comprise hundreds of mapping rules. Depending on the size and coverage of certain data sources, task fulfillment takes a considerable amount of time. During the live evaluation, the participants were asked to work on each of these 3 tasks exemplarily to be able to fill out the System Usability Scale questionnaire. They completed the tasks asynchronously and reported their results when they finished.

Configuration Files

For 3 of the 4 data integration module tasks, we asked the participants to use their own configuration file for analysis. These comprise rules to define how local concepts are mapped to concepts in the central data warehouse. The file format is XML. The configuration files are used by a data transformation and upload client software. Configuration files do not contain any instance data. By using real configuration files instead of an artificial example, we were able to test our app in a realistic scenario and identify faulty mappings. In addition, this setup allowed participants to work with familiar information.

Experimenter Notes

The experimenter completed a notes sheet alongside following the evaluation procedure. It was structured to contain one row per participant and the following columns: *Experience level*, *English level*, *age*, *profession*, *roles* (see the *Introduction* section), *evaluation date*, *training start timestamp*, *training finished timestamp*, *notes for training*. Each of the 3 modules contains the following columns: *module tasks* (stating whether tasks were solved successfully), *module finished timestamp*, *notes for module*, *timestamp module questionnaire filled*.

System Usability Scale Questionnaires

The questionnaires handed to the participants contained 10 usability questions defined in the System Usability Scale. They were put into a Microsoft Excel sheet with one row for each question and columns for values of 0 to 4. The final score for the 10 questions was calculated within the sheet. The participants were handed one sheet per evaluated module.

Quantitative Analysis Sheet

A spreadsheet was used to collect the scores per participant and module to calculate the quantitative analysis parameters, that is, *range from*, *range to*, *mean score*, and *SD*. These 4 parameters were additionally calculated with respect to the participant's experience level, using the following formula:

$$\text{Score weighted by experience} = \text{score} - 4 \times (\text{experience level}-1)$$

Given an experience level from 1 to 5, the score weighted by experience differs by up to 16 points, which corresponds to previous findings [35]. In addition to scores from the questionnaires, the corresponding experience levels, and the calculated values, no participant-related information was put into the sheet.

Setting

To evaluate our web app, we decided to interact with the participants remotely (participants were not invited to a local test laboratory) and synchronously (the evaluator and participant executed the test session in real time). We made one exception for a very time-consuming task type, which certain participants completed asynchronously. This method appeared to be the most efficient in terms of preparation effort, travel time, and risk of SARS-CoV-2 infection. Its suitability was shown in a comprehensive study: Bastien [36] summarized multiple studies stating that remote evaluations yield comparable results with a local laboratory evaluation. Although he found that automatic recording of every user interaction with the app can provide more insights about the app's usability, the setup is very time-consuming and would only be rational for larger participant numbers. The participants were approached in April and May of 2020. Data collection took place in May and June of 2020. Data analysis was conducted in July of 2020.

As a communication platform, we used the GoToMeeting web conference software by LogMeIn [37]. It allows participants to dial in via phone or software app. The latter also offers screen sharing capabilities, which all but one participant with technical issues were able to use.

Sampling

The target audience of CoMetaR is experts who contribute to the task of data integration as data providers, data managers, or data coordinators. Our implementation of CoMetaR is dedicated to lung research. Therefore, in this evaluation, we included members of the German Center for Lung Research and collaborating organizations. The included participants should cover a wide range of roles and responsibilities. These characteristics determine the module that they can work on effectively. For example, data managers who load data into a data warehouse have a data integration configuration file and can use the data integration module. The core module is relevant to all the user roles. In contrast, the provenance module is mostly relevant for data coordinators and data managers, whereas the data integration module is mostly relevant for data managers and data providers. In addition to their user role, profession, age, and English level, we also asked for the participants' experience with the app. English and experience levels were measured on a scale of 1 to 5.

Bastien [36] cited studies showing that most usability problems can be found in 5-15 participants. As Virzi [38] showed, only 4 to 5 participants were needed to identify about 80% of all usability issues, and this number is enough to reveal the most severe issues. Therefore, we planned to recruit at least 5 participants for each aspect of the web app. In total, we approached 13 potential participants, of whom 12 agreed to participate.

Ethical Considerations

All methods were performed in accordance with the relevant guidelines and regulations. This study was granted an exemption from requiring ethics approval by the ethics committee of the Faculty of Medicine at the Justus-Liebig-University in Giessen, Germany. Informed consent for participation in the study was obtained from all the participants.

All patient-related data were recorded anonymized. It covers age, profession, role, evaluated modules, English level, and experience with the app. The data were further coarsened using age classes of 10 years to prevent participant reidentification.

Procedure and Data Collection

Overview

Before any evaluation, we performed a screen sharing–supported training specific to the respective user's roles, regardless of previous experiences with the app. The goal of this training was to provide participants with equal basic knowledge about the web app's structure and functionality. We asked for the participants' previous experiences with the system, which may influence the evaluation outcome [35]. After the training, the participants shared their screens and completed the tasks given by the evaluator. After using the module, the participants filled out the System Usability Scale questionnaire. This also gave them the chance for retrospection and a short dialogue with the experimenter, potentially revealing more usability issues.

Instructions

After giving each participant introductory training regarding the app's functionalities, they had the option to ask questions and clarify misunderstandings. Following, for each tested module, they were asked to fulfill each task one by one. The tasks were communicated via speech. The experimenter asked the participants to verbalize their thoughts during the evaluation and reminded them whenever they forgot. After the participant solved the tasks for a module, the experimenter asked them to fill out the usability questionnaire we sent them previously via email. Furthermore, they were invited to participate in a retrospective dialogue, again noting the findings.

Role of the Experimenter

The experimenter played a passive role. During the evaluation, he was not supposed to speak besides reminding the participant to verbalize their thoughts. In cases where the participants were stuck, the experimenter gave hints to lead to the information that had to be received from the app. Meanwhile, the experimenter completed the structured notes sheet documenting the participants' verbalized thoughts, spontaneous reactions, and their app use behavior, focusing on the previously mentioned usability categories [32].

Recording and Transcription

The traditional think aloud method requires recording the entire evaluation session and the following transcription. As mentioned in the study design, we did not record sessions because transcription occurred during the session.

Analysis

Quantitative Analysis

For quantitative analysis, we calculated aggregated scores (*range from*, *range to*, *mean score*, and *SD*) based on the System Usability Scale questionnaires. We additionally calculated the same aggregations factoring in the experience level. This adjustment is motivated by previous findings, which show that usability scores vary up to 16 points based on the participant's experience level [35]. For example, a user with no experience (level 1) has the same base and adjusted score, whereas a user with a score of 70 and experience level 4 (of 5) has an adjusted score of 58. By calculating these moderated scores, we hope to obtain better insights into the app's usability, especially regarding entry barriers. All calculations were performed using Excel (see *Materials* section). We omitted subgroup analysis by English level, age, and profession as our sample size was too small.

Qualitative Analysis

We conducted a thematic analysis of the information gathered during the evaluations to identify usability issue patterns and to present a descriptive account of users' experiences. After familiarization with all notes, we went through all notes again and generated usability issue statements. We followed a latent approach, which means that we interpreted the data to create statements that were more meaningful. For example, task 2 asked the participants to indicate the properties of the spirometry parameter *FEV1*. In one case, a participant used the search function and entered *Spiro FEV1*, which led to no results (a note in the experimenter's structured notes file). Our conclusion is not that our app is unable to find a specific pattern but that users expect a more powerful search functionality, as is known from bigger internet companies (theme). After generating usability issue themes, we combined similar statements and reviewed them by checking if all notes were still well-represented by these statements. These were then assigned to 1 of the 7 usability categories described in ISO 9241-110 [32]: suitability for the task, conformity with user expectations, suitability for learning, suitability for individualization, self-descriptiveness, controllability, and error tolerance. The categorization was performed by the same person who underwent the evaluation sessions with all participants. Afterward, these groupings were discussed internally with another expert and potentially adjusted.

Software

For documentation and analysis, we used only Microsoft Excel and Microsoft Word.

Quality Assurance

The System Usability Scale questionnaire consists of 10 questions, 5 of which stated a positive usability and 5 of them stated negative usability. As some questions include negations, we assumed a possible misinterpretation. Therefore, we immediately checked each questionnaire for outliers and inquired when we identified potential misinterpretations. When inquiring, we again pointed out that we do not insist on better scores but on valid answers.

We wanted to ensure correct and comprehensive categorization, as well as unambiguous wording for qualitative analysis. A second person who was familiar with the study design and aspects of usability checked all categorizations. The resulting tables are the results of in-depth dialogues.

Results

Participants

All participants in this evaluation currently work for or in collaboration with the German Center for Lung Research. Their operation areas and responsibilities vary, but all contribute to the data integration task. Table 1 shows the details of all the 12 participants. They vary in age (28-63 years), experience with the system (1-4 on a scale of 1-5), English level (2-5 on a scale of 1-5), and profession (medical documentalists, medical informatics specialists, graduated biologists, bioinformatics specialists, study coordinators, and data managers).

Table 1. Characteristics of the 12 participants including age, experience level, English level, profession, user roles, and tested modules.

Characteristics	Participants											
	A	B	C	D	E	F	G	H	I	J	K	L
Age (years)	30-40	30-40	30-40	40-50	50-60	60-70	30-40	50-60	30-40	50-60	60-70	20-30
Experience level (1-5)	3	3	4	2	4	3	3	3	3	1	2	4
English level (1-5)	3	3	4	3	4	4	4	5	3	3	2	4
Profession	MD[a]	DM[b]	MI[c]	SC[d]	MD	GB[e]	MI	DM	DM	MD	MD	BI[f]
Has role data manager	✓[g]	✓	✓	✓				✓	✓	✓	✓	✓
Has role data provider	✓	✓	✓			✓						
Has role data coordinator	✓	✓	✓		✓	✓	✓					
Tested core module	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Tested provenance module	✓	✓	✓		✓	✓	✓	✓	✓	✓		✓
Tested data integration module	✓	✓	✓	✓					✓			

[a] MD: medical documentalist.

[b] DM: data manager.

[c] MI: medical informatics specialist.

[d] SC: study coordinator.

[e] GB: graduate biologist.

[f] BI: bioinformatics specialist.

[g] Characteristic present.

Quantitative Analysis

Time Expenditure

The training took between 10 and 30 minutes, depending on how many modules were presented and how many questions the participants had. After training, for task completion, the core module took between 8 and 26 (average 14, SD 6) minutes. The provenance module took between 3 and 20 (average 9, SD 5) minutes. The configuration module took between 21 and 51 (average 37, SD 12) minutes. Regarding the latter, we did not include the time spent asynchronously to complete the tasks.

Usability

Each participant solved the tasks of one or more CoMetaR modules (core module n=12, provenance module n=10, data integration module n=5). Subsequently, they completed one System Usability Scale questionnaire separately for each module. According to Bangor et al [39], a mean score of 50.9 or higher can be seen as *OK*, a mean score of 71.4 or higher is *Good*, and a mean score of 85.5 or higher is *Excellent*; a mean score of 70 or higher indicates that the interface is acceptable. The System Usability Scale score results are presented in Table 2. For *weighted by experience*, we subtracted up to 16 points based on the user’s own perceived experience.

Table 2. Aggregated System Usability Scale scores.

Module and score type	Values, mean (SD; range)
Core module	
Usability score	81.5 (9.1; 60.0-92.5)
Weighted by experience	73.8 (7.8; 60.0-84.5)
Provenance module	
Usability score	72.3 (16.0; 37.5-90.0)
Weighted by experience	63.9 (15.20; 37.5-79.5)
Data integration module	
Usability score	81.0 (9.9; 65.0-92.5)
Weighted by experience	73.0 (9.9; 57.0-84.5)

Functional Suitability

All the participants successfully solved all given tasks. In total, 12 participants solved 48 core module tasks, 10 participants solved 20 provenance module tasks, and 5 participants solved 20 data integration module tasks. In the case of task 2, 2 participants did not find the correct tree node and needed a hint. During the provenance module tasks, 1 participant lost track because he loaded too much information from multiple modules into the tree. He needed a hint to reset the app to solve task 5. In total, 97% (85/88) of all tasks were solved independently.

Qualitative Analysis

Our thematic analysis led to 24 usability issue themes, which covered all functional inadequacies and complications identified during the experiment. We grouped these themes into the 7 categories described in ISO 9241-110 (Textboxes 1-5). As the app does not offer possibilities for individualization, the respective category *suitability for individualization* does not appear in this evaluation. None of the observed issues were assigned to the category *controllability*.

Textbox 1. Issues in the category Suitability for the Task.

Core module

- Using the search function for *FEV1* shows more than 100 results because it is used as criterion for many diseases. Most of the results are located in the comorbidities-subtree.
- The help window does not help with task 2.

Provenance module

- The mouse-over tooltip of upload bars sometimes distracts and overlays other bars.
- Changing the selection of upload bars leads to changes in the concept tree. The system gives insufficient feedback that these changes were applied.

Textbox 2. Issues in the category Conformity with User Expectations.

Core module

- The search function only searches for fixed substrings and does not behave comparably to a mighty World Wide Web search engine. This might lead to incorrect conclusions whether a concept is part of the metadata.
- The users expected the fixed headings for the currently displayed subtree to be interactive.

Provenance module

- The provenance module disappears when clicking a tree element and the element's core information are shown instead.

Textbox 3. Issues in the category Suitability for Learning.

Core module

- An element’s change history is part of the core module and not the provenance module.
- Structural information for elements (added, moved, or removed) are not explicitly displayed in the element’s history (last changes).
- The number of search matches is not the number of matched concepts but of all matched attributes.
- Some annotations like *added* have rectangular representation in the minimap or outline and round-cornered representation in the tree.

Provenance module

- The structural annotations (added, moved, or removed) refer to the selected provenance timespan and not only to the selected uploads.
- It is not intuitive that a moved element’s old and new concept tree position are both selected when clicking one of them.

Textbox 4. Issues in the category Self-Descriptiveness.

Core module

- Many people search the code for *Forced Expiratory Volume in 1 Second* in the *Logical Observation Identifiers Names and Codes (LOINC)*–description instead of the concept’s core information.
- For some users, it is not intuitively clear that details for a tree node are shown when clicking them.
- Symbols in the tree are not explained through a legend, but only mouse-over tooltips.
- The minimap or outline next to the scrollbar is not intuitive for users that are not familiar with such.
- The scroll bar is differently styled than a standard scroll bar and might not instantly be recognized as such.

Provenance module

- For some users, it is not noticeable whether an upload was selected.
- The function of the *load all changes* button is not clear.
- The temporal order (left to right or right to left) of multiple uploads on the same day is not clear.

Data integration module

- For elements with more than one configuration rule, it is not intuitive that the rules are applied from top to bottom order.

Textbox 5. Issues in the category Error Tolerance.

Core module, provenance module, and data integration module

- Activating multiple modules and searches leads to an overload of information in the concept tree.
- Loading too many information into the tree and expanding many of affected tree elements leads to high central processing unit (CPU) use.

Discussion

Principal Findings

In total, 12 participants took part in the evaluation of up to 3 modules of the CoMetaR web app, and each participant completed up to 10 tasks; 97% (85/88) of all tasks were solved independently and successfully. The core module and data integration module both obtained a mean usability score of 81, which proves good and nearly excellent usability. For inexperienced users, we estimated a mean usability score of 73, which proves good and acceptable usability. The provenance module has a mean usability score of approximately 72, which implies good and acceptable usability. For inexperienced provenance module users, we estimated a mean usability score of 63, which indicates unacceptable usability. We identified 24 issues with the app, which we grouped into 5 usability categories based on ISO 9241-110. From our point of view, of particular note are (1) information displayed in the concept tree can be overwhelming, especially if information from multiple modules is shown at once. (2) For many users, the provenance module and its functionalities are not accessible. The number of options, such as filtering by timespan or upload package, demand an extensive introduction and learning period. (3) The search functionality can output far more hits than expected because every literal information about concepts is considered. Some sort of categorization or filtering may be useful.

Strengths and Limitations

The strength of our study design is the relationship between effort and outcome. Although we omitted the step of recording audio and video of each session, we found a considerable compilation of usability issues and clear quantitative categorization of our tested modules owing to the System

Usability Scale questionnaire. All testing sessions were performed by a single experimenter. For thematic analysis, an additional scientist was consulted.

Retrospectively, we identified 4 problems regarding the evaluation methodology. The web conference software used in this evaluation was always visible and, in some cases, overlapped crucial information in the browser window. Second, one person tried to participate via an Apple product and was not able to establish screen sharing because of missing technical literacy. The third problem concerns communicational logistics, specifically around task instructions being communicated verbally by the evaluator. Some participants missed important aspects of the tasks because they were inattentive or started solving the tasks before the instruction was finished. Finally, some tasks were not formulated in sufficient detail. For example, for task 5, a participant thought it would be sufficient to read the respective upload description, but we expected them to list all changes explicitly in detail.

We did not record audio and video, for which reason we probably missed single verbalizations and observations. Thus, we cannot claim that our list of usability issues is complete at 100%, which arguably is never the case. In addition, the experimenter already filtered information during the test sessions, which might have biased the qualitative analysis outcome. We still assume that we found most usability issues, especially the most severe ones, because the experimenter was able to follow every action throughout all sessions without difficulty.

As all tasks were performed in our production environment, the upload history and thus the collection of added, moved, or removed or modified concepts varied. This may have led to differing results among the participants. We assumed that these differences were negligible in the usability evaluation.

Comparison With Previous Work

In 2009, considering 317 web apps, Bangor et al [39] found that web apps have a mean usability score of 68.2, which confirms the above-average usability of our app. Owing to increased awareness regarding usability, these values might have changed, but we did not find a more recent usability score meta-analysis. To the best of our knowledge, our approach to calculate another score for inexperienced users has not been done before. It allows the assessment of usability scores for inexperienced or new users even though some participants already have experience with the app.

Regarding the think aloud method, it is usual to record and transcribe all user sessions. Other studies show that this consumes a considerable amount of time and labor, which is often done by multiple scientists. In addition, we did not count code quantities within a transcript, as this is often done in a thematic analysis. We adopted the highest-level themes from an ISO standard instead of creating them ourselves.

Implications and Future Work

After evaluating our app, we are able to improve it by addressing all found usability issues. This will, in the first place, improve research in the field of lung research because lung research–specific metadata availability and accessibility will be improved. This app has already been considered by other German Centers for Health Research. We hope to be able to generally improve the field of health research.

Second, we applied a methodology that allows the usability evaluation of metadata management apps with a considerably low effort in time and labor. In an adapted form, this method can be applied to similar apps. Although the first 4 tasks of our evaluation are specific to the field of lung research concerning content, their content-agnostic intention is to check if basic information can be retrieved from the app. This includes the existence and findability of concepts (task 1), identification of a concept's annotations (task 2), its development over time (task 3), and the export of information about a unit of concepts (task 4). The application programming interface for the data integration module is specific to our data integration configuration file format, but the tasks represent the crucial steps to be taken to verify such a configuration file. The next step for this project could be the application of this evaluation method to comparable apps to approve its reliability and to find common usability issues.

We also hope that the findings of our qualitative analysis raise other developers' awareness of possible shortcomings in their own apps. For example, they might also plan to visually annotate concepts in the concept tree, in which case we highly recommend not displaying too much information at once.

A potential alternative or addition to the think aloud method with a thematic approach could be a heuristic evaluation performed by usability experts. The advantages and disadvantages of both methods were researched by Yen and Bakken [30].

We experienced issues with the web conference software, whose control panel sometimes overlapped crucial information on the user display. For further remotely and synchronously performed evaluations, we recommend ensuring that all relevant web app content is always visible, for example, by choosing different conference software.

We found that the assumed average usability score for inexperienced users was approximately 8 points lower than the original average score. This implies, on the one hand, that entry barriers exist within the app. On the other hand, these barriers can at least partly be overcome with experience. Measuring such a score might be of special interest for apps that provide a more efficient alternative to existing methods of information retrieval. Entry barriers may lead to rapid rejection of the entire software.

Conclusions

Our goal was to find usability issues of the CoMetaR web app and to measure its usability as perceived by real users. We identified 24 issues, which will be starting points for app improvement. On average, the app was assessed as good and in parts nearly excellent in terms of usability. Our method proved effective and efficient in terms of effort and outcome. Future research should improve our app and evaluate similar solutions. We invite other researchers interested in evaluating biomedical metadata repositories to adapt our methodology.

All source codes are publicly accessible under GitHub [40]. The production instance of the German Center for Lung Research metadata repository is publicly accessible [41].

Acknowledgments

The German Center for Lung Research (German: Deutsches Zentrum für Lungenforschung) is funded by the German Federal Ministry of Education and Research (German: Bundesministerium für Bildung und Forschung). Marc Griffiths proofread the paper as a native English speaker.

All data generated or analyzed during this study are included in this published paper.

Authors' Contributions

MRS developed the collaborative metadata repository software, which was evaluated in this study. MRS and RWM elaborated on the study design, including the composition of tasks. MRS performed the interviews with all participants and interpreted the data. RWM and AG substantively revised the study during all steps.

Conflicts of Interest

None declared.

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
2. Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, et al. FAIR data sharing: the roles of common data elements and harmonization. J Biomed Inform 2020 Jul;107:103421 [FREE Full text] [doi: 10.1016/j.jbi.2020.103421] [Medline: 32407878]
3. Hume S, Chow A, Evans J, Malfait F, Chason J, Wold JD, et al. CDISC SHARE, a global, cloud-based resource of machine-readable CDISC standards for clinical and translational research. AMIA Jt Summits Transl Sci Proc 2018 May 18;2017:94-103 [FREE Full text] [Medline: 29888049]
4. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009 Jul;37(Web Server issue):W170-W173 [FREE Full text] [doi: 10.1093/nar/gkp440] [Medline: 19483092]
5. Sansone S, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, FAIRsharing Community. FAIRsharing as a community approach to standards, repositories and policies. Nat Biotechnol 2019 Apr;37(4):358-367 [FREE Full text] [doi: 10.1038/s41587-019-0080-8] [Medline: 30940948]
6. Pampel H, Vierkant P, Scholze F, Bertelmann R, Kindling M, Klump J, et al. Making research data repositories visible: the re3data.org Registry. PLoS One 2013 Nov 4;8(11):e78080 [FREE Full text] [doi: 10.1371/journal.pone.0078080] [Medline: 24223762]
7. Lenzerini M. Data integration: a theoretical perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2002 Presented at: SIGMOD/PODS02: International Conference on Management of Data and Symposium on Principles Database and Systems; Jun 3 - 5, 2002; Madison Wisconsin. [doi: 10.1145/543613.543644]
8. Zhang H, Guo Y, Li Q, George TJ, Shenkman E, Modave F, et al. An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. BMC Med Inform Decis Mak 2018 Jul 23;18(Suppl 2):41. [doi: 10.1186/s12911-018-0636-4] [Medline: 30066664]
9. Stathias V, Koleti A, Vidović D, Cooper DJ, Jagodnik KM, Terryn R, et al. Sustainable data and metadata management at the BD2K-LINCS Data Coordination and Integration Center. Sci Data 2018 Jun 19;5:180117 [FREE Full text] [doi: 10.1038/sdata.2018.117] [Medline: 29917015]
10. Dugas M, Meidt A, Neuhaus P, Storck M, Varghese J. ODMedit: uniform semantic annotation for data integration in medicine based on a public metadata repository. BMC Med Res Methodol 2016 Jun 01;16:65 [FREE Full text] [doi: 10.1186/s12874-016-0164-9] [Medline: 27245222]
11. Ong TC, Kahn MG, Kwan BM, Yamashita T, Brandt E, Hosokawa P, et al. Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading. BMC Med Inform Decis Mak 2017 Sep 13;17(1):134 [FREE Full text] [doi: 10.1186/s12911-017-0532-3] [Medline: 28903729]
12. Pecoraro F, Luzi D, Ricci FL. Designing ETL tools to feed a data warehouse based on electronic healthcare record infrastructure. Stud Health Technol Inform 2015;210:929-933. [Medline: 25991292]
13. Post AR, Krc T, Rathod H, Agravat S, Mansour M, Torian W, et al. Semantic ETL into i2b2 with Eureka!. AMIA Jt Summits Transl Sci Proc 2013 Mar 18;2013:203-207 [FREE Full text] [Medline: 24303265]

14. Post AR, Pai AK, Willard R, May BJ, West AC, Agravat S, et al. Metadata-driven clinical data loading into i2b2 for Clinical and Translational Science Institutes. AMIA Jt Summits Transl Sci Proc 2016 Jul 20;2016:184-193 [FREE Full text] [Medline: 27570667]
15. Nadkarni P, Marenco L. Chapter 2 - data integration: an overview. In: Methods in Biomedical Informatics: A Pragmatic Approach. Cambridge: Academic Press; 2014.
16. Rahm E, Bernstein P. A survey of approaches to automatic schema matching. The VLDB J 2001;10:334-350. [doi: 10.1007/s007780100057]
17. Stöhr MR, Günther A, Majeed RW. Provenance for biomedical ontologies with RDF and Git. Stud Health Technol Inform 2019 Sep 03;267:230-237. [doi: 10.3233/SHTI190832] [Medline: 31483277]
18. Stöhr MR, Günther A, Majeed RW. Verifying data integration configurations for semantical correctness and completeness. Stud Health Technol Inform 2019 Sep 03;267:66-73. [doi: 10.3233/SHTI190807] [Medline: 31483256]
19. Kadioglu D, Breil B, Knell C, Lablans M, Mate S, Schlue D, et al. Samply.MDR - A metadata repository and its application in various research networks. Stud Health Technol Inform 2018;253:50-54. [Medline: 30147039]
20. Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, et al. Portal of medical data models: information infrastructure for medical research and healthcare. Database (Oxford) 2016 Feb 11;2016:bav121 [FREE Full text] [doi: 10.1093/database/bav121] [Medline: 26868052]
21. Stöhr MR, Majeed RW, Günther A. Using RDF and Git to realize a collaborative metadata repository. Stud Health Technol Inform 2018;247:556-560. [Medline: 29678022]
22. Miller E. An introduction to the resource description framework. Bul Am Soc Inf Sci Tech 2005 Jan 31;25(1):15-19. [doi: 10.1002/bult.105]
23. Pastor-Sanchez J, Martínez-Mendez F, Rodríguez-Muñoz J. Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. Inf Res 2009 Dec;14(4) [FREE Full text]
24. Weibel SL, Koch T. The Dublin Core Metadata Initiative. D-Lib Magazine 2000 Dec;6(12) [FREE Full text] [doi: 10.1045/december2000-weibel]
25. ISO/TS 21526 Health informatics - Metadata repository requirements (MetaRep). International Organization for Standardization. 2019. URL: https://www.iso.org/standard/71041.html [accessed 2021-11-16]
26. Halilaj L, Grangel-Gonzalez I, Coskun G, Auer S. Git4Voc: Git-based versioning for collaborative vocabulary development. In: Proceedings of the 2016 IEEE Tenth International Conference on Semantic Computing (ICSC). 2016 Presented at: 2016 IEEE Tenth International Conference on Semantic Computing (ICSC); Feb 3-5, 2016; Laguna Hills, California. [doi: 10.1109/icsc.2016.44]
27. Arndt N, Radtke N, Martin M. Distributed collaboration on RDF datasets using Git. In: Proceedings of the 12th International Conference on Semantic Systems. 2016 Presented at: SEMANTiCS 2016: 12th International Conference on Semantic Systems; Sep 12 - 15, 2016; Leipzig Germany. [doi: 10.1145/2993318.2993328]
28. SPARQL Protocol And RDF Query Language (SPARQL). W3C Semantic Web. URL: https://www.w3.org/TR/2013/REC-sparql11-query-20130321/ [accessed 2021-11-16]
29. Stöhr MR, Majeed RW, Günther A. Metadata import from RDF to i2b2. Stud Health Technol Inform 2018;253:40-44. [Medline: 30147037]
30. Yen P, Bakken S. A comparison of usability evaluation methods: heuristic evaluation versus end-user think-aloud protocol - an example from a web-based communication tool for nurse scheduling. AMIA Annu Symp Proc 2009 Nov 14;2009:714-718 [FREE Full text] [Medline: 20351946]
31. Reen GK, Muirhead L, Langdon DW. Usability of health information websites designed for adolescents: systematic review, neurodevelopmental model, and design brief. J Med Internet Res 2019 Apr 23;21(4):e11584 [FREE Full text] [doi: 10.2196/11584] [Medline: 31012856]
32. ISO 9241-110 Ergonomics of human-system interaction - Part 110: interaction principles. International Organization for Standardization. 2020. URL: https://www.iso.org/standard/75258.html [accessed 2021-11-16]
33. Tullis T, Stetson J. A comparison of questionnaires for assessing website usability. In: Proceedings of the Usability Professionals' Association Conference. 2004 Presented at: Usability Professionals' Association Conference; Jun 7-11, 2004; Minneapolis, Minnesota, USA.
34. Brooke J. SUS: a 'Quick and Dirty' usability scale. In: Usability Evaluation In Industry. Boca Raton, Florida: CRC Press; 1996.
35. McLellan S, Muddimer A, Peres S. The effect of experience on System Usability Scale ratings. J Usability Stud 2012;7(2):56-67 [FREE Full text]
36. Bastien JC. Usability testing: a review of some methodological and technical aspects of the method. Int J Med Inform 2010 Apr;79(4):e18-e23. [doi: 10.1016/j.ijmedinf.2008.12.004] [Medline: 19345139]
37. GoToMeeting. LogMeIn. URL: https://www.gotomeeting.com/ [accessed 2021-05-07]
38. Virzi RA. Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough? Hum Factors 2016 Nov 23;34(4):457-468. [doi: 10.1177/001872089203400407]
39. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4(3):114-123.

40. Collaborative Metadata Repository (CoMetaR) Code Repository. GitHub. URL: https://github.com/dzl-dm/cometar [accessed 2021-11-16]

41. Collaborative Metadata Repository (CoMetaR) Web Application. Stöhr MR. URL: https://data.dzl.de/cometar [accessed 2021-11-16]

Abbreviations

CoMetaR: collaborative metadata repository
DC: Dublin Core
DZL: Deutsches Zentrum für Lungenforschung
FAIR: Findability, Accessibility, Interoperability, Reusability
FEV1: Forced Expiratory Volume in 1 Second
RDF: Resource Description Framework
SKOS: Simple Knowledge Organization System
SPARQL: SPARQL Protocol and Resource Description Framework Query Language

Edited by C Lovis; submitted 10.05.21; peer-reviewed by D Kadioglu, H Storf, A Blatch-Jones, C Schüttler, M Storck; comments to author 10.07.21; revised version received 13.08.21; accepted 11.10.21; published 29.11.21.

Please cite as:
Stöhr MR, Günther A, Majeed RW
The Collaborative Metadata Repository (CoMetaR) Web App: Quantitative and Qualitative Usability Evaluation
JMIR Med Inform 2021;9(11):e30308
URL: https://medinform.jmir.org/2021/11/e30308
doi:10.2196/30308
PMID:34847059

Original Paper

The Relationship Between Electronic Health Record System and Performance on Quality Measures in the American College of Rheumatology's Rheumatology Informatics System for Effectiveness (RISE) Registry: Observational Study

Nevin Hammam[1], MD, PhD; Zara Izadi[1], PhD; Jing Li[1], MPH; Michael Evans[1], MSc; Julia Kay[1], BA; Stephen Shiboski[2], PhD; Gabriela Schmajuk[1,3,4], BSc, MD; Jinoos Yazdany[1], MD, MPH

[1]Division of Rheumatology, Department of Medicine, University of California, San Francisco, CA, United States

[2]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, United States

[3]Philip R Lee Institute for Health Policy Research, San Francisco, CA, United States

[4]San Francisco Veterans Affairs Medical Center, San Francisco, CA, United States

Corresponding Author:
Jinoos Yazdany, MD, MPH
Division of Rheumatology
Department of Medicine
University of California
P O Box 0811
Floor 03, Room 3301
San Francisco, CA, 94110
United States
Phone: 1 628 206 8618
Email: jinoos.yazdany@ucsf.edu

Abstract

Background: Routine collection of disease activity (DA) and patient-reported outcomes (PROs) in rheumatoid arthritis (RA) are nationally endorsed quality measures and critical components of a treat-to-target approach. However, little is known about the role electronic health record (EHR) systems play in facilitating performance on these measures.

Objective: Using the American College Rheumatology's (ACR's) RISE registry, we analyzed the relationship between EHR system and performance on DA and functional status (FS) quality measures.

Methods: We analyzed data collected in 2018 from practices enrolled in RISE. We assessed practice-level performance on quality measures that require DA and FS documentation. Multivariable linear regression and zero-inflated negative binomial models were used to examine the independent effect of EHR system on practice-level quality measure performance, adjusting for practice characteristics and patient case-mix.

Results: In total, 220 included practices cared for 314,793 patients with RA. NextGen was the most commonly used EHR system (34.1%). We found wide variation in performance on DA and FS quality measures by EHR system (median 30.1, IQR 0-74.8, and median 9.0, IQR 0-74.2), respectively). Even after adjustment, NextGen practices performed significantly better than Allscripts on the DA measure (51.4% vs 5.0%; *P*<.05) and significantly better than eClinicalWorks and eMDs on the FS measure (49.3% vs 29.0% and 10.9%; *P*<.05).

Conclusions: Performance on national RA quality measures was associated with the EHR system, even after adjusting for practice and patient characteristics. These findings suggest that future efforts to improve quality of care in RA should focus not only on provider performance reporting but also on developing and implementing rheumatology-specific standards across EHRs.

(JMIR Med Inform 2021;9(11):e31186) doi:10.2196/31186

KEYWORDS

rheumatoid arthritis; electronic health record; patient-reported outcomes; quality measures; electronic health record; disease activity; quality of care; performance reporting; medical informatics; clinical informatics

Introduction

The routine collection of disease activity (DA) and patient-reported outcomes (PROs) such as functional status (FS) in rheumatoid arthritis (RA) are nationally endorsed quality measures and an important component of tracking outcomes and improving care [1-3]. The process of collecting these assessments is essential to the implementation of a treat-to-target strategy, which has been shown to improve outcomes for patients with RA and decrease health care utilization [4]. Nevertheless, studies have identified that quality of care provided to patients with RA remains inconsistent [5,6].

Data from the early years of the American College of Rheumatology's (ACR's) national Rheumatology Informatics System for Effectiveness (RISE) Registry, an electronic health record (EHR)–enabled registry, showed that among more than 150,000 patients with RA, only 50% had an RA DA score or a FS recorded as structured EHR data. The use of structured data fields facilitates disease monitoring, data retrieval, and quality reporting, especially in comparison to storing information in free text fields (ie, clinical notes) alone [7]. Several groups have developed EHR-based applications to help support the documentation of DA and FS measures as structured fields [8-11], and some EHR vendors have incorporated similar tools into their foundation software. However, no studies have evaluated the relationship between the EHR system and performance on important quality measures on a national scale across rheumatology practices.

In this study, we use the ACR's RISE registry to analyze the relationship between major EHR systems used by US rheumatologists and performance on DA and FS quality measures. In addition, we report on the characteristics of EHRs with higher performance on these measures.

Methods

Data Source

RISE is a national registry that passively collects data from EHRs of participating practices, aggregates and centrally analyzes performance on quality measures in rheumatology [12]. RISE practices have an on-site "registry connector" that uploads EHR data to the RISE clinical data warehouse on a nightly basis. Data that are uploaded include information regarding patient diagnoses, medications, laboratory studies, and vital signs.

When a practice first joins RISE, practice personnel work with registry staff to map structured data elements to relevant quality measures. Occasionally, practices request data elements to be pulled from clinical notes. This is feasible when data are recorded in a highly reliable, "semi-structured" format. Data elements extracted from the EHR are used to calculate electronic quality measures for submission to Centers for Medicare & Medicaid Services (CMS) as part of national pay-for-performance programs. For example, patients with RA are identified using ICD codes to enter a denominator population. DA scores are extracted (usually from structured fields) to determine whether denominator patients meet the criteria established by quality measures. Practice, provider, and patient-level performance on quality measures is fed back to providers through a web-based dashboard. Patient-level EHR data that include all variables mentioned above, in addition to quality measure denominator and numerator information is provided to RISE Data Analytic Centers for analysis. Additional details on the structure and function of the RISE registry are described elsewhere [12].

As of December 2018, RISE held validated data from 1113 providers in 226 practices, representing approximately 32% of the US clinical rheumatology workforce. RISE can connect to most certified EHR systems in the United States; as of 2018, the registry could map to over 30 different EHRs used by rheumatologists. EHRs largely reflect those used by community rheumatologists, as academic medical centers (and therefore large EHR vendors such as Epic and Cerner) are underrepresented in the registry.

Study Population and Period

We analyzed data on individuals with RA seen in rheumatology practices participating in RISE between January 1, 2018, and December 31, 2018. Patients included in this study were 18 years of age or older and had ≥1 International Classification of Diseases Clinical Modification, Ninth or Tenth Revision (ICD-9-CM or ICD-10-CM) code for RA with at least 1 clinical face-to-face encounter in 2018. These inclusion criteria are based on the denominator definitions for the quality measures used to calculate the outcomes of interest (see section below).

Outcomes

We selected two measures—routine assessment of DA and FS—for patients with RA since these are key components of a treat-to-target approach and among the newer process measures introduced in the national pay scheme for performance programs, specifically for rheumatologists. The American College of Rheumatology is currently collaborating with the National Quality Forum (NQF) to develop outcome measures on the basis of DA and FS. Thus, it is especially important to understand variations in the collection of these measures and any potential factors influencing their documentation. Assessment and documentation of DA and FS outcomes were assessed at the practice-level by calculating the performance on NQF-endorsed quality measures. Performance on each measure was defined as follows: (1) DA: percentage of patients aged 18 years and older with a diagnosis of RA, whose DA was assessed using a standardized measurement tool at 50% or more face-to-face encounters for RA during the measurement period [13]; and (2) FS: percentage of patients aged 18 years and older with a diagnosis of RA, whose FS was assessed using a standardized measurement tool at least once during the measurement period [14]. The measurement period was the 12-month period between January 1 and December 31, 2018.

Covariates

Covariates included EHR system (NextGen, eClinicalWorks, GE Centricity, eMDs, Allscripts, Amazing Charts, Aprima, others included Lytec MD, Medent, Medisoft, Raintree System IC, MD office, Integrity, Carecloud, MedTrio, Greenway/Primesuite, iPatientCare, Prime Clinical System, MacPractice MD, IMS, SRS EHR, PrognoCIS, Cerner, Practice Fusion, DrChrono, Chart Maker Clinical, STI, American Medical Software, Athena Clinicals, Praxis EMR, RheumDocs, Greenway Intergy, Athena UniCharts, and ChartLogic) and a variety of practice and patient characteristics previously associated with measure performance. Practice characteristics included the number of providers within the practice; practice type (single-specialty group practice, solo practitioner, multi-specialty group practice, other clinical settings, and large health system); and geographical region in the United States (Northeast, Midwest, South, and West). Practice-level sociodemographic variables were calculated by aggregating the characteristics of patients included in the study and included the proportion of patients aged ≥65 years, females, non-White individuals, and of those with noncommercial insurance. Patient characteristics included age, sex, race/ethnicity, insurance type (private, Medicaid, Medicare, or other), and Charlson comorbidity index (CCI) score calculated in accordance with the Deyo modification based on codes reported at any time during the study period [15].

Practice Documentation Workflow Survey

To learn more about potential reasons for differences in performance on DA and FS measures across EHR systems, we also assessed documentation workflows among a subset of RISE practices. A survey was disseminated electronically using a commercial survey web application to the RISE practices' providers and administrators between November 11, 2020, and April 14, 2021. The survey included 9 questions (Multimedia Appendix 1), covering the topics of practice characteristics (3 questions), and EHR system–related factors, such as the presence of a rheumatology-specific module or dedicated structured fields for PROs, which might influence DA and FS documentation workflows (6 questions).

Statistical Analysis

We used descriptive statistics to summarize patient and practice characteristics. Multivariable linear regression was used to examine the independent effect of EHR system on practice-level performance. We used zero-inflated models when the occurrence of zeros for practice-performance was meaningful (27.7% for DA; 40.4% for FS). These models allow for modeling of overly dispersed data. The outcome variable for the zero-inflated analyses was the count (rate) of patients in a practice, who received recommended care [16]. Zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models, all adjusted for potential confounders, were compared using Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC) values, and log-likelihood to assess the goodness of fit. The incidence rate ratio (IRR) for each of the predictor variables in both count (rate) and logit parts of the model was reported along with 95% CIs and *P* values.

All the models (linear and zero-inflated) were adjusted for practice characteristics (including practice type, size, and geographical region) and patient case-mix (including patient age, sex, race, and insurance) since these variables have been previously shown to have a significant association with performance on rheumatology quality measures [5,17], and our goal was to isolate the impact of the EHR vendor on performance. To account for differences in case-mix, we adjusted for the aggregate characteristics of patients seen in the practice (proportion of patients aged ≥65 years, proportion of females, proportion of non-White patients, and the proportion of patients with noncommercial insurance). Missing values were included in the analyses as their own separate category without imputation. For all analyses, *P* values less than .05 were considered statistically significant.

All analyses were performed using STATA statistical software (version 16, StataCorp). This study was approved by the Western Institutional Review Board, Inc. as well as the Committee on Human Research at the University of California, San Francisco.

Results

Practice and Subject Characteristics

Our practice-level analysis included 314,793 individuals with RA who met the inclusion criteria across 220 practices in this study; 6 practices had no patients who met the inclusion criteria and were therefore excluded. Among all included patients (N=314,793), the mean age was 62.0 (SD 14.3) years, 76.1% were female, 67.7% were White, and 31.8% had private insurance. Most practices were single-specialty group practices (56.8%). NextGen was the most commonly used EHR system (34.1% of practices), followed by eClinicalWorks (14.5%) and Amazing Charts (9.5%) (Table 1).

Table 1. Characteristics of practices in the Rheumatology Informatics System for Effectiveness registry.

Practice characteristics	Practices (n=220)
Providers in practice, n (%)	
1-4	162 (73.6)
5-9	44 (20.0)
10-20	14 (6.4)
Practice type, n (%)	
Single-specialty group practice	125 (56.8)
Solo practitioner	62 (28.2)
Multi-specialty group practice	29 (13.2)
Health system	4 (1.8)
Electronic health record system, n (%)	
NextGen	75 (34.1)
eClinicalWorks	32 (14.5)
Amazing Charts	21 (9.5)
eMDs	11 (5.0)
GE Centricity	10 (4.5)
Allscripts	8 (3.6)
Aprima	8 (3.6)
Other[a]	55 (25.0)
US regions	
South	98 (44.5)
West	48 (21.8)
Northeast	42 (19.1)
Midwest	32 (14.5)
Practice-level patient characteristics, mean (SD)	
Proportion of patients aged ≥65 years	0.47 (0.10)
Proportion of female patients	0.77 (0.04)
Proportion of non-White patients	0.33 (0.26)
Proportion with noncommercial insurance	0.68 (0.21)

[a]"Other" electronic health systems included any system used in <2% of practices, including Lytec MD, Medent, Medisoft, Raintree System IC, MD office, Integrity, Carecloud, MedTrio, Greenway/Primesuite, iPatientCare, Prime Clinical System, MacPractice MD, IMS, SRS EHR, PrognoCIS, Cerner, Practice Fusion, DrChrono, Chart Maker Clinical, STI, American Medical Software, Athena Clinicals, Praxis EMR, RheumDocs, Greenway Intergy, Athena UniCharts, and ChartLogic.

Overall, median (IQR) practice-level performance on DA and FS quality measures was 30.1 (0, 74.8) and 9.0 (0, 74.2), respectively. Figures 1 and 2 demonstrate the differences in quality measure performance on the DA and FS measures for practices using different EHR systems.

Figure 1. Distribution of practice-level performance on the disease activity quality measure for patients with rheumatoid arthritis, stratified by the electronic health record system.

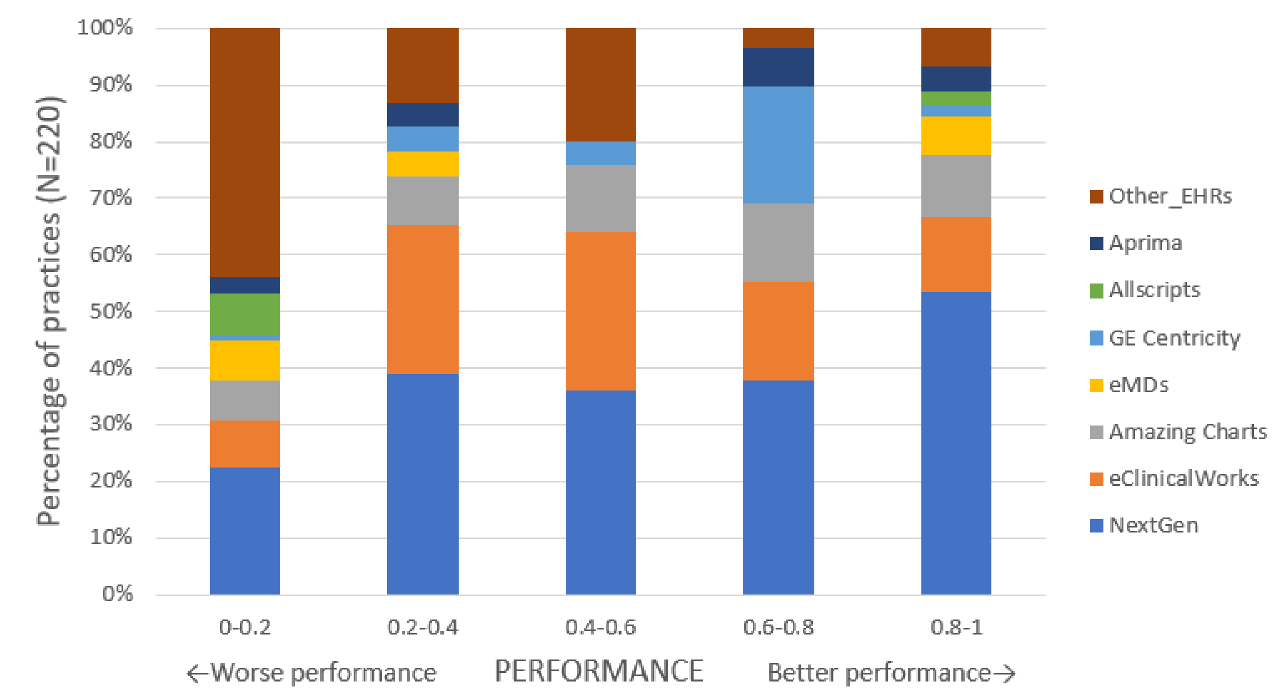
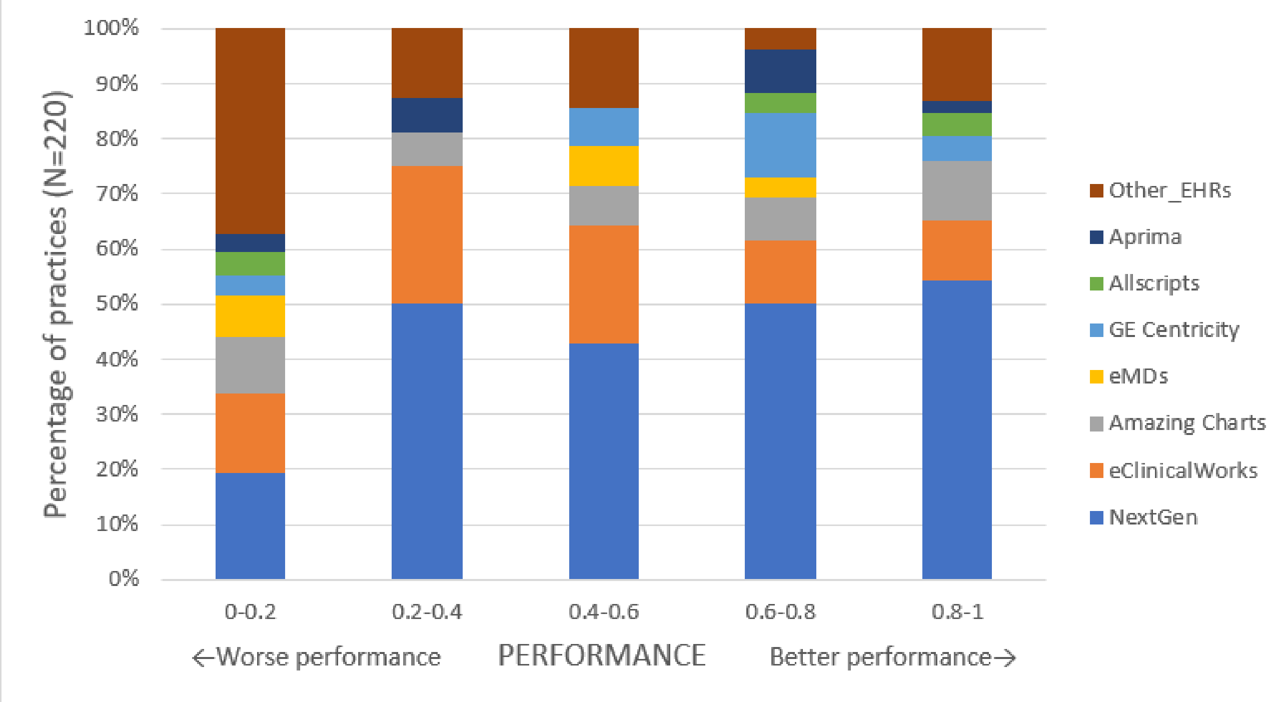


Figure 2. Distribution of practice-level performance on the functional status quality measure for patients with rheumatoid arthritis, stratified by the electronic health record system.



In unadjusted analyses, practices using NextGen showed significantly higher performance on DA and FS measures compared to practices using other EHR systems (Table 2). In multivariate linear regression analyses adjusting for practice characteristics and patient case-mix, practices that used NextGen had higher performance. Specifically, NextGen practices performed significantly better than Allscripts on the DA measure (51.4% vs 5.0%) and significantly better than eClinicalWorks and eMDs on the FS measure (49.3% vs 29.0% and 10.9%, respectively; Table 2). Full models with parameter estimates for practice and case-mix variables are included in Multimedia Appendix 2.

Table 2. Association of practice characteristics with measure performance, with marginal means estimated using multivariate regression models.

Electronic health record system	Disease activity measure performance				Functional status measure performance			
	Unadjusted performance, % (95% CI)	*P* value	Adjusted performance, % (95% CI)	*P* value	Unadjusted performance, % (95% CI)	*P* value	Adjusted performance, % (95% CI)	*P* value
NextGen	52.2 (44.5-59.9)	Reference	51.4 (42.3-60.5)	Reference	51.0 (42.9-59.2)	Reference	49.3 (39.4-59.3)	Reference
eClinicalWorks	45.7 (33.8-57.5)	.36	42.5 (30.0-54.9)	.29	30.3 (17.9-42.8)	.01	29.0 (15.4-42.6)	.03
Amazing Charts	46.7 (32.1-61.3)	.52	50.6 (35.3-65.8)	.93	32.1 (16.7-47.4)	.03	36.5 (19.8-53.2)	.22
eMDs	31.4 (11.3-51.6)	.06	29.8 (9.6-50.1)	.06	11.3 (0-32.6)	.001	10.9 (0-33.0)	.003
GE Centricity	62.0 (40.9-83.2)	.39	51.7 (30.1-73.4)	.99	48.6 (26.3-70.8)	.84	47.3 (23.8-70.8)	.88
Allscripts	11.4 (0-35.0)	<.001	5.0 (0-29.2)	.001	30.7 (5.8-55.6)	.13	28.1 (1.6-54.5)	.15
Aprima	46.1 (22.4-69.7)	.63	48.7 (24.6-72.8)	.84	33.6 (8.7-58.5)	.19	35.2 (8.9-61.5)	.34
Other[a]	13.2 (4.1-22.2)	<.001	17.3 (8.1-26.6)	<.001	15.0 (5.5-24.5)	<.001	16.8 (6.7-27.0)	<.001

[a] Other electronic health records included any system used in <2% of practices, including Lytec MD, Medent, Medisoft, Raintree System IC, MD office, Integrity, Carecloud, MedTrio, Greenway/Primesuite, iPatientCare, Prime Clinical System, MacPractice MD, IMS, SRS EHR, PrognoCIS, Cerner, Practice Fusion, DrChrono, Chart Maker Clinical, STI, American Medical Software, Athena Clinicals, Praxis EMR, RheumDocs, Greenway Intergy, Athena UniCharts, and ChartLogic. Adjusted models were adjusted for practice characteristics and patient case-mix.

Marginal means were estimated using multivariate regression models. Confidence intervals of <0 were truncated at 0.

Because we found a significant number of practices with zero performance (27.7% for DA; 40.4% for FS), we also used zero-inflated models to analyze the association between EHR systems and DA and FS documentation (Multimedia Appendix 3). Zero-inflated negative binomial models revealed that the differences in performance across EHRs were driven largely by the practices with absent documentation of DA and FS (zero performance): although we found no differences in the count portion of the ZINB model across EHRs, there were significant differences between NextGen versus other EHRs in the logit portion of the model. For example, practices using Allscripts had approximately a 2.5 times higher rate of having zero performance on DA and FS compared to practices that used NextGen (Table 3). On the other hand, practices that used GE Centricity were less likely to have zero performance on DA than practices that used NextGen (*P*<.01).

Table 3. Adjusted zero-inflated negative binomial models examining the main effect of electronic health record systems on the number of patients with rheumatoid arthritis who received recommended care.

Electronic health record system	Disease activity measure performance		Functional status measure performance	
	Incidence rate ratio[a], ratio (95% CI)	*P* value	Incidence rate ratio, ratio (95% CI)	*P* value
Count Model				
NextGen	Reference	Reference	Reference	Reference
eClinicalWorks	0.84 (0.59-1.18)	.31	0.84 (0.55-1.31)	.45
GE Centricity	0.88 (0.61-1.27)	.50	1.10 (0.66-1.82)	.71
eMDs	0.81 (0.44-1.48)	.49	0.67 (0.20-2.27)	.52
Allscripts	0.37 (0.07-1.98)	.25	1.08 (0.72-1.63)	.72
Amazing Charts	1.23 (0.78-1.92)	.37	1.34 (0.80-2.25)	.27
Aprima	1.16 (0.79-1.71)	.46	1.27 (0.78-2.07)	.35
Others[b]	0.81 (0.50-1.32)	.40	1.15 (0.78-1.69)	.49
Zero Inflated Model				
NextGen	Reference	Reference	Reference	Reference
eClinicalWorks	0.21 (–1.27 to 1.69)	.78	1.80 (0.72 to 2.89)	*<.001* [c]
GE Centricity	–19.01 (–20.08 to –17.94)	*<.001*	1.47 (–0.11 to 3.05)	.07
eMDs	1.92 (0.41 to 3.43)	*.01*	2.87 (1.40 to 4.34)	*<.001*
Allscripts	2.48 (0.85 to 4.11)	*.003*	2.82 (1.17 to 4.48)	*<.001*
Amazing Charts	1.02 (–0.38 to 2.42)	.16	2.21 (1.03 to 3.40)	*<.001*
Aprima	1.97 (0.28 to 3.65)	*.02*	2.31 (0.70 to 3.93)	*.01*
Other[b]	3.17 (2.13 to 4.21)	*<.001*	3.37 (2.36 to 4.39)	*<.001*

[a] In the count model, the incidence rate ratios represent the rate of having patients who received recommended care compared to NextGen; in the zero inflated model, the incidence rate ratios represent the rate of having zero performance compared to NextGen. Incidence rate ratios were adjusted for practice characteristics and patient case-mix.

[b] Other electronic health record systems included any system used in <2% of practices, including Lytec MD, Medent, Medisoft, Raintree System IC, MD office, Integrity, Carecloud, MedTrio, Greenway/Primesuite, iPatientCare, Prime Clinical System, MacPractice MD, IMS, SRS EHR, PrognoCIS, Cerner, Practice Fusion, DrChrono, Chart Maker Clinical, STI, American Medical Software, Athena Clinicals, Praxis EMR, RheumDocs, Greenway Intergy, Athena UniCharts, and ChartLogic.

[c] Italicized *P* values are statistically significant.

Finally, among 40 (18.2%) practices with survey responses, NextGen was the most commonly used EHR system (37.5%), followed by eClinicalWorks (22.5%) and Amazing Charts (7.5%); other EHR systems accounted for 32.5%. The majority of the responding practices using NextGen (93.3%) reported that they relied on structured data fields for DA and FS quality measure documentation in the EHR. Conversely, the vast majority of the non-NextGen practices reported that clinicians documented DA and FS in clinical notes. After the survey was closed, we additionally queried survey respondents to understand local workflows. For example, we found that NextGen includes rheumatology-specific templates that facilitate documentation of RA outcomes and functionality to track this information over time. In contrast, those using Amazing Charts enter DA and FS measures in a semistructured way (ie, in the same section of the note for every patient). This workflow, although it departs from the structured fields used by most NextGen practices, allows the registry vendor to manually extract these data for national performance reporting but is not amenable to tracking outcomes over time.

Discussion

Principal Findings

Although quality measures are often used to evaluate the performance of individual clinicians or health systems, the impact of health information technology on performance remains extremely understudied. We used a unique data source, the ACR's RISE registry, which captures data from rheumatology practices across the United States, to investigate the relationship between performance on nationally endorsed RA quality measures and the EHR system used by practices.

We found that after adjusting for both practice characteristics and patient case-mix, performance in practices using some EHR systems was consistently higher; the EHRs with the highest performance generally had rheumatology-specific templates or modules in their foundation software, which facilitated collection and tracking of key RA outcomes. These findings raise important questions about the role of EHR vendors in creating software that facilitates high quality of care in rheumatology.

In both rheumatology and more general practice, studies that formally assess the impact of health information technology systems, including EHRs, on quality measure performance are limited. In a prior study using the RISE registry, we found that NextGen practices were able to improve performance on quality measures more rapidly over time than practices with other EHR systems [17]. Literature is emerging to support the notion that EHR systems can be an important factor in quality of care; for example, one study investigated a large group of primary care practices with different EHRs and their frequency of unsafe prescribing of cyclosporine, tacrolimus, and diltiazem and found important differences across EHRs [18]. In the case of RISE practices, NextGen practices reported the availability of rheumatology-specific templates in the EHR foundation software to capture key RA outcomes, facilitating quality measurement and disease-tracking. Practices using EHRs that lack this feature are much less likely to document effectively, although some of them have managed to find other methods (eg, semistructured data collection in clinical notes) for increasing accurate registry data measures for use in quality measure calculations. However, this type of documentation is less conducive to chronic disease management since it renders longitudinal tracking of disease outcomes using the EHR challenging.

Using current technology, documentation of key RA outcomes in structured EHR data fields remains the most feasible way to ensure accurate data capture and quality measurement. However, in the future, it is possible that extraction of information from clinical notes will become easier. The application of advanced approaches using text searches, natural language processing algorithms, or machine learning to extract information about disease outcomes directly from clinical notes could become more feasible in the future [19], although these strategies have largely not been demonstrated to be reliable in nonresearch settings to date and are often very cost prohibitive [20]. On the other hand, our study illustrates that is feasible in the short-term for EHR systems to modify their foundation software to include content that facilitates rheumatology practice, including capture of key RA outcomes such as DA and FS.

Further, although the idea is provocative, we think it is time for health care quality measurement to consider the range of factors that influence performance. Using a broader quality measurement paradigm, EHR vendors, in addition to individual physicians or health systems, should share in incentives or penalties associated with quality measure performance for chronic diseases such as RA [9,21,22]. Although in theory, a marketplace with multiple competitive EHR systems might have resulted in lower prices and increased functionality across all systems, in reality, the costs of switching EHR systems is high, and the risk of downtime or loss of historical data makes changing EHR vendors difficult. A shared incentive model could encourage EHR vendors to implement tools to support quality measurement and improvement. Others have explored this idea of shared responsibility for performance, particularly in the realm of patient safety [23]. For example, poorly designed user interfaces can reduce clinicians' access to key data needed to ensure patient safety [24]. In a model designed around shared responsibility, EHR vendors would be incentivized to incorporate a user-centered process in designing and building software to support clinicians in meeting key quality metrics. Whether around patient safety or management of chronic diseases like RA, this would require EHR vendors to develop ongoing strategies to track the usability of their product and to set up systems for innovation and continuous quality improvement of software [25]. Public and private agencies could be charged with creating programs to incentivize ongoing usability-testing and to track resulting measure performance [26]. Such a model would, of course, require the full collaboration of practices and clinicians, who would need to incorporate such usability testing into their workflows and be willing to alter existing workflows to take advantage of software improvements.

Strengths and Limitations

Our study has important strengths. Data were collected passively from different EHR systems and includes all patients with RA who met criteria for the denominator of the quality measure at each practice, thus greatly reducing the risk of selection bias inherent in other study designs. Moreover, the large number of practices and EHR systems represented in the RISE registry provides a unique view of a range of rheumatology practices using different EHR systems across the United States. Along with these strengths, this study has some important limitations. Information obtained from the EHR reflects care documented rather than care delivered. Clinicians may have assessed DA or FS for their patients but failed to document them or documented them in ways that are not easily retrievable by the registry; however, since our intention was to examine capture of data in EHRs, which are used to quality measure performance, this limitation does not impact our conclusions. A second, important limitation, is that we were limited to studying the EHRs of practices that participate in RISE. The registry currently has limited coverage of some EHRs with a greater market share among academic practices in the United States (eg, Epic and Cerner) and covers only an estimated 32% of the US clinical rheumatology workforce. Further research is needed to assess the relationships between EHR systems and quality measure performance, especially in academic settings.

Conclusions

In summary, this study shows a strong relationship between the EHR system used by practices and performance on DA and FS quality measures for RA. Future research should investigate whether features of EHRs, which facilitate documentation and tracking of RA outcomes, facilitate improved outcomes among patients with RA over time. Developing rheumatology-specific standards across EHRs may promote routine collection of RA measures, which, in turn, could improve RA outcomes.

Acknowledgments

The authors thank Tracy Johannson and Lindsay Lewis at the American College of Rheumatology for assistance with electronic health record (EHR) survey and all those who participated in the surveys performed as part of this project. This work was funded by grant R18HS025638 from the Agency for Healthcare Research and Quality (AHRQ). JY and GS are supported by the Russell/Engleman Medical Research Center for Arthritis. JY is supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) P30 AR070155 and K24AR074534. The data presented here were supported by the American College of Rheumatology's RISE Registry. The views expressed represent those of the authors and do not necessarily represent the views of the American College of Rheumatology, AHRQ, or NIAMS. This study protocol was prepared by NH and the data analysis was executed by the RISE data analytic center at University of California San Francisco (UCSF).

Conflicts of Interest

None declared.

Multimedia Appendix 1

The questionnaire used in the patient-reported outcomes documentation workflow survey.

[DOCX File , 16 KB - medinform_v9i11e31186_app1.docx]

Multimedia Appendix 2

Association of practice characteristics with measure performance, using multivariate linear regression models.

[DOCX File , 22 KB - medinform_v9i11e31186_app2.docx]

Multimedia Appendix 3

Comparisons of zero-inflated Poisson and zero-inflated negative binomial models' characteristics on count patient who received recommended care.

[DOCX File , 14 KB - medinform_v9i11e31186_app3.docx]

References

1. NQF-Endorsed Measures for Musculoskeletal Conditions. National Quality Forum. 2015. URL: http://www.qualityforum.org/Publications/2015/01/NQF-Endorsed_Measures_for_Musculoskeletal_Conditions.aspx [accessed 2020-08-19]
2. Hobbs KF, Cohen MD. Rheumatoid arthritis disease measurement: a new old idea. Rheumatology (Oxford) 2012 Dec;51 Suppl 6:vi21-vi27. [doi: 10.1093/rheumatology/kes282] [Medline: 23221583]
3. Gilek-Seibert K, Prescott K, Kazi S. Outcome assessments in rheumatoid arthritis. Curr Rheumatol Rep 2013 Nov;15(11):370. [doi: 10.1007/s11926-013-0370-y] [Medline: 24072601]
4. Coulter A, Ellins J. Effectiveness of strategies for informing, educating, and involving patients. BMJ 2007 Jul 07;335(7609):24-27 [FREE Full text] [doi: 10.1136/bmj.39246.581169.80] [Medline: 17615222]
5. Schmajuk G, Trivedi AN, Solomon DH, Yelin E, Trupin L, Chakravarty EF, et al. Receipt of disease-modifying antirheumatic drugs among patients with rheumatoid arthritis in Medicare managed care plans. JAMA 2011 Feb 02;305(5):480-486 [FREE Full text] [doi: 10.1001/jama.2011.67] [Medline: 21285425]
6. Desai SP, Yazdany J. Quality measurement and improvement in rheumatology: rheumatoid arthritis as a case study. Arthritis Rheum 2011 Dec;63(12):3649-3660 [FREE Full text] [doi: 10.1002/art.30605] [Medline: 22127687]
7. Yazdany J, Bansback N, Clowse ME, Collier D, Laws K, Liao K, et al. Practice-Level Variation in Quality of Care in the Acr's Rheumatology Informatics System for Effectiveness (RISE) Registry. 2016 Presented at: 2016 ACR/ARHP Annual Meeting; November 11-16, 2016; Washington, DC URL: https://acrabstracts.org/abstract/practice-level-variation-in-quality-of-care-in-the-acrs-rheumatology-informatics-system-for-effectiveness-rise-registry/ [doi: 10.1136/annrheumdis-2017-eular.5640]
8. Chernitskiy V, DeVito A, Neeman N, Sehgal N, Yazdany J. Integrating Collection of Rheumatoid Arthritis Disease Activity and Physical Function Scores into an Academic Rheumatology Practice to Improve Quality of Care. 2014 Presented at: 2014 ACR/ARHP Annual Meeting; November 14-19, 2014; Boston, MA URL: https://tinyurl.com/y8cccst8 [doi: 10.1093/rheumatology/kev088.097]
9. Wells M, Sadun R, Jayasundara M, Holdgate N, Mohammad S, Weiner J, et al. Sustained Improvement in Documentation of Disease Activity Measurement As a Quality Improvement Project at an Academic Rheumatology Clinic. 2016 Presented at: 2016 ACR/ARHP Annual Meeting; November 11-16, 2016; Washington, DC URL: https://tinyurl.com/3p33k3wb
10. Collier DS, Kay J, Estey G, Surrao D, Chueh HC, Grant RW. A rheumatology-specific informatics-based application with a disease activity calculator. Arthritis Rheum 2009 Apr 15;61(4):488-494 [FREE Full text] [doi: 10.1002/art.24345] [Medline: 19333976]

11. Newman ED, Lerch V, Billet J, Berger A, Kirchner HL. Improving the quality of care of patients with rheumatic disease using patient-centric electronic redesign software. Arthritis Care Res (Hoboken) 2015 Apr;67(4):546-553 [FREE Full text] [doi: 10.1002/acr.22479] [Medline: 25417958]
12. Yazdany J, Bansback N, Clowse M, Collier D, Law K, Liao KP, et al. Rheumatology Informatics System for Effectiveness: A National Informatics-Enabled Registry for Quality Improvement. Arthritis Care Res (Hoboken) 2016 Dec;68(12):1866-1873 [FREE Full text] [doi: 10.1002/acr.23089] [Medline: 27696755]
13. Anderson J, Caplan L, Yazdany J, Robbins ML, Neogi T, Michaud K, et al. Rheumatoid arthritis disease activity measures: American College of Rheumatology recommendations for use in clinical practice. Arthritis Care Res (Hoboken) 2012 May;64(5):640-647 [FREE Full text] [doi: 10.1002/acr.21649] [Medline: 22473918]
14. Barber CEH, Zell J, Yazdany J, Davis AM, Cappelli L, Ehrlich-Jones L, et al. 2019 American College of Rheumatology Recommended Patient-Reported Functional Status Assessment Measures in Rheumatoid Arthritis. Arthritis Care Res (Hoboken) 2019 Dec;71(12):1531-1539 [FREE Full text] [doi: 10.1002/acr.24040] [Medline: 31709771]
15. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol 1992 Jun;45(6):613-619. [doi: 10.1016/0895-4356(92)90133-8] [Medline: 1607900]
16. Atkins DC, Baldwin SA, Zheng C, Gallop RJ, Neighbors C. "A tutorial on count regression and zero-altered count models for longitudinal substance use data": Correction to Atkins et al. (2012). Psychol Addict Behav 2013;27(2):379-379. [doi: 10.1037/a0033147]
17. Izadi Z, Schmajuk G, Gianfrancesco M, Subash M, Evans M, Trupin L, et al. Rheumatology Informatics System for Effectiveness (RISE) Practices See Significant Gains in Rheumatoid Arthritis Quality Measures. Arthritis Care Res (Hoboken) 2020 Sep 16. [doi: 10.1002/acr.24444] [Medline: 32937026]
18. MacKenna B, Bacon S, Walker AJ, Curtis HJ, Croker R, Goldacre B. Impact of Electronic Health Record Interface Design on Unsafe Prescribing of Ciclosporin, Tacrolimus, and Diltiazem: Cohort Study in English National Health Service Primary Care. J Med Internet Res 2020 Oct 16;22(10):e17003 [FREE Full text] [doi: 10.2196/17003] [Medline: 33064085]
19. Ford E, Carroll J, Smith H, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016 Sep;23(5):1007-1015 [FREE Full text] [doi: 10.1093/jamia/ocv180] [Medline: 26911811]
20. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. J Biomed Inform 2015 Oct;57:28-37 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.010] [Medline: 26187250]
21. Gandrup J, Li J, Izadi Z, Gianfrancesco M, Ellingsen T, Yazdany J, et al. Three Quality Improvement Initiatives and Performance of Rheumatoid Arthritis Disease Activity Measures in Electronic Health Records: Results From an Interrupted Time Series Study. Arthritis Care Res (Hoboken) 2020 Feb;72(2):283-291 [FREE Full text] [doi: 10.1002/acr.23848] [Medline: 30740931]
22. Newman E, Sharma T, Meadows A, Brown J, Rowe M, Vezendy S. Rheumatoid Arthritis Quality Measures – Automated Display of Care Gaps and Capture of Physician Decision Making at the Clinic Visit. 2015 Presented at: 2015 ACR/ARHP Annual Meeting; November 6-11, 2015; San Francisco, CA URL: https://tinyurl.com/m59a2nhh
23. Sittig DF, Belmont E, Singh H. Improving the safety of health information technology requires shared responsibility: It is time we all step up. Healthc (Amst) 2018 Mar;6(1):7-12 [FREE Full text] [doi: 10.1016/j.hjdsi.2017.06.004] [Medline: 28716376]
24. Horsky J, Kuperman GJ, Patel VL. Comprehensive analysis of a medication dosing error related to CPOE. J Am Med Inform Assoc 2005;12(4):377-382 [FREE Full text] [doi: 10.1197/jamia.M1740] [Medline: 15802485]
25. Middleton B, Bloomrosen M, Dente MA, Hashmat B, Koppel R, Overhage JM, American Medical Informatics Association. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. J Am Med Inform Assoc 2013 Jun;20(e1):e2-e8 [FREE Full text] [doi: 10.1136/amiajnl-2012-001458] [Medline: 23355463]
26. Singh H, Classen D, Sittig D. Creating an oversight infrastructure for electronic health record-related patient safety hazards. J Patient Saf 2011 Dec;7(4):169-174 [FREE Full text] [doi: 10.1097/PTS.0b013e31823d8df0] [Medline: 22080284]

Abbreviations

ACR: American College of Rheumatology
AHRQ: Agency for Healthcare Research and Quality
AIC: Akaike’s Information Criterion
BIC: Bayesian Information Criterion
CCI: Charlson comorbidity index score
CMS: Centers for Medicare & Medicaid Services
DA: disease activity
EHR: electronic health record
FS: functional status
ICD: International Classification of Diseases

IRR: Incidence rate ratio
NQF: National quality forum
PRO: patient-reported outcome
RA: rheumatoid arthritis
RISE: Rheumatology Informatics System for Effectiveness
ZINB: zero-inflated negative binomial
ZIP: zero-inflated Poisson

Edited by C Lovis; submitted 14.06.21; peer-reviewed by H Salmasian; comments to author 25.07.21; accepted 24.09.21; published 12.11.21.

Please cite as:
Hammam N, Izadi Z, Li J, Evans M, Kay J, Shiboski S, Schmajuk G, Yazdany J
The Relationship Between Electronic Health Record System and Performance on Quality Measures in the American College of Rheumatology's Rheumatology Informatics System for Effectiveness (RISE) Registry: Observational Study
JMIR Med Inform 2021;9(11):e31186
URL: https://medinform.jmir.org/2021/11/e31186
doi:10.2196/31186
PMID:34766910

Original Paper

Clinical Impact of an Analytic Tool for Predicting the Fall Risk in Inpatients: Controlled Interrupted Time Series

Insook Cho[1,2], PhD; In sun Jin[3], PhD; Hyunchul Park[4], MBA; Patricia C Dykes[2,5], PhD

[1]Nursing Department, College of Medicine, Inha University, Incheon, Republic of Korea

[2]The Center for Patient Safety Research and Practice, Division of General Internal Medicine, Brigham and Women's Hospital, Boston, MA, United States

[3]Department of Nursing, National Health Insurance Service Ilsan Hospital, Gyeonggi-do, Republic of Korea

[4]Graduate School of Information & Telecommunications, Konkuk University, Seoul, Republic of Korea

[5]Harvard Medical School, Boston, MA, United States

Corresponding Author:
Insook Cho, PhD
Nursing Department
College of Medicine
Inha University
100 Inha-ro, namu-gu
Incheon, 22212
Republic of Korea
Phone: 82 01042323943
Fax: 82 32 874 8201
Email: insook.cho@inha.ac.kr

Abstract

Background: Patient falls are a common cause of harm in acute-care hospitals worldwide. They are a difficult, complex, and common problem requiring a great deal of nurses' time, attention, and effort in practice. The recent rapid expansion of health care predictive analytic applications and the growing availability of electronic health record (EHR) data have resulted in the development of machine learning models that predict adverse events. However, the clinical impact of these models in terms of patient outcomes and clinicians' responses is undetermined.

Objective: The purpose of this study was to determine the impact of an electronic analytic tool for predicting fall risk on patient outcomes and nurses' responses.

Methods: A controlled interrupted time series (ITS) experiment was conducted in 12 medical-surgical nursing units at a public hospital between May 2017 and April 2019. In six of the units, the patients' fall risk was assessed using the St. Thomas' Risk Assessment Tool in Falling Elderly Inpatients (STRATIFY) system (control units), while in the other six, a predictive model for inpatient fall risks was implemented using routinely obtained data from the hospital's EHR system (intervention units). The primary outcome was the rate of patient falls; secondary outcomes included the rate of falls with injury and analysis of process metrics (nursing interventions that are designed to mitigate the risk of fall).

Results: During the study period, there were 42,476 admissions, of which 707 were for falls and 134 for fall injuries. Allowing for differences in the patients' characteristics and baseline process metrics, the number of patients with falls differed between the control (n=382) and intervention (n=325) units. The mean fall rate increased from 1.95 to 2.11 in control units and decreased from 1.92 to 1.79 in intervention units. A separate ITS analysis revealed that the immediate reduction was 29.73% in the intervention group (z=–2.06, P=.039) and 16.58% in the control group (z=–1.28, P=.20), but there was no ongoing effect. The injury rate did not differ significantly between the two groups (0.42 vs 0.31, z=1.50, P=.134). Among the process metrics, the risk-targeted interventions increased significantly over time in the intervention group.

Conclusions: This early-stage clinical evaluation revealed that implementation of an analytic tool for predicting fall risk may to contribute to an awareness of fall risk, leading to positive changes in nurses' interventions over time.

Trial Registration: Clinical Research Information Service (CRIS), Republic of Korea KCT0005286; https://cris.nih.go.kr/cris/search/detailSearch.do/16984

(JMIR Med Inform 2021;9(11):e26456) doi:10.2196/26456

KEYWORDS

clinical effectiveness; data analytics; event prediction; inpatient falls; process metrics

Introduction

Background

Inpatient falls are preventable adverse events that are the top 10 sentinel events in hospitals. Up to 1 million fall events occur annually in the United States, and the average cost of each event has been estimated at $7900–$17,099 (2019 USD) [1,2]. On average, ~400-700 falls occur annually in Korean tertiary academic hospitals [3-5].

Despite the availability of a considerable body of literature on fall prevention and reduction, falls remain a difficult, complex, and common problem that consume a great deal of time, attention, and mitigation efforts among nurses in practice [6,7]. Considering the studies on inpatient falls, most falls are preventable through tailored interventions and universal fall precautions [8]. However, fall prevention efforts are hindered by the inability to accurately estimate the risk of falling [9,10]. Several risk assessment tools developed using heuristic approaches have been widely used to estimate fall risk in practice. However, evidence regarding the efficacy of those tools is lacking [11,12], potentially resulting in a high false-positive rate and consequently increased burden on nurses. In addition, rating fall risk without identifying the underlying source uses nursing time but does not inform preventative interventions [13]. Our clinical observations reveal that nurses frequently tend to rely only on several universal precautions, not considering risk factors [14]. Implementation of cognitive, toileting-related, or sensory- and sleep-related assessments and interventions was rare.

The increased adoption of electronic health record (EHR) systems over the past decade has stimulated the development of predictive fall risk models using machine learning techniques, which are reported to exhibit better predictive performance than the existing fall risk assessment tools alone [15-18]. However, most of these models have not been validated in multiple settings, and their implementation is restricted by their use of aggregated data by hospital admission rather than by patient-days. None of these models have been evaluated prospectively to assess their performance or their impact on nursing practice. Nursing predictive analytics can include information regarding the likelihood of a future patient event through risk prediction models, which incorporate multiple predictor variables obtained automatically from the EHR. If such models are integrated into EHR systems, nurses can prospectively obtain information to inform their decision making on fall prevention intervention planning.

In this study, we used the prediction model that was developed in our previous study [18]. This model was designed to use nursing process data from EHRs and to consider nurses' fall prevention workflow. Automatic and manual chart reviews were performed to identify all positive events in the retrospective data. The aim of this prospective study was to determine the effect of a predictive fall risk analytic tool on fall outcomes in patients admitted to 12 medical surgical units in South Korea, as well as their impact on nurses' responses. This study hypothesized that providing nurses with information about patients' likelihood of falling within 24 hours of admission, based on data routinely captured in EHRs, would enable nurses to provide risk-targeted interventions and contribute to a reduction in patient fall rates.

Development of an Inpatient Fall Risk Prediction Model

This research team previously reported on the development of a fall risk prediction model [18]. Briefly, concepts of fall risk factors and preventive care were identified using two international practice guidelines [10,19] and two implementation guidelines [20,21] on preventing inpatient falls. Two standard vocabularies, the Logical Observation Identifiers Names and Codes [22] and the International Classification for Nursing Practice [22,23], were used to represent the concepts in the prediction model, which was then itself represented using a probabilistic Bayesian network.

The model was tested in two study cohorts obtained from two hospitals with different EHR systems and nursing vocabularies. The model concepts were mapped to local data elements of each EHR system, and two implementation models were developed for a proof-of-concept approach, followed by cross-site validation. The EHR data included in the model were demographics, administrative information, medications, Korean patient classification based on nursing needs, the fall risk assessment tool, and nursing fall risk prevention processes, including assessments and interventions. The two implementation models exhibited error rates of 11.7% and 4.87%, with *c* statistics of 0.96 and 0.99, respectively. The model performed 27% and 34% better than the existing Hendrich II tool [24] and the St. Thomas' Risk Assessment Tool in Falling Elderly Inpatients (STRATIFY) system [25], respectively.

Clinical Implementation of the Intelligent Nursing @ Safety Improvement Guide of Health information Technology System

The validation site model was implemented at a 900-bed public hospital in the metropolitan area of Seoul (Republic of Korea) that used STRATIFY to assess fall risks for all inpatients. The project, named Intelligent Nursing @ Safety Improvement Guide of Health information Technology (IN@SIGHT), was designed as a platform to support analytic tools as part of the infrastructure of a hospital EHR system, starting with a fall prediction analytic tool. The fall prediction analytic tool was integrated into the locally developed EHR system that had been in use for more than 10 years. The tool was deployed in 6 targeted nursing units (intervention group) on April 5, 2017, and all 204 nurses at those units automatically received the prediction results on a daily basis. This implementation process involved the chief of the Nursing Department, unit managers, unit champions, personnel of the Department of Medical Informatics, and the Patient Safety Committee. For 3 months before system deployment, three sessions of education on the IN@SIGHT system were provided to the intervention group, followed by peer-to-peer education provided by unit champions.

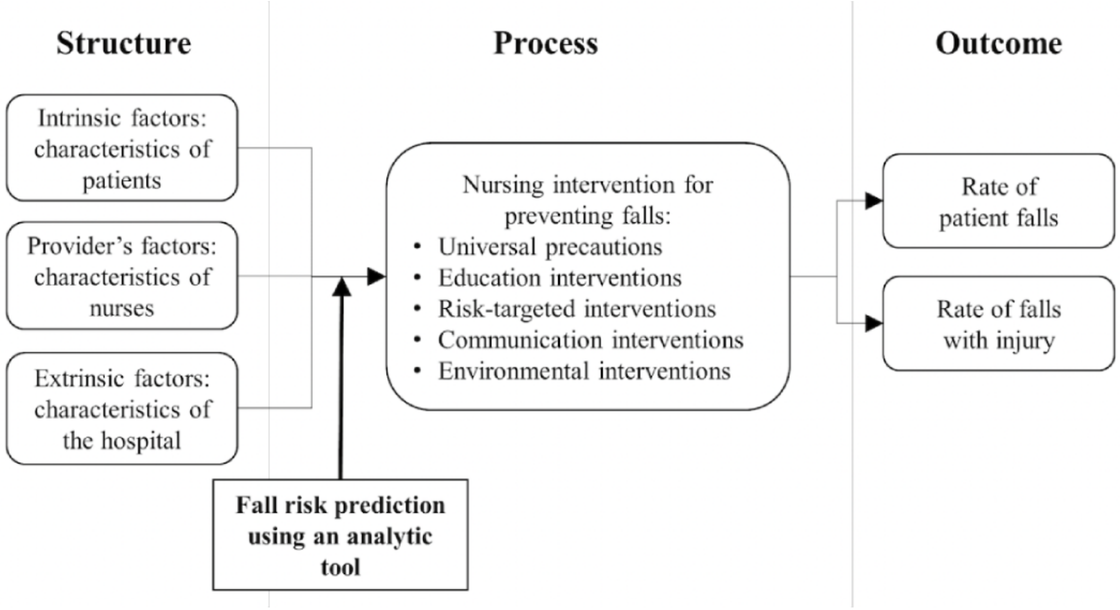
The Nursing Department decided to replace the existing STRATIFY with the analytic tool during this quasi-experimental study. The original model was customized by replacing the six data elements of STRATIFY with proxy data elements in the EHRs. The adjusted model, consisting of 40 nodes and 68 links, had an error rate of 9.3%, a spherical payoff of 0.92, and a *c* statistic of 0.87. Related work processes were redefined, and the existing fall prevention documentation screen of the EHRs was modified. The hospital decided to deliver the risk information in dichotomized format, with at-risk and no-risk categories at a cutoff point of 15%, which provided a high specificity of 89.4%. The analytic tool triggered an "at-risk" alert on the EHR system when the user selected an at-risk patient.

Methods

Study Framework and Objectives

A study framework was developed based on a nursing role effectiveness model (Figure 1) [26]. The original model was based on the structure-process-outcome design of the Donabedian quality care model but was reformulated for this empirical testing, focusing on nurses' independent roles in the process component. We assumed that the characteristics of the patients, nurses, and hospital were fixed because the study involved a single institution and the same medical-surgical units. The hypothesis being tested was that the intervention of fall risk prediction would affect the appropriateness of multifactorial interventions and would be followed by changes in outcome.

Figure 1. Conceptual framework of the study.


Structure
Process
Outcome
Intrinsic factors: characteristics of patients
Provider's factors: characteristics of nurses
Extrinsic factors: characteristics of the hospital
Nursing intervention for preventing falls:
• Universal precautions
• Education interventions
• Risk-targeted interventions
• Communication interventions
• Environmental interventions
Rate of patient falls
Rate of falls with injury
Fall risk prediction using an analytic tool

In accordance with the aim of this study, the impact of an electronic analytic tool for fall risk prediction on patient outcomes and nurses' responses was explored by addressing the following specific research questions:

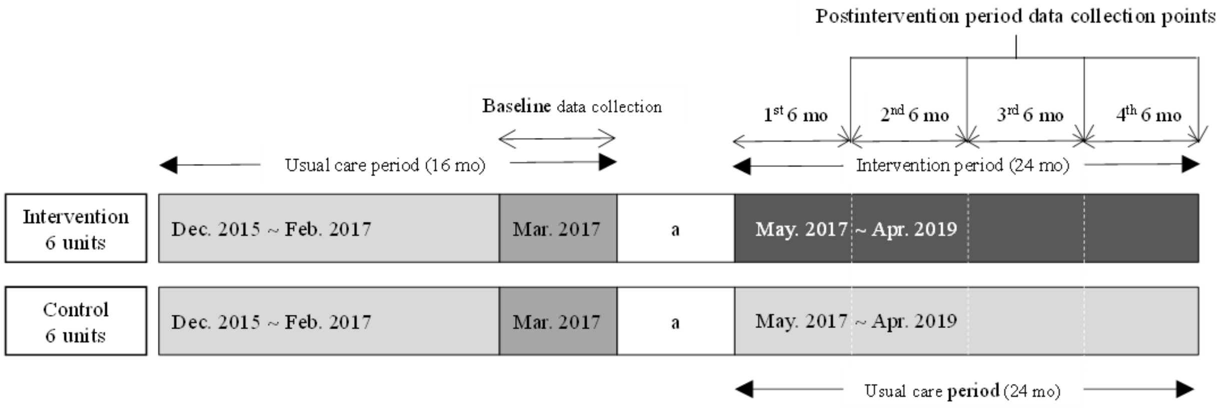
1. Did the predictive analytic tool influence the quality of nursing care as assessed using outcome indicators?
2. Did the predictive analytic tool affect nursing fall prevention activities provided to patients?
3. How did the effects change over time?

Study Design and Setting

This nonrandomized controlled trial used an interrupted time series (ITS) design. To control for bias due to time-varying confounders, such as other quality improvement (QI) initiatives occurring in parallel with the intervention and other events, the 12 medical-surgical units were selected and allocated to 1 of 2 groups using pairs of units matched according to the known fall rates and unit characteristics for individual units (Figure 2). All of the nurses and eligible patients participated in this study between May 1, 2017 and April 30, 2019. The patients met the following criteria: age ≥ 18 years and admitted to the hospital for >1 day in departments other than pediatrics, psychiatrics, obstetrics, and emergency care. The preintervention period was set at 16 months, which was the maximum retrospective time window. The 12 nursing units' nurse staffing ratios were changed at the time due to a policy for comprehensive nursing service in the Korean government's national health insurance. The postintervention period was 24 months. Process metrics, which measure the delivery of fall risk mitigation interventions by nurses to patients, were analyzed every 6 months.

This study was approved by the hospital's ethical review board (IRB no. NHIMC 2016-08-005). A waiver of informed consent was granted by the IRB due to the QI nature of the intervention, thus enabling the inclusion of all patients and nurses in the participating units. This study followed the Transparent Reporting of Evaluations with Nonrandomized Designs (TREND) reporting guidelines [27].

Figure 2. Nonequivalent-group design of the study.



[a] One-month clinical implementation or rest period.
Monthly rate of patient falls was collected through the whole period.
Monthly rate of falls with injury was collected only at the postintervention period.
Baseline data collection and postintervention period data collection include the data of patient characteristics, admission information, and process metrics.

Intervention

Nurses in the intervention units received 24-hour fall risk prediction results for each patient every morning. These results could be overridden based on the nurses' clinical judgment, such as when patients were receiving treatments, procedures, operations, or fall related high-risk drugs or whether they suffered a fall, seizure, or syncope. The fall risk predictions were created by the analytic tool using the data collected within the past 24 hours. For missing data, a priori values from the day before were assigned first, and then a replacement was used: a mean value for continuous variables and a modal value for categorical variables. Nurses in the intervention units used the STRATIFY risk assessment tool only on the day of admission. When an at-risk patient was selected by nurses in the EHR system, they received an alert once each shift informing them that the patient was at risk and were guided to a care plan screen that listed pertinent interventions ordered by priority according to the patient's risk factors. Nurses in the control units used only STRATIFY to assess fall risk according to their individual clinical judgment. They were able to manually open the same care plan window through menu navigation but received no alerts for at-risk patients.

Outcome Measures

The primary outcome was the overall rate of patient falls per 1000 patient-days during the study period, as defined by the National Database of Nursing Quality Indicator (NDNQI) outcome metrics of the American Nurses Association [28,29]:

> *A patient fall is a sudden, unintentional descent, with or without injury to the patient, that results in the patient coming to rest on the floor, on or against some other surface (e.g., a counter), on another person, or on an object (e.g., a trash can). NDNQI counts only falls that occur on an eligible inpatient unit that reports falls. When a patient rolls off a low bed onto a mat or is found on a surface where you would not expect to find a patient, this is considered a fall. If a patient who is attempting to stand or sit falls back onto a bed, chair, or commode, this is only counted as a fall if the patient is injured. All unassisted and assisted falls... are to be reported, including falls attributable to physiological factors such as fainting (known as physiological falls).*

The secondary outcomes were the overall rate of falls with injury, and process metrics. The rate of falls with injury was also measured using the aforementioned NDNQI definition. Process metrics were defined according to the Institute for Healthcare Improvement definition as "process indicators that measure compliance with key components of evidence-based prevention" [30]. Methods for identifying and defining key components of fall prevention are described elsewhere [31]. In brief, nursing activities identified by international guidelines on preventing falls are categorized into 17 components; of these, 7 nursing intervention components were used in this study. Process metrics were used to determine whether nursing behaviors independently affected patient outcomes. Each process metric measured the proportion of at-risk patients who were provided with targeted interventions. For example, all hospitalized patients are expected to be assessed for fall risk factors within 24 hours of admission, and at-risk patients are expected to receive risk-targeted interventions within 24 hours of their risk designation.

Data Collection

Monthly rates of patient falls were collected from 16 months before the experiment started (the preintervention period) from the hospital's quality assurance department to provide a baseline reference for comparisons. However, monthly rates of falls with injury before the experiment were not comparable due to differences in the criteria used to calculate them; only severe injuries were used as a sentinel event at the hospital. For process metrics, 1 month of data from before the experiment were collected as a baseline. During the study, data on patient demographics and medications, nursing activities, STRATIFY data, and administrative information were collected from the EHR system, and fall data were collected from the hospital's quality assurance department. To monitor and minimize the underreporting rate noted previously [31,32], the Nursing Department provided education to all units on the principles of reporting and documentation, and they provided monthly chart reviews and feedback.

Sample Size and Statistical Analysis

The study hypothesis was that the fall rate would be reduced by 15% during the 24-month implementation of the prediction program. We conservatively estimated the required sample size based on previous research [18] by assuming a fall rate in the control group of 2.0 per 1000 patient-days, an average of 15,000 patient-days per unit over 12 months, and an average 1700 admissions. The required number of falls in the control group was calculated using a Poisson distribution: $D_0 = z^2(\theta + 1)/\theta(\log_e \theta)^2$ [7]. We applied z=2.0; detecting a rate ratio (θ) of 0.85 between groups at the 5% significance level with a statistical power of 80% required 610 falls, which corresponded to a 24-month period for the 12 units.

The participant characteristics were compared using chi-square tests for categorical variables and *t* tests for continuous variables. The primary outcome of the rate of patient falls was compared by the controlled ITS, incorporating the control series analysis and the uncontrolled ITS [33]. We fit negative binomial models, including a lagged dependent variable to control for serial autocorrelation and monthly dummy variables, to generate seasonal fixed effects in each model. Each model included three variables to measure the relationship between time and patient fall rates: (1) a continuous variable to represent the underlying temporal trends, (2) a dummy variable for dates after May 1, 2017, to determine the change in fall rate related to the intervention, and (3) a continuous time variable beginning on that date to represent the change in slope. The coefficients of the second and third variables indicated whether the intervention had immediate and ongoing effects on the fall rate, respectively. The Student *t* test and a comparative time series analysis were conducted to analyze the rate of falls with injury, and chi-square analysis was used for the comparison of process metrics between groups.

Results

Patient Characteristics

This study involved 42,476 admissions of 40,345 unique patients in 12 units, corresponding to 362,805 patient-days in nursing units across both the control and intervention groups. In total, 2131 patients (5.02% of all admissions) were admitted to both an intervention and a control unit at different times. The patient characteristics differed significantly between the two groups (Table 1). Compared with the intervention units, the control units were characterized by older patients, a longer stay, fewer female patients, and more patients with a fall history at admission; rates of secondary diagnoses and surgical procedures were also higher. Approximately half of the patients in the intervention group had a respiratory or digestive disease or any form of cancer, while control patients had a greater diversity of primary diagnoses.

Table 1. Characteristics of patients in the intervention and control groups.

Variable	Intervention (n=24,336)	Control (n=18,140)	*P* value
Primary medical diagnosis, n (%)			
Respiratory or digestive disease	6150 (25.21)	3472 (19.14)	<0.001
Cancer	5990 (24.61)	2382 (13.13)	<0.001
Symptom or injury	2784 (11.44)	2561 (14.12)	<0.001
Cardiovascular disease	995 (4.09)	3096 (17.07)	<0.001
Benign tumor	860 (3.53)	211 (1.16)	<0.001
Infectious disease	514 (2.11)	388 (2.14)	<0.001
Neurologic disease	182 (0.75)	597 (3.29)	<0.001
Other[a]	6861 (28.19)	5433 (29.95)	<0.001
Other variables			
Age (years), mean (95% CI)	61.45 (61.23-61.67)	65.30 (65.05-65.54)	<0.001
Length of stay (days), mean (95% CI)	7.96 (7.91-8.00)	9.25 (9.13-9.37)	<0.001
Sex (female), n (%)	12,512 (51.41)	9053 (49.91)	0.002
History of fall at admission, n (%)	2873 (11.88)	4138 (23.58)	<0.001
Secondary diagnoses, n (%)	10,641 (43.73)	9361 (51.60)	<0.001
History of surgical procedures, n (%)	2483 (10.20)	8575 (47.27)	<0.001

[a] Including genitourinary, musculoskeletal, eye, ear, and skin diseases.

Primary Outcome: Rate of Patient Falls

There were 325 fall events in the intervention group and 382 in the control group. The mean monthly rate of falls decreased from 1.92 to 1.79 in the intervention group and increased from 1.95 to 2.11 in the control group**.** Controlled ITS analysis revealed that the postintervention versus preintervention change in the incidence rate ratio of the fall rate was −0.10 (SE 0.04, *P*=.014). There was no seasonal effect.

Due to the significant differences in patient characteristics between the control and intervention groups, we conducted separate before versus after comparisons between a period of time postintervention and the same period of time preintervention. In the intervention group, there was a significant reduction in the rate of falls of 29.73% (0.57 falls per 1000 patient-days) immediately postintervention (SE 0.14, *P*=.039). During the preintervention period, the slope exhibited a slightly decreasing trend (SE 0.08, *P*=.344), and after the intervention, the slope increased slightly but not significantly so (slope=0.01, SE 0.01, *P*=.059; Table 2). In the control group, there was a nonsignificant reduction in the rate of falls of 16.58% (0.16 falls per 1000 patient-days; SE 0.13, *P*=.20). The slope before the intervention increased (change in slope=0.08, SE 0.72, *P*=.292), while after the intervention, the slope increased slightly (change in slope=0.01, SE 0.01, *P*=.057).

Table 2. Results of interrupted time series analysis of rates of patient falls.

Group	Preintervention period trend	Change immediately after introduction of intervention	Postintervention period trend
Intervention	−0.07 (−0.22 to 0.08)	−0.30 (−0.58 to –0.14)[a]	0.01 (<−0.01 to 0.02)
Control	0.08 (−0.07 to 0.22)	−0.17 (−0.42 to 0.09)	0.01 (<−0.01 to 0.02)

[a] *P*=.04.

Data are rate ratio (95% CI) values.

Secondary Outcomes: Fall With Injury Rates and Process Metrics

During the intervention period, the mean monthly injury rate per 1000 patient-days was 0.42 in the intervention group and 0.31 in the control group. The comparative time series analysis revealed a nonsignificant increase in the rate ratio of 0.18 (*z*=1.50, *P*=.134).

Regarding process metrics, fall risk assessment was not conducted in almost three-quarters of patient-days in the control group, while in the intervention group, fall risk assessment was conducted on 100% of patient-days (Table 3). During the intervention period, the frequency of at-risk days was almost 40% in the control group but ranged from 24.5% to 34.6% in the intervention group. There was a high rate of implementation of a fall risk tool within 24 hours of hospital admission in both groups, although rates fluctuated over time in the control group. Rates of assessment of injury risk factors were assessed in all patients in the intervention group; these data were not available for the control group. Universal fall precautions and fall prevention education were provided to most patients in the control group consistently throughout the study period. Rates of implementation of communication and environmental interventions were initially significantly better in the control group than in the intervention group; however, those for the intervention group increased over time and had caught up with the control group by the third observation point. Although the rate of risk-targeted interventions incrementally increased in both groups, the intervention group showed better adherence than the control group at the fourth observation point (29.5% vs 18.1%, *P*<.001).

Table 3. Temporal changes in process metrics in the control and intervention groups.

Item	Baseline (1 month)	First 6 months of intervention	Second 6 months of intervention	Third 6 months of intervention	Fourth 6 months of intervention
Base information					
Patient-days	8254 vs 4207[a]	45,133 vs 31,675	46,403 vs 39,733	44,418 vs 44,741	42,553 vs 43,161
Days on which no risk assessment performed, %	72.5 vs 73.4[b]	0 vs 72.6	0 vs 77.1	0 vs 71.7	0 vs 79.8
At-risk days, %	43.0 vs 42.1[b]	24.5 vs 43.5[c]	31.4 vs 38.6[c]	32.7 vs 42.9[c]	34.6 vs 41.5[c]
Process metrics: patients assessed within 24 hours of hospital admission, %					
Use of a fall risk tool	99.3 vs 98.6 [b]	100.0 vs 99.2[c]	100.0 vs 70.8[c]	100.0 vs 95.3[c]	100.0 vs 98.8[c]
Injury risk factors (ABCs[d])	0 vs 0[b]	100.0 vs 0	100.0 vs 0	100.0 vs 0	100.0 vs 0
Process metrics: at-risk patients who received within 24 hours of risk identification, %					
Universal precautions[e]	86.1 vs 100.0[c]	69.7 vs 78.9[c]	88.8 vs 99.9[c]	37.8 vs 99.9[c]	91.2 vs 99.9[c]
Education interventions[e]	86.1 vs 100.0[c]	69.7 vs 78.9[c]	88.8 vs 99.9[c]	33.1 vs 98.1[c]	79.6 vs 97.8[c]
Risk-targeted interventions	<0.01 vs <0.01	<0.01 vs <0.01	<0.01 vs <0.01	12.5 vs 13.3[b]	29.5 vs 18.1[c]
Communication interventions[e]	61.7 vs 79.4[c]	87.6 vs 99.9[c]	76.0 vs 81.1[c]	30.2 vs 38.7[c]	66.2 vs 66.7[b]
Environmental interventions[e]	61.7 vs 79.4[c]	87.6 vs 99.9[c]	76.0 vs 81.1[c]	39.5 vs 54.9[c]	76.7 vs 76.0[b]

[a] All data shown as intervention group versus control group.

[b] Not significant.

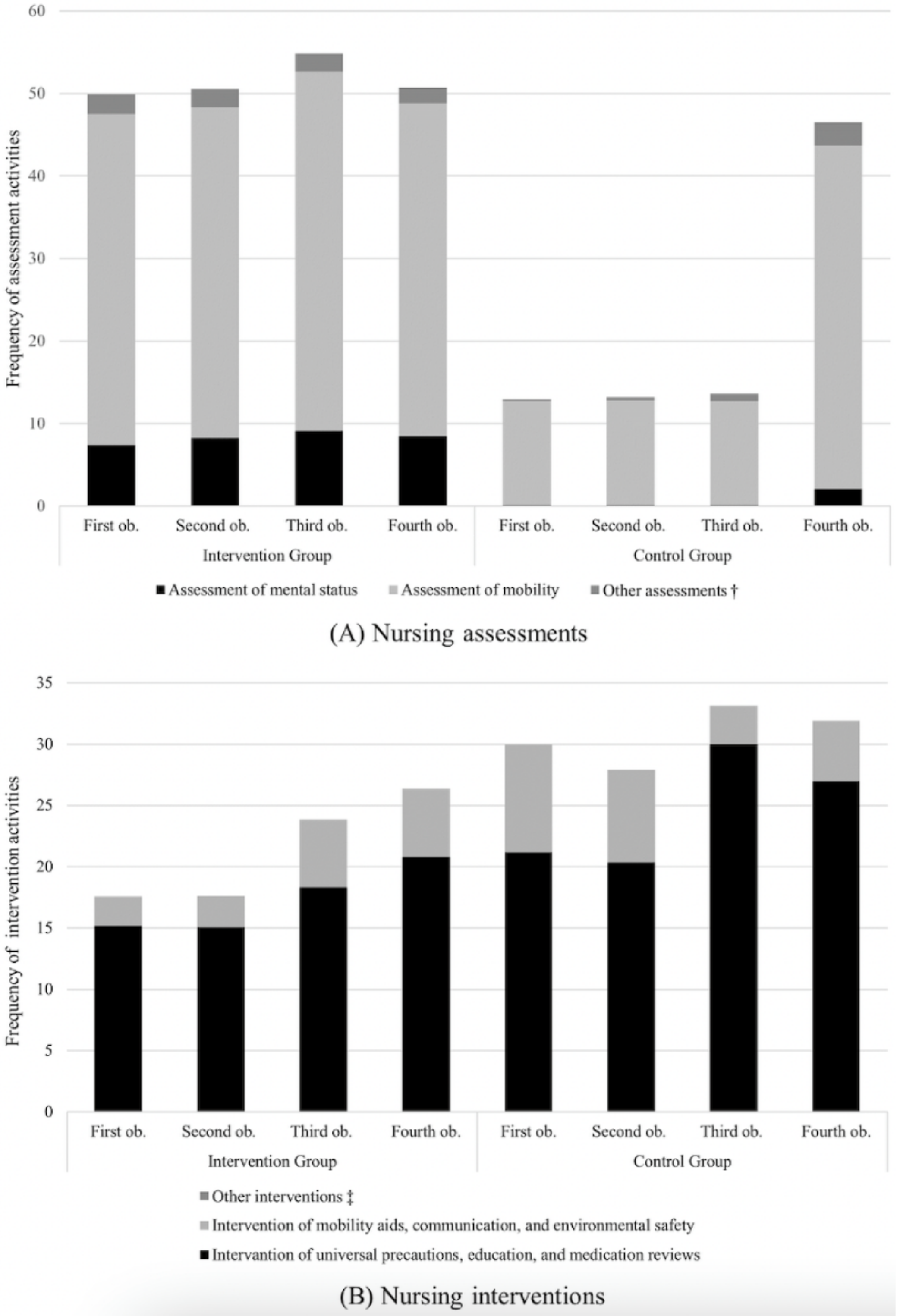
[c] *P*<.001.

[d] ABCs: age, bone health, anticoagulants, and current surgery (function that was performed automatically in the intervention group).

[e] Data collection not categorized in detail from baseline to the second observation point.

For the care components of nursing assessments, nurses in the intervention group performed various observation types, such as mental status, cognitive function, communication ability, and incontinence, including mobility, at each observation point (Figure 3A), while those in the control group appeared to focus largely on mobility assessments, the frequency of which suddenly increased at the last observation point. Universal precautions, education, and medication reviews were the most common interventions in both groups (Figure 3B). Although the frequency of interventions was lower in the intervention group than in the control group, there was a steady increase over time.

Figure 3. Changes in nursing assessments (A) and interventions (B) according to care components. ob.: observation point; †includes assessments of cognitive function, communication ability, gait status, incontinence, sleep pattern, and use of constraints; ‡includes interventions of toileting aids and for impaired mental and cognitive function, impaired sensory function, and sleep disturbance.



Discussion

Principal Findings

Implementation of an electronic analytic tool designed to predict fall risk was associated with reduced fall rates among inpatients at a public hospital in South Korea. However, comparison with the control group should be considered with caution due to notable differences in patient characteristics between the two groups. There was no significant difference in the rate of falls with injury between the control and intervention groups. Use of the electronic analytic tool was feasible, and it was accepted by nurses and improved the completion of risk assessments. Moreover, the process metrics for multifactorial and risk-targeted interventions for at-risk days were lower in the intervention group but increased over time. These findings suggest that although the effectiveness of an electronic analytic tool may be limited, it has potential as an aid to help nurses make informed clinical decisions.

The main challenges in this study were threefold: (1) random assignment of patients to the study groups was not possible; (2) it was not possible to control for co-interventions or external events at the hospital that may have affected the outcome, including QI activities; and (3) nurses' understanding of the analytic tool developed by a machine learning approach was not assessed. These issues were managed by selecting only medical-surgical units and assigning patients according to the particular characteristics of each unit. A controlled ITS design was adopted to control for time-varying confounders. Finally, the development and validation process of the predictive model and the mechanism of chaining joint probabilities of a Bayesian network were introduced via user education sessions. However, during the study, the research team confronted additional issues that made interpretation of the results challenging. Discussion on these issues is valuable for future research into risk prediction and alerting in real-world settings.

The fall rates of 1.79 and 2.11 in the intervention and control groups, respectively, in this study were lower than previously reported rates of 2.08-4.18 for an intervention study involving a cluster randomized controlled trial (RCT) in four urban US hospitals [34], 3.05 for a cluster RCT in Australia [7], and 2.80 for a US intervention study [35]. However, differences in the patient populations and in the structural elements at the facilities preclude direct comparison [36]. The low fall incidence rate in this study allowed us to observe changes in nursing behaviors over a 24-month follow-up period. A fall prevention intervention will not be effective if it does not influence nurse behaviors. We focused on how the analytic tool can influence nursing behaviors in order to ensure that interventions that are beneficial to patients are routinely provided. Our findings revealed that the intervention group performed more multifactorial patient assessments than the control group; however, the interventions in both groups were limited. Most of the preventive components involved education and medication review, which is perhaps unsurprising since these precautions are routinely applied to all inpatients regardless of their fall risk. Interventions associated with toileting, impaired mental and cognitive function, impaired sensory function, and sleep disturbance were rarely observed in both groups.

According to international guidelines for preventing falls [10,19-21], multifactorial assessment of risks and multifactorial, risk-targeted interventions are basic components of fall prevention strategies. Application of the analytic tool in this study ensured that risk factors were monitored daily for each patient in the intervention group and that alerts were delivered to their nurses via the hospital EHR system. A large increase in data-seeking and data-gathering activities was observed during the first 6 months of observation, whereas notable increases in overall interventions and risk-targeted interventions appeared 12 and 18 months later, respectively. This suggests that adoption of this new approach and its processes by nurses was time dependent and stepwise, in line with the findings of surveys conducted repeatedly during the study period [37]. Those surveys revealed that some nurses reported neutral or even slightly negative attitudes and experiences at the beginning of the study. However, the proportion of negative responses gradually decreased over time. These findings can be understood in terms of the non-adoption, abandonment, scale-up, spread, and sustainability (NASSS) framework [38] to explain the success of technology-supported health or social care programs. Staff members are often initially more concerned about threats to their scope of practice or to the safety and welfare of patients, leading them to initially gather more information about risks. A previous qualitative exploration study [39] that used one-on-one and focus group interviews to investigate nurses' perception of predictive information and how they act upon it found that nurses attempt to gather more information from other sources and review more detailed predictions during periods of uncertainty. Time delays in adoption and changing of behaviors are expected, given that predictive information is relatively new to nurses. The other relevant domain of the NASSS framework is the readiness of a hospital for a predictive analytic tool. The understanding and support, antecedent conditions, and level of readiness for a novel tool at the board level might influence the uptake time by nurses and the internal drivers for scaling up the tool.

Study Limitations

This study had limitations. The control group patients had more comorbidities that rendered them more vulnerable to falls than the intervention group. They were on average 4 years older, had a hospital stay that was 1.3 days longer, and had a greater history of falls. These variables are known important covariates [19], and we did not balance these covariates in the ITS experiment. The differences in these covariates between the two study groups may be attributable to an ascertainment bias issue; it is possible that rather than there being a true reduction in fall rates in the intervention group, more patients at a lower risk of falls were included in that group. Evaluation of the baseline data suggests that the nurses in the control group delivered significantly more fall-preventive interventions to their patients than did those in the intervention group, including more additional risk assessments, universal precautions, educational interventions, and communication and environmental interventions. Thus, control group patients were both more likely to fall and to receive more fall-preventive interventions from nurses. It is unclear how these counterbalancing factors interact and how they may have impacted the outcomes of this study; however, it can be assumed that the greater provision of interventions appears to have contributed to the reduced fall risk in the control group.

The temporal changes in process metrics and nursing activities can provide important clues as to the overall impact of this trial. In a previous study [18], we found that the analytic tool predicted about 20% of patient at-risk days, which was about a half of the rate classified using STRATIFY (~40%-50% of patient-days as at-risk days). The actual rate of falls in the hospital was much lower, at around 0.2% of patient-days. We assumed that more precise up-to-date predictions of fall events would decrease the nurses' burden on redundant interventions induced by false-positive warnings from STRATIFY. The analytic tool approach did not affect the universal fall precautions, but risk-targeted interventions, education, communication, and environmental interventions significantly increased compared with the control group, which remained at a steady state. These findings are meaningful, given that multifactorial interventions, including risk-targeted interventions, prevent anticipated physiologic falls, which are responsible for more than 70% of inpatient falls [34,40]. These process metrics revealed slow but explicit changes in nursing interventions, which indicates that the processes underlying care elements had changed and we could expect subsequent improvement in patient outcomes [41]. Continuous measurement and analysis of process metrics informed our understanding of the effects of interventions on patient outcomes and our interpretation of the effects of confounding, which has rarely been accounted for in previous studies [7,34,42].

Study Design Limitations

The design of this study had several limitations that impacted the interpretation of its findings. First, due to the unexpected differences in baseline characteristics between the intervention and control groups, robust conclusions could not be drawn regarding comparison of the primary outcome between them. Future studies should implement matching techniques, such as propensity score matching [43] or synthetic control approaches [44], to ensure balance between known covariates. Second, implementation of the intervention at a single site over a long study period introduced several challenges that could have reduced the effects of this study trial. One challenge was an unexpected event at the hospital whereby one nursing unit in each group moved to a new location 1 month after study initiation, and nurse staffing was thus reorganized due to the physical reconstruction of the hospital buildings. The fall rate markedly increased for several months in that intervention unit compared with the other five units in the group. However, the control group unit that was relocated showed only a slight increase compared with the other units in their group. The relocations were accompanied by changes in staff nurses and in the medical diagnoses of patients, both of which may have increased the burden on nurses and induced the sudden increase in the fall rate at the unit. Another unexpected event was the routinization of hourly nursing rounds to all inpatients mandated by the hospital's safety committee during the final intervention period. This may have accounted for the sudden increase in nursing assessments observed in the control group. In addition, conducting this study at a single hospital may have an indirect effect on the control units. The unit managers of the control group were also involved in the QI initiatives of this study, along with those of the intervention units. This could have caused a contamination effect, whereby the managers of the control group learned about the study intervention and decided to adopt it for their own units. Third, we were unable to compare the injury fall rates between the pre- and postintervention periods; therefore, the impact of the analytic tool on the rate of falls with injury remains unknown.

Inpatient fall prevention is a difficult and complex issue, for which there is little high-quality evidence [7,45]. Even after taking into account the study limitations, the findings of this early-stage evaluation of an analytic tool demonstrated that the interaction between the tool and nurses was adequate and the tool may have influenced nurses' decisions on preventive interventions. The analytic tool developed herein represents a potential new approach for patient-level risk surveillance and for improving the efficacy of interventions at the system level. The findings and challenges discussed herein will contribute to improving further research on risk prediction and alerting in real-world settings.

Conclusions

This was an early-stage clinical evaluation of a nursing predictive analytic application designed to forecast patient fall events in real time and at the point of care to improve outcomes and reduce costs. The effectiveness of the electronic analytic tool was supported only by the before-after comparison, not by the intervention-control comparison. Nurses were amenable to using the tool in practice, and over the course of the study, there were meaningful changes in process metrics, leading to more multifactorial and risk-targeted interventions to prevent patient falls.

Acknowledgments

We thank Chihang Kim and Jisun Cho of the Nursing Department and Yunjeong Choi of the medical information department at Ilsan Hospital for helping us conduct this study administratively and technically. We also appreciate the clinical staff of the Nursing Department and graduate students involved in data collection, review, and analysis. This study was supported by grants from the Korea Healthcare Technology R&D Project, Ministry for Health and Welfare (no. HI17C0807), the National Research Foundation of Korea (no. NRF-2019R1A2C2007583), and the Ministry of Trade, Industry and Energy of Korea (no. 20004861).

Conflicts of Interest

None declared.

References

1. Spetz J, Brown D, Aydin C. The economics of preventing hospital falls: demonstrating ROI through a simple model. J Nurs Adm 2015;45(1):50-57. [doi: 10.1097/NNA.0000000000000154] [Medline: 25479175]

2. Wong CA, Recktenwald AJ, Jones ML, Waterman BM, Bollini ML, Dunagan WC. The cost of serious fall-related injuries at three midwestern hospitals. Jt Comm J Qual Patient Saf 2011;37(2):81-87. [doi: 10.1016/s1553-7250(11)37010-9] [Medline: 21939135]

3. Lee E, Ahn M, Kim Y, Jo I, Jang D. Development and Effects of Fall Prevention Model: Safety for Utilizing Medical Big Data Based Artificial Intelligence. Seoul, Pub of Korea: Korean Society of Medical Informatics (KOSMI); 2018 Presented at: Proceeding of 2018 Fall KOSMI Conference; November 23-24, 2018; Jeonju-si, Korea URL: http://www.kosmi.org/
4. Cho M, Lee H. Factors associated with injuries after inpatient falls in a tertiary hospital. J Korean Clin Nurs Res 2017;23(2):202-210.
5. Lee JY, Jin Y, Piao J, Lee S. Development and evaluation of an automated fall risk assessment system. Int J Qual Health Care 2016;28(2):175-182. [doi: 10.1093/intqhc/mzv122] [Medline: 26851379]
6. Lopez KD, Gerling GJ, Cary MP, Kanak MF. Cognitive work analysis to evaluate the problem of patient falls in an inpatient setting. J Am Med Inform Assoc 2010;17(3):313-321 [FREE Full text] [doi: 10.1136/jamia.2009.000422] [Medline: 20442150]
7. Barker AL, Morello RT, Wolfe R, Brand CA, Haines TP, Hill KD, et al. 6-PACK programme to decrease fall injuries in acute hospitals: cluster randomised controlled trial. BMJ 2016;352:h6781 [FREE Full text] [doi: 10.1136/bmj.h6781] [Medline: 26813674]
8. Dykes PC, Burns Z, Adelman J, Benneyan J, Bogaisky M, Carter E, et al. Evaluation of a patient-centered fall-prevention tool kit to reduce falls and injuries: a nonrandomized controlled trial. JAMA Netw Open 2020;3(11):e2025889 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.25889] [Medline: 33201236]
9. The Joint Commission. Sentinel Event Alert 55: Preventing Falls and Fall-Related Injuries in Health Care Facilities. 2015. URL: http://www.jointcommission org/assets/1/18/SEA_55 pdf [accessed 2019-01-15]
10. Falls in older people: assessing risk and prevention. National Institute for Health and Care Excellence. 2013. URL: https://www.nice.org.uk/guidance/cg161/resources/falls-in-older-people-assessing-risk-and-prevention-pdf-35109686728645 [accessed 2021-11-08]
11. Cameron I, Dyer S, Panagoda C, Murray G, Hill K, Cumming R, et al. Interventions for preventing falls in older people in care facilities and hospitals. Cochrane Database Syst Rev 2018;9:CD005465 [FREE Full text] [doi: 10.1002/14651858.CD005465.pub4] [Medline: 30191554]
12. Park SH, Kim EK. Systematic review and meta-analysis for usefulness of fall risk assessment tools in adult inpatients. Korean Journal of Health Promotion 2016;16(3):180-191. [doi: 10.15384/kjhp.2016.16.3.180]
13. Dykes P, Khasnabish S, Burns Z, Adkison L, Alfieri L, Bogaisky M, et al. Development and validation of a fall prevention efficiency scale. J Patient Saf 2021:2021. [doi: 10.1097/PTS.0000000000000811] [Medline: 33480645]
14. Suh M, Cho I. Effectiveness of nursing care provided for fall prevention: survival analysis of nursing records in a tertiary hospital. Jpn J Nurs Sci 2021;18(2):e12403. [doi: 10.1111/jjns.12403] [Medline: 33448157]
15. Lindberg DS, Prosperi M, Bjarnadottir RI, Thomas J, Crane M, Chen Z, et al. Identification of important factors in an inpatient fall risk prediction model to improve the quality of care using EHR and electronic administrative data: A machine-learning approach. Int J Med Inform 2020;143:104272 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104272] [Medline: 32980667]
16. Jung H, Park H, Hwang H. Improving prediction of fall risk using electronic health record data with various types and sources at multiple times. Comput Inform Nurs 2020;38(3):157-164. [doi: 10.1097/CIN.0000000000000561] [Medline: 31498252]
17. Oshiro CES, Frankland TB, Rosales AG, Perrin NA, Bell CL, Lo SHY, et al. Fall ascertainment and development of a risk prediction model using electronic medical records. J Am Geriatr Soc 2019;67(7):1417-1422. [doi: 10.1111/jgs.15872] [Medline: 30875089]
18. Cho I, Boo E, Chung E, Bates DW, Dykes P. Novel approach to inpatient fall risk prediction and its cross-site validation using time-variant data. J Med Internet Res 2019;21(2):e11505 [FREE Full text] [doi: 10.2196/11505] [Medline: 30777849]
19. Registered Nurses' Association of Ontario (RNAO). Preventing Falls and Reducing Injury from Falls (4th ed.). 2017. URL: https://rnao.ca/sites/rnao-ca/files/bpg/FALL_PREVENTION_WEB_1207-17.pdf [accessed 2019-01-15]
20. Reducing Falls and Injuries from Falls Getting Started Kit: Evidence Update. 2015. URL: http://www.patientsafetyinstitute.ca/en/toolsResources/Documents/Interventions/Reducing%20Falls%20and%20Injury%20from%20Falls/Falls%20Evidence%20update%202018-01.PDF [accessed 2019-01-15]
21. Ganz D, Huang C, Saliba D, Miake-Lye I, Hempel S, Ensrud K. Preventing falls in hospitals: a toolkit for improving quality of care. Ann Intern Med 2013;158(5 Pt 2):390-396.
22. Vreeman DJ, McDonald CJ, Huff SM. LOINC®: a universal catalog of individual clinical observations and uniform representation of enumerated collections. Int J Funct Inform Personal Med 2010;3(4):273-291 [FREE Full text] [doi: 10.1504/IJFIPM.2010.040211] [Medline: 22899966]
23. World Health Organization. Classifications: International Classification for Nursing Practice (ICNP). 2018. URL: http://www.who.int/classifications/icd/adaptations/icnp/en/ [accessed 2019-09-03]
24. Hendrich A. How to try this: predicting patient falls. Using the Hendrich II Fall Risk Model in clinical practice. Am J Nurs 2007;107(11):50-58; quiz 58. [doi: 10.1097/01.NAJ.0000298062.27349.8e] [Medline: 18075342]
25. Oliver D, Britton M, Seed P, Martin FC, Hopper AH. Development and evaluation of evidence based risk assessment tool (STRATIFY) to predict which elderly inpatients will fall: case-control and cohort studies. BMJ 1997;315(7115):1049-1053 [FREE Full text] [doi: 10.1136/bmj.315.7115.1049] [Medline: 9366729]

26. Irvine D, Sidani S, Hall LM. Linking outcomes to nurses' roles in health care. Nurs Econ 1998;16(2):58-64, 87. [Medline: 9592519]
27. Des Jarlais DC, Lyles C, Crepaz N, TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health 2004;94(3):361-366. [doi: 10.2105/ajph.94.3.361] [Medline: 14998794]
28. National Database of Nurisng Quality Indicators (R). Guidelines For Data Collection and Submission On Patient Falls Indicator. South Bend, IN: Press Ganey; 2020.
29. American Nurses Association. Guideline for Data Collection on the American Nurses Association's National Quality Forum Endorsed Measures: Nursing Care Hours per Patient Day; Skill-Mix; Falls; Falls with Injury. Kansas City: KU School of Nursing; 2012.
30. Boushon B, Nielsen G, Quigley P, Rutherford P, Taylor J, Shannon D. How-to Guide: Reducing Patient Injuries from Falls. Cambridge, MA: Institute for Healthcare Improvement; 2012. URL: http://www.IHI.org/ [accessed 2018-05-01]
31. Cho I, Boo E, Lee S, Dykes P. Automatic population of eMeasurements from EHR systems for inpatient falls. J Am Med Inform Assoc 2018;25(6):730-738 [FREE Full text] [doi: 10.1093/jamia/ocy018] [Medline: 29659868]
32. Hill A, Hoffmann T, Hill K, Oliver D, Beer C, McPhail S, et al. Measuring falls events in acute hospitals-a comparison of three reporting methods to identify missing data in the hospital reporting system. J Am Geriatr Soc 2010;58(7):1347-1352. [doi: 10.1111/j.1532-5415.2010.02856.x] [Medline: 20487077]
33. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol 2017;46(1):348-355 [FREE Full text] [doi: 10.1093/ije/dyw098] [Medline: 27283160]
34. Dykes PC, Carroll DL, Hurley A, Lipsitz S, Benoit A, Chang F, et al. Fall prevention in acute care hospitals: a randomized trial. JAMA 2010;304(17):1912-1918 [FREE Full text] [doi: 10.1001/jama.2010.1567] [Medline: 21045097]
35. Dykes PC, Duckworth M, Cunningham S, Dubois S, Driscoll M, Feliciano Z, et al. Pilot testing fall tips (tailoring interventions for patient safety): a patient-centered fall prevention toolkit. Jt Comm J Qual Patient Saf 2017;43(8):403-413. [doi: 10.1016/j.jcjq.2017.05.002] [Medline: 28738986]
36. Bouldin E, Andresen E, Dunton N, Simon M, Waters T, Liu M. Falls among adult patients hospitalized in the United States: prevalence and trends. J Patient Saf 2013;9(1):13. [doi: 10.1097/pts.0b013e3182699b64] [Medline: 23143749]
37. Cho I, Jin I. Responses of staff nurses to an EMR-based clinical decision support service for predicting inpatient fall risk. Stud Health Technol Inform 2019;264:1650-1651. [doi: 10.3233/SHTI190579] [Medline: 31438275]
38. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
39. Jeffery A, Kennedy B, Dietrich M, Mion L, Novak L. A qualitative exploration of nurses’ information-gathering behaviors prior to decision support tool design. Appl Clin Inform 2017;08(03):763-778. [doi: 10.4338/aci-2017-02-ra-0033] [Medline: 32847152]
40. Oliver D, Healey F, Haines TP. Preventing falls and fall-related injuries in hospitals. Clin Geriatr Med 2010;26(4):645-692. [doi: 10.1016/j.cger.2010.06.005] [Medline: 20934615]
41. Mant J. Process versus outcome indicators in the assessment of quality of health care. Int J Qual Health Care 2001;13(6):475-480. [doi: 10.1093/intqhc/13.6.475] [Medline: 11769750]
42. Cumming RG, Sherrington C, Lord SR, Simpson JM, Vogler C, Cameron ID, et al. Cluster randomised trial of a targeted multifactorial intervention to prevent falls among older people in hospital. BMJ 2008;336(7647):758-760. [doi: 10.1136/bmj.39499.546030.be] [Medline: 18332052]
43. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011;46(3):399-424 [FREE Full text] [doi: 10.1080/00273171.2011.568786] [Medline: 21818162]
44. Linden A. Combining synthetic controls and interrupted time series analysis to improve causal inference in program evaluation. J Eval Clin Pract 2018;24(2):447-453. [doi: 10.1111/jep.12882] [Medline: 29356225]
45. DiBardino D, Cohen ER, Didwania A. Meta-analysis: multidisciplinary fall prevention strategies in the acute care inpatient population. J Hosp Med 2012;7(6):497-503. [doi: 10.1002/jhm.1917] [Medline: 22371369]

Abbreviations

EHR: electronic health record
IN@SIGHT: Intelligent Nursing @ Safety Improvement Guide of Health information Technology
ITS: interrupted time series
NASSS: non-adoption, abandonment, scale-up, spread, and sustainability
NDNQI: National Database of Nursing Quality Indicator
QI: quality improvement
STRATIFY: St. Thomas’ Risk Assessment Tool in Falling Elderly Inpatients
TREND: Transparent Reporting of Evaluations with Nonrandomized Designs

Edited by R Kukafka; submitted 12.12.20; peer-reviewed by M Stein, J op den Buijs, H Jin, C Fincham; comments to author 16.03.21; revised version received 15.04.21; accepted 08.10.21; published 25.11.21.

Please cite as:
Cho I, Jin IS, Park H, Dykes PC
Clinical Impact of an Analytic Tool for Predicting the Fall Risk in Inpatients: Controlled Interrupted Time Series
JMIR Med Inform 2021;9(11):e26456
URL: https://medinform.jmir.org/2021/11/e26456
doi:10.2196/26456
PMID:34626168

Original Paper

Health Information Needs of Young Chinese People Based on an Online Health Community: Topic and Statistical Analysis

Jie Wang[1,2], PhD; Xin Wang[3], PhD; Lei Wang[1], MS; Yan Peng[1], PhD

[1]School of Management, Capital Normal University, Beijing, China

[2]State key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

[3]Department of Electrical and Computer Engineering, The State University of New York at Stony Brook, Stony Brook, NY, United States

Corresponding Author:
Jie Wang, PhD
School of Management
Capital Normal University
No. 56 Xisanhuan North Rd, Haidian District
Beijing, 100089
China
Phone: 86 01068901018
Email: wangjie@cnu.edu.cn

Abstract

Background: The internet has been widely accessible and well accepted by young people; however, there is a limited understanding of the internet usage patterns and characteristics on issues related to health problems. The contents posted on online health communities (OHCs) are valuable resources to learn about youth's health information needs.

Objective: In this study, we concurrently exploited statistical analysis and topic analysis of online health information needs to explore the distribution, impact factors, and topics of interest relevant to Chinese young people.

Methods: We collected 60,478 health-related data sets posted by young people from a well-known Chinese OHC named xywy.com. Descriptive statistical analysis and correlation analysis were applied to find the distribution and influence factors of the information needs of Chinese young people. Furthermore, a general 4-step topic mining strategy was presented for sparse short texts, which included sentence vectorization, dimension reduction, clustering, and keyword generation.

Results: In the Chinese OHC, Chinese young people had a high demand for information in the areas of gynecology and obstetrics, internal medicine, dermatology, plastic surgery, and surgery, and they focused on topics such as treatment, symptoms, causes, pathology, and diet. Females accounted for 69.67% (42,136/60,478) and young adults accounted for 87.44% (52,882/60,478) of all data. Gender, age, and disease type all had a significant effect on young people's information needs and topic preferences (P<.001).

Conclusions: We conducted comprehensive analyses to discover the online health information needs of Chinese young people. The research findings are of great practical value to carry out health education and health knowledge dissemination inside and outside of schools according to the interests of youth, enable the innovation of information services in OHCs, and improve the health literacy of young people.

(JMIR Med Inform 2021;9(11):e30356) doi:10.2196/30356

KEYWORDS

information needs; young people; online health community; topic analysis

Introduction

Background

To live a healthy life, people may pay greater attention to the information related to physical and mental health, disease, nutrition, and health protection. Heath information can guide health and clinical behaviors [1,2], and the availability of the internet makes it convenient to retrieve health-related information [3]. According to the search behavior report on popular science needs of Chinese citizens in 2018 [4], health and medical science rank the first in the search index among the popular science topics concerned, with a search proportion of 66.8%. The large number of users and the convenience of information access make online health communities (OHCs) one of the most important sources for searching and exchanging health-related information, experiences, advice, support, and opinions [5]. The large-scale sharing of health information also makes OHCs a valuable and abundant source of data for addressing public health questions [6]. Therefore, user-generated content in OHCs is one of the most direct and convenient ways of learning the topics of interest for users [7].

Young people are the future and the hope of all nations, thus promoting the health of young people is an important part of the strategy of a healthy China. Youth aged between 10 to 19 years face a range of health risks and this age is an important developmental period when health behaviors, values, and attitudes are established; these are often carried into adulthood [8]. According to the definition of young people from World Health Organization (WHO), we defined those between 10 and 25 years of age as young people, and within this group, those between 18 and 25 years old as young adults and those between 10 and 17 years old as minors, to provide a deeper understanding of the characteristics of health-related internet usage for this important age group [9].

Although the internet is widely accessible and well accepted by young people, there is a limited understanding of internet usage patterns and characteristics on issues related to health problems [10]. Despite the importance, little progress has been made to meet the need of providing online health information. Research on young people's online health mostly rely on data collected from questionnaire surveys or interviews, with the number of data samples being fewer than 1000 [11-13]. These can hardly be expected to represent the actual information needs of young people. The related data analyses have been mostly based on basic statistics and correlation of questionnaire data and interview data [13-15], and few studies have been performed with the aim of understanding the user-generated content through natural language processing (NLP) techniques to discover the topics and interests of youth.

The analysis of content of online health information, however, is very hard. The user-generated question and answer text data in OHC is often short in length and sparse in content, and the sparsity in short-text documents poses great challenges for topic analysis. Classic topic models such as latent Dirichlet allocation [16] and probabilistic latent semantic analysis [17] fail to work effectively due to the lack of word co-occurrence patterns in each short document [18,19]. Another feasible way to realize topic analysis for short text is based on word embedding models, such as Word2Vec [20]. However, such models usually use static coding methods and only consider the local information of the text. Without the overall information, this method cannot distinguish feature words by context semantics [21]. In addition, because of the sparsity, the feature vector cannot represent the semantics of short text well.

Related Work

In this section, we summarize the related work that investigated the online health information need of young people, including the work on data collection, data analysis methods, and the discovered topics.

The growth of the internet has made health information more accessible than ever before [22]. For young people, the daily internet access rate is generally high, and the internet has become an important resource to support their self-care and health-related activities and services [10].

Many studies have been made to understand the online health information needs of young people. The data collection approaches used include questionnaire survey, interview, and web crawler collection [23-25]. The corresponding data analysis methods are also different. For the questionnaires and survey data, descriptive statistical analysis, correlation analysis, and multiple logistic regression analysis are generally applied [11,12,14]. For interview data, many studies use content analysis and statistical analysis [13,15]. Recently, with the increase of user-generated content from OHCs, social media, and health service websites, some researchers have begun to collect data through web crawler and to develop text mining techniques, such as topic analysis and sentiment analysis, to discover user health information needs [26-28]. For example, text mining technology was used to analyze the pregnancy data of MedHelp in OHCs, and the adopted and unused answers were classified with a support vector machine–radial basis function kernel classification algorithm [26]. Based on the extracted information of 1000 consultation records from one OHC, the features of the health information needs of patients with hypertension were explored by content analysis and clustering analysis [28].

A variety of studies have been conducted to find the topics of interest of young people from online health information [10]. The results indicate that most online health information is closely related to the self-development of young people. The topics include daily health-related issues [29-31], physical growth [13], mental health [32,33], sexual and reproductive health [34-36], and physiological diseases [34,37]. Daily health-related issues, such as exercise and nutrition, beauty and skin care, fitness and diet, flu, and infection draw significant attention from young people [29]. They also use internet information on symptoms and treatment options for physiological diseases, such as arthritis or diabetes, and may turn to alternative sources according to the topic [34,37]. Young people who experience mental health issues often seek help and information related to their psychosocial health and advice from peers or doctors online [24,34]. For sexual health issues, both males and females are likely to look for information and help about such sensitive topics [24,34]. The internet has become a major resource for young people in supporting their self-care and health-related activities and services. The actual needs of young people may vary across different countries or different age groups [11,12,38,39].

Although many studies have been made on the online health information of young people, the number of samples for most is small and does not adequately reflect the general needs of youth. Moreover, previous studies have generally not been based on user-generated content nor have they used NLP technology to develop further research.

Objective

To fill the gap of current research, this paper presents a framework with a set of techniques to analyze online health information of interest to youth in China. The main contributions of this paper are the following.

We propose a topic analysis scheme to extract information from short-text messages in 4 steps: sentence vectorization, dimension reduction, clustering, and keyword generation. We used the advanced pretrained Siamese network model sentence-BERT (SBERT) to generate high-quality sentence vectors and principal component analysis (PCA) to reduce the vector dimension for more effective clustering. These techniques can be extended to apply to other topic extraction tasks based on short texts from the internet.

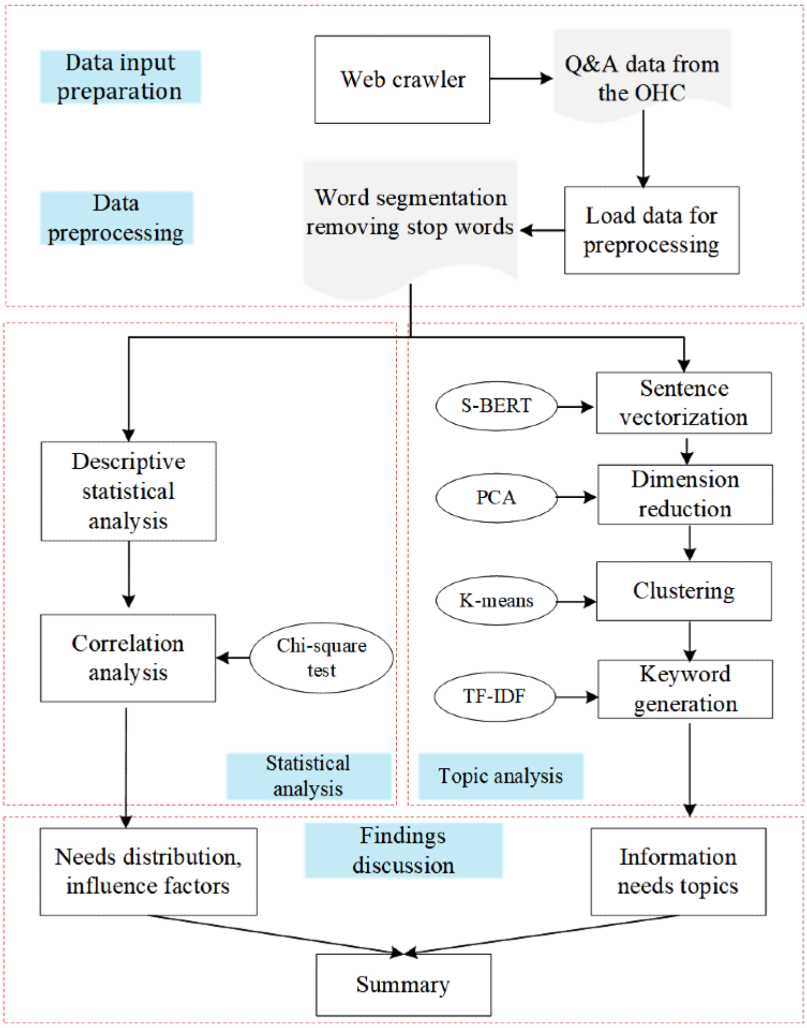
Concurrently exploiting statistical analysis and topic analysis, we also explored the distribution, impact factors, and topics of interest based on the online health information of Chinese young people posted on a popular Chinese OHC. The research findings are of great practical value to carry out health education and health knowledge dissemination inside and outside schools according to the interests of youth, enable the innovation of information services in OHCs, and improve the health literacy of young people.

Methods

Study Design

The overall research framework is displayed in Figure 1. It was divided into 4 major steps: input data preparation, data preprocessing, data analysis, and findings discussion. Among these, the data analysis consisted of 2 parts: statistical analysis and topic analysis. The statistical analysis part applied descriptive statistical analysis and correlation analysis to find the distribution and major factors related to Chinese youth’s information needs. Meanwhile, the topic analysis part used a 4-step strategy to mine the topics of specific diseases. In the 4-step topic extraction strategy, the first step used the representative pretrained language model SBERT to realize the sentence vectorization. The PCA algorithm was then used to reduce the vector dimension to improve the clustering efficiency and accuracy. After the optimal number of clusters was determined by the silhouette coefficient, a *k*-means clustering algorithm was adopted to get *k* clusters, term frequency–inverse document frequency (TF-IDF) was applied to acquire the keywords of each cluster, and the information needs topics were generated.

Figure 1. Research framework. OHC: online health community; PCA: principal component analysis; Q&A: question and answer; SBERT: sentence-BERT; TF–IDF: term frequency–inverse document frequency.



Input Data Preparation

Our data set was collected by a web crawler from a popular Chinese OHC named xywy.com, which allows users to publish health-related questions in many different disease categories. xywy.com is one of the OHCs that had explored and implemented medical and health services in China earlier. As the pioneer of OHC, its completeness and accuracy of information content are widely recognized [40].

A total of 60,478 question and answer messages posted by young users from June 1, 2019, to June 1, 2020, were collected as input data. Each message contained a set of tags, including user gender, age, question time, department affiliation, question title, question content, and doctor's responses.

Data Preprocessing

There were 2 important steps in data preprocessing: word segmentation and removing stop words. As the data source in this study was closely related to medical and health terms, the accuracy of word segmentation could be improved by combining them with a Chinese medical thesaurus. In this study, the Jieba library and Chinese medical thesaurus, CMesh [41], were used together to facilitate the word segmentation.

Removing the stop words that convey little useful meaning can reduce the dimension of the feature space [42]. Therefore, after applying the Baidu stop-word table, we removed all stop words, including articles, conjunctions, pronouns, and linking verbs.

Topic Extraction Strategies

We created a set of questions about a specific disease, Q= $\{q_1, q_2, \dots q_{|Q|}\}$. It contained |Q| questions, and q_i was the i-th question in Q. For topic extraction from Q, we needed to first cluster questions in Q into k clusters C_1, C_2, $\dots C_k$, and then generate N key words to provide the topic of cluster C_j.

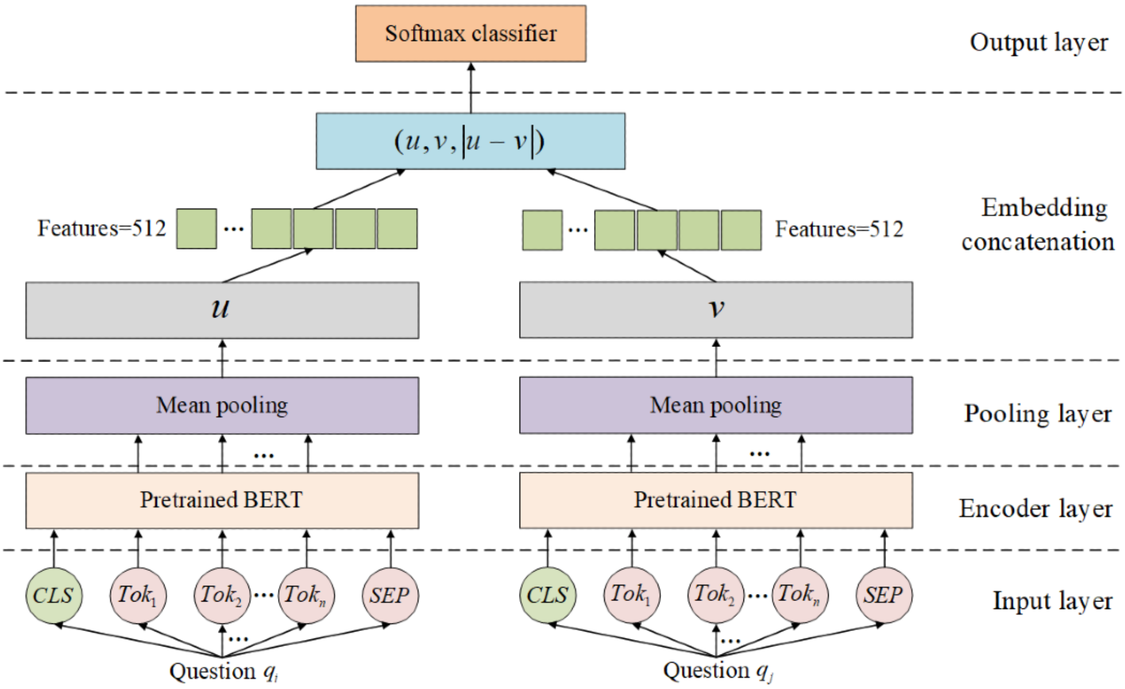
Sentence Vectorization

To extract topics from Q, the first thing was to represent the short question text data q_i in Q with an appropriate form to calculate the distance between question texts. As mentioned earlier, standard topic models and general word embedding methods were not suitable for this task. Therefore, we applied an effective pretrained NLP model in this step.

BERT [43] is now widely used in various NLP tasks. However, the sentence representation generated by BERT is not efficient for a clustering purpose. As BERT requires 2 sentences to be entered into the model at the same time for information interaction when calculating semantic similarity, it results in a significant computational overhead, and experiments [44] have shown the results to usually be even worse than those of some word-embeddings models.

Instead, we chose the improved pretrained model SBERT [44] to generate sentence vectors for the question text in Q. As shown in Figure 2, SBERT used Siamese network structure to generate semantically meaningful sentence vector representations. In the input stage, sentences q_i and q_j were each encoded by pretrained BERT. After that, the 2 sentences were normalized through a pooling layer to obtain the fixed-length vectors u and v. After this, the (u, v, |u-v|) concatenated by u, v, and |u-v| was passed through the softmax layer to acquire the classification labels of the 2 sentence vectors, where |u-v| denoted the element-wise difference between u and v. SBERT directly used the cosine similarity to compare the similarity between 2 sentence vectors, which greatly improved the speed of inference while maintaining accuracy.

Figure 2. The procedure of sentence vectorization based on sentence-BERT. CLS: a sign placed at the beginning of a sentence for subsequent classification tasks; SEP: a sign placed between 2 sentences to distinguish them; TOK: token embedding.



Dimension Reduction for Sentence Vectors

After punctuation and invalid symbols were removed, q_i in Q had the average length of 45 Chinese characters. For each q_i in Q, SBERT generated a 512-dimension vector. A higher dimension causes more computation overhead and prevents the cluster algorithm from achieving better results on a relatively large input data set. We thus chose to use the PCA technique to reduce the vectors dimension, which is an effective method to process, compress, and extract information based on the covariance matrix of variables.

To reduce the dimension of *nm*-dimensional vector matrix $R^{n\times m}$ generated by SBERT, we first calculated the eigenvalues and eigenvectors of the correlation matrix R of $R^{n\times m}$, and then $R^{n\times m}$ was projected to the eigenvector space $R^{n\times k}$ corresponding to the first *k*-dominant eigenvalues whose cumulative contribution rate was λ. That is, the original vector was reduced from the *m*-dimension to the *k*-dimension.

Sentence Vector Clustering

For topic extraction, we first clustered sentence vectors output by SBERT. In this step, it is necessary to measure the distance (or similarity) between 2 sentence vectors and determine the number of clusters to form.

In this study, all the sentence vectors generated by SBERT had the same length, and the cosine distance [45] was used to measure the similarity between 2 sentence vectors. *k*-means clustering algorithm was then adopted to get *k* clusters, with each cluster being a topic for a specific disease.

The clustering number *k* had an important influence on the clustering results of the *k*-means algorithm, and we used the silhouette coefficient [46] to evaluate the clustering effect, which combined 2 factors, cluster intracohesion and cluster interdissimilarity.

Generation Of Keywords

Keywords needed to be generated to describe the topics of interest in different clusters. The representation method based on frequent values has often been used because it reduces the text dimension and has a better effect [47]; we thus applied the TF-IDF algorithm [48] to extract keywords from the clusters results.

For each cluster C_j, TF-IDF was used to calculate the importance of words in C_j, and key words were selected based on the importance level of words. After the high word frequency in a cluster and the low text frequency in the disease question set were combined, the top *N* words with high word importance levels were selected to generate the topic words for a certain disease, and *N* was the user setting parameter. Thus, the topics of information needed for a certain disease were generated according to the topic words of each cluster.

Results

We conducted a set of experiments over the real user-generated data set crawled online to reveal the distribution, influence factors, and topics of interest of Chinese young people.

The models and algorithms in this paper were programmed based on Python 3.6 (Python Software Foundation) under the deep learning framework PyTorch 1.5.1 and TensorFlow 1.14.

To evaluate the clustering effect for short-text based on the sentence vectors generated by BERT and SBERT, we selected 8701 samples from the whole data set that had disease labels from different departments. Experiment results showed that the clustering effect was significantly improved by SBERT, with adjusted Rand index [49], adjusted mutual information [50], and Fowlkes and Mallows index [51] evaluation metric values of 32.1%, 28.6%, and 25.1% higher than those of BERT, respectively.

The Results of Statistical Analysis

The distribution of Chinese young people's interests based on the collected data of the health information is shown in Figure 3. Based on the percentage of the question data, the needs were mainly concentrated in gynecology and obstetrics, internal medicine, dermatology, plastic surgery, and surgery.

Statistical data indicated that the ratio of female to male gender distribution was about 100:116.9 in China [52]; however, the ratio of the number of questions raised by female to male users in our collected data was about 229.72 (n=42,136) to 100 (n=18,342), which showed that the young female users were more willing to use OHC for health consultation than were the male users. The results of the chi-square test between gender and departments showed that there were significant differences in health interest areas between different genders (X^2_1=17004.9; P<.001). As shown in Figure 4, the information needs of female users were mainly concentrated on the departments of gynecology and obstetrics, internal medicine, plastic surgery, and dermatology, while those of males were mainly focused on internal medicine, dermatology, andrology, and surgery. It could also be seen that both male and female young people tended to use the OHC to get help about sex-related issues, with females being more concerned about plastic and cosmetic issues and men being more concerned about surgical issues.

In terms of age distribution, the willingness to access health information from the OHC appeared to increase as age increased. Young adults aged 18 to 25 years were the main group of young users in the OHC, accounting for 87.44% (52,882/60,478) of the total number. The results of the chi-square test between age and departments showed that there were significant differences in the health information need areas of young people at different ages (X^2_1=4437.6; P<.001). The department distribution of the needs at different ages is shown in Figure S1 in Multimedia Appendix 1. The young adults' needs were mainly concentrated on gynecology and obstetrics, internal medicine, dermatology, and plastic surgery, which was basically consistent with the overall distribution of needs. The information needs of minors were mainly in the areas of internal medicine, gynecology and obstetrics, dermatology, pediatrics, and preventive health care.

Figure 3. Distribution of Chinese young peoples' health information needs.

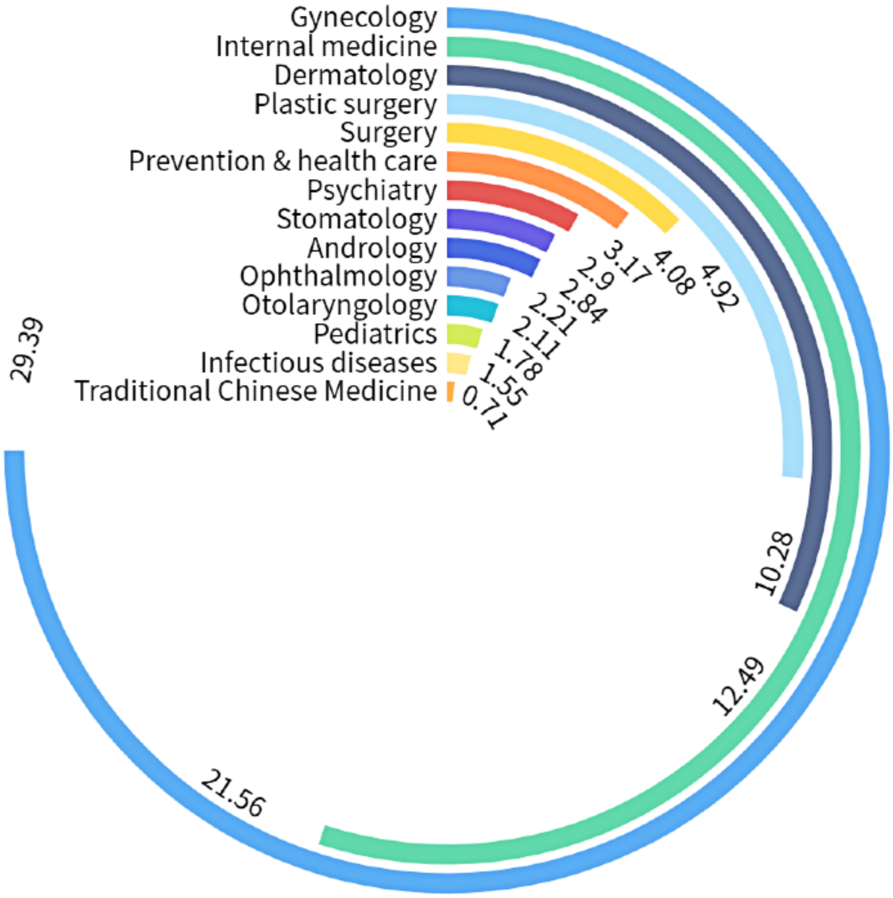
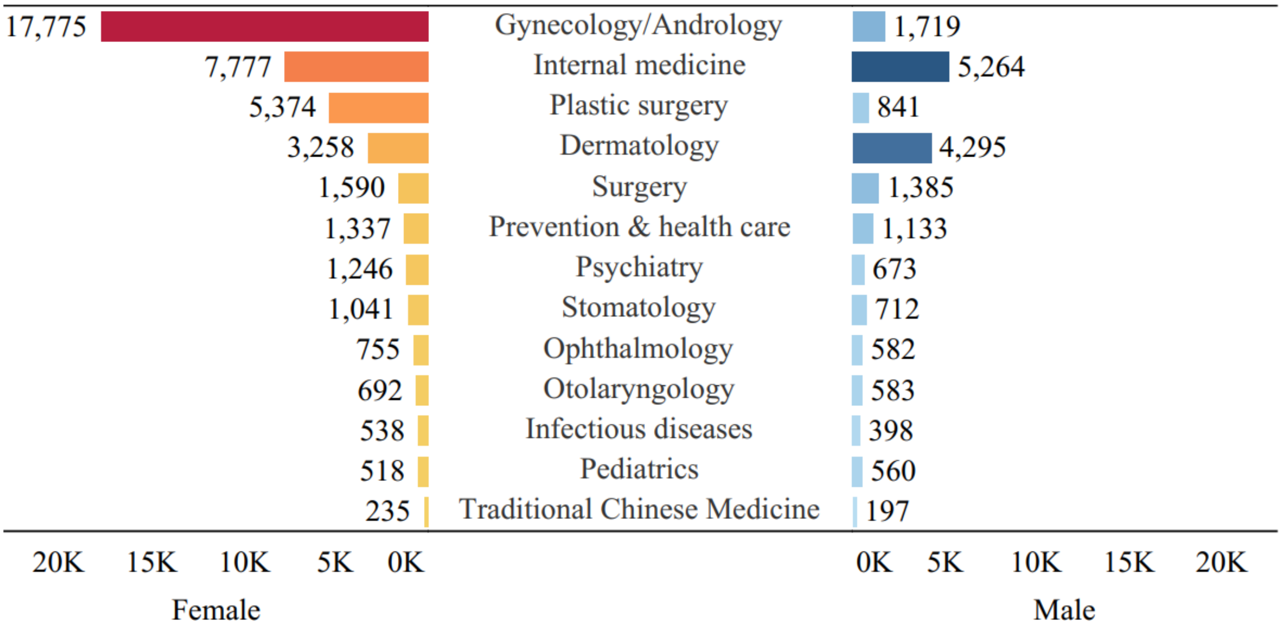


Figure 4. Health information needs distribution of Chinese young people by gender.



The Results of Topic Analysis

To further explore the topics of interests of Chinese young people related to health information, we first selected 5 representative diseases, including irregular menstruation, influenza, vitiligo, weight loss, and depression. After applying our 4-step topic extraction strategy, keywords were generated and topics were extracted for each selected disease. The top *N* key words of each cluster ranked by the word importance were selected to characterize the topics. Unless otherwise specified, *N* was set to 10 in this study.

To better understand the experiment results, a table and word cloud were used to display the topic extraction results. The final topic extraction results of menstrual irregularities and weight loss are presented in Table 1 and Table 2, the topics of vitiligo are presented in Table S1 in Multimedia Appendix 1, while the results of influenza and depression respect are shown in the form of word cloud in Figure S2 and Figure S3 in Multimedia Appendix 1, respectively. The keywords here eliminated words such as “vitiligo,” and “ influenza,” and other disease name words, such as well as “how” and “what,” along with other meaningless words.

Table 1. Topic extraction results for menstrual irregularities.

Topic	Frequency, n (%) (N=1400)	Concrete content	Keywords (top 5)
Treatment	625 (44.64)	Consult for treatment of menstrual irregularities and medications	Delay, how to treat, how to do, dysmenorrhea, causes
Pathology	84 (6.00)	Consult for types, etiology, and pathology of menstrual irregularities	What is going on, menstruation, bleeding, causes, brown
Symptom	242 (17.29)	Consult for signs, symptoms, and tests of menstrual irregularities	Delay, leucorrhea, examination, symptoms, feelings
Pregnancy	189 (13.50)	Counseling whether menstrual irregularities are associated with pregnancy	Pregnancy, have sexual intercourse, birth control pills, boyfriend, safety period
Diet	260 (18.57)	Consult for menstrual irregularities, dietary contraindications and precautions, and suitability of certain foods	What to eat, food, conditioning, diet, brown sugar

Table 2. Topic extraction results for weight loss.

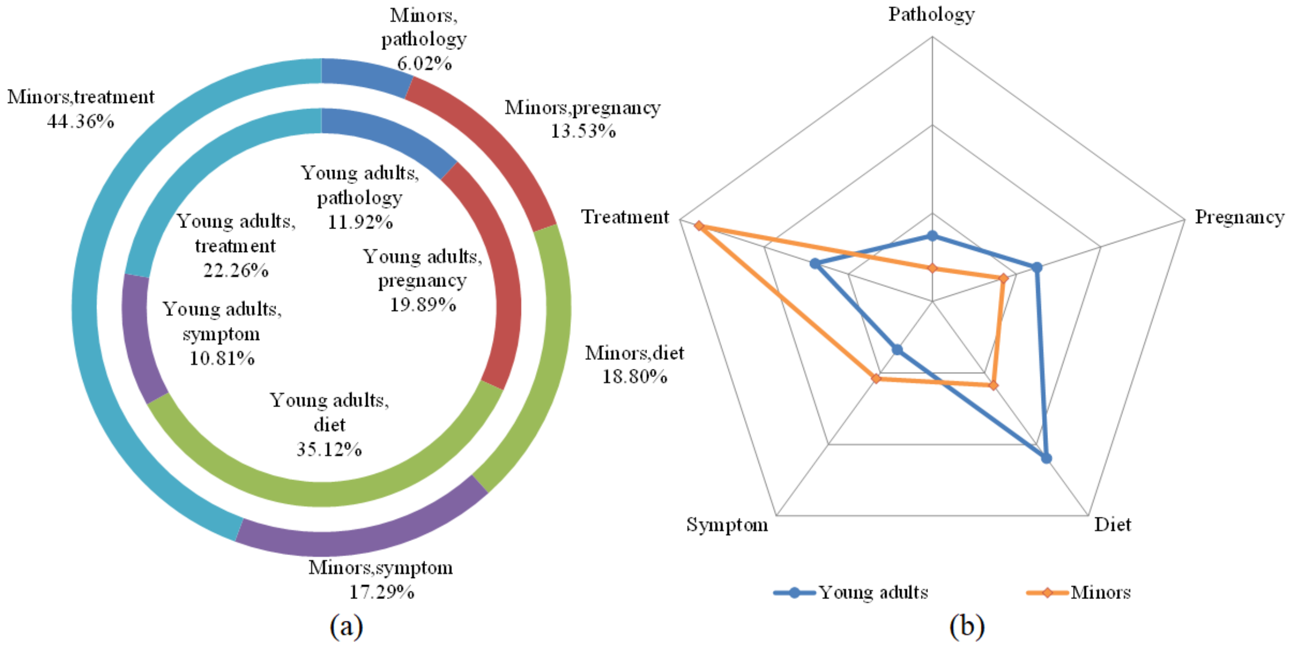
Topic	Frequency, n (%) (N=3381)	Concrete content	Keywords (top 5)
Diet	2371 (70.11)		
Coarse grain	411 (12.16)	Counseling on coarse grain cereals that help lose weight, as well as consumption effects	Potatoes, sweet potatoes, corn, red beans, oats
Fruits and vegetables	487 (14.40)	Counseling on fruits and vegetables that help you lose weight and how they work	Apples, fruits, bananas, cucumbers, bitter gourd
Beverages	353 (10.44)	Counseling on various types of beverages that help with weight loss and how well they work	Yogurt, honey water, milk, diet tea, coffee
Weight loss recipes	1120 (33.13)	Counseling on healthy recipes that help to lose weight	What to eat, how to eat, food, effect, dieting
Surgery	614 (18.16)	Counseling on various surgical weight loss methods, effects, and costs.	Diet, treatment, liposuction, thin face pin, surgery, lipolysis, effect
Pathology	396 (11.71%)	Counseling on the causes of obesity and weight loss methods	Obesity, getting fat, causing, sweets, why

Overall, the topics for all types of diseases were mainly focused on treatment, symptoms, pathology, and diet. For irregular menstruation, influenza, and vitiligo, young people were most concerned about the topic of treatment. Unlike other diseases for which users were mainly concerned about the treatment method, patients with vitiligo were also concerned about the treatment cost and location of treatment. Young people consulting on weight loss were most concerned about the role of diet in weight loss, including the information on how to choose diet recipes and the types of roughage grains, fruits and vegetables, and beverages that help with weight loss. In contrast to other physiological disorders, young people under the depression department were not concerned about the diet topic. They were more interested in symptoms than in treatment. The results of the chi-square test between the disease type and the information needs topic showed that there were significant differences in the topic of information needs between young people with physical and psychological disorders (X^2_1=2591.7; P<.001).

The gender distribution of vitiligo and influenza is shown in Figure S4 in Multimedia Appendix 1. The chi-square tests between young people's gender and information need topics in the data of influenza (X^2_1=113.7; P<.001), vitiligo (X^2_1=100.6; P<.001), weight loss (X^2_1=49.0; P<.001), and depression (X^2_1=88.7; P<.001) all indicate that there were significant differences in topics of interest in young people of different gender. In all 4 diseases, male young people asked more questions on the treatment topics and pathology topics than did females, and female young people asked more questions on the diet topics than did males.

In Figure 5, we used the distribution map and radar chart to show the information needs topic distribution of different ages for menstrual irregularities. The chi-square tests of irregular menstruation (X^2_1=44.4; P<.001), influenza (X^2_1=81.1; P<.001), vitiligo (X^2_1=64.2; P<.001), weight loss (X^2_1=157.5; P<.001), and depression (X^2_1=30.2; P<.001) indicated that there were significant differences in health information needs of young people of different ages. In all the 5 selected diseases, young adults asked a higher proportion of questions on the topic of diet than did minors, indicating that young adults were more concerned about the topic of diet than were minors.

Figure 5. (a) Distribution map of menstrual irregularities topic distribution at different age stages. (b) Radar chart of menstrual irregularities topic distribution at different age stages.



Discussion

Principal Findings

There are four principal findings in this study. First, Chinese young people's interests on online health information are mainly distributed in the following areas in descending order: gynecology and obstetrics, internal medicine, dermatology, plastic surgery, and surgery. It is worth noting that sexual and reproductive health issues are a concern of both Chinese male and female young people. The development of sexual organs and the awakening of sexual consciousness during adolescence likely lead young people to pay greater attention to health issues related to sexual organs and the reproductive system, including sexual organ development, urinary infection, and menstrual irregularities. However. because of the cultural background and relatively poor sex education level in China, young people are often shy to talk about sexual problems offline, so they hope to receive helpful information online [53]. Young people have a high level of oil secretion in their skin, and bad habits and dietary habits are common, which not only affect the health of their skin but also the aesthetics of their appearance, so health information related to skin problems and cosmetic surgery is also urgently sought out by young people.

Second, most young people in Chinese OHCs are female; this is because there are significant gender differences in the level of health awareness and attention to health information among young people in social media [54], with females having a higher level of health awareness and attention to health information than males. Moreover, gender is an important factor affecting the need for health-related information for Chinese young people. Male young people are more concerned about treatment and pathology topics than are females, and female young people are more concerned about diet topics than are males. This is the same conclusion as that found by previous studies [55], where male young people were significantly lower than females both in terms of their level of dietary health and awareness of a healthy diet.

Third, young adults aged 18 to 25 year are the main group of young users in OHCs. This is because the number of young adults using the internet and OHCs is much larger than that of minors aged 10 to 17 years. As a group that has initially left the family and entered society, young adults lack parental care and help in health issues and are more willing to seek help in OHCs. There are also significant age differences in young people's health information needs in OHCs. Compared with young adults, the interest in gynecology and obstetrics is lower while the interest in pediatrics is higher among minors, and this difference is mainly determined by the developmental stage. Furthermore, young adults are more concerned with the topic of diet than are minors. This is because primary and secondary schools in China do not currently provided adequate dietary health education [56], but as young people grow older, the channels for dietary health education expand, their knowledge of dietary health increases, and their awareness of the importance of healthy eating rises.

Finally, for Chinese young people, the information needs mainly focus on treatment, symptoms, etiology, pathology, and diet, whereas less attention is paid to the topic of prevention. Meanwhile, there are significant differences for different disease types. For physiological diseases, such as irregular menstruation, influenza, and vitiligo, young people pay most attention in OHCs to the treatment to understand the treatment methods, costs, and hospital-related information. For mental diseases, such as depression, they are most concerned about the topic of symptoms, hoping the OHC can help them to judge whether they have the condition or not. This is because young people lack knowledge about psychological health and have difficulty in self-judging mental illness, so more young people are eager to seek help from doctors by describing their symptoms in OHCs to determine whether they have a psychological illness [57].

Theoretical Implications

Based on real user-generated content, this study applied a web crawler, NLP, and statistical analysis technologies to comprehensively analyze Chinese youth's online information needs. This study attempted to reduce the deficiencies of the related literature, whose limitations included small research samples and relatively simple data analysis methods.

To deal with the challenge of mining topics from a massive collection of sparse short text from the internet, we used a general 4-step topic mining strategy. Using an advanced pretraining model, SBERT, and PCA dimension reduction, we generated high-quality clusters for extracting the topic of health information needs. From a technical point of view, this scheme provides a good method of topic analysis for short texts collected from the internet. Furthermore, with minor changes, such as removing word segmentation in the data preprocessing step, it can be extended to apply to other similar tasks using English-language data from websites.

Our study also found that there were significant differences between Chinese and other countries' youth in the distribution and topics of online health information needs, which may have important implications for other researchers by providing data support and a basis for further research on differentiation.

Practical Implications

Many practical implications could be derived from this study. First, the education of disease prevention for young people should be strengthened. The topic mining results of various diseases showed that youth pay the least attention to the topic of disease prevention, which indicates that schools, families, and internet health and service platforms including OHCs should pay more attention to the education and guidance of disease prevention for youth.

As mentioned earlier, sexual and reproductive health were one of the most concerning fields for Chinese young people. Therefore, it is necessary to improve network management to guide youth in treating and understanding sex-related information on the internet. An effective way is to establish professional and authoritative sexual health–related knowledge platforms to provide scientific information to young people at different stages of development.

Moreover, the information service mode of OHCs requires innovation. At present, most of the information service models of Chinese OHCs are centered on the aggregation and organization of health information resources, which ignores the needs of users to some extent and is challenged in providing accurate service. Therefore, databases on user's health information needs should be established based on the results from mining and analysis of their actual interests. Based on the OHCs’ service platform, information matching and precision service of the information resources and information needs databases should be realized, which will provide personalized information and health services for youth and other users.

Limitations

This study has some limitations. First, the experimental data were collected from only a single website, and thus the data source setting was substantially limited. In future studies, we plan to collect a larger data set from different OHCs to ensure the research results are more comprehensive and reliable. Second, although the presented framework showed good results in topic mining tasks for short texts from the internet, there is still much room for improvement related to the clustering tasks in specialized domains, and our future work will integrate expertise in specific domains into the model to improve its performance.

Conclusions

In this study, we conducted statistical analysis and topic analysis of online health information to explore the distribution, impact factors, and topics of interests of Chinese young people. A general topic analysis strategy using the pretraining model SBERT was proposed to extract high-quality topics based on large-scale sparse short texts from the internet. The research findings are helpful for health education departments to understand the real health-related needs of young people, carry out targeted education, and improve young people’s health literacy, and may be useful for OHCs to innovate and improve information service.

Acknowledgments

This study was funded by the Beijing Natural Science Foundation(#1202020) and the Open Foundation of State Key Laboratory of Networking and Switching Technology (#SKLNST-2020-1-19). The funding body was not involved in the study design, data collection, data analysis, or writing of the manuscript.

Authors' Contributions

All authors were involved in the design of the study. JW led the drafting of the manuscript with assistance from XW, LW, and YP. LW implemented and tested the software used to collect data and perform the analyses. XW and YP supervised the project and revised the manuscript for important intellectual content. All authors read and agreed with the analysis and the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Related work.
[DOCX File , 878 KB - medinform_v9i11e30356_app1.docx]

References

1. Yuelin L, Wenjuan C. A review of the studies on health information seeking behavior overseas. Library and Information Service 2012;19:128.
2. Shi J, Li C, Qian Y, Zhou L, Zhang B. Information needs of domestic and international hcqa users an empirical analysis. Data Analysis and Knowledge Discovery 2019;3(5):2019-2010. [doi: 10.11925/infotech.2096-3467.2018.0813]
3. Rowlands IJ, Loxton D, Dobson A, Mishra GD. Seeking health information online: association with young Australian women's physical, mental, and reproductive health. J Med Internet Res 2015 May 18;17(5):e120 [FREE Full text] [doi: 10.2196/jmir.4048] [Medline: 25986630]
4. China Association for Science and Technology. The search behavior report on popular science needs of Chinese citizens in 2018. URL: http://www.kepuchina.cn/more/201904/t20190416_1041674.shtml [accessed 2021-07-13]
5. Johnston AC, Worrell JL, Di Gangi PM, Wasko M. Online health communities: an assessment of the influence of participation on patient empowerment outcomes. Info Technology & People 2013 May 31;26(2):213-235. [doi: 10.1108/ITP-02-2013-0040]
6. Conway M, Hu M, Chapman WW. Recent advances in using natural language processing to address public health research questions using social media and consumer generated data. Yearb Med Inform 2019 Aug;28(1):208-217 [FREE Full text] [doi: 10.1055/s-0039-1677918] [Medline: 31419834]
7. Liu J, Kong J, Zhang X. Study on differences between patients with physiological and psychological diseases in online health communities: topic analysis and sentiment analysis. Int J Environ Res Public Health 2020 Feb 26;17(5):1508 [FREE Full text] [doi: 10.3390/ijerph17051508] [Medline: 32111045]
8. McKinnon KA, H Y Caldwell P, Scott KM. How adolescent patients search for and appraise online health information: A pilot study. J Paediatr Child Health 2020 Aug;56(8):1270-1276. [doi: 10.1111/jpc.14918] [Medline: 32479676]
9. Orientation programme on adolescent health for health care providers. World Health Organization. URL: https://extranet.who.int/iris/restricted/handle/10665/42868 [accessed 2021-07-13]
10. Park E, Kwon M. Health-related internet use by children and adolescents: systematic review. J Med Internet Res 2018 Apr 03;20(4):e120 [FREE Full text] [doi: 10.2196/jmir.7731] [Medline: 29615385]
11. Kim SU, Martinović I, Katavić SS. The use of mobile devices and applications for health information: A survey of Croatian students. Journal of Librarianship and Information Science 2019 Oct 22;52(3):880-894. [doi: 10.1177/0961000619880937]
12. Hassan S, Masoud O. Online health information seeking and health literacy among non-medical college students: gender differences. J Public Health (Berl.) 2020 Mar 09:1-7. [doi: 10.1007/s10389-020-01243-w]
13. Gaskin G, Anoshiravani A. Internet access and attitudes toward online personal health information among detained youth. Journal of Adolescent Health 2012 Feb;50(2):S91. [doi: 10.1016/j.jadohealth.2011.10.239]
14. Wartella E, Rideout V, Montague H, Beaudoin-Ryan L, Lauricella A. Teens, health and technology: a national survey. MaC 2016 Jun 16;4(3):13-23. [doi: 10.17645/mac.v4i3.515]
15. Patterson SP, Hilton S, Flowers P, McDaid LM. What are the barriers and challenges faced by adolescents when searching for sexual health information on the internet? Implications for policy and practice from a qualitative study. Sex Transm Infect 2019 Sep;95(6):462-467 [FREE Full text] [doi: 10.1136/sextrans-2018-053710] [Medline: 31040251]
16. Blei DM, NG AY, Jordan MI, Lafferty J. Latent Dirichlet allocation. Journal of Machine Learning Research 2003;1:993-1022. [doi: 10.1162/jmlr.2003.3.4-5.993]
17. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci 1990 Sep;41(6):391-407. [doi: 10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9]
18. Hong L, Davison B. Empirical study of topic modeling in Twitter. 2010 Jul 25 Presented at: The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; July 25-28, 2010; Washington DC, USA. [doi: 10.1145/1964858.1964870]
19. Rashid J, Shah SMA, Irtaza A. Fuzzy topic modeling approach for text mining over short text. Information Processing & Management 2019 Nov;56(6):102060. [doi: 10.1016/j.ipm.2019.102060]
20. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint 2013 [FREE Full text]
21. Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint 2014 [FREE Full text]
22. McMullan RD, Berle D, Arnáez S, Starcevic V. The relationships between health anxiety, online health information seeking, and cyberchondria: Systematic review and meta-analysis. J Affect Disord 2019 Feb 15;245:270-278. [doi: 10.1016/j.jad.2018.11.037] [Medline: 30419526]

23. Alghamdi ES, Alqarni AS, Bakarman MM, Moniem Mukhtar A, Bakarman MA. Use of internet health information among students in Jeddah, Saudi Arabia: a cross-sectional study. GJHS 2019 Apr 03;11(5):51. [doi: 10.5539/gjhs.v11n5p51]
24. Fergie G, Hunt K, Hilton S. What young people want from health-related online resources: a focus group study. J Youth Stud 2013 Aug;16(5):579-596 [FREE Full text] [doi: 10.1080/13676261.2012.744811] [Medline: 24748849]
25. Song J, Song TM, Seo D, Jin JH. ata mining Data mining of web-based documents on social networking sites that included suicide-related words among Korean adolescents. J Adolesc Health 2016 Dec;59(6):668-673. [doi: 10.1016/j.jadohealth.2016.07.025] [Medline: 27693129]
26. Prabha MS, Sarojini B. Online healthcare information adoption assessment using text mining techniques. Mobile Netw Appl 2019 Apr 19;24(4):1160-1165. [doi: 10.1007/s11036-019-01253-3]
27. Oh S, Zhang Y, Park MS. Cancer information seeking in social question and answer services: identifying health-related topics in cancer questions on Yahoo! Answers. Information Research-an International Electronic Journal 2016 Sep;21(3):718 [FREE Full text]
28. Luo A, Xin Z, Yuan Y, Wen T, Xie W, Zhong Z, et al. Multidimensional feature classification of the health information needs of patients with Hypertension in an online health community through analysis of 1000 patient question records: observational study. J Med Internet Res 2020 May 29;22(5):e17349 [FREE Full text] [doi: 10.2196/17349] [Medline: 32469318]
29. Henderson E, Keogh E, Rosser B, Eccleston C. Searching the internet for help with pain: adolescent search, coping, and medication behaviour. Br J Health Psychol 2013 Feb;18(1):218-232. [doi: 10.1111/bjhp.12005] [Medline: 23126577]
30. Manganello JA, Sojka CJ. An exploratory study of health literacy and African American adolescents. Comprehensive Child and Adolescent Nursing 2016 Jul 20;39(3):221-239. [doi: 10.1080/24694193.2016.1196264]
31. Wetterlin FM, Mar MY, Neilson EK, Werker GR, Krausz M. eMental health experiences and expectations: a survey of youths' Web-based resource preferences in Canada. J Med Internet Res 2014 Dec 17;16(12):e293 [FREE Full text] [doi: 10.2196/jmir.3526] [Medline: 25519847]
32. Lal S, Nguyen V, Theriault J. Seeking mental health information and support online: experiences and perspectives of young people receiving treatment for first-episode psychosis. Early Interv Psychiatry 2018 Jun 26;12(3):324-330. [doi: 10.1111/eip.12317] [Medline: 26810026]
33. Pretorius C, Chambers D, Coyle D. Young people's online help-seeking and mental health difficulties: systematic narrative review. J Med Internet Res 2019 Nov 19;21(11):e13873 [FREE Full text] [doi: 10.2196/13873] [Medline: 31742562]
34. Buhi ER, Daley EM, Fuhrmann HJ, Smith SA. An observational study of how young people search for online sexual health information. J Am Coll Health 2009 Sep;58(2):101-111. [doi: 10.1080/07448480903221236] [Medline: 19892646]
35. Magee JC, Bigelow L, Dehaan S, Mustanski BS. Sexual health information seeking online: a mixed-methods study among lesbian, gay, bisexual, and transgender young people. Health Educ Behav 2012 Jun 13;39(3):276-289. [doi: 10.1177/1090198111401384] [Medline: 21490310]
36. Martin S. Young People'S Sexual Health Literacy: Seeking, Understanding, and Evaluating Online Sexual Health Information. University of Glasgow. 2017. URL: https://theses.gla.ac.uk/8528/ [accessed 2021-07-13]
37. Fergie G, Hilton S, Hunt K. Young adults' experiences of seeking online information about diabetes and mental health in the age of social media. Health Expect 2016 Dec 08;19(6):1324-1335 [FREE Full text] [doi: 10.1111/hex.12430] [Medline: 26647109]
38. Duduciuc AC. Online health information seeking during adolescence: a quantitative study regarding Romanian teenagers. SCECO. Economic Edition 2015 Dec 29(22):89-95. [doi: 10.29358/sceco.v0i22.329]
39. Montagni I, Cariou T, Feuillet T, Langlois E, Tzourio C. Exploring digital health use and opinions of university students: field survey study. JMIR Mhealth Uhealth 2018 Mar 15;6(3):e65 [FREE Full text] [doi: 10.2196/mhealth.9131] [Medline: 29549071]
40. Li J, Tang J, Yen DC, Liu X. Disease risk and its moderating effect on the e-consultation market offline and online signals. ITP 2019 Aug 05;32(4):1065-1084. [doi: 10.1108/itp-03-2018-0127]
41. Chinese medical subject headings. Institute of Medical Information, Chinese Academy of Medical Sciences. URL: http://cmesh.imicams.ac.cn/index.action?action=index&noMsg=1 [accessed 2021-07-13]
42. Abdi A, Shamsuddin SM, Hasan S, Piran J. Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. Information Processing & Management 2019;56(4):1245-1259. [doi: 10.1016/j.ipm.2019.02.018]
43. Devlin J, Chang M, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint 2018 [FREE Full text]
44. Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 2019 Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; November 03- 07，2019; Hong Kong, China p. 1908.10084. [doi: 10.18653/v1/D19-1410]
45. Sun J, Jie L, Lianyu Z. Clustering algorithms research. Journal of Software 2008 Jun 30;19(1):48-61. [doi: 10.3724/sp.j.1001.2008.00048]
46. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 1987 Nov;20:53-65. [doi: 10.1016/0377-0427(87)90125-7]

47. Zhang W, Yoshida T, Tang X, Wang Q. Text clustering using frequent itemsets. Knowledge-Based Systems 2010 Jul;23(5):379-388. [doi: 10.1016/j.knosys.2010.01.011]
48. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management 1988 Jan;24(5):513-523. [doi: 10.1016/0306-4573(88)90021-0]
49. Vinh N, Epps J, Bailey J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. The Journal of Machine Learning Research 2010;11:2837-2854 [FREE Full text]
50. Rojas TJ, Santos PM, Mora M. New version of davies-bouldin index for clustering validation based on cylindrical distance. 2013 Presented at: 32nd International Conference of the Chilean Computer Science Society (SCCC); Nov 11-15, 2013; Temuco, Chile p. 49-53. [doi: 10.1109/sccc.2013.29]
51. Nguyen X, Epps J, Bailey J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? 2009 Presented at: ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning; June 14-18, 2009; Montreal, QC, Canada p. 1073-1080. [doi: 10.1145/1553374.1553511]
52. The 45th China Statistical Report on Internet Development. China Internet Network Information Center. URL: http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/202004/P020200428596599037028.pdf [accessed 2021-07-13]
53. Wang X. , Chang, C. The possibility of online education applying in sex education among adolescents. Chinese Journal of Health Education 2019;35(08):739-743. [doi: 10.16168/j.cnki.issn.1002-9982.2019.08.014]
54. Gao J. Research on the Influence of Social Media Health Information Attention on College Students' Health Behavior and Cognition. CNKI. 2017. URL: http://cdmd.cnki.com.cn/Article/CDMD-10475-1017231135.htm
55. Li F. An investigation report on the dietary health of College Students. Journal of Hebei Institute of Socialism 2009;02:93-96. [doi: 10.3969/j.issn.1009-6981.2009.02.034]
56. Huan C. Absence and make up of nutrition policy in youth health promotion. Journal of Hebei Institute of Socialism 2018;24:5-10. [doi: 10.3969/j.issn.1006-9577.2018.03.001]
57. Guo S, Hao Z, Zhao Y. Correlative research among adolescents' health literacy, psychological behavioral problems, and help-seeking attitude and willingness. Chinese Nursing Research 2017;31(16):1951-1955. [doi: 10.3969/j.issn.1009-6493.2017.16.010]

Abbreviations

OHC: online health community
NLP: natural language processing
PCA: principal component analysis
TF-IDF: term frequency–inverse document frequency
SBERT: sentence-BERT
WHO: World Health Organization

Edited by C Lovis; submitted 12.05.21; peer-reviewed by S Xu, S Wei; comments to author 06.07.21; revised version received 30.08.21; accepted 25.09.21; published 08.11.21.

Please cite as:
Wang J, Wang X, Wang L, Peng Y
Health Information Needs of Young Chinese People Based on an Online Health Community: Topic and Statistical Analysis
JMIR Med Inform 2021;9(11):e30356
URL: https://medinform.jmir.org/2021/11/e30356
doi:10.2196/30356
PMID:34747707

Original Paper

Active Annotation in Evaluating the Credibility of Web-Based Medical Information: Guidelines for Creating Training Data Sets for Machine Learning

Aleksandra Nabożny[1], MSc; Bartłomiej Balcerzak[2], PhD; Adam Wierzbicki[2], Prof Dr Hab; Mikołaj Morzy[3], PhD, DSc; Małgorzata Chlabicz[4], MD, PhD

[1]Department of Software Engineering, Gdańsk University of Technology, Gdańsk, Poland

[2]Polish-Japanese Academy of Information Technology, Warsaw, Poland

[3]Faculty of Computing and Telecommunications, Poznan University of Technology, Poznań, Poland

[4]Department of Population Medicine and Lifestyle Diseases Prevention, Medical University of Białystok, Białystok, Poland

Corresponding Author:
Aleksandra Nabożny, MSc
Department of Software Engineering
Gdańsk University of Technology
11/12 Gabriela Narutowicza St
Gdańsk, 80-233
Poland
Phone: 48 602327778
Email: aleksandra.nabozny@pja.edu.pl

Abstract

Background: The spread of false medical information on the web is rapidly accelerating. Establishing the credibility of web-based medical information has become a pressing necessity. Machine learning offers a solution that, when properly deployed, can be an effective tool in fighting medical misinformation on the web.

Objective: The aim of this study is to present a comprehensive framework for designing and curating machine learning training data sets for web-based medical information credibility assessment. We show how to construct the annotation process. Our main objective is to support researchers from the medical and computer science communities. We offer guidelines on the preparation of data sets for machine learning models that can fight medical misinformation.

Methods: We begin by providing the annotation protocol for medical experts involved in medical sentence credibility evaluation. The protocol is based on a qualitative study of our experimental data. To address the problem of insufficient initial labels, we propose a preprocessing pipeline for the batch of sentences to be assessed. It consists of representation learning, clustering, and reranking. We call this process active annotation.

Results: We collected more than 10,000 annotations of statements related to selected medical subjects (psychiatry, cholesterol, autism, antibiotics, vaccines, steroids, birth methods, and food allergy testing) for less than US $7000 by employing 9 highly qualified annotators (certified medical professionals), and we release this data set to the general public. We developed an active annotation framework for more efficient annotation of noncredible medical statements. The application of qualitative analysis resulted in a better annotation protocol for our future efforts in data set creation.

Conclusions: The results of the qualitative analysis support our claims of the efficacy of the presented method.

(JMIR Med Inform 2021;9(11):e26065) doi:10.2196/26065

KEYWORDS

active annotation; credibility; web-based medical information; fake news

Introduction

Background

In 2020 and 2021, the world has not been fighting only a pandemic; more precisely, it has been fighting both a pandemic and an infodemic [1]. The spread of COVID-19 has been accompanied by an equally unfortunate and dangerous spread of misinformation such as fake news linking the COVID-19 epidemic to 5G technology [2]. Disinformation has influenced other disease outbreaks such as the measles outbreak in Germany that involved more than 570 reported measles cases and caused infant deaths [3]. This study suggests that there exist numerous similar examples. From anticholesterol treatment to psychiatry—potentially harmful noncredible medical content on varied topics proliferates on the web.

Web-based information related to health and medicine is a large and influential category of web content, to the extent that the term *Dr Google* has been coined. The case of health-related web content is interesting from the point of view of informatics not only because medical information is highly specialized and written using domain-specific vocabulary, but also because medical information on the web is often misinterpreted or taken out of context. Health-related fake news reports often rely on factually correct medical statements such as the antiseptic effect of silver ions, which translates into a false belief in the universal effectiveness of colloidal silver for treating any disease. Debunking health-related web content requires not only expertise but also awareness of the possible effects of misinterpreted information. The breadth of specialized medical knowledge, coupled with the impact of context on fake news debunking, increases the difficulty of the problem of medical fake news detection.

Fully automated methods are currently not mature enough to detect medical fake news with sufficient accuracy. A realistic system for detecting and debunking medical fake news needs to keep medical experts in the loop. However, such an approach is not scalable because medical experts and health professionals cannot allocate sufficient time to handle the volume of misinformation spreading on the web. Another issue is that, in general, compared with credible medical content, noncredible medical web content is sparse. Assuming a real human–assisted system for assessing the credibility of medical statements, statistically, out of 100 assessed statements, the expert will catch no more than 20 unreliable items (as shown by our data collection experiment). The purpose of our work is to create an automatic tool to maximize the number of potentially noncredible sentences to be verified in the first place. The sentences are then reordered so that the most noncredible content shows up first to be annotated by a human judge. In such a way, we can optimize medical experts' time and efficacy when annotating medical information. Even if only a portion of potentially noncredible sentences gets annotated by the expert, it will include the most suspicious content.

We propose to use a method called active annotation. It dramatically improves the use of annotators' time. Active annotation implements a highly efficient human-in-the-loop component for augmented text annotation. The main idea behind active annotation is to use an unsupervised machine learning method (grouping of sentences into clusters based on sentence similarity) to organize the training data to suggest annotation labels for human annotators. When active annotation is used, the work of human annotators (medical experts) is focused on difficult noncredible medical statements. In addition, because the annotators process clusters of semantically similar sentences, our method significantly reduces the cost of cognitively expensive context switching. However, it is the annotators who decide the final labeling of the data.

The method proposed in this paper extends currently known active annotation methods by a cluster-ranking procedure that ensures that medical experts first see the content clusters that are most likely to contain noncredible content. This approach allows us to speed up the discovery of noncredible content. In our view, the process of detecting and debunking medical misinformation will never stop, and therefore a method that optimizes the use of medical experts' is of essential importance.

To test our method, we conducted an experiment with the participation of medical experts. They were asked to evaluate the credibility of medical and health-related Web content. The result of the experiment is a large data set that contains numerous examples of medical misinformation. We conducted an explorative and qualitative analysis of this data set, searching for patterns of similarity among the different examples of medical misinformation. The result of this analysis (which included an in-depth case study of misinformation related to cholesterol therapy with statins) was the discovery of distinct narratives of medical misinformation. We believe that these narratives are general in nature and will be of great use for detecting medical misinformation in the future.

Our direct experiences with the annotation team dictate a set of rules that have been formalized as a strict protocol for medical text annotation. Most importantly, we noted that the annotators tended to use external contexts extensively when annotating data. This, in turn, led to incoherent annotation labels across the data set and a divergence between the notions of statement credibility and statement truthfulness. We share our experience and present an annotation protocol that we have used to mitigate some of the annotation problems.

The original contributions presented in our paper are as follows:

- An annotation schema, an annotation protocol, and a unique annotated data set comprising 10,000 sentences taken from web-based content on medical issues, labeled by medical experts as credible, noncredible, or neutral. The entire data set is available in a public repository [4].
- A method for ranking sentences submitted to medical experts for labeling. Our active annotation method increases the likelihood that medical experts will discover noncredible sentences and thus optimizes the use of medical experts' time.
- A qualitative analysis of the labeled data set. We discovered 4 distinct narratives (both syntactic and semantic) present in the noncredible statements. We believe that these narratives can be further used to discern noncredible statements in areas of medicine other than the areas covered by our data set.

Literature Review

Health literacy is a rising concern, especially during the COVID-19 pandemic. However, research shows that more than half of the population struggles with making proper judgments and taking decisions in everyday life concerning their health [5]. Moreover, studies from the United States, Europe, and Australia [6,7] have found that web-based health information is written above the average reading level of adults. There is clearly the need for external tools or strategies to support laypersons in assessing the credibility of web-based health information. Expert fact-checking is one of the proposed strategies [8] because short-format refutational medical expert fact-checks have proven to be free from the *backfire effect* [9] (the *backfire effect* has been described in the study by Nyhan and Reifler [10]). Research shows that using expert sources to correct health misinformation in social media permanently corrects users' false beliefs.

The related work on the general news media domain [11] demonstrates that a credible source can promote false information and vice versa. Technological innovation in the fight against disinformation, as the authors argue, should go beyond discrediting noncredible sources of information and should instead promote more careful information consumption [11]. The literature has reported on successful machine learning models that classify entire articles or information sources [12,13]. Of note, these models can easily overfit (ie, obtain high classification accuracy for publications from media outlets present in the training set but fail to generalize to previously unseen media outlets). The possible performance drop in classifying fake news from previously unseen sources has been examined in the literature [12]. The study by Afsana et al [14] is, to the best of our knowledge, the most accurate classification model for assessing the quality of web-based health information. The authors declare accuracy ranging from 84% to 90% varied over 10 criteria. The model includes source-level and article-level features. The relationship of the described criteria with credibility remains an open research question.

The assessment of the veracity of individual claims contained in open-domain news articles is an emerging and fast-growing field of research. The scope of activities includes the creation of data sets containing the claims collected from fact-checking websites, such as MultiFC [15], Liar [16], and Truth of Varying Shades [17], and the existing solutions are based on a variety of approaches, from semi-automatic knowledge graph creation [18] to choosing check-worthy claims and comparing them against verified content (ClaimBuster) [19]. The open-domain solutions or solutions used in journalism [20] are not easily transferable to the medical domain.

The MedFact system [21] is a stand-alone web-based health information consumption support system. In MedFact, the user is automatically equipped with relevant trusted sources during web-based discussions.

State-of-the-art information retrieval models [22,23] forms part of the fully and semi-automatic fact-checking systems. A combination of such systems' judgments and human judgments has been successfully applied in the study by Ghenai and Mejova [24] for the specific case of capturing the spread of rumors regarding the Zika virus. Our goal is to test the combination of an unsupervised machine learning model with a human-in-the-loop approach as a robust tool to support the assessment of the credibility of web-based medical statements.

The quality assessment coding scheme for lay medical articles had been proposed in the 1990s under the Discern handbook project [25] and as Health on the Net (HON) principles. However, the guidelines have to comply with the rapidly evolving web-based reality; thus, new tools and updates are designed every few years, such as the Ensuring Quality Information for Patients (2004) [26] tool, Evidence-Based Patient Information (2010) [27], and Good Practice Guidelines for Health Information (2016) [28], to name a few. Keselman et al [29] propose different credibility assessment criteria based on 25 web-based articles regarding type 2 diabetes. These criteria (objectivity, emotional appeal, promises, and certainty) can be automatically captured by language models and lexicon-based machine learning. Work on web-based journalism has developed good practices that can also be used by medical experts in credibility evaluation. Medical practitioners who directly communicate medical information to patients can observe their reactions and subsequent actions and therefore have a special agency in credibility evaluation.

Successful application of machine learning models requires the annotation of vast corpora of medical information. However, this annotation is prohibitively expensive given the required expertise of the annotators and their limited capacity. Active annotation is a technique that facilitates large-scale data annotation by providing an auxiliary ranking of sentences that should be manually annotated by medical experts and by expediating labeling of other sentences to the underlying machine learning model. In this study, we are particularly inspired by the approach presented by Marinelli et al [30]. The authors propose initially dividing text documents into separate clusters, selecting pivot documents (k-closest documents to the center of each cluster), and generating a tentative label for the cluster. Next, a small set of text documents is selected and presented to human annotators with a proposed label and a binary annotation decision (to accept or reject the label). The authors claim that in many applications, obtaining a full annotation schema before annotation may be difficult and turning the annotation task into a binary question–answering task significantly speeds up the process [30].

Language Modeling

The term *language model* is confusing because it serves as an umbrella term for different concepts. As a general rule, a language model is a way in which textual content (tokens, words, sentences, paragraphs, and documents) is represented. Historically, text documents have been represented using 2 prevalent models: the bag-of-words model (where a document is represented simply as the set of words appearing in the document) and the one-hot encoding model (where a document is represented by a binary vector of a length equal to the size of the vocabulary and each position in the vector encodes the presence or absence of a word in the document). The most consequential limitation of these models was the inability to capture the semantic similarity between words. For instance, if a document contained the word *diabetes* and another document contained the word *insulin*, there was no straightforward way of deciding that the documents shared a common topic. This limitation has been abruptly neutralized with the advent of word embeddings. Word embeddings are dense continuous vector representations of words from a given vocabulary, which means that each word is assigned a unique vector whose elements are arbitrary numbers. Unlike one-hot encoding vectors where each vector has a length equal to the size of the vocabulary, word embedding vectors have, at most, several hundred dimensions. The vectors are trained on the text corpus to capture various semantic relationships among words. For instance, words such as *apple*, *pear*, and *orange* appear close to each other in the vector space because part of their representation encodes the notion of being a fruit. Analogically, the distance between the words *Russia* and *Moscow* is similar to the distance between the words *Great Britain* and *London* because the difference between the respective word vectors encodes the notion of a capital city.

Since the seminal work of Mikolov et al [31], word embeddings have revolutionized the field of natural language processing. After the initial success of the *word2vec* algorithm, numerous alternatives have been introduced: Global Vector embeddings trained through matrix factorization [32], embeddings trained on sentence dependency parse trees [33], embeddings in the hyperbolic space [34], subword embeddings [35], and many more. The common feature of these embeddings is the static assignment of dense vector representations to words. Each word receives the same embedding vector, irrespective of the context in which the word appears in a sentence. These static embeddings can be used to create representations for larger text units such as sentences, paragraphs, and documents. However, static embeddings are inherently unable to capture the intricacies hidden in the structure of the language and encoded in the context in which each word appears. Consider these 2 sentences: "A photo reveals significant damage to the tissue" and "Please do not throw used tissues into the toilet." The word *tissue* will receive the same vector although the context allows disambiguation of the meaning of the word.

To mitigate this limitation, modern language models depend on deep neural network architectures to calculate accurate, context-dependent word and sentence embeddings. First, context-dependent language models used either the long short-term memory network architecture [36] or gated recurrent unit networks [37] to capture contextual dependencies among the words appearing in a sentence. In other words, unlike static word embeddings, context-dependent language models calculate an embedding word vector based on the context (ie, words surrounding the embedded words). In the aforementioned example, the word *tissue* would receive 2 different vector representations: in the first sentence, the vector for the word *tissue* would be much closer to the vectors of words such as *skin* or *cell*; in the second sentence, the vector for the word *tissue* would be closer to the vector of the word *handkerchief*. These early recurrent architectures, however, suffered from performance drawbacks, and in 2018 they were replaced by transformer architecture [38]. This architecture allowed the training of much better embeddings, such as Google's Universal Sentence Encoder [39] or the (infamous) Generative Pre-trained Transformer 3 [40].

The current state-of-the-art language model, Bidirectional Encoder Representations from Transformers (BERT) [41], produces continuous word vector representations by training the neural network using 2 parallel objectives: guessing the masked word in a sentence (ie, trying to predict the word based on the context) and deciding whether 2 sentences appear one after another. Given such training objectives, the network applies similar weights to the nodes regarding input words that appear in a similar context. Sentence-BERT (sBERT) [42] is a straightforward extension of the original BERT architecture for creating sentence embeddings. This model is based on Siamese BERT networks [43] (2 identical models trained simultaneously) that are fine-tuned on the Natural Language Inference and Semantic Textual Similarity tasks. The model serves as an encoder for sentences. The encoder calculates vector representations of sentences so that semantically similar sentences have low cosine distance in the latent embedding space. This is both more efficient and produces semantically richer sentence representations than simply averaging the vectors of words that appear in each sentence.

Methods

Presentation of 3 Steps

To validate the efficacy of the active annotation approach, we need to create a data set of sentences on medical topics gathered from the Web, after which we need to obtain credibility evaluations of these sentences from medical experts. We need to propose methods for selecting sentences from the Web, annotating of these sentences by medical experts, and organizing these sentences into a processing pipeline to use the experts' time and attention most efficiently. These 3 steps we elaborate on in this section.

Data Selection

We performed annotation on a data set of 247 articles collected manually from various eHealth websites. The data set consists of more than 10,000 sentences. All documents were annotated by medical professionals sentence by sentence. The sentences constitute a stratified sample of source texts of varying credibility. We first discussed the most problematic topics of specific medical fields with the medical practitioners. Next, we manually searched for articles that presented contradicting views regarding these topics. These topics include the following:

1. Pediatrics:
   - Children's antibiotics consumption (432 sentences)
   - Children's steroids consumption (701 sentences)
   - Vaccination (1262 sentences)
   - Dietary interventions for children with autism (431 sentences)
   - Food allergy testing (1401 sentences)
2. Psychiatry:
   - Effectiveness of psychiatric medication and electroconvulsive therapy (2272 sentences)
3. Cardiology:

- Benefits of statin therapy in treating cardiovascular disease (CVD; 2029 sentences)
- Dietary interventions for heart health improvement (423 sentences)
- Benefits of consumption of antioxidants (694 sentences)

4. Gynecology:
   - Benefits of cesarean section over natural birth (359 sentences)
   - Selective serotonin reuptake inhibitor consumption during pregnancy (169 sentences)
   - Aspirin consumption during pregnancy (257 sentences)

Our collection of web-based health-related and medical articles reflects topics potentially causing controversy and misinformation among patients.

Methodology of Selecting Source Websites

The source websites were selected as follows. First, we asked each medical practitioner 2 questions:

1. "In your medical practice, what kind of false beliefs and rumors do you encounter when interacting with patients?"
2. "The truthfulness of which facts do you have to prove to your patients most often?"

The answers to these questions served as the basis for manually creating web queries. To create a data set of web medical articles addressed to laypersons, we submitted these queries to the Google search engine and then manually selected sources. The full list of these queries is listed in Multimedia Appendix 1. The manual collection was supported by the HON browser plugin (HON tag–certified webpages). As a result, 12.6% (31/247) of the extracted articles originated from HON-certified sources. The remaining 87.4% (216/247) come from domains such as the following:

- Large news media outlets (eg, *The Guardian*, *The New York Times*, and BBC)
- Q&A forums, both general and topic-specific (eg, "Quora", "Yahoo", "community.babycenter.com")
- Parenting blogs (eg, "scarymommy.com")
- Uncertified health portals (eg, "choosingwisely.org", "practo.com", and "heartuk.org.uk")
- Advertising websites for medical supplements and medical testing (eg, "everlywell.com", "yorktest.com", and "naturesbest.co.uk/antioxidants")

The full list of data sources is available in Multimedia Appendix 2.

In this study, we consider a sentence as the unit of consistent information that undergoes credibility assessment. According to Wikipedia [44], "a sentence is a set of words that in principle tells a complete thought." Thus, unless a sentence is highly complex, we can assume that the segmentation of a document into sentences is the easiest way to automatically extract single statements. To be precise, a single sentence may contain several statements. We have also observed that expert annotators tend to focus on statements rather than entire sentences when labeling data. However, we do not have a robust method of statement demarcation. In addition, most sentences contain a single main statement; thus, we decided to make the sentence the atomic unit of annotation and classification.

An additional reason for focusing on single sentences is the phenomenon of shrinking attention. Recent studies suggest that, over recent decades, collective attention spans are becoming shorter across all domains of culture, including the web [45]. It is debatable as to what the underlying cause of this phenomenon is. The most likely explanations suggest the impact of the rapid acceleration in the rate of production and consumption of information. Given finite attention resources, this inevitably leads to more cursory interaction with information. It is possible that this phenomenon also affects the consumption of health-related information, which only exacerbates the problem of the ubiquitousness of medical fake news on the web.

Expert Annotators

In all, 9 medical professionals took part in the experiment: 2 cardiologists, 1 gynecologist, 3 psychiatrists, and 3 pediatricians. All the experts had completed 6 years of medical studies, followed by a 5-year specialization program that culminated in a specialization examination. The experts were paid for a full day of work (approximately 8 hours each). Of the 9 experts, 8 (89%) had at least 10 years of clinical experience. The gynecologist was a resident physician; we accepted his participation in the experiment because of his status as a PhD candidate in medicine. Of the 3 psychiatrists, 1 (33%) held a PhD degree in medical sciences. The experts were allowed to browse certified medical information databases throughout the experiment. Each expert evaluated the credibility of content within their specialization (cardiology, gynecology, psychiatry, or pediatrics).

Annotation Protocol

Our goal is to create a rich and diverse corpus of medical sentences assessed and labeled in terms of their credibility by medical experts. To obtain reliable and comparable credibility evaluations, the experts participating in our study were supported by a detailed annotation protocol.

The medical experts evaluated the credibility of sentences with the following set of labels and the corresponding instruction:

- CRED (credible): the sentence is reliable; does not raise major objections; contains verifiable information from the medical domain
- NONCRED (not credible): the sentence contains false or unverifiable information; contains persuasion contrary to current medical recommendations; contains outdated information
- NEU (neutral): the sentence does not contain factual information (eg, it is a question); is not related to medicine

The experts were asked to base their answers mostly on their experience, knowledge, and intuition, but they were also allowed to use an external database that they would usually use in the course of their medical practice. The main direction provided to the experts was to focus on the patient's alleged perception of the information. The control question stated as follows: "If the patient asked you if he or she should trust this statement, would you say yes or no?"

In addition, we collected the following information for each sentence:

- Time needed for evaluation (in milliseconds)
- (Optional) Reason for evaluating the sentence as noncredible
- Number of surrounding sentences needed to understand the context of the sentence being evaluated

Examples of credible sentences from the *cholesterol and statins* topic include the following:

> *Lp(a), the worst cholesterol, is a number most doctors don't measure.*
>
> *Monitoring cholesterol levels is crucial because individuals with unhealthy cholesterol levels typically do not develop specific symptoms.*
>
> *Non-communicable chronic disease is now the biggest killer on the planet.*

Examples of noncredible sentences include the following:

> *For the remaining 90% of the population, the total cholesterol had no predictive value.*
>
> *It seems likely that fear of fat is unreal, based on a carry-on of the cholesterol fear.*
>
> *Most people don't need to cut down on the cholesterol that's found in these foods.*

Examples of neutral sentences include the following:

> *Seven [research items] found no link between LDL cholesterol and cardiovascular mortality.*
>
> *These perspectives won't make headlines and they won't appeal to those who want a simple and definite answers.*
>
> *This is not why I went to medical school.*

Impact of Sentence Context on Credibility Evaluation

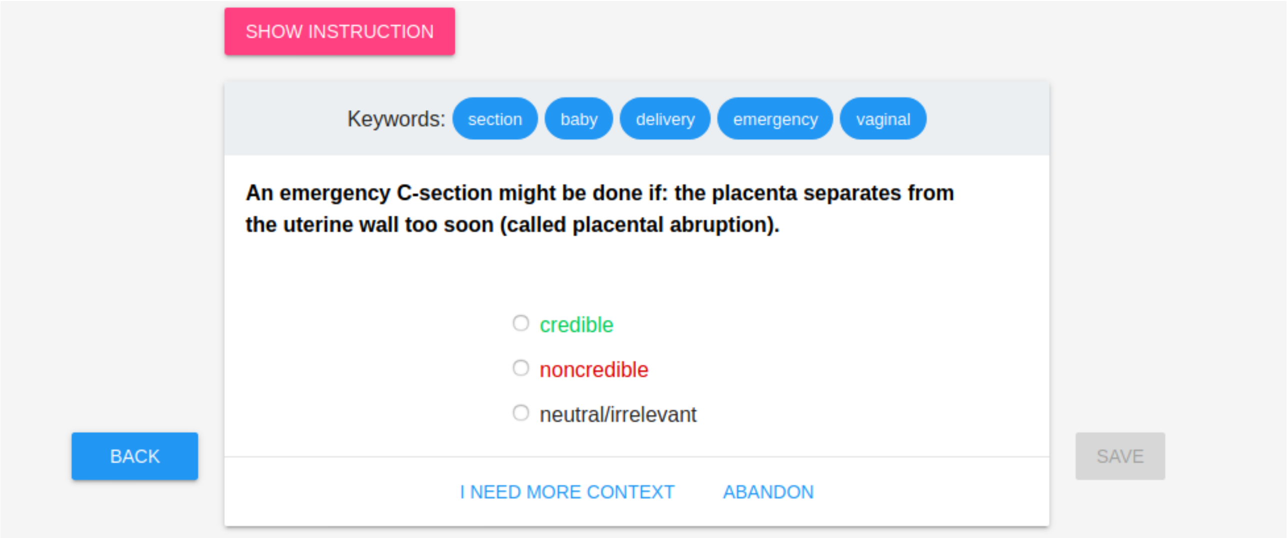
Table 1 shows how many sentences required additional *m*-surrounding sentences to provide the context for annotation. When focusing on noncredible statements, more than 71.27% (1377/1932) of the sentences were self-explanatory, 26.6% (514/1932) of the sentences required a single sentence of context, and less than 2.17% (42/1932) of the sentences required 2 or more sentences of context. Thus, we conclude that our choice of the sentence as the unit of information is justified.

Table 1. Number of surrounding sentences (*m*) needed to understand the context and evaluate the credibility of a sentence for all data, only credible subset, only noncredible subset, and only neutral subset (n=10,649).

m	All data, n (%)	Credible subset, n (%)	Noncredible subset, n (%)	Neutral subset, n (%)
0	8565 (80.43)	4955 (80.07)	1377 (71.27)	2233 (88.3)
1	1958 (18.39)	1165 (18.83)	514 (26.6)	279 (11.03)
2	107 (1)	57 (0.92)	34 (1.76)	16 (0.63)
3	12 (0.11)	5 (0.08)	6 (0.31)	1 (0.04)
<3	8 (0.07)	6 (0.1)	2 (0.05)	0 (0)

For the annotation process, we used the software developed specifically for this experiment. During the experiment, the medical expert could not see the context of the whole document while annotating a sentence. However, we provided the most relevant keywords collected from the rest of the document. Keywords were extracted using the methods described in the study by Nabożny et al [46]. A single task is shown in Figure 1.

Figure 1. Annotation interface: single sentence view.

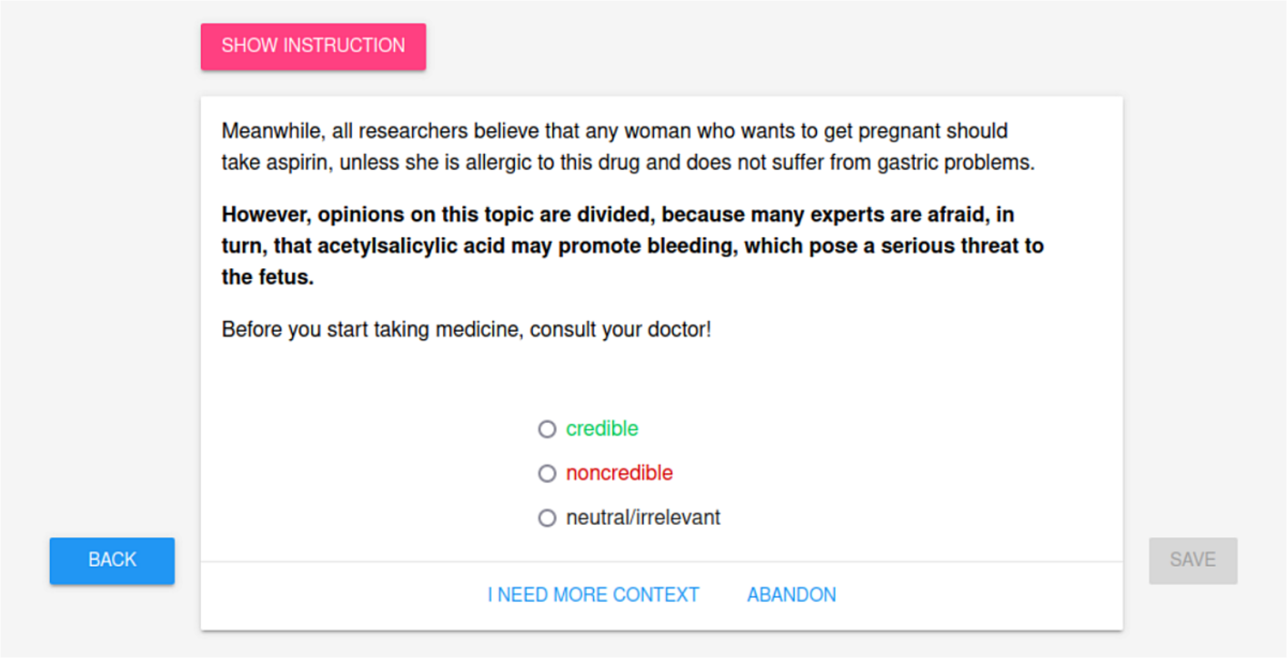


An emergency C-section might be done if: the placenta separates from the uterine wall too soon (called placental abruption).

○ credible
○ noncredible
○ neutral/irrelevant

If the medical expert decided that a sentence could not be assessed because of insufficient context (despite visible keywords), they could display the preceding and succeeding sentences in the annotation view, as shown in Figure 2. Each medical expert was asked to annotate approximately 1000 randomly chosen sentences. Whenever the medical expert labeled a sentence as noncredible, they were asked to provide the reason for their decision. To avoid the effect of intentionally skipping the NONCRED label to complete the task quicker, providing the reason was optional, and the expert could also choose an explanation from a set of tags prepared beforehand.

Figure 2. Annotation interface: sentence in context view.



Meanwhile, all researchers believe that any woman who wants to get pregnant should take aspirin, unless she is allergic to this drug and does not suffer from gastric problems.

However, opinions on this topic are divided, because many experts are afraid, in turn, that acetylsalicylic acid may promote bleeding, which pose a serious threat to the fetus.

Before you start taking medicine, consult your doctor!

○ credible

○ noncredible

○ neutral/irrelevant

The set of possible explanations prepared in advance included the following:

- The sentence contains argumentation that is weak or irrelevant, given the context of the subject being discussed.
- The sentence contains an encouragement to act inconsistently with current medical knowledge.
- The author of this sentence shows signs of the lack of substantive knowledge or is not objective.
- The sentence is an anecdote or a rumor.
- The sentence is an advertisement of an unproven drug or substance or an unproven therapy.
- The sentence cites research that was conducted on a small sample.
- The sentence contains invalid numerical data.
- The sentence contains outdated information.
- The sentence is incomprehensible or grammatically incorrect.

Most of the annotation was conducted in controlled laboratory conditions. The experts were performing annotation tasks in the presence of a supervisor who was conducting the experiment. At any time, the medical experts had access to the detailed instruction (definitions of each label) and could also ask the supervisor for assistance. The experts completed 70% of the tasks in controlled conditions, and the rest were completed with web-based assistance within a few days after the conclusion of the laboratory experiment.

Sentence Processing Pipeline Using Clustering and Reranking

Inspired by the active learning paradigm, we designed an assessment loop for medical sentence credibility. The core idea of the active annotation approach is to augment annotation efforts by 2 mechanisms:

- *Clustering*:
  Semantically similar sentences are automatically grouped into clusters. The process of clustering uses sentence-embedding representation. Each sentence is represented as a vector computed by the language model. As each sentence is a vector, mathematical measures of a distance can be used, such as the Euclidean distance or the cosine distance. We use the k-means algorithm to divide sentences into clusters. K-means is a simple iterative procedure where clustered items (in our case, vectors representing sentences) are assigned to the closest of k points representing cluster centers (also known as centroids). After assigning each item to the nearest centroid, the positions of the centroids are updated to reflect the geometric mean of assigned items. Finally, items are reassigned to the nearest centroid, and the procedure is repeated until no more reassignments are possible. The resulting clustering maximizes the similarity among the items assigned to a cluster and at the same time minimizes the similarity among the items assigned to different clusters. In other words, if 2 sentences are assigned to the same cluster, the distance between their vector representations is small, which in turn means that the sentences are

semantically similar (because semantic similarity is the criterion of embedding vector training). When human annotators are presented with sentences from a cluster, they process sentences that share a common topic. This reduces the cognitive workload of human annotators because they do not have to switch contexts between annotated sentences.

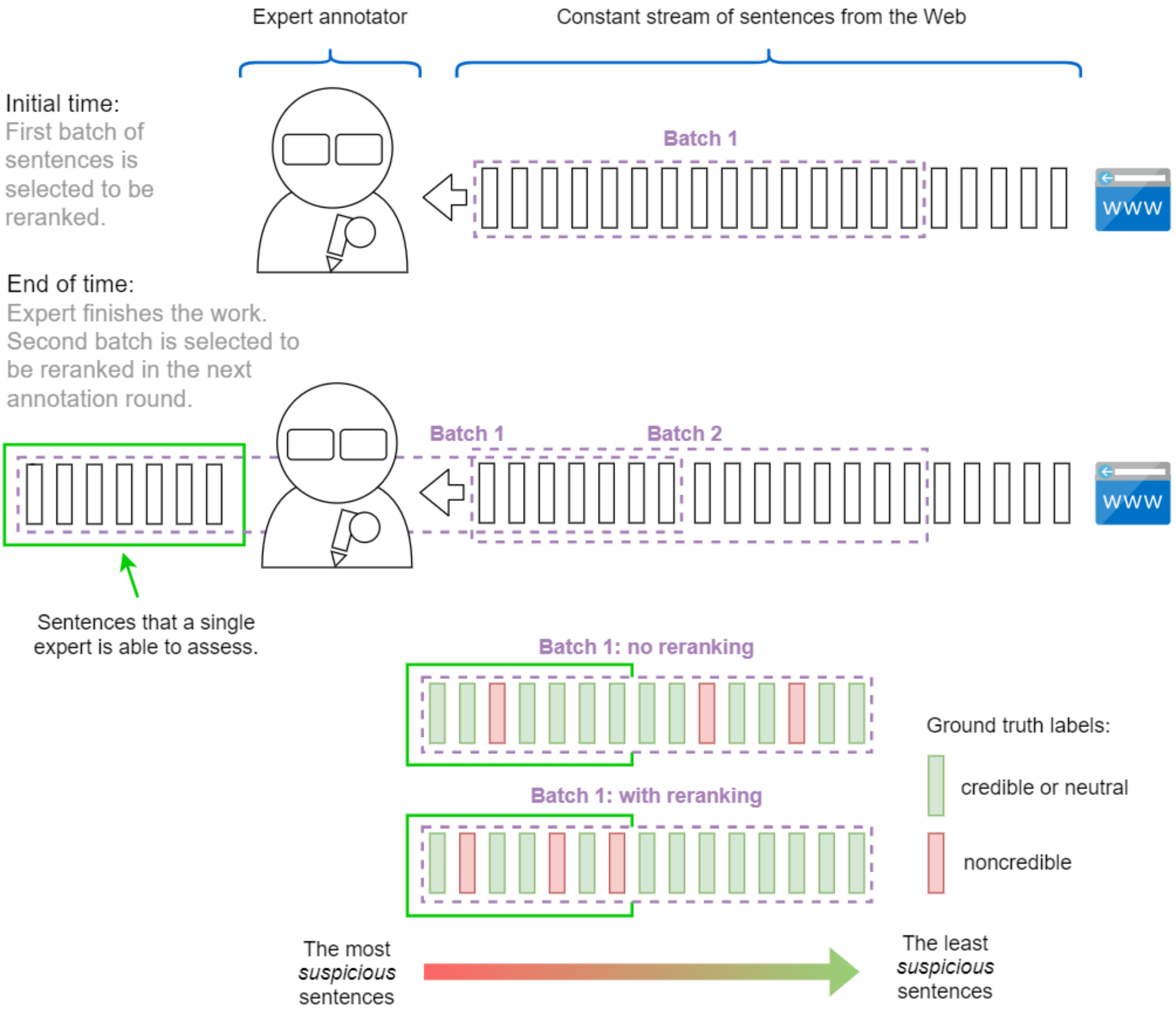
- *Reranking*:
  Noncredible statements are moved to the top of the ranking. Human annotators are required to identify noncredible statements; thus, every time human annotators are presented with a credible or neutral sentence, they may consider it to be a waste of their precious time. By combining sentence embeddings and clustering, we push sentences that are close to the already labeled noncredible sentences to the top of the ranking, prioritizing these sentences for the next round of manual annotation.

In the active annotation process, the following steps are performed in the assessment loop:

1. Sentences from the corpus are encoded by the language model to produce sentence embeddings.
2. The k-means clustering algorithm [47] is applied, and the top *k* sentences nearest to the cluster center are chosen for initial human annotation. We use the elbow method [48] to find the number of clusters (which represents the number of distinct topics in the corpus).
3. Medical experts annotate selected sentences.
4. The algorithm reranks all sentences based on the distribution of labels within clusters.
5. Medical experts annotate sentences from the top of the ranking, triggering another reranking procedure.

The general idea behind reranking is presented in Figure 3.

Figure 3. Sentence reranking: general idea.

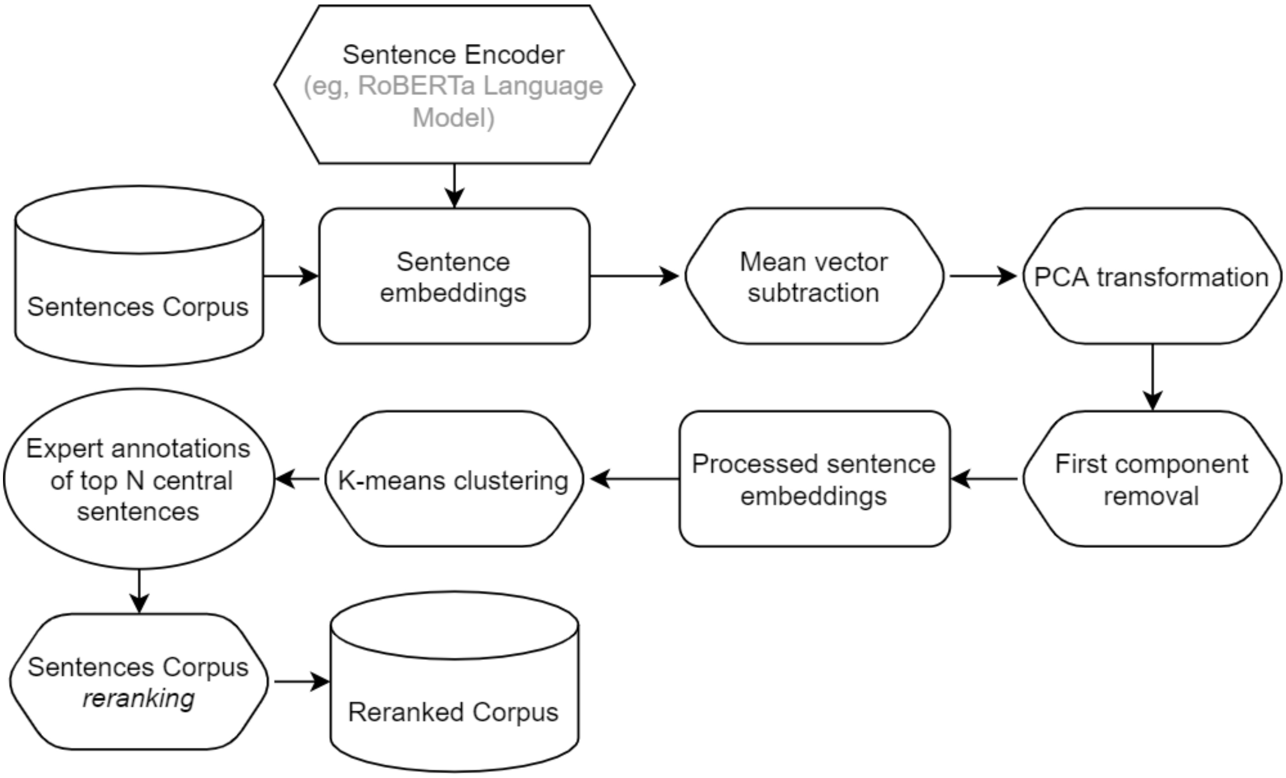

Expert annotator
Constant stream of sentences from the Web
Initial time:
First batch of sentences is selected to be reranked.
Batch 1
WWW
End of time:
Expert finishes the work. Second batch is selected to be reranked in the next annotation round.
Batch 1
Batch 2
WWW
Sentences that a single expert is able to assess.
Batch 1: no reranking
Batch 1: with reranking
Ground truth labels:
credible or neutral
noncredible
The most suspicious sentences
The least suspicious sentences

Step 4 is crucial to the method. First, we find clusters with a large proportion of labeled noncredible statements. During initial iterations of the method, only a small fraction of sentences are manually labeled, but the clustering step groups semantically similar sentences; therefore, we expect that many sentences belonging to a cluster with predominantly noncredible labels also would turn out to be noncredible. In step 5, more sentences are manually labeled, providing a better approximation of the true distribution of labels within clusters. By repeating steps 4 and 5, we annotate more and more sentences, prioritizing the annotation of noncredible sentences.

For sentence embeddings computations, we use the sBERT modification Robustly Optimized BERT Pretraining Approach where embeddings are calculated based on the same model as BERT but with slightly different training objectives and hyperparameters [49]. We also use a simple preprocessing technique where we subtract the mean and exclude the first principal component from each embedding vector [50,51] (principal component analysis transformation). The assumption behind this step is that the first principal component encodes syntactic rules of the grammar of the sentences without contributing to their semantics. The removal of the first component strips sentence vectors of grammar and leaves only the part of the vector where the meaning is encoded.

Figure 4 presents the overview of the sentence processing pipeline.

Figure 4. Processing pipeline. PCA: principal component analysis; RoBERTa: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach.



The key component of the pipeline is the clustering and reranking strategy. For reranking, we perform 2-level sorting. The first sorting is applied to clusters, and the second sorting reorders sentences within clusters. We rank clusters based on the proportions of credible, noncredible, and neutral labels in the top *m* most central sentences. Our scoring formula penalizes clusters with a significant proportion of credible sentences. At the same time, it rewards clusters with a significant proportion of noncredible sentences. This strategy enables us to push most of the noncredible sentences to the top of the ranking, thus positioning them at the top of the queue for medical expert evaluation.

Let p(c), p(n), and p(u) denote the probability that a random sentence is credible, noncredible, or neutral, respectively. This probability is computed by manually annotating *m* most central sentences in the cluster. The cluster score is defined as follows:

$$score@k = 1/e^{-(p[n]-p[c])} + 1/w^{p(u)+1} \quad \textbf{(1)}$$

The first component of the formula is the sigmoid function with the difference between p(n) and p(c) as the argument. If the difference is positive, which means that there is an advantage of noncredible proportion over credible, the sigmoid function gives results close to 1 (the bigger the difference, the closer to 1). If the difference is negative, the sigmoid value tends toward zero. The second component of the formula is the parametrizable function, which enables giving proper scoring weight to p(u). For example, given w=1.5, it orders clusters with p(n)=0.4 and p(c)=0.3 below clusters with p(n)=0.5 and p(c)=0.4. Without the second component, both clusters would receive the same score.

The intracluster ranking of sentences is performed based on the distance of sentences from the center of the cluster, with more central sentences placed at the top of the ranking. The distance is measured as the cosine distance in the latent embedding space. The final ranking of all sentences is obtained by first ordering all clusters in the decreasing order of score@k and, next, by reordering sentences within each cluster by the growing distance from the center of the cluster.

Results

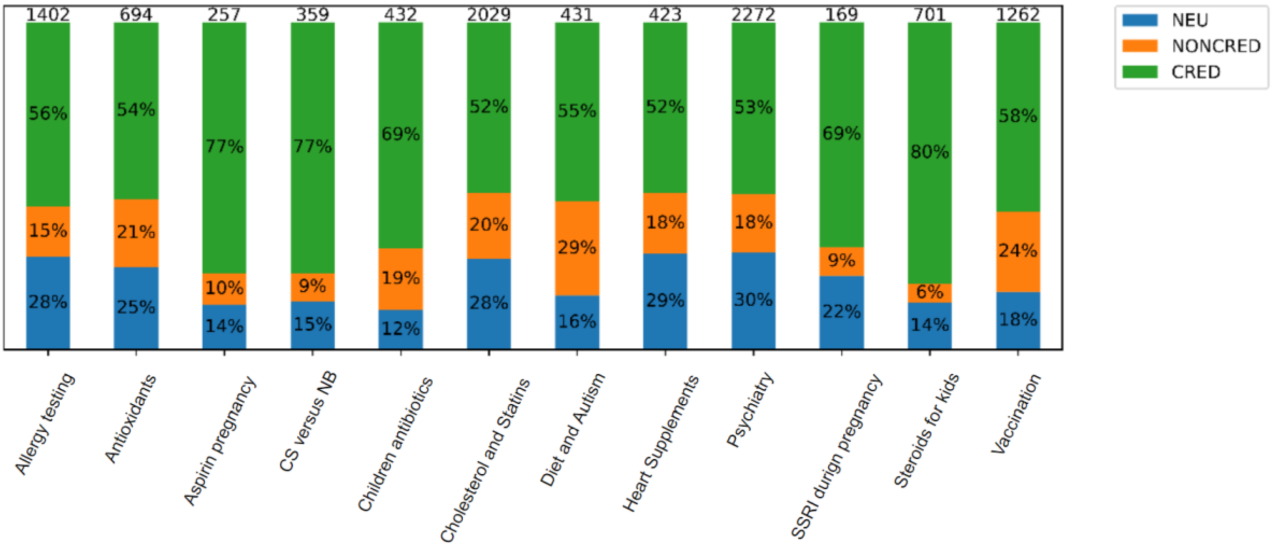
Overview

We used the method described in the previous section to create an annotated data set. We now describe the results. First, we present the data set statistics. Next, we depict the effect of our sentence pipelining method on the effectiveness of the medical experts' time allocation. Subsequently, we conduct a qualitative analysis of the credible and noncredible sentences, focusing on a single topic.

Distribution of Labels Within the Data Set

The distribution of labels (CRED, NONCRED, and NEU) for each topic is shown in Figure 5. Distribution varies for each topic but within a certain range. For example, the CRED label is always at least two times more frequent than the NONCRED label and significantly more frequent than the NEU label. The NEU label applies to no more than 30% (3195/10,649) of the sentences in all topics, which leads us to the conclusion that, regardless of the topic, more than 59.99% (6389/10,649) of the statements warrant credibility checking.

Figure 5. Distribution of credible, noncredible, and neutral sentence labels within topics. CS: cesarean section; CRED: credible; NB: natural birth; NEU: neutral; NONCRED: noncredible; SSRI: selective serotonin reuptake inhibitor.



Although the articles were explicitly picked so that they reflect potentially controversial topics, the proportion of noncredible sentences was generally small. Taking into account the alarm-raising calls of the medical experts, we can conclude that even a small contribution of noncredible content throughout the web has a substantial influence on the formation of people's views.

Justification for Using the Lift Measure

We have chosen the lift measure to evaluate the effectiveness of our method. Throughout the qualitative analysis, it became apparent that semantic similarity measures retrieved from neural language models lose important information encoded in annotations. Our objective is to optimize medical experts' time by focusing their attention on statements that are possibly noncredible. Using the lift measure, we determined the relative time savings by indicating how many more noncredible sentences a medical expert would see by reviewing a given percentage of the entire sentence corpus using our ranking. The lift measure specified for each ranking percentile is defined as follows:

lift@p = N/p × *recall@p***(2)**

where *p* is the percentile, *N* is the total number of sentences in the corpus, and *recall@p* defines, for a given percentile *p* of the ranking, how many noncredible statements have been included in the *p*th percentile of the ranking.

The key parameter of our method is *m*, the number of top sentences in a cluster for manual annotation. We tested our method on a full data set (all topics merged) for 3 *m* values, each of which is listed in Table 2. In Table 3, we present the lift results for the separate topic of *cholesterol and statins*. The baseline value for lift is 1. Thus, we can interpret the results as follows: the number by which a given value exceeds 1 tells us how many more noncredible sentences medical experts would discover at a given corpus percentile when using the reranking procedure. For example, when reviewing 20% of the full corpus, medical experts would discover 29% more noncredible sentences if the batch were to be reranked using the *m* value of 5 than without applying the procedure.

Table 2. Lift results for the full data set. *m* is the number of top sentences from each cluster to be manually reviewed.

lift@*m*	Number of clusters	Batch percentile			
		1% (approximately 100 sentences)	10% (approximately 1000 sentences)	20%	40%
lift@5	200	1.36	*1.36* [a]	*1.29*	*1.17*
lift@10	130	1.23	1.31	1.3	*1.17*
lift@15	100	*1.49*	1.27	1.22	1.16

[a] The best performing set of parameters for a given batch percentile is italicized.

Table 3. Lift results for the cholesterol and statins topic. *m* is the number of top sentences from each cluster to be manually reviewed.

lift@*m*	Number of clusters	Batch percentile			
		1% (approximately 20 sentences)	10% (approximately 200 sentences)	20%	40%
lift@5	40	1.75	1.24	1.26	1.27

The number of clusters for each experiment is chosen based on 2 criteria: the elbow method [48] and the proportion of sentences to be manually reviewed. The latter should not exceed 15% of the batch. Let us take Table 3 as an example: we delegate 5 × 40 = 200 top sentences from each cluster to be manually reviewed by the experts. These 200 sentences out of the approximately 2000 sentences in the *cholesterol and statins* topical category make up 10% of the set. It means that by gathering initial labels from only 10% of the sentences from the topical corpus, we can obtain significant (eg, 27% in the 40th percentile) savings of experts' time during text annotation sessions.

Zooming in on a Topical Cluster: Case Study of Statins

We conducted a case study in the subdomain of cholesterol and statins. We did this to gain insight into the process of credibility evaluation and the nature of noncredible medical sentences. The focus on a single topic was dictated by the size and diversity of our data set. Presenting an in-depth qualitative analysis of the entire data set would take too much space. The following is a qualitative analysis of all sentences labeled noncredible by the experts in the selected topic.

Brief Introduction to the Topic of Statin Use

Numerous epidemiological studies, Mendelian randomization studies, and randomized controlled trials have consistently demonstrated a relationship between the absolute changes in plasma low-density lipoprotein (LDL) and the risk of atheromatous CVD. The inverse association between plasma high-density lipoprotein and the risk of CVD is among the most consistent and reproducible associations in observational epidemiology. Higher plasma Lp(a) concentrations are associated with an increased risk of CVD, but it appears to be a much weaker risk factor for most people than LDL cholesterol [52]. Commonly, plasma cholesterol is used to calculate cardiovascular risk, whereas LDL is used to evaluate the achieving of target values according to the estimated cardiovascular risk.

Hypercholesterolemia (dyslipidemia with an increased levels of circulating cholesterol) is not the only factor responsible for the development of CVD, but also obesity, poor diet, lack of physical activity, smoking, and high blood pressure (hypertension). To prevent CVD, physicians recommend that patients quit smoking; eat a diet in which approximately 30% of the calories come from fat, choosing polyunsaturated fats and avoiding saturated fats and trans fats; reduce high blood pressure; increase physical activity; and maintain their weight within normal limits [53].

Hydroxymethylglutaryl-coenzyme A reductase inhibitors (statins) lower cholesterol synthesis. Statins represent the cornerstone for the treatment of hypercholesterolemia and in the prevention of CVD, although muscle-related side effects have strongly limited patients' adherence and compliance [53]. The evidence in support of muscle pain caused by statins is in some cases equivocal and not particularly strong. The reported symptoms are difficult to quantify and rarely is it possible to establish a causal link between statins and muscle pain. In randomized controlled trials, statins have been well tolerated, and muscle pain–related side effects were similar to those caused by placebo. An exchange of statins may be beneficial, although all statins have been associated with muscle pain. In some patients, a reduction of dose is worth trying, especially in primary prevention [54]. Statins have been linked also to digestive problems, mental fuzziness, and glucose metabolism, and they may rarely cause liver damage. The influence of the diabetogenic action of statins is still unclear. Despite these observations, the CVD preventive benefit of statin treatment outweighs the CVD risk associated with the development of new diabetes [55]. There is good evidence that statins given late in life to people at risk for vascular disease do not prevent cognitive decline or dementia [56]. Statins can cause transient elevation of liver enzymes, which has led to the unnecessary cessation of these substances prematurely [57]. Coenzyme Q10 (CoQ10) is widely used as a dietary supplement, and one of its roles is to act as an antioxidant. Decreased levels have been shown in diseased myocardium and in Parkinson disease. Farnesyl pyrophosphate is a critical intermediate for CoQ10 synthesis, and blockage of this mechanism may be important in statin myopathy. Supplementation with CoQ10 has been reported to be beneficial in treating hypertension, statin myopathy, heart failure, and problems associated with chemotherapy; however, this use of CoQ10 as a supplement has not been confirmed in randomized controlled clinical trials [58].

In conclusion, recent analyses and randomized controlled trials have been published confirming that the cardiovascular benefits of statin therapy in patients for whom it is recommended by current guidelines greatly outweigh the risks of side effects [59]. The Cholesterol Treatment Trialists Collaboration meta-analysis showed that for each 1 mmol/L reduction in LDL, major vascular events (myocardial infarction, coronary artery disease death, or any stroke or coronary revascularization) were reduced by 22% and total mortality was reduced by 10% over 5 years [59].

Extracting Categories From Raw Data

Our data set contains 1986 unique sentences about cholesterol and statins. Of the 1986 sentences, 1041 (52.42%) were labeled by medical experts as credible, 551 (27.74%) as neutral, and 394 (19.84%) as noncredible. We have reviewed the compliance of the assessments in the noncredible class with the annotation protocol. As a result, of the 394 noncredible annotations, 72 (18.3%) were discarded as noncompliant. The following are some examples of sentences erroneously annotated as noncredible:

> *“Why are they putting patient lives at risk?” Sentence is a question and should be labeled as neutral.*
>
> *“Researchers chose 30 studies in total to analyze.” Sentence does not contain any medical terms and should be labeled as neutral.*
>
> *“They [statins] work by blocking an enzyme called HMG-CoA reductase, which makes your body much slower at synthesizing cholesterol.” Sentence contains factually true statement and should be labeled as credible.*

Finally, of the 1986 sentences, we identified 322 (16.21%) as noncredible. We extracted 18 claim categories, which represented 61.5% (198/322) of all noncredible sentences. The process of claim category extraction involved the following steps:

1. The annotator examined all the sentences from the noncredible class one by one.
2. If a sentence matched an already existing category, it was assigned to that category; otherwise, a new category was created.
3. After processing all the sentences, categories with only 1 sentence were merged into a Miscellaneous category that contained the remaining 29.5% (95/322) of the noncredible sentences.

We also compared the compliance of the extracted claim categories with current medical guidelines and knowledge. The category counts are presented in Table 4, and these categories are listed and explained in Table 5

Table 4. The number of occurrences of a particular claim category within the *cholesterol* and *statins* subset of sentences.

Claim category	Number of occurrences	Is related claim factually incorrect?	Is category based on the content or on the form?
Miscellaneous	95	N/A[a]	Form
(stat) Side effects	43	Yes	Content
(chol) Not an indicator of CVD[b] risk	25	Yes	Content
Diet as good as drugs	22	Yes	Form
(chol) Too low is harmful	18	Yes	Content
Lifestyle changes are enough	15	Yes	Content
Big pharma	14	Yes	Content
Inflammation theory	14	Yes	Content
(stat) Cause diabetes	13	Yes	Content
(stat) Not needed	10	Yes	Content
(chol) Makes cells and protects nerves	8	No	Content
(stat) Not effective	7	Yes	Content
(stat) Prescription based solely on (chol) level	7	Yes	Content
Detailed data	7	N/A	Form
(stat) Cause cognitive impairment	6	Yes	Content
(stat) Not studied enough	6	Yes	Content
High HDL[c] neutralizes high LDL[d]	6	No	Content
Harmful CoQ10[e] loss	4	Yes	Content
(chol) Consumption not an issue	3	Yes	Content
Lifestyle versus statins	2	Yes	Content
No liver function monitoring	2	Yes	Content

[a] N/A: not applicable.

[b] CVD: cardiovascular disease.

[c] HDL: high-density lipoprotein.

[d] LDL: low-density lipoprotein.

[e] CoQ10: Coenzyme Q10.

Table 5. Claim category and explanations of claim categories extracted manually from all noncredible sentences from the *cholesterol* and *statins* topic.

Claim category	Claim explanation
(stat) Side effects	Statins' side effects outweigh the benefits
(chol) Not an indicator of CVD[a] risk	Total cholesterol is not an indicator of CVD
Diet as good as drugs	Aggregation of different dietary interventions to lower cholesterol, triglycerides, or sugars
(chol) Too low is harmful	Too low cholesterol level is harmful
Lifestyle changes are enough	People can lower cholesterol level just by developing good habits and eating a proper diet
Big pharma	People (eg, physicians and pharmaceutical company workers) make considerable profit through prescribing statins
Inflammation theory	It is inflammation that causes CVD, not excessive cholesterol level; cholesterol is an effect, not a cause
(stat) Cause diabetes	Statins increase the risk of diabetes
(stat) Not needed	Statins are given to healthy people who do not need them
(chol) Makes cells and protects nerves	Cholesterol produces hormones that make body cells and protect nerves
(stat) Not effective	Statins do not fulfill their role in reducing the risk of CVD
(stat) Prescription based solely on (chol) level	Statin prescription is based solely on total cholesterol level
Detailed data	Sentences contain detailed data, for example, "LDL[b] cholesterol level should not exceed 200 md/dL"
(stat) Cause cognitive impairment	Statin consumption causes different forms of cognitive impairment (including memory loss and slow information processing)
(stat) Not studied enough	Statins' effectiveness is not studied enough
High HDL[c] neutralizes high LDL	HDL is a so-called good cholesterol, whereas LDL is a so-called bad cholesterol; high levels of the former neutralize negative consequences of high levels of the latter
Harmful CoQ10[d] loss	Statin-related CoQ10 loss is harmful
(chol) Consumption not an issue	People should not worry about cholesterol consumption
Lifestyle versus statins	Lifestyle changes are more effective ways to prevent CVDs than statin consumption
No liver function monitoring	Monitoring of liver function tests is no longer recommended in patients on statin therapy
Miscellaneous	None of the above

[a] CVD: cardiovascular disease.
[b] LDL: low-density lipoprotein.
[c] HDL: high-density lipoprotein.
[d] CoQ10: Coenzyme Q10.

Of the 322 noncredible sentences, 198 (61.5%) fall into specific claim categories. Most of the categories have at least 6 examples that spread across different documents. We have designated categories with only 2 or 3 occurrences as separate because the entire noncredible class is relatively small and finding even a few similar sentences may indicate that the claim is being duplicated on the web.

Of the 95 sentences that did not fall into any claim category, we identified 9 (9%) that bear the hallmarks of a conspiracy theory, 7 (7%) containing reasoning based on anecdotal evidence, and 9 (9%) containing misleading statistical reporting:

- Conspiracy theory (referring to groups of interests such as prostatin vs antistatin researchers): "Ironically, prostatin researchers themselves are the ones who are guilty of cherry-picking."
- Anecdotal evidence: "What's worse, my doctor has never asked if I smoke cigarettes, exercise regularly, or eat a healthy diet."
- Misleading statistical evidence: "OK, maybe the benefits of taking a statin are small, but many smart doctors say a reduction of five-tenths or six-tenths of 1% is worthwhile."

As part of qualitative analysis, we compared 2 sets of clusters: automatically created versus manually created. We were able to select sentences that contain similar words and statements but differ in the narrative details that skewed the experts' judgments. We have identified 4 types of false and misleading narratives that occur frequently in the noncredible class. These narratives are as follows:

1. Slippery slope: The sentence is factually true, but the consequences of the presented fact are exaggerated. Example:

> *Hence, while the drug might synergise with a statin to prevent a non-fatal (or minor) heart attack, it seems to increase the risk of some other equally life-threatening pathology, resulting in death.*

> *Cholesterol also helps in the formation of your memories and is vital for neurological function.*

2. Hedging: The sentence is factually incorrect, but there is a part of it that softens the overtone of the presented statement. Example:

> *However, cholesterol content should be less of a concern than fat content.* [CRED]
>
> *Coenzyme Q10 supplements may help prevent statin side effects in some people, though more studies are needed to determine any benefits of taking it.* [CRED]
>
> *The FDA warns on statin labels that some people have developed memory loss or confusion while taking statins.* [CRED]

3. Suggested negative consequences: The sentence is mostly factually true, but given the context of the expert's experience, there is a risk that the presented information may lead the patient to act contrary to current medical guidelines. Examples:

> *For starters, statin drugs deplete your body of coenzyme Q10 (CoQ10), which is beneficial to heart health and muscle function.*
>
> *Cholesterol is a waxy, fatty steroid that your body needs for things like: cell production.*

4. Twisting words: the presence of a single word changes the overtone of the sentence. Examples:

> *Statins may slightly increase the risk for Type 2 diabetes, a condition that can lead to heart disease or stroke. [CRED]*
>
> *For example, it may be enough to eat a nutritious diet, exercise regularly, and avoid smoking tobacco products. [NONCRED]*
>
> *versus*
>
> *Eating a healthy diet and doing regular exercise can help lower the level of cholesterol in your blood.* [CRED]

Discussion

Principal Findings

The results of our experiments show that applying the active annotation paradigm for credibility assessment in the medical domain produces measurable gains in terms of the use of medical experts' time. Active annotation allows us to raise the number of noncredible statements annotated by medical experts by 30% on average, within a fixed time and monetary budget. Annotation of medical information cannot be crowdsourced because it requires the deep and broad domain knowledge of medical experts and their time is expensive. We regard the problem of prohibitively expensive annotation costs as the main obstacle to the broad use of machine learning models in the evaluation of the credibility of web-based medical resources. Our proposal is a step toward a significant lowering of these costs.

However, there is still room for improvement. Our qualitative analysis shows that most of the noncredible sentences can be classified into a limited number of categories. The subset of approximately 200 noncredible sentences from the *cholesterol and statins* subdomain can be divided into 18 categories, each representing approximately one false statement. These 18 categories fall into 61.5% (198/322) of the total number of all sentences labeled in full accordance with the annotation protocol. This indicates the importance of precise semantic clustering. More accurate clustering helps to detect noncredible sentences faster. It also enables the tagging of clusters with topic-related labels by nonexperts for later reviewing by medical experts and, as a result, the even more useful sentence ranking. In other words, it might be possible to use crowdsourcing to some extent during preprocessing and include an expert in the loop in the main annotation pipeline, further reducing the annotation costs.

Another conclusion that we drew from the qualitative analysis concerns the precision of the semantic similarity measure based on sentence embeddings. The method captures well the overall theme of the sentence but often misses the stance of the presented claim. This error is understandable because the stance in the medical domain is often expressed through subtle sentence modifications, as listed in the *Results* section. Sentence embeddings also struggle with finding a good representation of the form of the sentence—whether it is a supposition, a question, or a statement. Recognition of the form of the sentence can improve the accuracy of classification of neutral sentences that do not require medical expert annotation.

Finally, the qualitative analysis has revealed 4 distinct narratives present in noncredible sentences. Although our analysis was limited to the topic of cholesterol and statins, we feel that these narratives are more general in nature and may apply broadly to false medical information on other topics. If this hypothesis is confirmed, it may be possible to develop machine learning models for these narratives (eg, a model searching for instances of hedging expressions or words capable of twisting the stance of the sentence). Tagging these narratives during credibility annotation may not only increase the precision of sentence classifiers built upon such data sets, but, most importantly, also help disambiguate experts' labeling process.

Conclusions and Future Work

With the web quickly becoming one of the primary sources of the first medical information for the general public [60], the ability to distinguish between credible and noncredible information is indispensable. Financial interests of the alternative medicine community, combined with the rising distrust of the medical establishment, produce voluminous corpora of medical information of questionable quality. Of note, too many people fall prey to medical misinformation because it becomes increasingly harder to tell credible content from harmful deceit.

A possible solution to the problem of medical information source credibility is external certification. In our experiments, we correlated medical experts' labels with HON labels. The certification certainly works because only 18% (240/1333) of the sentences originating from HON-certified websites were classified by our experts as noncredible. However, obtaining the certificate is not simple, the certification process is long, and the entire framework does not scale well. This scalability problem demonstrates the bottleneck of any approach used for checking the credibility of medical content—the availability and time of medical professionals who need to be involved in the evaluation. In our work, we have taken the approach of optimizing the use of the time spent by experts on credibility evaluation of medical web content. The main goal of our future work will be the improvement and extension of this approach using active annotation and active learning methods.

In contrast, an ambitious goal would be to replace medical experts' evaluations with an automated credibility evaluation system. Such a system would use advanced natural language processing and machine classification algorithms. The results of our research demonstrate the challenges that would need to be overcome to make this possible.

The computational linguistic community is currently divided into 2 opposing camps: those who attribute *understanding of meaning* to language models and those who do not [61]. Despite the recent successes of modern language models such as Generative Pre-trained Transformer 3, the evidence seems to support a more cautious position. Indeed, a language model trained only on the form (raw text) cannot capture the true meaning of the text. The meaning, in this context, should be understood as the relationship between the linguistic form and the communicative intent of the speaker.

Our case goes beyond the learning of the meaning of sentences. As we have shown in this paper, there is an additional layer of complexity introduced by the notion of credibility of a statement to a user. Many machine learning solutions focus on the identification of factual flaws when addressing misinformation. However, fact-checking is not enough in the medical information domain. Often one encounters fake news and disinformation woven around factually true statements. We have seen time and time again medical experts using contextual information when assigning labels denoting sentence credibility. Most often they would take into account the most probable course of action taken by a patient who consumes medical information. Because of this mechanics of annotation, the relationship between sentence credibility and sentence truthfulness becomes ambiguous, further complicating the shape of the decision boundary between credible and noncredible medical statements.

This observation leads us to an important conclusion about the design of information-processing pipelines for medical content credibility evaluation. The first step is the compilation of large, high-quality data sets for machine learning model training. The active annotation approach presented in this paper allows doubling the number of sentences annotated by medical experts per cost unit (time or monetary). This, in turn, results in larger and more comprehensive training data sets. As a side effect, active annotation produces topical clusters of sentences, which can be used in 2 ways: (1) by allowing nonexpert annotators (whose time is far less expensive) to preprocess large batches of sentences to be reviewed by medical experts and (2) by reducing the cognitive stress of expert annotators due to the removal of context switching.

These 2 effects combined can further enhance the annotation process and increase the volume of annotated data. We also plan to extend the scope of the data set by covering more topics and providing more annotations.

The second step toward the support of medical content credibility evaluation would be the investigation of statistical models' efficacy for automatic classification of medical sentences as either credible or noncredible. Having an accurate classifier of medical sentence credibility, we might develop machine-assisted methods for finding consensus among human annotators, for example, by correlating human annotations with the confidence scores of the classifier. Finally, we would like to pursue active annotation in the light of 2 frameworks. Bayesian reasoning provides a set of tools for modeling individual annotators' beliefs about annotated data. Expectation maximization, in contrast, allows finding the best approximations (or maximum a posteriori estimates) of the unknown point credibility scores from empirical data. We see several possibilities of including the active annotation step in the iterative processes of Bayesian inference or expectation maximization.

Acknowledgments

The research leading to these results has received funding from the European Economic Area Financial Mechanism 2014-2021 (project registration number: 2019/35/J/HS6/03498).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Queries used to retrieve articles.
[DOCX File , 13 KB - medinform_v9i11e26065_app1.docx]

Multimedia Appendix 2
List of article URLs.
[DOCX File , 36 KB - medinform_v9i11e26065_app2.docx]

References

1. 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] [Medline: 32113495]
2. 5G conspiracy theories prosper during the coronavirus pandemic. Snopes. URL: https://www.snopes.com/news/2020/04/09/5g-conspiracy-theories-prosper-during-the-coronavirus-pandemic [accessed 2021-11-01]
3. Jablonka A, Happle C, Grote U, Schleenvoigt BT, Hampel A, Dopfer C, et al. Measles, mumps, rubella, and varicella seroprevalence in refugees in Germany in 2015. Infection 2016 Dec;44(6):781-787. [doi: 10.1007/s15010-016-0926-7] [Medline: 27449329]
4. Medical credibility corpus. GitHub. URL: https://github.com/alenabozny/medical_credibility_corpus [accessed 2021-11-17]
5. Sørensen K, Pelikan JM, Röthlin F, Ganahl K, Slonska Z, Doyle G, HLS-EU Consortium. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). Eur J Public Health 2015 Dec;25(6):1053-1058 [FREE Full text] [doi: 10.1093/eurpub/ckv043] [Medline: 25843827]
6. Keleher H, Hagger V. Health literacy in primary health care. Aust J Prim Health 2007 Jul 15;13(2):24-30 [FREE Full text] [doi: 10.1071/PY07020]
7. Cheng C, Dunn M. Health literacy and the internet: a study on the readability of Australian online health information. Aust N Z J Public Health 2015 Aug 25;39(4):309-314. [doi: 10.1111/1753-6405.12341] [Medline: 25716142]
8. Trethewey SP. Strategies to combat medical misinformation on social media. Postgrad Med J 2020 Jan 15;96(1131):4-6 [FREE Full text] [doi: 10.1136/postgradmedj-2019-137201] [Medline: 31732511]
9. Ecker UK, O'Reilly Z, Reid JS, Chang EP. The effectiveness of short-format refutational fact-checks. Br J Psychol 2020 Feb 02;111(1):36-54 [FREE Full text] [doi: 10.1111/bjop.12383] [Medline: 30825195]
10. Nyhan B, Reifler J. When corrections fail: the persistence of political misperceptions. Polit Behav 2010 Mar 30;32(2):303-330. [doi: 10.1007/s11109-010-9112-2]
11. Horne BD, Gruppi M, Adali S. Trustworthy misinformation mitigation with soft information nudging. In: Proceedings of the 2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). 2019 Presented at: 2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA); Dec 12-14, 2019; Los Angeles, CA, USA URL: https://ieeexplore.ieee.org/document/9014346 [doi: 10.1109/tps-isa48467.2019.00039]
12. Dhoju S, Rony MM, Kabir MA, Hassan N. Differences in health news from reliable and unreliable media. In: Proceedings of the WWW '19: The Web Conference. 2019 Presented at: WWW '19: The Web Conference; May 13-17, 2019; San Francisco USA. [doi: 10.1145/3308560.3316741]
13. Fernández-Pichel M, Losada DE, Pichel JC, Elsweiler D. Reliability prediction for health-related content: a replicability study. In: Advances in Information Retrieval. Cham: Springer; 2021.
14. Afsana F, Kabir MA, Hassan N, Paul M. Automatically assessing quality of online health articles. IEEE J Biomed Health Inform 2021 Feb;25(2):591-601. [doi: 10.1109/jbhi.2020.3032479] [Medline: 33079686]
15. Augenstein I, Lioma C, Wang D, Lima LC, Hansen C, Hansen C, et al. MultiFC: a real-world multi-domain dataset for evidence-based fact checking of claims. 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: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 2019; Hong Kong, China URL: https://aclanthology.org/D19-1475 [doi: 10.18653/v1/d19-1475]
16. Wang WY. "Liar, Liar Pants on Fire": a new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2017 Presented at: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); Jul 2017; Vancouver, Canada URL: https://aclanthology.org/P17-2067 [doi: 10.18653/v1/p17-2067]
17. Rashkin H, Choi E, Jang JY, Volkova S, Choi Y. Truth of varying shades: analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; Sep, 2017; Copenhagen, Denmark. [doi: 10.18653/v1/d17-1317]
18. Tchechmedjiev A, Fafalios P, Boland K, Gasquet M, Zloch M, Zapilko B, et al. ClaimsKG: a knowledge graph of fact-checked claims. In: The Semantic Web – ISWC 2019. Cham: Springer; Oct 2019.
19. Hassan N, Arslan F, Li C, Tremayne M. Toward automated fact-checking: detecting check-worthy factual claims by ClaimBuster. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017 Presented at: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13 - 17, 2017; Halifax NS Canada. [doi: 10.1145/3097983.3098131]
20. Karmakharm T, Aletras N, Bontcheva K. Journalist-in-the-loop: continuous learning as a service for rumour analysis. 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): System Demonstrations. 2019 Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations; Nov 2019; Hong Kong, China. [doi: 10.18653/v1/d19-3020]

21. Samuel H, Zaïane O. MedFact: towards improving veracity of medical information in social media using applied machine learning. In: Advances in Artificial Intelligence. Cham: Springer International Publishing; 2018.
22. Yilmaz Z, Yang W, Zhang H, Lin J. Cross-domain modeling of sentence-level evidence for document retrieval. 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: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov, 2019; Hong Kong, China URL: https://aclanthology.org/D19-1352 [doi: 10.18653/v1/d19-1352]
23. Chen D, Zhang S, Zhang X, Yang K. Cross-lingual passage re-ranking with alignment augmented multilingual BERT. IEEE Access 2020 Dec 1;8:213232-213243. [doi: 10.1109/access.2020.3041605]
24. Ghenai A, Mejova Y. Catching Zika fever: application of crowdsourcing and machine learning for tracking health misinformation on Twitter. In: Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI). 2017 Presented at: IEEE International Conference on Healthcare Informatics (ICHI); Aug 23-26, 2017; Park City, UT, USA. [doi: 10.1109/ichi.2017.58]
25. Shepperd S, Charnock D. Why DISCERN? Health Expect 1998 Nov 04;1(2):134-135 [FREE Full text] [doi: 10.1046/j.1369-6513.1998.0112a.x] [Medline: 11281867]
26. Moult B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. Health Expect 2004 Jun;7(2):165-175 [FREE Full text] [doi: 10.1111/j.1369-7625.2004.00273.x] [Medline: 15117391]
27. Bunge M, Mühlhauser I, Steckelberg A. What constitutes evidence-based patient information? Overview of discussed criteria. Patient Educ Couns 2010 Mar;78(3):316-328. [doi: 10.1016/j.pec.2009.10.029] [Medline: 20005067]
28. Working Group GPGI. Good practice guidelines for health information. Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen 2016;110-111:e1-e8 [FREE Full text] [doi: 10.1016/j.zefq.2016.01.004]
29. Keselman A, Smith CA, Murcko AC, Kaufman DR. Evaluating the quality of health information in a changing digital ecosystem. J Med Internet Res 2019 Feb 08;21(2):e11129 [FREE Full text] [doi: 10.2196/11129] [Medline: 30735144]
30. Marinelli F, Cervone A, Tortoreto G, Stepanov EA, Di Fabbrizio G, Riccardi G. Active annotation: bootstrapping annotation lexicon and guidelines for supervised NLU learning. Proc Interspeech 2019:574-578. [doi: 10.21437/interspeech.2019-2537]
31. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013 Presented at: Proceedings of the 26th International Conference on Neural Information Processing Systems; Dec 5 - 10, 2013; Lake Tahoe Nevada. [doi: 10.5555/2999792.2999959]
32. 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: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Oct, 2014; Doha, Qatar URL: https://aclanthology.org/D14-1162 [doi: 10.3115/v1/d14-1162]
33. Levy O, Goldberg Y. Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2014 Presented at: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); Jun, 2014; Baltimore, Maryland URL: https://aclanthology.org/P14-2050 [doi: 10.3115/v1/p14-2050]
34. Nickel M, Kiela D. Poincaré embeddings for learning hierarchical representations. arXiv. 2017. URL: https://arxiv.org/abs/1705.08039 [accessed 2021-11-17]
35. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. ArXiv.org. 2017 Dec. URL: https://arxiv.org/abs/1607.04606 [accessed 2021-11-17]
36. 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] [Medline: 9377276]
37. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014 Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP); Oct, 2014; Doha, Qatar URL: https://aclanthology.org/D14-1179 [doi: 10.3115/v1/d14-1179]
38. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. 2017. URL: https://arxiv.org/abs/1706.03762 [accessed 2021-11-17]
39. Cer D, Yang Y, Kong S, Hua N, Limtiaco N, St. John R, et al. Universal sentence encoder. arXiv. 2018. URL: https://arxiv.org/abs/1803.11175 [accessed 2011-11-17]
40. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. ArXiv.org. 2020. URL: https://arxiv.org/abs/2005.14165 [accessed 2021-11-17]
41. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. ArXiv.org. 2018. URL: https://arxiv.org/abs/1810.04805 [accessed 2021-11-17]
42. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. ArXiv.org. 2019. URL: https://arxiv.org/abs/1908.10084 [accessed 2021-11-17]

43. Chicco D. Siamese neural networks: an overview. Methods Mol Biol 2021;2190:73-94. [doi: 10.1007/978-1-0716-0826-5_3] [Medline: 32804361]
44. Sentence (linguistics). Wikipedia. URL: https://en.wikipedia.org/wiki/Sentence_(linguistics) [accessed 2021-11-17]
45. Lorenz-Spreen P, Mønsted BM, Hövel P, Lehmann S. Accelerating dynamics of collective attention. Nat Commun 2019 Apr 15;10(1):1759 [FREE Full text] [doi: 10.1038/s41467-019-09311-w] [Medline: 30988286]
46. Nabożny A, Balcerzak B, Koržinek D. Enriching the context: methods of improving the non-contextual assessment of sentence credibility. In: Web Information Systems Engineering – WISE 2019. Cham: Springer; 2019.
47. Krishna K, Murty MN. Genetic K-means algorithm. IEEE Trans Syst Man Cybern B Cybern 1999;29(3):433-439. [doi: 10.1109/3477.764879] [Medline: 18252317]
48. Yuan C, Yang H. Research on K-Value selection method of K-means clustering algorithm. J 2019 Jun 18;2(2):226-235. [doi: 10.3390/j2020016]
49. Liu Y, Myle O, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. ArXiv.org. 2019. URL: https://arxiv.org/abs/1907.11692 [accessed 2021-11-17]
50. Raunak V. Simple and effective dimensionality reduction for word embeddings. ArXiv.org. 2017. URL: https://tinyurl.com/vf29aah8 [accessed 2021-11-17]
51. Mu J, Bhat S, Viswanath P. All-but-the-top: simple and effective postprocessing for word representations. ArXiv.org. 2017. URL: https://arxiv.org/abs/1702.01417 [accessed 2021-11-17]
52. Alhmoud EN, Barazi R, Fahmi A, Abdu A, Higazy A, Elhajj M. Critical appraisal of the clinical practice guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular Riskuropean Society of Cardiology (ESC) and European Atherosclerosis Society (ESC/EAS) 2019 guidelines. J Pharm Health Serv Res 2020 Nov;11(4):423-427. [doi: 10.1111/jphs.12371]
53. Ferri N, Corsini A. Clinical pharmacology of statins: an update. Curr Atheroscler Rep 2020 Jun 03;22(7):26. [doi: 10.1007/s11883-020-00844-w] [Medline: 32494971]
54. Pergolizzi Jr JV, Coluzzi F, Colucci RD, Olsson H, LeQuang JA, Al-Saadi J, et al. Statins and muscle pain. Expert Rev Clin Pharmacol 2020 Mar 27;13(3):299-310. [doi: 10.1080/17512433.2020.1734451] [Medline: 32089020]
55. Yandrapalli S, Malik A, Guber K, Rochlani Y, Pemmasani G, Jasti M, et al. Statins and the potential for higher diabetes mellitus risk. Expert Rev Clin Pharmacol 2019 Sep 31;12(9):825-830. [doi: 10.1080/17512433.2019.1659133] [Medline: 31474169]
56. McGuinness B, Craig D, Bullock R, Passmore P. Statins for the prevention of dementia. Cochrane Database Syst Rev 2016 Jan 04(1):CD003160. [doi: 10.1002/14651858.CD003160.pub3] [Medline: 26727124]
57. Shrestha A, Mulmi A, Munankarmi R. Statins and abnormal liver enzymes. S D Med 2019 Jan;72(1):12-14. [Medline: 30849222]
58. Saha SP, Whayne TF. Coenzyme Q-10 in human health: supporting evidence? South Med J 2016 Jan;109(1):17-21. [doi: 10.14423/SMJ.0000000000000393] [Medline: 26741866]
59. Cholesterol Treatment Trialists' (CTT) Collaboration, Baigent C, Blackwell L, Emberson J, Holland LE, Reith C, et al. Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials. Lancet 2010 Nov 13;376(9753):1670-1681 [FREE Full text] [doi: 10.1016/S0140-6736(10)61350-5] [Medline: 21067804]
60. Sun Y, Zhang Y, Gwizdka J, Trace CB. Consumer evaluation of the quality of online health information: systematic literature review of relevant criteria and indicators. J Med Internet Res 2019 May 02;21(5):e12522 [FREE Full text] [doi: 10.2196/12522] [Medline: 31045507]
61. Bender EM, Koller A. Climbing towards NLU: on meaning, form, and understanding in the age of data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul, 2020; Online p. 5185-5198. [doi: 10.18653/v1/2020.acl-main.463]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
CoQ10: Coenzyme Q10
CVD: cardiovascular disease
HON: Health on the Net
LDL: low-density lipoprotein
sBERT: sentence–Bidirectional Encoder Representations from Transformers

Edited by C Lovis; submitted 26.11.20; peer-reviewed by W Xu, M Jordan-Marsh, A Hidki, S Nagavally; comments to author 09.02.21; revised version received 29.03.21; accepted 24.09.21; published 26.11.21.

Please cite as:
Nabożny A, Balcerzak B, Wierzbicki A, Morzy M, Chlabicz M
Active Annotation in Evaluating the Credibility of Web-Based Medical Information: Guidelines for Creating Training Data Sets for Machine Learning
JMIR Med Inform 2021;9(11):e26065
URL: https://medinform.jmir.org/2021/11/e26065
doi:10.2196/26065
PMID:34842547

Original Paper

eHealth Literacy and Beliefs About Medicines Among Taiwanese College Students: Cross-sectional Study

Chiao Ling Huang[1*], PhD; Chia-Hsun Chiang[2*], PhD; Shu Ching Yang[2*], PhD

[1]Faculty of Education, Department of Educational Information Technology, East China Normal University, Shanghai, China

[2]Intelligent Electronic Commerce Research Center, Institute of Education, National Sun Yat-Sen University, Kaohsiung, Taiwan

[*]all authors contributed equally

Corresponding Author:
Shu Ching Yang, PhD
Intelligent Electronic Commerce Research Center
Institute of Education
National Sun Yat-Sen University
70 Lienhai Rd
Kaohsiung, 80424
Taiwan
Phone: 886 75251521
Fax: 886 75255892
Email: shyang@mail.nsysu.edu.tw

Abstract

Background: Good eHealth literacy and correct beliefs about medicines are beneficial for making good health care decisions and may further influence an individual's quality of life. However, few studies have discussed these two factors simultaneously. Moreover, gender differences are associated with health literacy and beliefs about medicines. Therefore, it is important to examine the multiple relationships between college students’ eHealth literacy and beliefs about medicines, as well as gender differences.

Objective: This study aims to (1) examine the multiple relationships between eHealth literacy and beliefs about medicines and (2) analyze gender differences in eHealth literacy and beliefs about medicines with Taiwanese college students.

Methods: We used a paper-and-pencil questionnaire that included age, gender, 3-level eHealth literacy, and beliefs about medicines to collect data. In total, 475 data points were obtained and analyzed through independent *t* tests and canonical correlation analyses.

Results: The *t* test (t_{473}=3.73; *P*<.001; t_{473}=–2.10; *P*=.04) showed that women had lower functional eHealth literacy and more specific concerns about medicines than men. Canonical correlation analyses indicated that the first and second canonical correlation coefficients between eHealth literacy and beliefs about medicines reached a significant level, implying that a multivariate relationship indeed existed.

Conclusions: These findings reveal that women in Taiwan have lower functional eHealth literacy and stronger concerns about medicines than men. In addition, students with higher eHealth literacy have more positive perceptions of and beliefs about medicines.

(JMIR Med Inform 2021;9(11):e24144) doi:10.2196/24144

KEYWORDS

beliefs about medicines; college student; eHealth literacy; health education

Introduction

Beliefs about medicines refer to a concept related to the cognitive representation of medicines. Such beliefs are used to evaluate individuals’ beliefs regarding the necessity of and problems associated with prescribed medication and their more general beliefs about the overuse and detrimental effects of medicines [1]. According to the theory of self-regulation, beliefs about medicine are related to the decision to take medicine [1]. Studies have shown that nonadherence to medication regimens is more likely to occur among those with negative views regarding their medication, and that these views are accompanied by stronger concerns about potential harm, which caused them to believe that taking medicines is harmful in general [2,3].Therefore, helping people accurately understand medicines' use can be a key element in reducing misunderstandings regarding medication and further building positive attitudes toward medication.

In addition to directly help people establish correct notions of medication use, enhancing individuals' health literacy may be another useful approach because the process of forming beliefs about medicines may depend on information seeking and processing. Health literacy is the capability of individuals to access, comprehend, and effectively utilize health-related information, and it is a critical factor in the disease management and health promotion arena [4]. When patients request more information and written information is clear and easily understood, patients are assisted in improving their knowledge of and adherence to treatment [5]. Additionally, studies have shown that patients with poor health literacy are more inclined to incorrectly interpret labels and health information [6], more often possess negative emotionality that affect adherence [7], are more likely to believe that medications are necessary, and are more concerned about the possible side effects of their medications [8]. Namely, individuals with poor health literacy are less likely to comprehend information about diseases and medicines and have negative beliefs about medicines.

The internet has become a popular way to search for, obtain, and share health-related information over the last decade. Nevertheless, the abilities needed to collect and evaluate information through the internet differ from those needed to use books and other hard copies. In the cyberworld, individuals must be equipped with the ability, resources, and motive to seek, understand, and evaluate web-based health information [9]. eHealth literacy, which consists of functional, interactive, and critical levels [10], is defined as the capability to acquire needed and correct health information through the internet and further use this information to resolve health questions or make health decisions [11]. Apparently, the need for eHealth literacy is becoming increasingly important, as it is an indicator of how an individual applies web-based health information in his/her life [12]. Past studies have found that individuals with limited functional eHealth literacy may consult doctors frequently and that those with higher critical eHealth literacy can more effectively utilize health services [13]. However, research clarifying the multivariate correlations between eHealth literacy and beliefs about medicines is relatively scant. Furthermore, most previous studies on beliefs about medicines have focused on chronic disease patients [1,3,8,14] rather than normal samples, creating an academic gap.

In Taiwan, one can quickly and conveniently find information about medicinal drugs and disorders via the internet [15]. Before starting college, students' health literacy and beliefs about medicines are mostly influenced by parents and teachers. However, autonomous learning is encouraged in higher education. In this learning atmosphere, students have more opportunities to seek information on the internet and develop strong independent thinking and self-care abilities. Therefore, eHealth literacy and beliefs about medicines that students develop during college may well influence whether they will make good health care decisions in adulthood. Moreover, gender may play different roles in health literacy and beliefs about medicines. Females are reportedly more likely to have lower health literacy [16-18], stronger concerns about medicines [14], firmer beliefs about the necessity of taking medicine [19], and firmer beliefs that medications are overprescribed [20] than males.

Therefore, to obtain a comprehensive understanding of this topic and to address the lack of other samples in previous studies, with a focus on health education in Taiwan, the present study aims to examine the multiple relationships between college students' eHealth literacy and beliefs about medicines and to analyze gender differences. Specifically, we attempted to answer the following questions: What is the relationship between eHealth literacy and beliefs about medicines among college students? Further, what is the effect or role of gender in these two concepts? Based on the extant literature, we propose the following hypotheses:

1. Taiwanese college students with higher eHealth literacy are more likely to have positive perceptions of and beliefs about medicines.
2. Taiwanese women are more likely to have lower eHealth literacy and stronger beliefs about medicines than men.

Methods

Study Design and Participants

We collected 2 samples from Taiwan; 1 sample was used for pretesting, and the other sample was used for the formal study. Pretesting was employed to validate the appropriateness of the research instrument used in this study by performing an exploratory factor analysis (EFA) to confirm the construct validity. The formal study was used to analyze the relation between college students' eHealth literacy and their beliefs about medicines and the associated gender differences.

During pretesting, a convenience sampling approach was used to recruit students to participate in this investigation. We contacted our acquaintances who teach at other universities to help us promote this research and distribute the questionnaires. In total, 199 data points were returned and analyzed in the pretest study. A stratified random sampling method was adopted for the formal study. Specifically, we divided Taiwan into 3 regions and then used a computer to randomly select schools. For the selected schools, we contacted their instructors and asked whether they were willing to assist with distributing the questionnaire. Ultimately, 500 questionnaires were distributed, and 475 were returned. Among the 475 responses, missing values for each question did not exceed 2. We used the series mean to replace these missing data.

Measures

eHealth Literacy Scale

The participants' eHealth literacy was measured by the eHealth Literacy Scale (eHLS) [21]. The eHLS measures functional (3 items; eg, "I find the online health information difficult to understand"), interactive (4 items; eg, "I can locate health information efficiently through search engines"), and critical literacy (5 items; eg, "I will think about whether the online health information applies to my situation"). Functional eHealth literacy refers to basic competency in reading and writing web-based health information. Interactive eHealth literacy refers to the communication and social competencies used to consume information in a web-based social multimedia environment. Critical eHealth literacy involves people's cognitive competency in appraising, judging, or evaluating web-based information relevant to health [21]. According to Chiang et al’s [21] report, item analysis, EFA, and confirmatory factor analysis (CFA) were employed to determine the reliability and validity of eHLS. Specifically, the results of item analysis revealed that the comparisons between extreme measures ranged from 3.93 to 7.31 (*P*<.001), and the coefficient of correlation ranged from 0.70 to 0.85 (*P*<.01). The EFA and CFA results revealed that the Kaiser–Meyer–Olkin (KMO) measure was 0.83, the Bartlett test for sphericity was significant (*P*<.001), the factor loadings ranged from 0.53 to 0.90, the explained variance was 61.10%, the standardized factor loadings ranged from 0.60 to 0.86 (*P*<.001), composite reliability ranged from 0.75 to 0.84, and the average variance extracted for each dimension ranged from 0.50 to 0.52. In addition, for the goodness-of-fit indexes, χ^2_{51}=139.00, comparative fit index=0.96, root mean square error of approximation=0.06, and standardized root mean square residual=0.05. Each item in the eHLS was rated by the respondents on a 5-point Likert scale, with 1 indicating strong disagreement and 5 indicating strong agreement. Higher eHLS scores indicated that the participants had higher eHealth literacy, and all the variables were regarded as continuous variables. According to Sharma’s [22] guidelines, the reliability of a Likert-type rating scale can be obtained by computing the value of Cronbach α. The Cronbach α values obtained in our study sample were .82 (functional), .83 (interactive) and .87 (critical).

Beliefs About Medicines Scale

The Beliefs About Medicines Scale (BMS) was designed by the authors on the basis of Horne et al’s [1] scale. Three specialist professors in this field helped test the content validity of the BMS. We provided each professor with two sheets: one contains clear information about this scale, including this measurement’s purpose and each dimension’s definition, and the other contains an evaluation form with 2 options (inappropriate and appropriate) for each item. Experts were asked to complete the evaluation form and suggest modifications for items they rated inappropriate. Based on their suggestions, we revised and confirmed the BMS items until all experts were satisfied. The BMS includes specific and general sections measuring college students’ beliefs about medicines (Multimedia Appendix 1). The respondents rated each item in the BMS on a 5-point Likert scale, where 1 indicated strong disagreement and 5 indicated strong agreement, and all the variables were regarded as continuous variables. The results of the EFA (principal axis factors method with direct oblimin rotation) resulted in a 4-factor structure, which is the same as Horne et al’s [1] scale, and revealed that the factor loadings ranged from 0.54 to 0.88 and that the explained variance was 59.12%. Before conducting the EFA, we examined the results of the KMO and Bartlett sphericity tests to ensure that these data were appropriate for performing an EFA. According to Sharma’s [22] guidelines, the reliability of a Likert-type rating scale can be obtained by computing the value of Cronbach α. The Cronbach α values obtained in our study sample were .77 (specific-necessity, 4 items; eg, “I cannot live without my medicines”), .76 (specific-concern, 4 items; eg, “Taking medicines worries me”), .72 (general-overuse, 3 items; eg, “Most medicines are addictive”), and .84 (general-harm, 2 items; eg, “Doctors prescribe too many medicines”), respectively.

Data Analysis

In the pretest study, peer review was employed to test the content validity of the 13-item BMS, and an EFA was used to assess its construct validity. In the formal study, we calculated the internal consistency coefficients (α values) of each instrument and performed a descriptive statistical analysis, independent *t* tests, and a canonical correlation analysis to gain a better understanding of our samples and clarify the relationship among the research variables. Notably, all statistical analyses were performed using SPSS. A value of *P*<.05 was considered statistically significant.

Ethical Considerations

This study used an anonymous questionnaire to gather data, which is consistent with the government’s institutional review board rules for exempt review. All participation was voluntary and confidential. To ensure anonymity and confidentiality, we asked the students not to write any personal information on the questionnaire, and all the questionnaires were stored in a locked cabinet that only the researchers could access. Before distributing the questionnaires, the lecturer clearly informed the participants of the study aim and noted that they had the right to refuse to participate at any time without penalty; in addition, the participants were informed that their participation would not have any influence on their grades.

Results

Descriptive Statistics of eHealth Literacy and Beliefs About Medicines

Most of the students were under 22 years of age, except for 31 students whose ages ranged from 23 to 40 years, and the average age of our participants was 20.37 years. Table 1 presents the descriptive statistics of eHealth literacy and beliefs about medicines, showing that the college students basically had a moderate or high self-perceived level of eHealth literacy (all the means exceed 3). The participants also had low perceptions of the necessity of medicines (mean 1.78), low concerns about medications (mean 2.88), and low levels of the belief that medicines are harmful (mean 2.44) and overused (mean 2.50).

Table 1. Descriptive analysis of eHealth literacy and beliefs about medicines.

Attribute	Value, mean (SD)
eHealth literacy	
Functional	3.95 (0.77)
Interactive	3.73 (0.70)
Critical	3.81 (0.72)
Beliefs about medicines	
Specific-necessity	1.78 (0.72)
Specific-concerns	2.88 (0.94)
General-harm	2.44 (0.83)
General-overuse	2.50 (0.94)

Gender Differences in eHealth Literacy and Beliefs About Medicines

The *t* test results shown in Table 2 reveal that men had higher functional eHealth literacy than women (mean_{men} 4.07, $\text{mean}_{\text{women}}$ 3.81; t_{473}=3.73; *P*<.001) and that women had stronger concerns about medications than men (mean_{men} 2.79, $\text{mean}_{\text{women}}$ 2.97; t_{473}=–2.10; *P*=.04).

Table 2. *t* test for gender differences in eHealth literacy and beliefs about medicines.

Attributes	Score (men; n=250), mean (SD)	Score (women; n=225), mean (SD)	*t* test (*df*)	*P* value
eHealth literacy				
Functional	4.07 (0.77)	3.81 (0.75)	3.73 (473)	<.001
Interactive	3.77 (0.69)	3.69 (0.72)	1.22 (473)	.22
Critical	3.84 (0.72)	3.77 (0.72)	1.07 (473)	.29
Beliefs about medicines				
Specific-necessity	1.74 (0.72)	1.82 (0.71)	–1.10 (473)	.27
Specific-concerns	2.79 (0.93)	2.97 (0.95)	–2.10 (473)	.04
General-harm	2.40 (0.84)	2.47 (0.82)	–0.92 (473)	.36
General-overuse	2.46 (0.95)	2.55 (0.94)	–1.07 (473)	.28

Relationship Between eHealth Literacy and Beliefs About Medicines

The results of the canonical correlation analysis presented in Table 3 reveal that the first and second canonical correlation coefficients between eHealth literacy and beliefs about medicines were 0.28 and 0.15, respectively, which reached a significant level (*P*<.05).

The results of the first canonical variate indicated that the students with relatively high functional and interactive eHealth literacy were less concerned about the 4 aspects related to the medicines used. The results of the second canonical variate revealed that the students with relatively high critical eHealth literacy were less concerned about the specific necessity of the medicines used.

Table 3. Canonical correlation analysis of eHealth literacy and beliefs about medicines.

Attributes	First canonical variate	Second canonical variate
eHealth literacy set, r_s[a]		
Functional	*0.96*	–0.17
Interactive	*0.50*	0.07
Critical	0.48	*0.74*
Variance percentage	46.94	19.46
Redundancy	3.65	0.44
Beliefs about medicines set, r_s		
Specific-necessity	*–0.63*	*–0.69*
Specific-concerns	*–0.63*	0.34
General-harm	*–0.77*	0.37
General-overuse	*–0.60*	–0.07
Variance percentage	43.42	18.25
Redundancy	3.38	0.42
R_c[b]	0.28	0.15
R_c^2	0.08	0.02
P value	<.001	.03

[a] r_s: structure coefficients (canonical loadings) and absolute values of ≥0.50 are italicized.

[b] R_c: canonical correlation coefficients.

Discussion

Principal Findings

This study found that women had lower functional eHealth literacy and higher specific concern about medicines than men; hence, hypothesis 2 is only partially supported by our findings. Previous studies have found that female sex is associated with limited functional health literacy [16,17]. Researchers have argued that the lower level of health literacy among females was probably owing to their low educational level [17,18]. However, this study found that despite similar education levels, functional eHealth literacy may also have gender differences.

Functional literacy is closely linked to reading comprehension and numeracy skills [23]. Although reading comprehension competence is similar between the sexes [24], studies have found that females have poorer numeracy skills than males [25], including those with higher education [26]. Poorer numeracy may account for the phenomenon that female students were more likely to have inadequate functional eHealth literacy than male students, but this merits further study. In addition, overestimation and underestimation of eHealth literacy may have occurred because self-reporting was used for assessment. It is possible that functional eHealth literacy is similar between male and female college students. Future studies could develop a direct test of eHealth literacy and clarify whether a difference in eHealth literacy exists on the basis of gender among college students.

Consistent with Viktil et al [14], this study found that women worried more about the potential negative effects of medication than men. Physiological sex differences may result in a dissimilar incidence rate of disease and response to drug therapy [27]. Studies have shown that females are more likely to use gender-specific drugs and general medications [28] and have more adverse drug reactions than males [28,29]. In addition, it is worth noting that women’s adverse drug reactions might have been ignored in pharmaceutical research for many years. In contrast, men’s adverse drug reactions have gained more attention. For example, independent safety committees ceased Behre et al’s [30] study because the committees found that the injectable combination hormonal contraceptive for men had some side effects (ie, depression and mood changes) and the risks outweighed the potential benefits for the participants. Thus, women's concerns about medication are not simply owing to the physiological differences between the 2 sexes but also because pharmaceutical research has ignored the side effects of medication for women (or, in the absence of relevant experimental data, it is unclear what side effects these drugs have on women). In this case, if women need to take medication for treatment, they have no other choice but to adapt to the current medication, which may elicit women's concerns. Therefore, this finding’s result should be interpreted with caution.

Hypothesis 1 is also only partially supported by the results of the canonical correlation analysis. The first canonical result indicated that students with relatively high functional eHealth literacy were less concerned about the 4 aspects related to medicine. According to Taiwan’s Ministry of Health and Welfare, relevant indications should be printed on medication packaging, such as drug names, cautions, principal indications, and main adverse reactions [15]. Functional health literacy particularly emphasizes basic reading skills, which are used to address health information [10,31,32]. Lower functional health literacy is a barrier preventing patients from understanding their health conditions, such as diseases and proper treatment [33]. Previous studies have also shown that individuals with poor functional health literacy tend to have the attitude that drug therapy is requisite and are inclined to express concerns about the possible adverse reactions or sequela of their medications [8].

Therefore, compared to students with higher functional eHealth literacy, those with lower functional eHealth literacy might have poorer abilities to understand information related to their medications and thus believe that they need more personal medication and have more concerns about the potential negative effects of medicines; such students were also more inclined to believe that medicines are harmful and overused by physicians. Scholars have argued that illustrated medication information provides useful reinforcement [34], and for individuals with low health literacy, illustrations enhance their health knowledge and adherence to medications [35]. Thus, illustrated medication information could help students with low eHealth literacy build positive perceptions and beliefs about medicines.

Moreover, the results of the first set of canonical correlations showed that students with higher interactive eHealth literacy tended to have positive perceptions and beliefs about medicines. In Taiwan, people can conveniently and quickly search for information about medications and disorders through the internet, news, and magazines [15]. Interactive eHealth literacy identifies the communication and social competency that are employed in consuming health information in the web-based environment [13,21,36]. Students with adequate interactive eHealth literacy have the capability to collect information and extract meaning from various types of communications and further build positive beliefs about medicines. A study found that a summary information leaflet, regular health presentations and monthly meetings can effectively change patients' health education concept and positively influence their attitudes toward medicines [37]. Therefore, health education practitioners could conduct regular health education campaigns to build positive beliefs about medicines for students with low interactive eHealth literacy.

Finally, the results of the second set of canonical correlation analysis revealed that students with adequate critical eHealth literacy were less likely to have stronger perceptions of a need for medications to maintain their present and future health. Critical eHealth literacy refers to the most advanced cognitive competency that is used to critically appraise and evaluate web-based information relevant to health [13,21,36]. Given that web-based health information often lacks evidence and is incorrect or misinforming, patients are easily confused and likely to form inaccurate and negative beliefs about medications [38]. Cultivating students' critical health literacy is necessary because critical literacy can help them analyze health information and take an active role in addressing their health-related issues [39], thus giving college students lower perceptions of the need for medications to maintain their current and future health.

Limitations of This Study

Although this study contributes to our understanding of the relevant correlates of eHealth literacy and beliefs about medicines, it has some limitations. First, this study used a cross-sectional design. Thus, we gathered data at a single time point and could not determine the development of beliefs about medicines along with eHealth literacy. Second, since the 2 scales (eHLS and BMS) applied in this study did not have predictive validity information, these scales' utility may be questionable. We suggest that future studies should consider this issue, using more complete scales to address this concern. However, even with this flaw, the study provides a good starting point for further studies in general. Third, our participants may have been likely to overestimate or underestimate their eHealth literacy and beliefs about medicines owing to social desirability expectations. Future studies could adopt research methods other than self-reporting assessment to address this issue. Fourth, the sample in the current study was restricted by age to college students in Taiwan. The findings should be interpreted considering the sample's homogeneity. In particular, students' different demographic characteristics (eg, major, type of family, religious practices, and community) may influence beliefs about medicines and eHealth literacy. Future studies can use diverse samples and perform covariance analysis to control for these potentially influential factors. Finally, students' healthy skepticism of medications is not documented as part of this study. Healthy skepticism helps people wary of medicine information and gauges medications' value from a different perspective. Future studies could consider this variable.

Conclusions

To the best of our knowledge, this study is the first to explore the relationship among the 3 levels of eHealth literacy and 4 aspects of beliefs about medicines among college students, especially Taiwanese college students. Our study contributes not only to research but also to educational practice. Our results indicate that higher the eHealth literacy of Taiwanese college students, the more positive the perceptions and beliefs about medicines they held, thus providing some insights for health educational practitioners who could help college students build positive beliefs about medicines by promoting their eHealth literacy. In addition, this study found that women had lower functional eHealth literacy and stronger concerns about medicines than men. Therefore, health educators could develop gender-specific programs to improve women's functional eHealth literacy and reduce their concerns about medicines.

Acknowledgments

This research was financially supported by the Intelligent Electronic Commerce Research Center from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Each section of the Beliefs About Medicines Scale.
[DOCX File , 14 KB - medinform_v9i11e24144_app1.docx]

References

1. Horne R, Weinman J, Hankins M. The beliefs about medicines questionnaire: the development and evaluation of a new method for assessing the cognitive representation of medication. Psychol Health 1999 Jan;14(1):1-24. [doi: 10.1080/08870449908407311]
2. Chapman SCE, Horne R, Chater A, Hukins D, Smithson WH. Patients' perspectives on antiepileptic medication: relationships between beliefs about medicines and adherence among patients with epilepsy in UK primary care. Epilepsy Behav 2014 Feb;31:312-320 [FREE Full text] [doi: 10.1016/j.yebeh.2013.10.016] [Medline: 24290250]
3. Sweileh WM, Zyoud SH, Abu Nab'a RJ, Deleq MI, Enaia MI, Nassar SM, et al. Influence of patients' disease knowledge and beliefs about medicines on medication adherence: findings from a cross-sectional survey among patients with type 2 diabetes mellitus in Palestine. BMC Public Health 2014 Jan 30;14:94 [FREE Full text] [doi: 10.1186/1471-2458-14-94] [Medline: 24479638]
4. Chakraverty D, Baumeister A, Aldin A, Monsef I, Jakob T, Seven Ü, et al. Gender differences of health literacy in first and second generation migrants: a systematic review. Eur J Public Health 2019;29(4):ckz186.045. [doi: 10.1093/eurpub/ckz186.045]
5. Weinman J. Providing written information for patients: psychological considerations. J R Soc Med 1990 May;83(5):303-305 [FREE Full text] [Medline: 2380946]
6. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
7. Kalichman SC, Ramachandran B, Catz S. Adherence to combination antiretroviral therapies in HIV patients of low health literacy. J Gen Intern Med 1999 May;14(5):267-273 [FREE Full text] [doi: 10.1046/j.1525-1497.1999.00334.x] [Medline: 10337035]
8. Kale MS, Federman AD, Krauskopf K, Wolf M, O'Conor R, Martynenko M, et al. The association of health literacy with illness and medication beliefs among patients with chronic obstructive pulmonary disease. PLoS One 2015;10(4):e0123937 [FREE Full text] [doi: 10.1371/journal.pone.0123937] [Medline: 25915420]
9. Karnoe A, Kayser L. How is eHealth literacy measured and what do the measurements tell us? A systematic review. Knowl Manag E-Learn 2015;7(4):576-600. [doi: 10.34105/j.kmel.2015.07.038]
10. Nutbeam D. The evolving concept of health literacy. Soc Sci Med 2008 Dec;67(12):2072-2078. [doi: 10.1016/j.socscimed.2008.09.050] [Medline: 18952344]
11. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
12. Masilamani V, Sriram A, Rozario AM. eHealth literacy of late adolescents: credibility and quality of health information through smartphones in India. Comunicar: Revista Científica de Comunicación y Educación 2020 Jul 01;28(64):85-95. [doi: 10.3916/c64-2020-08]
13. Luo YF, Yang SC, Chen AS, Chiang CH. Associations of eHealth literacy with health services utilization among college students: cross-sectional study. J Med Internet Res 2018 Oct 25;20(10):e283 [FREE Full text] [doi: 10.2196/jmir.8897] [Medline: 30361201]
14. Viktil KK, Frøyland H, Rogvin M, Moger TA. Beliefs about medicines among Norwegian outpatients with chronic cardiovascular disease. Eur J Hosp Pharm 2014 Apr;21(2):118-120 [FREE Full text] [doi: 10.1136/ejhpharm-2013-000346] [Medline: 24683471]
15. Safety Medication Handbook. Taiwan Food and Drug Administration. 2017. URL: https://www.fda.gov.tw/tc/includes/GetFile.ashx?id=f636694183609501982 [accessed 2021-10-12]
16. Baron-Epel O, Balin L, Daniely Z, Eidelman S. Validation of a Hebrew health literacy test. Patient Educ Couns 2007 Jul;67(1-2):235-239. [doi: 10.1016/j.pec.2007.02.005] [Medline: 17386994]
17. Javadzade SH, Sharifirad G, Radjati F, Mostafavi F, Reisi M, Hasanzade A. Relationship between health literacy, health status, and healthy behaviors among older adults in Isfahan, Iran. J Educ Health Promot 2012;1:31 [FREE Full text] [doi: 10.4103/2277-9531.100160] [Medline: 23555134]
18. Amoah PA, Phillips DR. Socio-demographic and behavioral correlates of health literacy: a gender perspective in Ghana. Women Health 2020 Feb;60(2):123-139. [doi: 10.1080/03630242.2019.1613471] [Medline: 31092133]

19. Emilsson M, Gustafsson PA, Öhnström G, Marteinsdottir I. Beliefs regarding medication and side effects influence treatment adherence in adolescents with attention deficit hyperactivity disorder. Eur Child Adolesc Psychiatry 2017 May;26(5):559-571 [FREE Full text] [doi: 10.1007/s00787-016-0919-1] [Medline: 27848023]
20. Hong SH. Potential for physician communication to build favorable medication beliefs among older adults with hypertension: a cross-sectional survey. PLoS One 2019 Jan 7;14(1):e0210169 [FREE Full text] [doi: 10.1371/journal.pone.0210169] [Medline: 30615656]
21. Chiang CH, Yang SC, Hsu WC. Development and validation of the e-health literacy scale and investigation of the relationships between e-health literacy and healthy behavior among undergraduate students in Taiwan. Formosa J Mental Health 2015;28(3):389-420. [doi: 10.30074/FJMH.201509_28(3).0002]
22. Sharma B. A focus on reliability in developmental research through Cronbach's alpha among medical, dental and paramedical professionals. Asian Pac J Health Sci 2016;3(4):271-278 [FREE Full text] [doi: 10.21276/apjhs.2016.3.4.43]
23. Zegers CA, Gonzales K, Smith LM, Pullen CH, De Alba A, Fiandt K. The psychometric testing of the functional, communicative, and critical health literacy tool. Patient Educ Couns 2020 Nov;103(11):2347-2352. [doi: 10.1016/j.pec.2020.05.019] [Medline: 32622692]
24. Waldrop-Valverde D, Jones DL, Jayaweera D, Gonzalez P, Romero J, Ownby RL. Gender differences in medication management capacity in HIV infection: the role of health literacy and numeracy. AIDS Behav 2009 Feb;13(1):46-52 [FREE Full text] [doi: 10.1007/s10461-008-9425-x] [Medline: 18618237]
25. Mohammadi Z, Tehrani Banihashemi A, Asgharifard H, Bahramian M, Baradaran HR, Khamseh ME. Health literacy and its influencing factors in Iranian diabetic patients. Med J Islam Repub Iran 2015;29:230 [FREE Full text] [Medline: 26478888]
26. Cook R. Gender differences in adult numeracy skills: what is the role of education? Educ Res Eval 2018 Dec 11;24(6-7):370-393. [doi: 10.1080/13803611.2018.1540992]
27. Fan HM, Wu TW, Peng TR. Sex-related differences in pharmacotherapy. Formosa J Clin Pharm 2018;26(2):115-121. [doi: 10.6168/FJCP.201804_26(2).0004]
28. Rademaker M. Do women have more adverse drug reactions? Am J Clin Dermatol 2001;2(6):349-351. [doi: 10.2165/00128071-200102060-00001] [Medline: 11770389]
29. Zopf Y, Rabe C, Neubert A, Janson C, Brune K, Hahn EG, et al. Gender-based differences in drug prescription: relation to adverse drug reactions. Pharmacology 2009;84(6):333-339. [doi: 10.1159/000248311] [Medline: 19844133]
30. Behre HM, Zitzmann M, Anderson RA, Handelsman DJ, Lestari SW, McLachlan RI, et al. Efficacy and safety of an injectable combination hormonal contraceptive for men. J Clin Endocrinol Metab 2016 Dec;101(12):4779-4788. [doi: 10.1210/jc.2016-2141] [Medline: 27788052]
31. Netemeyer RG, Dobolyi DG, Abbasi A, Clifford G, Taylor H. Health literacy, health numeracy, and trust in doctor: effects on key patient health outcomes. J Consum Aff 2020;54(1):3-42. [doi: 10.1111/joca.12267]
32. Wills J, Sykes S, Hardy S, Kelly M, Moorley C, Ocho O. Gender and health literacy: men's health beliefs and behaviour in Trinidad. Health Promot Int 2020 Aug 1;35(4):804-811. [doi: 10.1093/heapro/daz076] [Medline: 31407795]
33. Kalichman SC, Benotsch E, Suarez T, Catz S, Miller J, Rompa D. Health literacy and health-related knowledge among persons living with HIV/AIDS. Am J Prev Med 2000 May;18(4):325-331. [doi: 10.1016/s0749-3797(00)00121-5]
34. Roberts NJ, Ghiassi R, Partridge MR. Health literacy in COPD. Int J Chron Obstruct Pulmon Dis 2008;3(4):499-507 [FREE Full text] [doi: 10.2147/copd.s1088] [Medline: 19281068]
35. Negarandeh R, Mahmoodi H, Noktehdan H, Heshmat R, Shakibazadeh E. Teach back and pictorial image educational strategies on knowledge about diabetes and medication/dietary adherence among low health literate patients with type 2 diabetes. Prim Care Diabetes 2013 Jul;7(2):111-118. [doi: 10.1016/j.pcd.2012.11.001] [Medline: 23195913]
36. Yang SC, Luo YF, Chiang CH. The associations among individual factors, eHealth literacy, and health-promoting lifestyles among college students. J Med Internet Res 2017 Jan 10;19(1):e15 [FREE Full text] [doi: 10.2196/jmir.5964] [Medline: 28073739]
37. Magadza C, Radloff SE, Srinivas SC. The effect of an educational intervention on patients' knowledge about hypertension, beliefs about medicines, and adherence. Res Social Adm Pharm 2009 Dec;5(4):363-375. [doi: 10.1016/j.sapharm.2009.01.004] [Medline: 19962679]
38. Linn AJ, van Weert JCM, Gebeyehu BG, Sanders R, Diviani N, Smit EG, et al. Patients' online information-seeking behavior throughout treatment: the impact on medication beliefs and medication adherence. Health Commun 2019 Nov;34(12):1461-1468. [doi: 10.1080/10410236.2018.1500430] [Medline: 30052088]
39. Mitchell B, Begoray D. Electronic personal health records that promote self-management in chronic illness. Online J Issues Nurs 2010;15(3) [FREE Full text]

Abbreviations

BMS: Beliefs About Medicines Scale
CFA: confirmatory factor analysis
EFA: exploratory factor analysis

eHLS: eHealth Literacy Scale
KMO: Kaiser–Meyer–Olkin

Edited by G Eysenbach; submitted 06.09.20; peer-reviewed by M Kaminski, I Mircheva, N Kaur, MDG Pimentel, R Krukowski; comments to author 07.10.20; revised version received 28.11.20; accepted 04.10.21; published 30.11.21.

Please cite as:
Huang CL, Chiang CH, Yang SC
eHealth Literacy and Beliefs About Medicines Among Taiwanese College Students: Cross-sectional Study
JMIR Med Inform 2021;9(11):e24144
URL: https://medinform.jmir.org/2021/11/e24144
doi:10.2196/24144
PMID:34851301

Original Paper

Visualizing Knowledge Evolution Trends and Research Hotspots of Personal Health Data Research: Bibliometric Analysis

Jianxia Gong[1], PhD; Vikrant Sihag[2], PhD; Qingxia Kong[3], PhD; Lindu Zhao[1], PhD

[1]School of Economics and Management, Southeast University, Nanjing, China

[2]Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, Netherlands

[3]Department of Technology and Operations Management, Erasmus University Rotterdam, Rotterdam, Netherlands

Corresponding Author:
Lindu Zhao, PhD
School of Economics and Management
Southeast University
No 2 Sipailou
Nanjing, 210096
China
Phone: 86 2583793776
Fax: 86 2583794731
Email: openmis@163.com

Abstract

Background: The recent surge in clinical and nonclinical health-related data has been accompanied by a concomitant increase in personal health data (PHD) research across multiple disciplines such as medicine, computer science, and management. There is now a need to synthesize the dynamic knowledge of PHD in various disciplines to spot potential research hotspots.

Objective: The aim of this study was to reveal the knowledge evolutionary trends in PHD and detect potential research hotspots using bibliometric analysis.

Methods: We collected 8281 articles published between 2009 and 2018 from the Web of Science database. The knowledge evolution analysis (KEA) framework was used to analyze the evolution of PHD research. The KEA framework is a bibliometric approach that is based on 3 knowledge networks: reference co-citation, keyword co-occurrence, and discipline co-occurrence.

Results: The findings show that the focus of PHD research has evolved from medicine centric to technology centric to human centric since 2009. The most active PHD knowledge cluster is developing knowledge resources and allocating scarce resources. The field of computer science, especially the topic of artificial intelligence (AI), has been the focal point of recent empirical studies on PHD. Topics related to psychology and human factors (eg, attitude, satisfaction, education) are also receiving more attention.

Conclusions: Our analysis shows that PHD research has the potential to provide value-based health care in the future. All stakeholders should be educated about AI technology to promote value generation through PHD. Moreover, technology developers and health care institutions should consider human factors to facilitate the effective adoption of PHD-related technology. These findings indicate opportunities for interdisciplinary cooperation in several PHD research areas: (1) AI applications for PHD; (2) regulatory issues and governance of PHD; (3) education of all stakeholders about AI technology; and (4) value-based health care including "allocative value," "technology value," and "personalized value."

(JMIR Med Inform 2021;9(11):e31142) doi:10.2196/31142

KEYWORDS

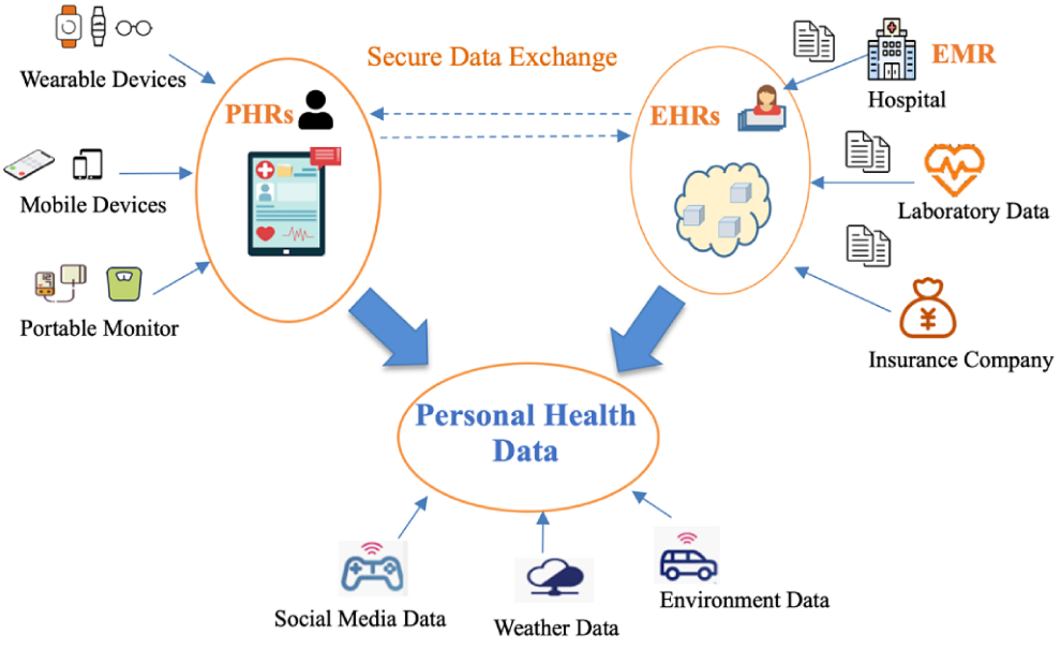
knowledge evolution trends; research hotspots; personal health data; bibliometrics

Introduction

Over the past 20 years, the use of patient medical information has rapidly increased in both clinical practice and research [1,2]. Improved access to personal health data (PHD), thanks to emerging technologies such as wearable devices, and mobile phones have improved health care delivery and physician–patient relationships, particularly for patients with noncommunicable chronic diseases [3]. PHD can play an important role in providing patient-centered rather than disease-centered health care by facilitating health care providers to learn about an individual's medical history and current health status [4-6]. At the same time, this data-driven approach is helping to provide cost-effective and high-quality health care—known as value-based health care [7]. It is expected that PHD will continue to transform the health care industry.

PHD includes both clinical data (eg, electronic medical records [EMRs], electronic health records [EHRs], personal health records [PHRs]) and nonclinical data (eg, sentiments, emotions, characteristics, and social media behavior) [2]. Figure 1 shows the relationship between EMR, EHR, PHR, and PHD. EMR files are real-time electronic files including only clinical records that have replaced paper files; these are usually not sent to other health care providers outside the treating hospital or clinic [8]. This transition to electronic records signifies a great digital transition in the health care industry. The standardization of EHR has provided a repository of health information that has greatly facilitated interoperability between different institutions [2]. EHR usually belongs to health care organizations [9] and cannot be easily transmitted between different organizations because of different data standards and health information systems. To overcome this limitation, PHR was generated [6]. PHRs are electronic records of health-related information that conform to national interoperability standards and can be drawn from multiple sources (eg, EHRs, laboratory test results, smartphones, and wearable devices), while being managed, shared, and controlled by the individual [10].

Figure 1. PHR, EHR, and EMR relationships. EHR: electronic health record; EMR: electronic medical record; PHR: personal health record.



Health care providers now have access to clinical data from EHR and patients' self-reported health data (eg, test results, medication lists, allergies) from PHRs. However, they do not have access to the patients' self-reported experiences, attitudes, feelings, and emotional states. The development of the internet of things and wearable devices means that PHD can also include nonclinical health-related data, such as daily physical activity and diets. Individuals are now sharing more and more detailed health information via social media platforms such as Twitter and through online health communities such as PatientsLikeMe [11]. Hill [12] defined PHD as any data related to an individual's health condition [12], while Plastiras and O'Sullivan [13] viewed PHD as health data generated by patients during their daily life. In this study, PHD is defined as data related to clinical and nonclinical well-being, including EMR, EHR, PHR, and environment and social media data. Incorporating broader nonclinical PHD such as emotions and feelings has been shown to enhance personalized health care delivery [14,15].

PHD research has gained attention in various fields, including computer science, bioinformatics, medicine, and public health. Searching for the keyword "personal health data" on Web of Science shows that relevant articles on PHD have increased greatly (Multimedia Appendix 1). Several systematic reviews have been published on different topics associated with PHD (Table 1). These include security and privacy problems associated with EHR [16], data types and standardization [6], facilitators and barriers to using EHR in the United States [17,18], barriers to data sharing [19], and ethical issues of data collection [20]. Others have investigated factors affecting the use of PHR and big data applications of PHD [11,21,22]. While the PHD research literature grows rapidly, some scholars acknowledged the value of presenting comprehensive landscape and topic evolution process of PHD publications for researchers in various disciplines, in which bibliometric as a quantitative analysis method can be useful. Some scholars analyzed the status and detected the high-frequency terms of EHR [23-26]. Wen et al [27] analyzed the production trends of publications on EHRs by countries from 2009 to 2015. Wang et al [28] used bibliometric methods to compare publication hotspots in EHRs from different periods among 6 countries. The recent articles by Qian et al [29] and Zhenni and Yuxing [30] applied social network analysis and topic modeling methods to explore the EHR publications in-depth to evaluate the publications trends and detect the frontiers. However, these were mainly aimed at a specific type of health data: EHR. Karampela et al [2] used a systematic mapping approach to present the publication channel, publication year, and major research topics to provide a more complete overview of PHD research. However, it is not clear what phase each topic is in, how each topic is progressing, what knowledge trends are evolving, and which topics will become research hotspots.

This study aims to examine the evolving trends and to detect the potential research hotspots of PHD by identifying, classifying, and clustering PHD research topics from 2009 to 2018. We used knowledge evolution analysis (KEA) with bibliometric techniques to review articles retrieved from the Web of Science database. This study traces the evolution of PHD using knowledge networks based on reference co-citation, keyword co-occurrence, and discipline co-occurrence. Revealing the interrelationships between PHD research topics will provide a solid framework for future research. Table 2 presents the key questions that will be answered in this study.

Table 1. Comparison of literature reviews.

Study	Research question	Sample size	Time range	Method
Archer et al [17]	PHRs[a] design, functionality, implementation, application, outcomes, and benefits	130	Unlimited-2010	Systematic review
Fernández-Alemán et al [16]	Security and privacy in EHRs[b]	49	2006-2011	Systematic review
Van Panhuis et al [19]	Barriers to data sharing	65	Unlimited-2013	Systematic review
Kruse et al [18]	Adoption factors of EHRs	31	2012-2015	Systematic review
Roehrs et al [6]	Data types, standards, profiles, goals, methods, functions, and architecture with PHRs	97	2008-2017	Systematic review
Yin et al [11]	Machine learning in online personal health data	103	2010-2018	Systematic review
Maher et al [20]	Ethical issues in passive data collection	48	Unlimited-2018	Systematic review
Abd-alrazaq et al [21]	Factors affecting the use of PHRs	97	2000-2018	Systematic review
Mehta and Pandit [22]	Big data analytics in PHD[c]	58	2013-2018	Systematic review
Wang et al [28]	Evolution of publication hotspots in EHRs	17,678	1957-2016	Bibliometric method
Wen et al [27]	Production trends of EHR	1803	1991-2005	Bibliometric method
Guo et al [23]	Status, hotspots of EHR	5095	2005-2010	Bibliometric method
Liang et al [24]	Status, directions of EHR	1262	1990-2013	Bibliometric method
Ruixian et al [25]	Status of EMR[d] in China	262	1999-2004	Bibliometric method
Zhenni and Yuxing [30]	Hot spots in EHR	13,438	1900-2019	Bibliometric method
Qian et al [29]	Landscape, hot topics, trends of EHRs	13,438	1900-2019	Bibliometric method
Lin et al [26]	Status of EMR research in China	1752	1999-2012	Bibliometric method
Karampela et al [2]	Publication source, publication year, research topic	246	Unlimited-2018	Systematic mapping study
This study	Knowledge evolution trajectory of PHD, including EHR, PHR, and EMR	8281	2009-2018	Bibliometric method

[a]PHR: personal health record.

[b]EHR: electronic health record.

[c]PHD: personal health data.

[d]EMR: electronic medical record.

Table 2. Mapping questions.

Question and ID	Mapping question	Rationale
MQ1[a]: References		
MQ1.1	How does the references co-citation network shape?	To understand the main topics and the development of research topics in PHD.[b]
MQ1.2	How has the knowledge cluster evolved?	To identify which PHD topic has the most longevity and the newest hotspot.
MQ1.3	What are the citation bursts of reference networks?	To explore the emerging PHD research topic characterized by articles.
MQ2: Keywords		
MQ2.1	What are the keyword bursts in recent years?	To explore the emerging research interests in PHD characterized by keywords.
MQ3: Disciplines		
MQ3.1	What does the discipline categories co-occurrence network shape?	To identify the trends of discipline categories that are involved in PHD.
MQ3.2	What are the discipline categories bursts?	To explore the discipline categories that increased abruptly in PHD.

[a] MQ: mapping question.

[b] PHD: personal health data.

Methods

Data Collection

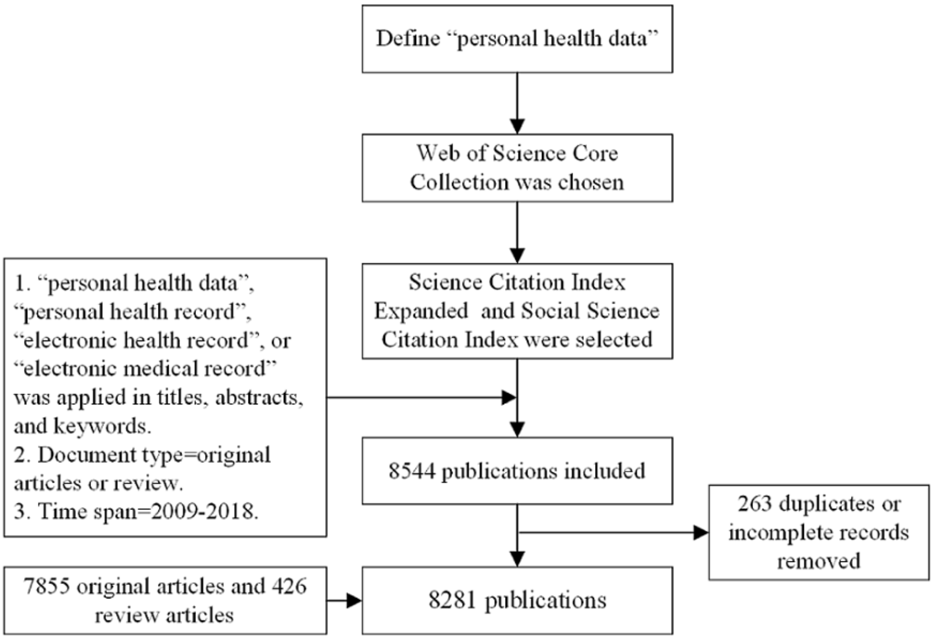
In 2009, the American Health Information Management Association launched a foundation program "Better health information for all" [2]. From then on, PHD research has developed greatly. Therefore, the time span for the retrieval is from 2009 to 2018 (The data collection was on March 8, 2019). In this review, we relied on scholarly publications in the Web of Science Core Collection, which covers over 21,000 science and social science journals and gives access to multiple databases that reference cross-disciplinary research. Web of Science has been long recognized as an ideal data source for bibliometric analysis.

To ensure the quality of the data set, we retrieved both original research articles and review articles from Science Citation Index Expanded and Social Science Citation Index. As there is no common definition for PHD, the following terms were searched in titles, abstracts, or keywords to identify PHD-related research in the Web of Science database: "personal health data", "personal health record", "electronic health record" or "electronic medical record". In Web of Science, the "Topic Search" function returns results in titles, abstracts, or keywords. Thus, the search query was defined as follows:

TS(Topic)=("personal health data" OR "personal health record" OR "electronic health record" OR "electronic medical record") AND DT(Document Types)=("Articles" OR "Review") AND PY(Year Published)=(2009-2018).

This search yielded 8544 publications. After eliminating publications with replicated or incomplete retrieval data, 8281 records were left, 7855 (94.86%) of which were original articles and 426 (5.14%) review articles. The data set selection process follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow (Figure 2).

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow chart of data selection.

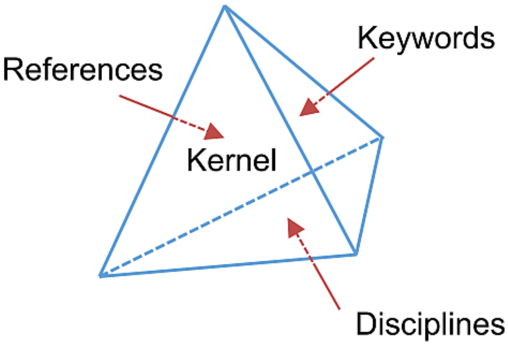

Define "personal health data"
Web of Science Core Collection was chosen
Science Citation Index Expanded and Social Science Citation Index were selected
1. "personal health data", "personal health record", "electronic health record", or "electronic medical record" was applied in titles, abstracts, and keywords.
2. Document type=original articles or review.
3. Time span=2009-2018.
8544 publications included
263 duplicates or incomplete records removed
7855 original articles and 426 review articles
8281 publications

Data Analysis

Overview

We used KEA to analyze the evolution of PHD research. The KEA followed a bibliometric approach, whereby each article is viewed as a knowledge resource. The relationships of various knowledge resources represent knowledge networks: reference co-citation, keyword co-occurrence, and discipline co-occurrence. These knowledge networks can be analyzed along the 3 dimensions of references, disciplines, and keywords using similarity-based clustering [31,32]. This combination of reference, keyword, and discipline networks represents a knowledge kernel, which is a three-dimensional space depicting the overall knowledge network of a research field (Figure 3). As such, the 3 knowledge networks present the evolution of a knowledge kernel along the 3 dimensions of references, disciplines, and keywords. Taken together, the 3 knowledge networks represent the knowledge evolution of a knowledge kernel. This approach is referred to as KEA. Besides, the burst detection technique was employed to identify emerging research hotspots.

Figure 3. 3D attributions of knowledge kernel.



An article typically cites and is cited by many others. To identify the interrelationships between articles, reference co-citation analysis is commonly used. Co-citation analysis can only categorize part of the cited literature in a research field, so keyword co-occurrence and discipline co-occurrence techniques were also used to reveal information on other key topics. These 3 techniques can help analyze the dynamics of a research field over time and are discussed in detail below.

Reference Co-citation Network

Small [33] defined co-citation as "the frequency with which two items of earlier literature are cited together by the later literature". The reference co-citation network was generated with a threshold of 4 or more co-citations [34], and the networks were divided into several clusters, with each network being labeled by terms extracted from the titles of the most representative citing articles [35]. This analysis shows how PHD research focus changes over time.

Keywords Co-occurrence Network

A list of predefined keywords represents the core idea of an article. Keyword co-occurrence refers to the statistical correlation between keywords that appear in the same article. A keyword co-occurrence network links keywords listed in the same article and presents the relationships between these keywords as a network map. The shortest distance between any 2 keywords that are not linked directly is viewed as the closeness of the 2 words [34]. The cluster formed by closely linked keywords represents a key subject domain of a research field. The burst detection algorithm shows how keywords emerge through frequency analysis to signify the most active PHD research hotspots over time [36].

Disciplines Co-occurrence Network

In this technique, each scientific article is assigned to 1 or more disciplines to calculate the statistical correlation between disciplines. When an article is assigned to 2 disciplines, these disciplines are related, and related disciplines combine to form a discipline co-occurrence network [37]. A burst detection algorithm can be used to detect the most active disciplines in PHD articles [36,38].

In this study, we used CiteSpace 5.2.R2, a bibliometric tool to analyze PHD articles [39].

Results

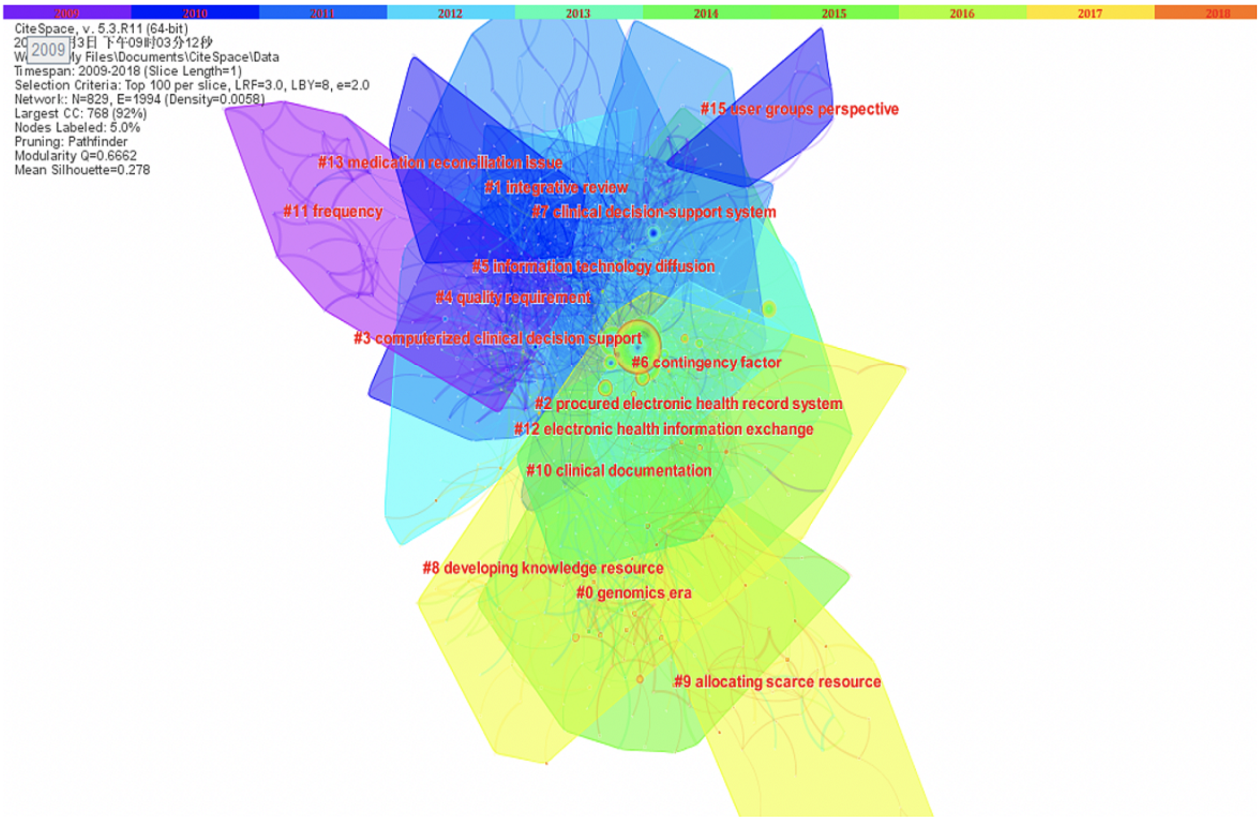
In the following sections, we present the KEA of references, disciplines, and keywords in the published PHD research.

Reference Co-citation Network

We constructed a co-citation network of the top 100 most cited articles each year from 2009 to 2018. Clustering was performed using the log-likelihood ratio method. The analysis identified 15 major clusters. Silhouette values ≥0.7 indicate high similarity among articles in the same cluster, while modularity Q values ≥0.6662 indicate high differences between clusters [34].

Figure 4 shows the evolution trajectory of the PHD knowledge kernel based on the reference co-citation network. The colored bars at the top of the figure represent different years. The corresponding colored curves represent co-citations occurring in that year. The size of a node depicted with the citation “tree rings” represents the number of times an article was cited [34]. The networks are further decomposed into clusters as tightly coupled references. Each cluster is labeled using terms extracted from noun phrases in titles.

Figure 4. Co-citation clusters of references (Modularity Q=0.6662, Mean Sihouette=0.278, Selection Criteria=Top 100 per slice).


2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
CiteSpace, v. 5.3.R11 (64-bit)
20 2009 月3日 下午09时03分12秒
W 2009 ly Files\Documents\CiteSpace\Data
Timespan: 2009-2018 (Slice Length=1)
Selection Criteria: Top 100 per slice, LRF=3.0, LBY=8, e=2.0
Network: N=829, E=1994 (Density=0.0058)
Largest CC: 768 (92%)
Nodes Labeled: 5.0%
Pruning: Pathfinder
Modularity Q=0.6662
Mean Silhouette=0.278
#15 user groups perspective
#13 medication reconciliation issue
#1 integrative review
#11 frequency
#7 clinical decision-support system
#5 information technology diffusion
#4 quality requirement
#3 computerized clinical decision support
#6 contingency factor
#2 procured electronic health record system
#12 electronic health information exchange
#10 clinical documentation
#8 developing knowledge resource
#0 genomics era
#9 allocating scarce resource

From Figure 4, we can see that the most popular PHD research topics changed over time. Before 2013, knowledge clusters such as clusters 3 (clinical decision support), 5 (information technology diffusion), and 2 (EHR system) mainly focused on medicine and technology. From 2013 onward, the focus shifted to health care resource allocation, such as clusters 8 and 9, focusing on developing knowledge resources and allocating scarce resources. A closer examination of clusters 8 and 9 can be found in Multimedia Appendix 2. It lists articles with coverage ≥9%, which represents the percentage of members in each cluster that articles cite. To some extent, these articles are the most representative articles of each cluster. For example, the articles focusing on developing knowledge resources for precision medicine [40], use of EHRs for clinical decision [41,42], and review of an integrated clinical decision support system [43] are the most representative articles of cluster 8 (developing knowledge resource). Likewise, the articles focusing on scarce resource allocating for heart disease [44], a population-level EHR cohort study [45], and data science application in critical care [46] are the most representative articles of cluster 9 (allocating scarce resource). The major clusters are described in detail in Table 3.

Table 3. Description of co-citation clusters.[a]

Mean year[b]	Cluster ID	Size[c]	Silhouette	Label (LLR[d])
2003	15	10	0.984	User groups perspective
2005	3	69	0.805	Clinical decision support
2005	5	62	0.815	Information technology diffusion
2005	10	31	0.777	Clinical documentation
2006	1	72	0.872	Integrative review
2006	13	18	0.947	Medication reconciliation issue
2007	4	64	0.838	Quality requirement
2007	6	61	0.809	Contingency factor
2010	2	70	0.804	EHR[e] systems
2010	7	61	0.847	Clinical decision support system
2010	12	30	0.833	Electronic health information exchange
2011	0	95	0.849	Genomic era
2011	11	31	0.936	Frequency
2013	8	51	0.893	Developing knowledge resource
2013	9	43	0.950	Allocating scarce resource

[a] The connected components in cluster 14 are less than the default value (K=25), so CiteSpace did not report 14 [39].

[b] The average year of the articles in a cluster.

[c] The number of articles in each cluster.

[d] LLR: log-likelihood ratio.

[e] EHR: electronic health record.

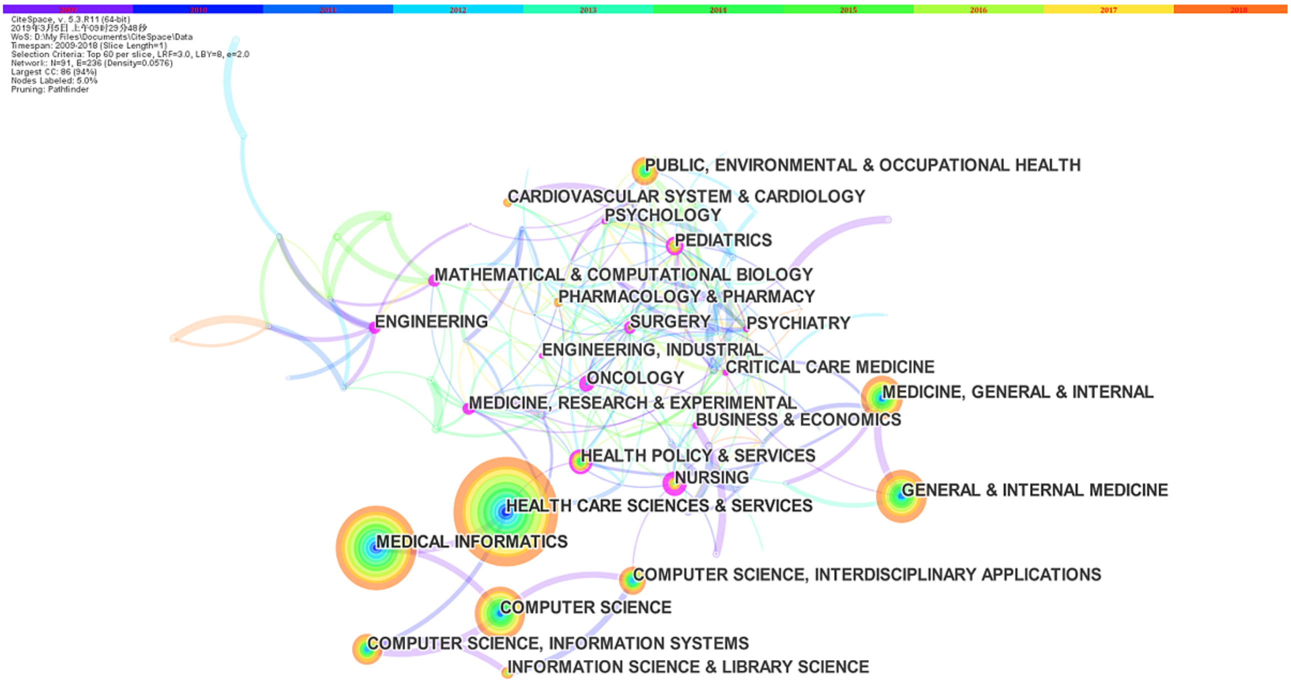
Keyword Co-occurrence Network

Multimedia Appendix 3 shows the keyword co-occurrence networks. Multimedia Appendix 4 shows the 56 keywords with the strongest burst out of 100 keywords that were frequently cited each year between 2009 and 2018. This was performed using the "burst detection" function in CiteSpace. In 2009, keywords with the strongest burst mainly focused on basic PHD issues (eg, privacy, physician order entry, and standard) and medical issues (eg, diabetes mellitus, heart disease, blood pressure). Between 2010 and 2013, the keywords clinical information system, database, ambulatory care, personal health record had the strongest burst. Since 2013, burst keywords included attitude and satisfaction, implying that PHD research evolved from focusing on technology- and medicine-centered perspectives to focusing on human-centered perspectives. The most recent burst keywords (eg, readmission, emergency department, usability) appear to be likely PHD research hotspots, focusing on efficiency and quality of health care resources.

Discipline Co-occurrence Network

Figure 5 shows the evolution trajectory of the PHD knowledge kernel based on discipline co-occurrence networks. The size of a node represents the number of articles in a specific discipline. The links between nodes show interdisciplinary collaborations. The colors of links show when a connection was made for the first time. The tree rings represent the co-occurrence history of a discipline. The color of a circle ring denotes the time of corresponding citations. The largest node was health care sciences, followed by medical informatics, general and internal medicine, and computer science, indicating that these are the mainstream disciplines in PHD studies. Nodes with high betweenness centrality (indicated by the purple rim) [35], including health policy and services, psychology, and business and economics, may be pivotal to the paradigm shift of PHD research.

Figure 5. Disciplines co-occurrence network (2009–2018) (Pruning=Pathfinder, Node=91, Density=0.0576, Selection Criteria= Top 60 per slice).


2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
CiteSpace, v. 5.3.R11 (64-bit)
2019年3月5日 上午09时29分48秒
WoS: D:\My Files\Documents\CiteSpace\Data
Timespan: 2009-2018 (Slice Length=1)
Selection Criteria: Top 60 per slice, LRF=3.0, LBY=8, e=2.0
Network: N=91, E=236 (Density=0.0576)
Largest CC: 86 (94%)
Nodes Labeled: 5.0%
Pruning: Pathfinder
PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH
CARDIOVASCULAR SYSTEM & CARDIOLOGY
PSYCHOLOGY
PEDIATRICS
MATHEMATICAL & COMPUTATIONAL BIOLOGY
PHARMACOLOGY & PHARMACY
ENGINEERING
SURGERY
PSYCHIATRY
ENGINEERING, INDUSTRIAL
CRITICAL CARE MEDICINE
ONCOLOGY
MEDICINE, GENERAL & INTERNAL
MEDICINE, RESEARCH & EXPERIMENTAL
BUSINESS & ECONOMICS
HEALTH POLICY & SERVICES
NURSING
GENERAL & INTERNAL MEDICINE
HEALTH CARE SCIENCES & SERVICES
MEDICAL INFORMATICS
COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS
COMPUTER SCIENCE
COMPUTER SCIENCE, INFORMATION SYSTEMS
INFORMATION SCIENCE & LIBRARY SCIENCE

Disciplines with the strongest burst are shown in Multimedia Appendix 5. Management was at the top of the list with a burst strength of 4.4358 between 2009 and 2011. Before 2013, most research hotspots, such as biochemistry and molecular biology, dentistry, and oral surgery and medicine, were medicine and biology disciplines. From 2013 to 2016, various technologies were combined into PHD research, including computer science (artificial intelligence [AI]) and medical laboratory technology. Since 2016, substance abuse and psychology disciplines have become more popular in PHD research. Psychology had a relatively high burst strength (6.5215) and appears to be a significant discipline for future research. Social sciences also had a strong burst (4.8105) for the longest time, making it a central focus of PHD research.

Discussion

Principal Findings

To the best of our knowledge, this is the first systematic review to show how PHD research has evolved and which research areas are potential hotspots. We examined the PHD knowledge kernel in 3 networks—reference co-citation, keyword co-occurrence, and discipline co-occurrence—to unveil how knowledge clusters evolved, which subjects are key, and which disciplines are being studied in PHD research. The proposed KEA framework can be extended to other similar interdisciplinary research areas. This is also the first study to focus on all types of PHD, including EMR, EHR, and PHR; previous reviews have focused on 1 type of health data. Lastly, this study included a large number of articles (8281 articles) and was not restricted to specific research questions or research types.

The reference co-citation network revealed that PHD research mainly focused on medicine and technology issues (eg, clinical decision systems) before 2013. From 2013 onward, the focus shifted toward developing knowledge resources and allocating scarce health care resources. The results also suggest that from 2013 onward, research communities have been actively seeking methods to make meaningful use of PHD. The overall trend of EHR research mirrors the previous finding of Qian et al [29] that EHR research has evolved from the adoption of EHR to higher-level application and integration of EHR. A well-cited publication is one from Blumenthal and Tavenner [47], which briefs about how EHR benefits patients and caregivers. Other studies have explored the benefits of clinical decision support systems based on EHR as well as barriers to using EHR [18,48,49]. Moreover, the application of PHD in medical research has evolved with technological development. At first, EHR-based clinical decision support systems were mainly used to diagnose and treat specific diseases such as diabetes and heart disease [50]. Later on, more effort was made to develop and systematically incorporate health care data to improve genomics and precision medicine [40].

The reference co-citation network also showed that the most active PHD knowledge cluster is developing knowledge resources and allocating scarce resources. This is supported by the analysis of the keywords that shows PHD studies focusing on emergency health care typically involve the application of the latest knowledge and use of scarce resources [44,46]. The co-citation analysis also demonstrated that the focus of PHD research is moving away from improving treatment decisions to optimizing resource distribution to different groups. This pertains to the allocative value of value-based health care, which aims to equalize resource allocation and improve health care outcomes between different groups [51], thereby improving health care services. In line with the aforementioned, AI applications have proven to be effective, especially in image interpretation [52,53] and diagnosis [54,55]. During the COVID-19 pandemic, the AI system played an important role in rapid early detection and diagnosis [56,57]. AI also can help in optimizing treatment regimens, prevention strategies, and allocation of scarce health resources to narrow down the inequality in health care, especially in resource-poor settings attributed to the shortage of human resources and medical devices [58]. These findings suggest that it is necessary to improve the equity in health resource allocation. Notably, value-based health care and AI applications should be given more attention.

The keyword co-occurrence analysis revealed that technical issues such as data privacy, data standardization, data quality, and interoperability between different information systems were studied first, which makes sense as these are initial and critical steps for using PHD. Data quality is important because it ensures the accuracy of the information provided. Interoperability between information systems is also important for information exchange. Privacy protection encourages people to share their health data. The importance of these technical issues has been well supported by other systematic reviews [6,16,59,60]. These findings suggest that adequate processes for collecting PHD are prerequisites for the utilization of PHD and more effort should be put in place at the initial stage of data standardization and optimizing interoperability.

The bursts in topics related to psychology and human factors (eg, attitude, satisfaction, education) indicate the switch from technology-centric issues to more human-centric issues in PHD studies. The study by Blumenthal [4] and Meier [61] showed that meaningful use of PHD requires more attention to education, attitude, and satisfaction of all the stakeholders. Patient satisfaction is critical for successful health care and depends on quality, communication, and interpersonal interactions with health care providers [62]. Moreover, as AI-based technology including machine learning, natural language processing, and artificial networks is integrated into health care more deeply, the “black box” algorithms have raised concerns about technology liability as well as patient and clinician trust [57,63]. Further research on regulatory issues and governance of PHD is therefore recommended.

Our findings also supported the unified theory of acceptance and use of technology [64], which comprises 4 key elements (ie, performance expectancy, effort expectancy, social influence, and facilitating conditions) that influence how we use technology. These elements are related to how humans interact with technology and make sure that technology creates value for patients, physicians, and administrators, which eventually improves satisfaction. As technologies (eg, AI, internet of things) are now widely used in health care, these issues are gaining more importance [65]. The aforementioned human factors reflect the notion of “personalized value,” another dimension of value-based health care, which emphasizes that every patient should be fully informed about the benefits and risks of treatments [66]. Therefore, the technology developer and health care institutions need to consider these human factors for the effective adoption of PHD-related technology.

The discipline co-occurrence analysis revealed the evolution of PHD research over various disciplines over the past 10 years with a more recent focus on computer science, including AI, machine learning, and deep learning. This agrees with the notion that computer science can increase the value of PHD [11,67]. Yin et al [11] reviewed the effectiveness of machine learning technology in personal health investigations based on online PHD [11], and Payrovnaziri et al [68] conducted a review of AI models that use EHR data. Hou et al [36] pointed out that AI could be used not only as a screening tool to interpret radiology images but also to interpret these images with greater consistency than humans can. Moreover, AI-based technology has the potential to improve efforts toward precision medicine. Tran et al [69] stated that AI technology leverages individual health data and data science to enhance prognosis, diagnosis, and rehabilitation. Regardless of the specific technique or function, the general aim of these technologies is to ease the shortage of human and device resources and optimize the allocation of scarce health care resources. This notion of effective technology application within PHD research presents another dimension of value-based health care known as “technical value” [70]. These findings suggest that all stakeholders should be educated about AI technology to promote value generation through PHD.

Overall, our results indicate that health data analytics should go beyond improving decision-making processes to providing better results for populations [71]. In line with this, PHD research is transitioning toward a more human-centric approach with a new focus on value-based health care: “allocative value,” “technology value,” and “personalized value” [70]. These findings indicate that PHD research has the potential to meet the triple aims of value-based health care in the future.

Limitations

There are some limitations to this review. First, the scope of the data is limited by the source (the Web of Science) and the search items used. This study did not use “sentiments,” “emotions,” and “social media data” for data set search, as they are not well-defined terminologies or keywords, which might bias the data set. An iterative query refinement would improve the quality of the data set, although the search strategy adequately met the study purpose. Second, the results present an overview of how structure and knowledge have evolved in PHD research; however, details on more specific research topics are lacking. Researchers need to explore this in detail using additional methods and other scholarly publications. Topics to address include health care inequity and cost-effective health care through joint efforts of professional health care networks and patient networks [72]. Third, the co-citation networks rely on citation relationships between articles. While some citations reflect a strong connectedness, other citations might reflect a weaker connectedness. Further research is needed to distinguish between different kinds of citations.

Conclusions

This study used KEA to review the evolution of PHD research and identify research hotspots. The results show that the focus of PHD research has evolved from medicine centric to technology centric, to human centric since 2009. PHD is applied to optimize the allocation of scarce health care resources and to improve the quality and efficiency of health care services. Moreover, AI-based technology is becoming more relevant in PHD research, and that this technology may be used to ease the shortage of human and device resources. Furthermore, PHD research is now paying more attention to topics related to psychology and human factors, such as education, attitude, and satisfaction of stakeholders. These findings indicate opportunities for interdisciplinary cooperation in several PHD research areas: (1) AI applications for PHD; (2) regulatory issues and governance of PHD; (3) education of all stakeholders about AI technology; (4) value-based health care including “allocative value,” “technology value,” and “personalized value.”

Acknowledgments

This work was funded by the National Natural Science Foundation of China (No. 71671039).

Authors' Contributions

All authors have made a substantial intellectual contribution to this study. JG, VS, QK, and LZ designed the study together. JG performed the database searches and data analysis. JG wrote the first draft of the manuscript with the support of VS, QK, and LZ. QK and VS commented on the draft and added to the revisions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The annual number of published articles on personal health data in Web of Science (2009–2018).
[DOCX File , 52 KB - medinform_v9i11e31142_app1.docx]

Multimedia Appendix 2
A list of articles that contributed to clusters #8 and #9.
[DOCX File , 19 KB - medinform_v9i11e31142_app2.docx]

Multimedia Appendix 3
Keywords co-occurrence network.
[DOCX File , 341 KB - medinform_v9i11e31142_app3.docx]

Multimedia Appendix 4
The 56 keywords with the strongest burst (2009-2018).
[DOCX File , 23 KB - medinform_v9i11e31142_app4.docx]

Multimedia Appendix 5
The 15 disciplines with the strongest burst (2009–2018).
[DOCX File , 18 KB - medinform_v9i11e31142_app5.docx]

References

1. Meier CA, Fitzgerald MC, Smith JM. eHealth: extending, enhancing, and evolving health care. Annu Rev Biomed Eng 2013;15:359-382. [doi: 10.1146/annurev-bioeng-071812-152350] [Medline: 23683088]
2. Karampela M, Ouhbi S, Isomursu M. Personal health data: A systematic mapping study. Int J Med Inform 2018 Oct;118:86-98. [doi: 10.1016/j.ijmedinf.2018.08.006] [Medline: 30153927]
3. Bietz MJ, Bloss CS, Calvert S, Godino JG, Gregory J, Claffey MP, et al. Opportunities and challenges in the use of personal health data for health research. J Am Med Inform Assoc 2016 Apr;23(e1):e42-e48 [FREE Full text] [doi: 10.1093/jamia/ocv118] [Medline: 26335984]
4. Blumenthal D. Launching HITECH. N Engl J Med 2010 Feb 04;362(5):382-385. [doi: 10.1056/NEJMp0912825] [Medline: 20042745]
5. Liu L, Stroulia E, Nikolaidis I, Miguel-Cruz A, Rios Rincon A. Smart homes and home health monitoring technologies for older adults: A systematic review. Int J Med Inform 2016 Jul;91:44-59. [doi: 10.1016/j.ijmedinf.2016.04.007] [Medline: 27185508]
6. 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]

7. Betancourt JR. In pursuit of high-value healthcare: the case for improving quality and achieving equity in a time of healthcare transformation. Front Health Serv Manage 2014;30(3):16-31. [Medline: 25291891]
8. Miller RH, Sim I. Physicians' use of electronic medical records: barriers and solutions. Health Aff (Millwood) 2004;23(2):116-126. [doi: 10.1377/hlthaff.23.2.116] [Medline: 15046136]
9. Madden JM, Lakoma MD, Rusinak D, Lu CY, Soumerai SB. Missing clinical and behavioral health data in a large electronic health record (EHR) system. J Am Med Inform Assoc 2016 Nov;23(6):1143-1149 [FREE Full text] [doi: 10.1093/jamia/ocw021] [Medline: 27079506]
10. Kahn JS, Aulakh V, Bosworth A. What it takes: characteristics of the ideal personal health record. Health Aff (Millwood) 2009;28(2):369-376. [doi: 10.1377/hlthaff.28.2.369] [Medline: 19275992]
11. Yin Z, Sulieman LM, Malin BA. A systematic literature review of machine learning in online personal health data. J Am Med Inform Assoc 2019 Jun 01;26(6):561-576 [FREE Full text] [doi: 10.1093/jamia/ocz009] [Medline: 30908576]
12. Segen JC. McGraw Hill Concise Medical Dictionary of Modern Medicine. New York, NY: McGraw-Hill Companies, Inc; 2002:353-354.
13. Plastiras P, O'Sullivan D. Exchanging personal health data with electronic health records: A standardized information model for patient generated health data and observations of daily living. Int J Med Inform 2018 Dec;120:116-125. [doi: 10.1016/j.ijmedinf.2018.10.006] [Medline: 30409336]
14. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
15. Hassanalieragh M, Page A, Soyata T, Sharma G, Aktas M, Mateos G, et al. Health Monitoring and Management Using Internet-of-Things (IoT) Sensing with Cloud-Based Processing: Opportunities and Challenges. New York, NY: IEEE; 2015 Jun Presented at: SCC '15: Proceedings of the 2015 IEEE International Conference on Services Computing; June 27, 2015 to July 2, 2015; New York, NY p. 285-292. [doi: 10.1109/scc.2015.47]
16. Fernández-Alemán JL, Señor IC, Lozoya PÁO, Toval A. Security and privacy in electronic health records: a systematic literature review. J Biomed Inform 2013 Jun;46(3):541-562 [FREE Full text] [doi: 10.1016/j.jbi.2012.12.003] [Medline: 23305810]
17. Archer N, Fevrier-Thomas U, Lokker C, McKibbon KA, Straus SE. Personal health records: a scoping review. J Am Med Inform Assoc 2011;18(4):515-522 [FREE Full text] [doi: 10.1136/amiajnl-2011-000105] [Medline: 21672914]
18. Kruse CS, Kothman K, Anerobi K, Abanaka L. Adoption Factors of the Electronic Health Record: A Systematic Review. JMIR Med Inform 2016 Jun 01;4(2):e19 [FREE Full text] [doi: 10.2196/medinform.5525] [Medline: 27251559]
19. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health 2014 Nov 05;14:1144 [FREE Full text] [doi: 10.1186/1471-2458-14-1144] [Medline: 25377061]
20. Maher NA, Senders JT, Hulsbergen AFC, Lamba N, Parker M, Onnela J, et al. Passive data collection and use in healthcare: A systematic review of ethical issues. Int J Med Inform 2019 Sep;129:242-247. [doi: 10.1016/j.ijmedinf.2019.06.015] [Medline: 31445262]
21. Abd-Alrazaq AA, Bewick BM, Farragher T, Gardner P. Factors that affect the use of electronic personal health records among patients: A systematic review. Int J Med Inform 2019 Jun;126:164-175. [doi: 10.1016/j.ijmedinf.2019.03.014] [Medline: 31029258]
22. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: A systematic review. Int J Med Inform 2018 Dec;114:57-65. [doi: 10.1016/j.ijmedinf.2018.03.013] [Medline: 29673604]
23. Guo H, Dai T, Hu H. Research Status, Hotspots and Trends of Electronic Health Records: Bibliometric Analysis Based on PubMed Database. China Digit Med 2011;8:e1. [doi: 10.3969/j.issn.1673-7571.2011.08.007]
24. Liang Z, Yong L, Rui Z, Tingting H, Jialin L. Bibliometrics on Electronic Health Records of Web of Science. Chinese J Evidence-Based Med 2013;13(11):1307-1312. [doi: 10.7507/1672-2531.20130225]
25. Ruixian Y, Yu C, Haoyu L. Bibliometric Analysis on Researches of Electronic Medical Records in China. Inf Res 2015;11(217):18-21. [doi: 10.3969/j.issn.1005-8095.2015.11.005]
26. Lin D, Liu J, Zhang R, Li Y, Huang T. [Application Status of Evaluation Methodology of Electronic Medical Record: Evaluation of Bibliometric Analysis]. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi 2015 Apr;32(2):350-356. [Medline: 26211253]
27. Wen H, Ho Y, Jian W, Li H, Hsu YE. Scientific production of electronic health record research, 1991-2005. Comput Methods Programs Biomed 2007 May;86(2):191-196. [doi: 10.1016/j.cmpb.2007.02.002] [Medline: 17400328]
28. Wang Y, Zhao Y, Dang W, Zheng J, Dong H. The Evolution of Publication Hotspots in Electronic Health Records from 1957 to 2016 and Differences Among Six Countries. Big Data 2020 Apr 01;8(2):89-106. [doi: 10.1089/big.2019.0024] [Medline: 32319801]
29. Qian Y, Ni Z, Gui W, Liu Y. Exploring the Landscape, Hot Topics, and Trends of Electronic Health Records Literature with Topics Detection and Evolution Analysis. IJCIS 2021;14(1):744. [doi: 10.2991/ijcis.d.210203.006]
30. Zhenni N, Yuxing Q. The Status, Hot Topics in the Field of Electronic Health Records: A Literature Review Based on Lda2vec. New York, NY: Association for Computing Machinery; 2020 Presented at: JCDL '20: Proceedings of the

ACM/IEEE Joint Conference on Digital Libraries in 2020; August 1-5, 2020; Virtual Event p. 479-480. [doi: 10.1145/3383583.3398572]

31. Liu L, Mei S. Visualizing the GVC research: a co-occurrence network based bibliometric analysis. Scientometrics 2016 Aug 20;109(2):953-977. [doi: 10.1007/s11192-016-2100-5]
32. Boyack KW, Small H, Klavans R. Improving the accuracy of co-citation clustering using full text. J Am Soc Inf Sci Tec 2013 Jul 19;64(9):1759-1767. [doi: 10.1002/asi.22896]
33. Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents. J. Am. Soc. Inf. Sci 1973 Jul;24(4):265-269. [doi: 10.1002/asi.4630240406]
34. Chen C, Dubin R, Kim MC. Emerging trends and new developments in regenerative medicine: a scientometric update (2000 - 2014). Expert Opin Biol Ther 2014 Sep;14(9):1295-1317. [doi: 10.1517/14712598.2014.920813] [Medline: 25077605]
35. Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. J. Am. Soc. Inf. Sci 2006 Feb 01;57(3):359-377. [doi: 10.1002/asi.20317]
36. Hou J, Yang X, Chen C. Emerging trends and new developments in information science: a document co-citation analysis (2009–2016). Scientometrics 2018 Mar 7;115(2):869-892. [doi: 10.1007/s11192-018-2695-9]
37. Liu Z, Yin Y, Liu W, Dunford M. Visualizing the intellectual structure and evolution of innovation systems research: a bibliometric analysis. Scientometrics 2015 Jan 22;103(1):135-158. [doi: 10.1007/s11192-014-1517-y]
38. Kleinberg J. Bursty and hierarchical structure in streams. Data Min Knowl Discov 2003;7(4):373-397. [doi: 10.1145/775047.775061]
39. Chen C. CiteSpace: A Practical Guide for Mapping Scientific Literature. Hauppauge, NY: Nova Science Publishers; 2016:26-39.
40. Hoffman JM, Dunnenberger HM, Kevin Hicks J, Caudle KE, Whirl Carrillo M, Freimuth RR, et al. Developing knowledge resources to support precision medicine: principles from the Clinical Pharmacogenetics Implementation Consortium (CPIC). J Am Med Inform Assoc 2016 Jul;23(4):796-801 [FREE Full text] [doi: 10.1093/jamia/ocw027] [Medline: 27026620]
41. Hicks JK, Dunnenberger HM, Gumpper KF, Haidar CE, Hoffman JM. Integrating pharmacogenomics into electronic health records with clinical decision support. Am J Health Syst Pharm 2016 Dec 01;73(23):1967-1976 [FREE Full text] [doi: 10.2146/ajhp160030] [Medline: 27864204]
42. Caraballo P, Bielinski S, St Sauver JL, Weinshilboum R. Electronic Medical Record-Integrated Pharmacogenomics and Related Clinical Decision Support Concepts. Clin Pharmacol Ther 2017 Aug 26;102(2):254-264. [doi: 10.1002/cpt.707] [Medline: 28390138]
43. Hinderer M, Boeker M, Wagner SA, Lablans M, Newe S, Hülsemann JL, et al. Integrating clinical decision support systems for pharmacogenomic testing into clinical routine - a scoping review of designs of user-system interactions in recent system development. BMC Med Inform Decis Mak 2017 Jun 06;17(1):81 [FREE Full text] [doi: 10.1186/s12911-017-0480-y] [Medline: 28587608]
44. Amarasingham R, Patel PC, Toto K, Nelson LL, Swanson TS, Moore BJ, et al. Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study. BMJ Qual Saf 2013 Dec;22(12):998-1005 [FREE Full text] [doi: 10.1136/bmjqs-2013-001901] [Medline: 23904506]
45. Koudstaal S, Pujades-Rodriguez M, Denaxas S, Gho JMIH, Shah AD, Yu N, et al. Prognostic burden of heart failure recorded in primary care, acute hospital admissions, or both: a population-based linked electronic health record cohort study in 2.1 million people. Eur J Heart Fail 2017 Sep;19(9):1119-1127 [FREE Full text] [doi: 10.1002/ejhf.709] [Medline: 28008698]
46. Sanchez-Pinto LN, Luo Y, Churpek MM. Big Data and Data Science in Critical Care. Chest 2018 Nov;154(5):1239-1248 [FREE Full text] [doi: 10.1016/j.chest.2018.04.037] [Medline: 29752973]
47. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010 Aug 05;363(6):501-504. [doi: 10.1056/NEJMp1006114] [Medline: 20647183]
48. Kruse CS, Kristof C, Jones B, Mitchell E, Martinez A. Barriers to Electronic Health Record Adoption: a Systematic Literature Review. J Med Syst 2016 Dec;40(12):252 [FREE Full text] [doi: 10.1007/s10916-016-0628-9] [Medline: 27714560]
49. Middleton B, Bloomrosen M, Dente MA, Hashmat B, Koppel R, Overhage JM, American Medical Informatics Association. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. J Am Med Inform Assoc 2013 Jun;20(e1):e2-e8 [FREE Full text] [doi: 10.1136/amiajnl-2012-001458] [Medline: 23355463]
50. 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 03;157(1):29-43 [FREE Full text] [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]
51. Gray M. Value based healthcare. BMJ 2017 Jan 27;356:j437. [doi: 10.1136/bmj.j437] [Medline: 28130219]
52. Lakhani P, Prater AB, Hutson RK, Andriole KP, Dreyer KJ, Morey J, et al. Machine Learning in Radiology: Applications Beyond Image Interpretation. J Am Coll Radiol 2018 Feb;15(2):350-359. [doi: 10.1016/j.jacr.2017.09.044] [Medline: 29158061]

53. Nichols JA, Herbert Chan HW, Baker MAB. Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophys Rev 2019 Feb;11(1):111-118 [FREE Full text] [Medline: 30182201]
54. Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. Cancer Lett 2020 Feb 28;471:61-71. [doi: 10.1016/j.canlet.2019.12.007] [Medline: 31830558]
55. Szolovits P, Patil RS, Schwartz WB. Artificial intelligence in medical diagnosis. Ann Intern Med 1988 Jan 01;108(1):80-87. [doi: 10.7326/0003-4819-108-1-80] [Medline: 3276267]
56. Jin C, Chen W, Cao Y, Xu Z, Tan Z, Zhang X, et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat Commun 2020 Oct 09;11(1):5088 [FREE Full text] [doi: 10.1038/s41467-020-18685-1] [Medline: 33037212]
57. Vaishya R, Javaid M, Khan IH, Haleem A. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr 2020;14(4):337-339 [FREE Full text] [doi: 10.1016/j.dsx.2020.04.012] [Medline: 32305024]
58. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Glob Health 2018;3(4):e000798 [FREE Full text] [doi: 10.1136/bmjgh-2018-000798] [Medline: 30233828]
59. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
60. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. Int J Med Inform 2008 May;77(5):291-304. [doi: 10.1016/j.ijmedinf.2007.09.001] [Medline: 17951106]
61. Meier C. A role for data: an observation on empowering stakeholders. Am J Prev Med 2013 Jan;44(1 Suppl 1):S5-11 [FREE Full text] [doi: 10.1016/j.amepre.2012.09.018] [Medline: 23195168]
62. Venkatesh V, Zhang X, Sykes TA. “Doctors Do Too Little Technology”: A Longitudinal Field Study of an Electronic Healthcare System Implementation. Information Systems Research 2011 Sep;22(3):523-546. [doi: 10.1287/isre.1110.0383]
63. Rigby M. Ethical dimensions of using artificial intelligence in health care. AMA J Ethics American Medical Association 2019;21(2):121-124. [doi: 10.1001/amajethics.2019.121]
64. Venkatesh V, Thong JYL, Xu X. Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. MIS Quarterly 2012;36(1):157. [doi: 10.2307/41410412]
65. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36 [FREE Full text] [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
66. Gray M, Jani A. Promoting Triple Value Healthcare in Countries with Universal Healthcare. Healthc Pap 2016;15(3):42-48. [Medline: 27009586]
67. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc 2018 Oct 01;25(10):1419-1428 [FREE Full text] [doi: 10.1093/jamia/ocy068] [Medline: 29893864]
68. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J Am Med Inform Assoc 2020 Jul 01;27(7):1173-1185 [FREE Full text] [doi: 10.1093/jamia/ocaa053] [Medline: 32417928]
69. Tran BX, Nghiem S, Sahin O, Vu TM, Ha GH, Vu GT, et al. Modeling Research Topics for Artificial Intelligence Applications in Medicine: Latent Dirichlet Allocation Application Study. J Med Internet Res 2019 Nov 01;21(11):e15511 [FREE Full text] [doi: 10.2196/15511] [Medline: 31682577]
70. Kerr DJ, Jani A, Gray SM. Strategies for Sustainable Cancer Care. Am Soc Clin Oncol Educ Book 2016;35:e11-e15 [FREE Full text] [doi: 10.1200/EDBK_156142] [Medline: 27249712]
71. van de Klundert J. Healthcare analytics: big data, little evidence. In: Tutorials in Operations Research. Catonsville, MD: The Institute for Operations Research and the Management Sciences (INFORMS); Nov 2016:1-22.
72. Patrício L, de Pinho NF, Teixeira JG, Fisk RP. Service Design for Value Networks: Enabling Value Cocreation Interactions in Healthcare. Service Science 2018 Mar;10(1):76-97. [doi: 10.1287/serv.2017.0201]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
EMR: electronic medical record
KEA: knowledge evolution analysis
PHD: personal health data
PHR: personal health record
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by G Eysenbach; submitted 11.06.21; peer-reviewed by Q Chen; comments to author 05.07.21; revised version received 17.08.21; accepted 17.09.21; published 01.11.21.

Please cite as:
Gong J, Sihag V, Kong Q, Zhao L
Visualizing Knowledge Evolution Trends and Research Hotspots of Personal Health Data Research: Bibliometric Analysis
JMIR Med Inform 2021;9(11):e31142
URL: https://medinform.jmir.org/2021/11/e31142
doi:10.2196/31142
PMID:34723823

Original Paper

Event Prediction Model Considering Time and Input Error Using Electronic Medical Records in the Intensive Care Unit: Retrospective Study

MinDong Sung[1*], MD; Sangchul Hahn[2*], MSc; Chang Hoon Han[3*], MD; Jung Mo Lee[3], MD; Jayoung Lee[2], MD, PhD; Jinkyu Yoo[2], MSc; Jay Heo[4], MSc; Young Sam Kim[5], MD, PhD; Kyung Soo Chung[5], MD, PhD

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea

[2]AITRICS. Inc, Seoul, Republic of Korea

[3]Division of Pulmonology, Department of Internal Medicine, National Health Insurance Service Ilsan Hospital, Goyang-si, Republic of Korea

[4]Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

[5]Division of Pulmonology, Department of Internal Medicine, Yonsei University Health System, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Kyung Soo Chung, MD, PhD
Division of Pulmonology
Department of Internal Medicine
Yonsei University Health System
50-1 Yonsei-ro, Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 2227 8308
Email: chungks78@gmail.com

Abstract

Background: In the era of artificial intelligence, event prediction models are abundant. However, considering the limitation of the electronic medical record–based model, including the temporally skewed prediction and the record itself, these models could be delayed or could yield errors.

Objective: In this study, we aim to develop multiple event prediction models in intensive care units to overcome their temporal skewness and evaluate their robustness against delayed and erroneous input.

Methods: A total of 21,738 patients were included in the development cohort. Three events—death, sepsis, and acute kidney injury—were predicted. To overcome the temporal skewness, we developed three models for each event, which predicted the events in advance of three prespecified timepoints. Additionally, to evaluate the robustness against input error and delays, we added simulated errors and delayed input and calculated changes in the area under the receiver operating characteristic curve (AUROC) values.

Results: Most of the AUROC and area under the precision-recall curve values of each model were higher than those of the conventional scores, as well as other machine learning models previously used. In the error input experiment, except for our proposed model, an increase in the noise added to the model lowered the resulting AUROC value. However, the delayed input did not show the performance decreased in this experiment.

Conclusions: For a prediction model that was applicable in the real world, we considered not only performance but also temporal skewness, delayed input, and input error.

(JMIR Med Inform 2021;9(11):e26426) doi:10.2196/26426

KEYWORDS

machine learning; critical care; prediction model; intensive care unit; mortality; AKI; sepsis

Introduction

Since intensive care resources are always limited, and better resource allocation leads to better outcomes [1], conventional scores such as Acute Physiology And Chronic Health Evaluation (APACHE) [2], Simplified Acute Physiology Score (SAPS) [3], and Mortality Probability Models (MPMs) [4] have been used to predict the outcome of patients admitted to the intensive care unit (ICU). However, because the status of patients in the ICU changes rapidly, predicting adverse events and clinical complications, which are a major cause of mortality and poor outcomes, can buy some time to intervene and change the natural disease course [5,6]. Although conventional scores are widely used, these scores use only the features of patients at admission, and there have been many attempts to develop prediction models using time-series data.

With the increased use of electronic medical records (EMRs) [7] and artificial intelligence (AI), many AI models have been developed to predict events in the health care domain [8], and the intensive care domain is no exception. Additionally, the ICU generates many different kinds of frequently measured data. Thus, many models have been developed with a focus on ICU data [9-13].

Previous models were developed using retrospective EMR data. To apply these models in the real world, two points should be considered. The model should know more than whether an event will occur within the predicted time frame. In most studies, the distribution of event occurrence within the follow-up time is skewed to one side [5,9]. We defined this phenomenon as “temporal skewness,” which means more true-positive samples occur when the prediction time is getting closer to the actual event onset time. In particular, the performance metrics of a rapid response team are directly linked to a guarantee of temporal dependence regarding treatment intervention feasibility, similar to the 1-hour bundle suggested by sepsis treatment guidelines. Second, medical record data are often entered incorrectly, delayed, or even frequently missed in the field during patient care [14]. These errors should affect any real-time prediction model. Even if humans were replaced by an internet of medical things (IoMT) sensor [15], these sensors can generate noise in the data and transactions can be delayed. Thus, the model should be robust in consideration of these input errors.

Therefore, the prediction model using EMRs should achieve the following: (1) correction of temporal skewness and (2) robustness against delayed input and data input errors. Herein, we developed a novel prediction model using deep learning techniques that can be clinically applied to achieve the two abovementioned points.

Methods

Study Participants and the Development Cohort

We retrospectively enrolled adult patients who were admitted to the ICU from 2013 to 2017 at the Severance Hospital, a tertiary academic medical center in South Korea that includes medical, medicosurgical, neurological, cardiac surgery recovery, coronary care units, and has a total of 200 ICU beds. Patient information was anonymized by replacing the in-hospital patient ID with a surrogate key and shifting time-related information, such as birth date and chart input time, by randomly chosen periods before the analysis. The study was approved by the institutional review board of Severance Hospital, Yonsei University Health System, Seoul, Korea (IRB 4-2017-0939) and Ilsan Hospital (NHIMC 2018-06-004-001). All methods were performed following the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines.

Model Development

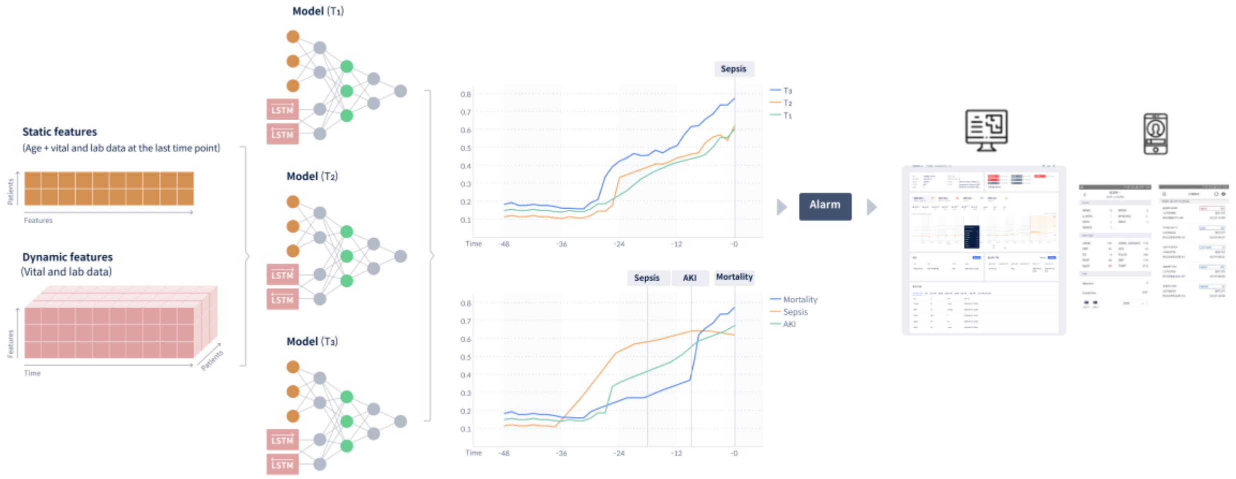
We developed prediction models for 3 events: mortality, sepsis, and acute kidney injury (AKI). These are considered major events in the ICU, and prediction of and interventions for these events will help change the clinical course of the patients. The model used 19 features: 6 vital signals, 11 laboratory tests, the Glasgow Coma Score (GCS), and age (see Multimedia Appendix 1). Two of the authors, who are intensive care specialists (KSC and CH), selected features that are widely and routinely used in general ICUs. We excluded any patients who were under the age of 18 years, who did not have at least one valid record with 1 of 5 vital signs (ie, pulse rate, systolic blood pressure, diastolic blood pressure, respiratory rate, and body temperature), and whose event time was after their ICU stay.

The events were identified by the following working definitions. *Mortality* was defined as an in-ICU death recorded in the EMR. According to the clinical surveillance definition [16], *sepsis* was defined as patients who had at least one concurrent acute organ dysfunction. Sepsis was indicated by the initiation of vasopressors or mechanical ventilation; elevated lactate level; or significant changes in the baseline creatinine level, bilirubin level, or platelet count within the 48 hours before or 24 hours after suspected serious infection. *Suspected serious infections* were defined by blood culture and sustained administration of new antibiotics. AKI, according to the Kidney Disease: Improving Global Outcomes (KDIGO) clinical practice guideline [17], was defined as follows: increase in serum creatinine level by 0.3 mg/dL within 48 hours, increase in serum creatinine level to 1.5 times the baseline level that was known or presumed to have occurred within the prior 7 days, or a decrease of 0.5 $mL \cdot kg^{-1} \cdot h^{-1}$ in the urine volume for 6 hours. The onset time of the AKI defined the time point at which the creatinine level was elevated.

Each prediction model was based on bidirectional long short-term memory (biLSTM) and designed as a binary classification model, which answers *yes* or *no* questions. The model used 2 types of data: (1) a dynamic feature, which is time-series data, and (2) a static feature. The sampling frequency of the dynamic feature was 1 hour. We used biLSTM for dynamic features and fully connected layers for static features. We connected the outputs from LSTM and fully connected layers and used them as an input for classification layers (Figure 1). The biLSTM layer has 20 hidden nodes. To train, we use Adam optimizer, a learning rate of 0.001, a batch size of 32, and maximum epochs of 300 (see Multimedia Appendix 2 for details). Additionally, to consider the time interval in which future events will occur, we set 3 future time points: T_1 (near future), T_2 (mid-term future), and T_3 (distant future). Considering the clinical circumstances and shift schedule of medical staff, each event has different time points: mortality and AKI were predicted 3, 6, and 12 hours in advance, and sepsis was predicted 2, 4, and 6 hours in advance. For the model predicting the event within T_i (i Î {1, 2, 3}), we preprocessed the data as positive and negative instances. Specifically, we randomly chose some of the time points within T_i from the event onset for positive instances and some of the time points within T_i from randomly chosen time points for negative instances.

After selecting the prediction times, we collected input features from the admission time to the prediction time. Since EMR data have missing data, to impute the missing data, we applied the carry-forward method if valid data existed before the missing time point. If there were no valid data before the missing time point, but valid data existed after the missing time point, we filled the missing value with normal values of the features. To reduce overfitting, we used L2 regularization to the weights of each layer and stopped the model early when the performance of the model for validation set did not improve 60 epochs in a row after the 100th epoch while training each model. To correct the imbalance in outcomes, we used balanced minibatch training.

Figure 1. Overview of the model structure. AKI: acute kidney injury; T1: near future; T2: mid-term future; T3: distant future.



Performance Measurements

We compared the model performance with other widely used scores and models. Model performance for mortality was compared to that of the APACHE-II and Sequential Organ Failure Assessment (SOFA) scores, and model performance for sepsis prediction was compared to that for the SOFA score. Although these scores are not gold standards for predicting events, the physician's decisions have been based on these scores. Therefore, we compared our models with these scores, as in previous studies [18-20]. Additionally, we compared our model with other popular machine learning models (eg, logistic regression and XGBoost) (see Multimedia Appendix 2 for details). However, there are no gold standard scores for AKI; therefore, we compared the model only with other machine learning models for AKI events. The prediction performance of the individual models was measured as the area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), specificity, and F_1 score with a fixed sensitivity of 0.85, as considered in a previous study [21].

Validation

The model was validated in two ways: First, 5-fold cross-validation was performed using the development cohort—the standard for evaluation of a machine learning algorithm. Then, the model was externally validated in the independent validation cohort. The validation cohort included patients who were admitted to the ICU of the National Health Insurance Corporation Ilsan Hospital, a secondary hospital run by the national insurance company, between January and December 2017.

Error and Delayed Input Experiment

The model robustness against entry error and delayed inputs was compared with the two machine learning models by measuring how much the AUROC and SD values were affected by adding noise. To test the robustness to error input, we added Gaussian noise at normalized features with specific ranges (ie, 1/1000, 1/200, 1/100, 1/20, 1/10, 1/2, and one of each feature scale) to randomly chosen data for two vital signs within 10% of the time sequence. Next, to compare with other machine learning models, we added noise on two randomly chosen vital signs. Additionally, we tested the robustness to the delayed input. To make delayed input errors, we deleted the data within specific hours (ie, 0-10 hours) for two randomly chosen vital signs; then, the deleted data were imputed with the carry-forward method.

All analyses were performed using Python (version 3.6.7) [22], and the model was built using the TensorFlow 1.14 [23] deep learning framework.

Data Availability

The datasets generated during and/or analyzed in this study are not publicly available owing to hospital regulations for electronic medical data but can be made available from the corresponding author upon reasonable request.

Results

Demographic Characteristics

A total of 21,732 and 2487 patients were included in the development and validation cohorts, of which 57.13% (n=12,416) and 56.49% (n=1405) were male participants, respectively. The mean participant age was 60.97 (SD 15.2) and 69.05 (SD 14.13) years in the development and validation cohorts, respectively. The prevalence of mortality, sepsis, and AKI was 783 (3.6%), 679 (3.12%), and 1978 (9.15%) in the development cohort and 209 (8.4%), 243 (9.77%), and 287 (11.54%) in the validation cohort Table 1.

Table 1. Demographic characteristics of the study cohorts.

Characteristic	Development cohort (n=21,732)	Validation cohort (n=2487)
Patients, n (%)	20,053 (92.27)	2362 (94.97)
Age in years, mean (SD)	60.97 (15.2)	69.05 (14.13)
Sex, male, n (%)	12,416 (57.13)	1405 (56.49)
Death, n (%)	783 (3.6)	209 (8.40)
Sepsis, n (%)	679 (3.12)	243 (9.77)
Acute kidney injury, n (%)	1,978 (9.1)	287 (11.54)
Length of ICU[a] stay (days), mean (SD)	3.23 (19.15)	2.99 (3.65)
APACHE II[b] score, mean (SD)	11.57 (5.04)	16.21 (7.25)
SOFA[c] score, mean (SD)	3.66 (3.01)	4.11 (1.04)
ICU admission[d]		
MICU[e]	3138 (14.44)	606 (24.37)
SICU[f]	4604 (21.19)	1141 (45.88)
CCU[g]	5172 (23.79)	740 (29.75)
HICU[h]	3335 (15.35)	—[i]
NCU[j]	5483 (25.23)	—

[a] ICU: intensive care unit.

[b] APACHE: Acute Physiology and Chronic Health Evaluation.

[c] SOFA: Sequential Organ Failure Assessment.

[d] Patient could have multiple admissions to the ICU; the sum of the types of ICU admissions exceeds 100%.

[e] MICU: medical intensive care unit.

[f] SICU: surgical intensive care unit

[g] CCU: critical care unit.

[h] HICU: high intensity care unit.

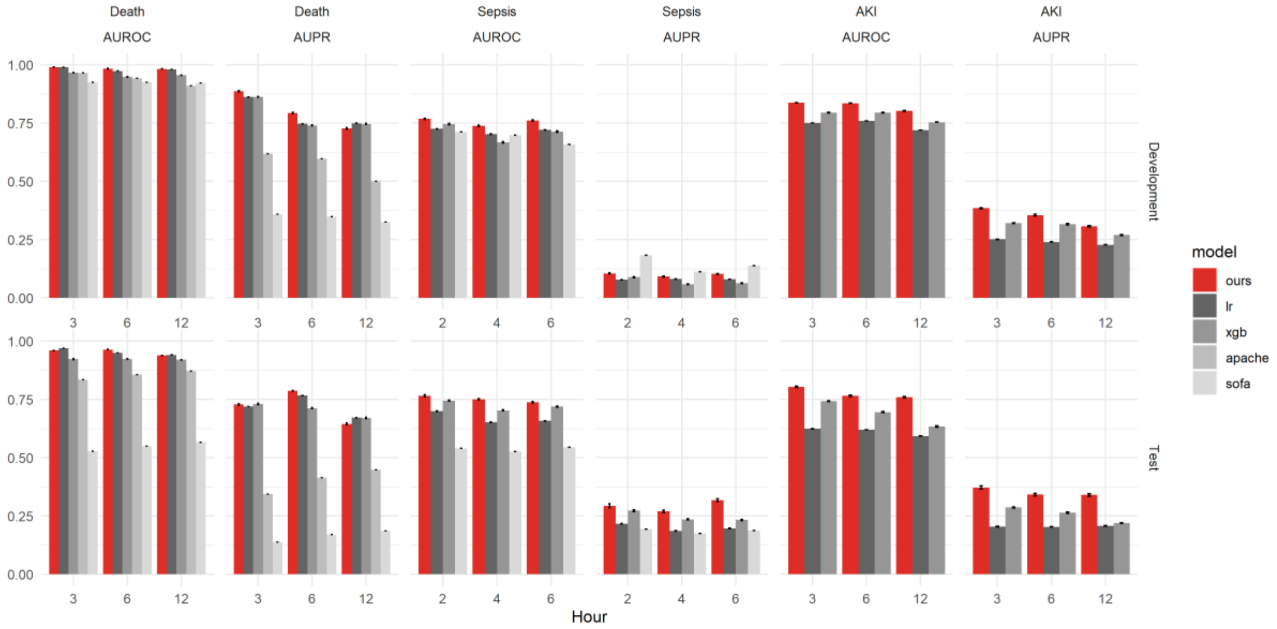
[i] Not available.

[j] NCU: neonatal care unit.

Model performance

For the development cohort, the AUROC values of the death prediction model 3, 6 and 12 hours in advance were 0.990, 0.984, and 0.982, respectively. For the validation cohort, the model achieved AUROC values of 0.960, 0.964, and 0.938 to predict mortality 3, 6, and 12 hours in advance, respectively. The AUPRC values of the death prediction model 3, 6, 12 hours in advance were 0.887, 0.794, and 0.727, respectively, in the development cohort, and 0.728, 0.786, and 0.645, respectively, in the test cohort. The model compared with APACHE-II, SOFA, logistic regression, and XGBoost models. Our model yielded a higher AUROC and AUPRC value than the other models, except a few points. Moreover, the AUROC values of sepsis prediction models 2, 4, and 6 hours in advance were 0.768, 0.739, and 0.761, respectively, in the development cohort and 0.766, 0.751, and 0.738, respectively, in the test cohort. The AUPRC values of sepsis prediction models 2, 4, and 6 hours in advance were 0.105, 0.092, and 0.103, respectively, in the development cohort and 0.294, 0.270, and 0.318, respectively, in the test cohort. These performances were significantly higher than those using the SOFA score (the gold standard medical score), logistic regression, and XGBoost models, except AUPRC values in the development cohort. Although the AUROC values of our models were higher than SOFA scores, AUPRC values were lower than SOFA scores. Finally, the AUROCs of the AKI prediction model 3, 6, and 12 hours in advance were 0.838, 0.836, and 0.802, respectively, in the development cohort and 0.804, 0.766, and 0.760, respectively, in the test cohort. The AUPRC values of AKI prediction model were 0.385, 0.356, and 0.307, respectively, in the development cohort and 0.372, 0.342, and 0.340, respectively, in the test cohort; these values were higher than those using the other two machine learning models (logistic regression and XGBoost; see Figure 2 and Multimedia Appendices 3 and 4).

Figure 2. AUROC and AUPR values of each model at each prediction hour. APACHE: Acute Physiology And Chronic Health Evaluation; area under the receiver operating characteristic curve; AUPR: area under the precision-recall curve; SOFA: Sequential Organ Failure Assessment; xgb: XGBoost.



Sensitivity to Error and Delayed Input

The individual models were evaluated by adding data errors as noise. AUROCs of all models except our proposed model were decreased by increasing the added noise. For example, in the mortality prediction model, when adding Gaussian noise with a feature range, the AUROC of our model dropped to 0.0004 (SD 0.002), whereas it was 0.270 (SD 0.0530) for the logistic regression model, and 0.0732 (SD 0.0442) for the XGBoost model, respectively. Other models show similar results. However, in the delayed input experiments, the mean differences in the AUROC between the original and delayed input data were almost 0 in the validation cohort (see Figure 3 and Multimedia Appendix 5).

As shown in Figure 4, each graph shows how each model works. In the mortality prediction model, 12 hours before the event, the alarm is turned on with only the 12-hour model. As the event nears its time, the alarm is turned on with the 6-hour and 3-hour models, sequentially. Other events show similar results. Because each event model predicts different time windows, the models' prediction can overcome temporal skewness, although there were slight time differences between actual events and predictions.

Figure 3. Changes in AUROC when data errors and delayed inputs were simulated. AKI: acute kidney injury; AUROC: area under the receiver operating characteristic curve; LR: logistic regression; T1: near future; T2: mid-term future; T3: distant future; XGB: XGBoost.

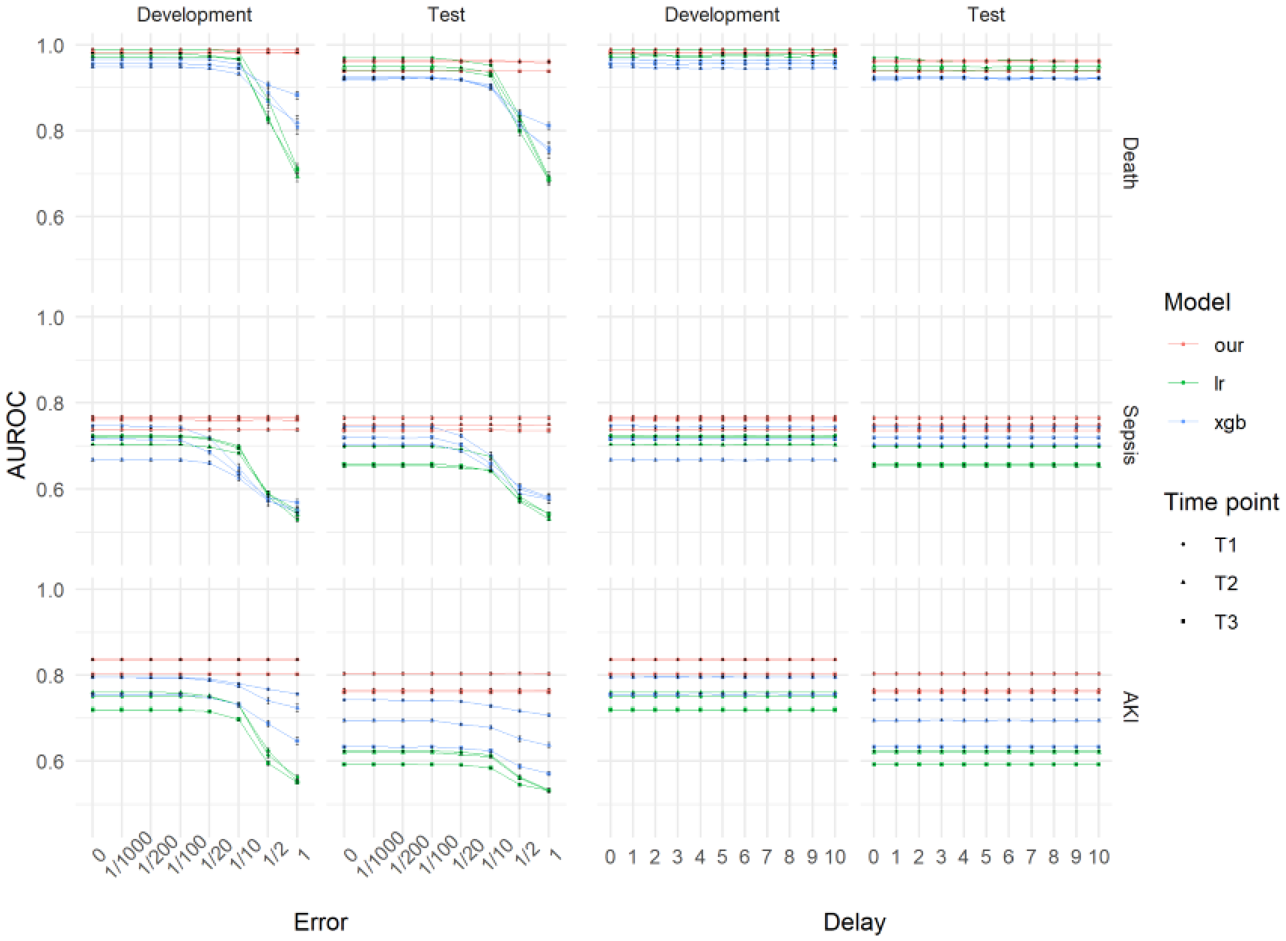
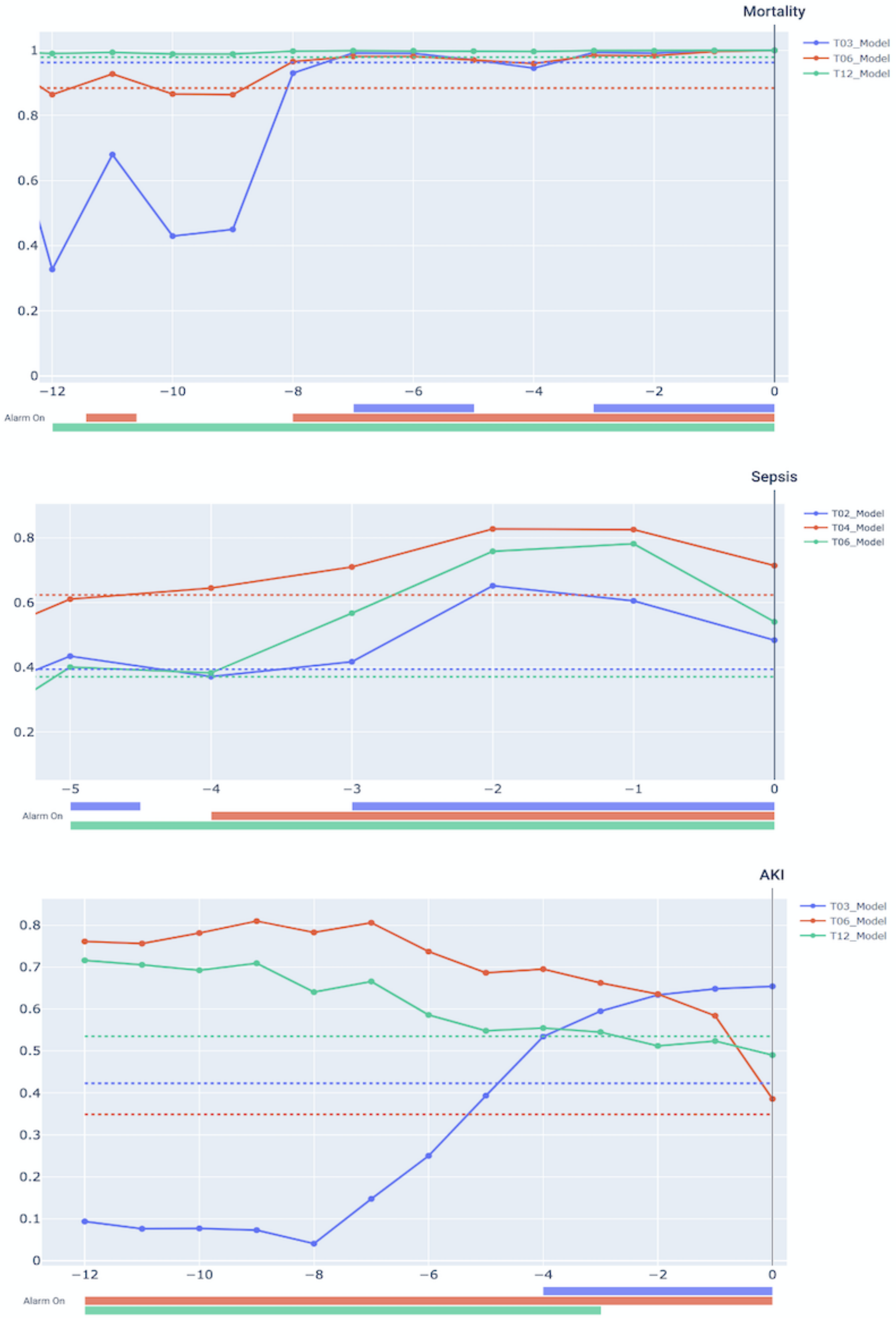


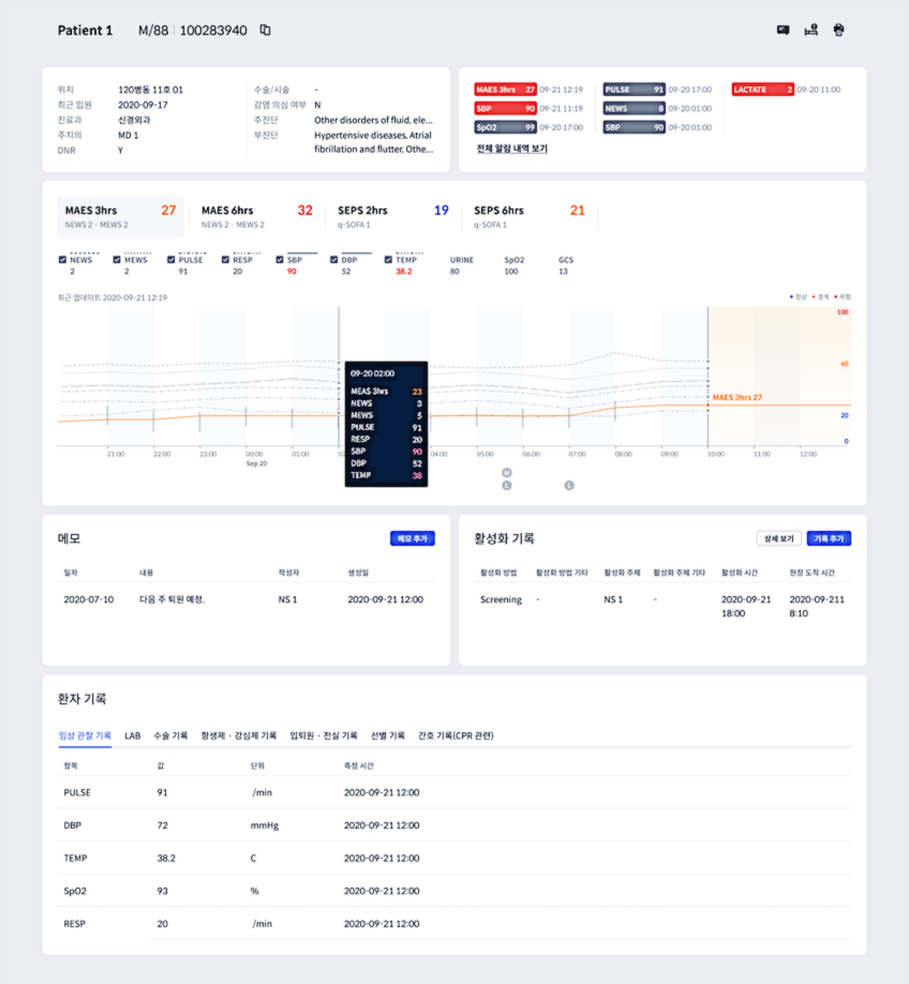
Figure 4. An illustrative example of the prediction of the models. The solid line indicates each model's score. The dotted line indicates the threshold of each model which set by a sensitivity of 0.85 in (A) Mortality (B) Sepsis (C) AKI. AKI: acute kidney injury.



Deployment

These models have been implemented in tertiary and secondary hospitals in Korea. Figure 5 shows a screenshot of the application used to deploy the models.

Figure 5. Screenshot of the application on which the model was deployed.



Patient 1 M/88 | 100283940

위치	120병동 11호 01	수술/시술	-
최근 입원	2020-09-17	감염 의심 여부	N
진료과	신경외과	주진단	Other disorders of fluid, ele...
주치의	MD 1	부진단	Hypertensive diseases, Atrial fibrillation and flutter, Othe...
DNR	Y		

MAES 3hrs 27 09-21 12:19 | PULSE 91 09-20 17:00 | LACTATE 2 09-20 11:00
SBP 90 09-21 11:19 | NEWS 8 09-20 01:00
SpO2 99 09-20 17:00 | SBP 90 09-20 01:00
전체 알람 내역 보기

MAES 3hrs 27 NEWS 2 · MEWS 2 | MAES 6hrs 32 NEWS 2 · MEWS 2 | SEPS 2hrs 19 q-SOFA 1 | SEPS 6hrs 21 q-SOFA 1

☑ NEWS 2 ☑ MEWS 2 ☑ PULSE 91 ☑ RESP 20 ☑ SBP 90 ☑ DBP 52 ☑ TEMP 38.2 URINE 80 SpO2 100 GCS 13

최근 업데이트 2020-09-21 12:19

09-20 02:00
MEAS 3hrs 23
NEWS 3
MEWS 5
PULSE 91
RESP 20
SBP 90
DBP 52
TEMP 38

메모 메모 추가

일자	내용	작성자	생성일
2020-07-10	다음 주 퇴원 예정.	NS 1	2020-09-21 12:00

활성화 기록 상세 보기 기록 추가

활성화 방법	활성화 방법 기타	활성화 주체	활성화 주체 기타	활성화 시간	현장 도착 시간
Screening	-	NS 1	-	2020-09-21 18:00	2020-09-211 8:10

환자 기록

임상 관찰 기록 LAB 수술 기록 항생제 · 강심제 기록 입퇴원 · 전실 기록 선별 기록 간호 기록(CPR 관련)

항목	값	단위	측정 시간
PULSE	91	/min	2020-09-21 12:00
DBP	72	mmHg	2020-09-21 12:00
TEMP	38.2	C	2020-09-21 12:00
SpO2	93	%	2020-09-21 12:00
RESP	20	/min	2020-09-21 12:00

Discussion

Principal Findings

This study demonstrated the prediction models for events in the ICU that consider not only whether the event occurs but also in which time intervals it would occur. By considering all three models that predict the event at different time intervals, physicians can infer *when* the event would occur. Additionally, the robustness of the model was tested by simulated data errors and delayed input. All models showed similar robustness to delayed input; however, only our proposed model was deemed robust to input errors.

The labeling of the outcomes events is one of the most important things in supervised learning, such as these models. Mortality was defined by the EMR-recorded mortality data. However, for sepsis, according to the Sepsis-3 definition [24], the time point at when the infection was suspected, and organ failure began needs to be known. To overcome this issue, Rhee et al [16] proposed a definition of sepsis for clinical surveillance. Nemati et al [21] defined sepsis similarly except for some time intervals because all the definitions were based on the Sepsis-3 definition. The AKI definition depends on serum creatinine levels. In addition to mortality, since AKI and sepsis were defined by a laboratory test, the event label could be incorrect. This point could make the performance of the two models poorer than that of the mortality prediction model.

Because of this working definition, there was a difference in sepsis prevalence in the two cohorts: the mortality rate was 3.12% and 10.04%, and the sepsis prevalence was 3.12% and 11.17% in the development and validation cohorts, respectively. This is probably because the surgical ICU patients comprised a larger proportion in the development cohort than in the validation cohort. This resulted in the APACHE and SOFA scores of the validation cohort being higher than those of the development cohort.

Many studies have attempted to predict events in the ICU. For instance, Hyland et al [9] developed a model to predict circulatory failure in the ICU. Additionally, circulatory failure in the ICU was assessed using a gradient boosting method with the Shapley Additive Explanations (SHAP) value. The model calculates scores every 5 minutes to predict the risk of circulatory failure within the next 8 hours, and it has an AUROC of 0.90. However, because the model was developed as a within-setting model, it is not clear how long it will take for the event to occur. The model only predicts whether the event will occur within 8 hours, even though the event could occur after only 1 hour. Meyer et al [10] predicted mortality, bleeding, and the need for renal replacement therapy 24 hours after cardiothoracic surgery; the AUROCs for these events were 0.87, 0.95, and 0.96, respectively. Even though the model predicted real-time events, the outcome was fixed-time events. Nemati et al predicted sepsis in the ICU using 65 features, including EMR and high-resolution bedside monitoring data; their model yielded AUROCs of 0.82, 0.81, 0.80, and 0.79 for predicting sepsis 4, 6, 8, and 12 hours in advance, respectively. The model was based on the Weibull-Cox regression model, considering within-setting timepoints. To overcome the temporal skewness of the model, Kim et al [11] developed a model that predicts the time point of in-hospital cardiac arrest using a character-level gated recurrent unit with a Weibull distribution. They assumed that the temporal skewness conformed to the Weibull distribution and then predicted the time point at which the distribution indicated the maximum value. Our data also showed temporal skewness of the positive events. When plotted the event-prediction time with each group, most of the predicted true event was found to occur near the real event occurrences (Multimedia Appendix 6). This phenomenon can be shown in other time prediction models. The temporal skewness is important when the model is applied in the real world. When physicians receive an alarm from the model, the working time—that is, the time between the alarm alert and the real event—should be enough to intervene disease progression.

The mortality and AKI predict model showed that the nearer the prediction time was to the event time, the higher the AUROC value was. However, the analyses pertaining to sepsis events, showed the 6 hours in advance prediction model worked better than the 2 hours in advance prediction model. This might be because the definition of sepsis is more subjective than that of other events.

Most robustness assessments of previous models have focused on generalization to any data input. For example, weight decay [25] and the early stopping method [26] are well-known approaches that make the model more robust. However, in this study, we focused on robustness to error and delayed input. All the models showed robustness to the delayed input. This may be because the carry-forward method (used to impute the deleted data) resulted in the delayed input data not being considerably different from the original data, unlike the noise-add experiment. However, the error input experiment showed that our models were more robust than other models (Figure 3). Although we randomly selected two vital signs in the error input experiment, we performed the sensitive analysis by selecting specific pairs of vital signs and adding noise only to those pairs. The mean differences between the original model and the noise added model considered on a scale of 1 were less than 0.003. The performance was still similar such that vital signs were selected, and noises were added (Multimedia appendix 7). Moreover, unlike the time-series model that requires values from time windows, the non–time-series model needs one abstracted value. It seems that making values abstract can lead to higher robustness than the time-series model. However, the non–time series models yielded lower AUROC values than those of our models, except the sepsis prediction model with test dataset (Multimedia Appendix 8). This finding suggested that the time-series model yielded a higher performance and was more robust to the error and delayed input (see Figure 3) than the non–time series models. This can be explained by the fact that the time-series model learned from all the time-series features rather than one time-series representative value.

To the best of our knowledge, this is the first attempt to evaluate the robustness of the model against delayed input and input error. There was no metric for the robustness of an error and delayed input. Thus, the AUROC variation—that is, the mean difference—was used to evaluate robustness when noise was added, or the input was delayed.

Limitations

There are some limitations to our study. First, we could not consider the correlation between each event. For example, both mortality and AKI can be caused by sepsis. However, in this model, each event was considered an independent outcome. Further research should be performed to predict these correlated outcomes. Second, we evaluated input error and delayed input by adding simulated noise to retrospective data. In addition, the model works in the real world. To evaluate these points, a prospective study should be performed. Third, the input features were selected manually; however, these few variables are commonly used in ICUs worldwide to predict patient outcomes. According to survey on sepsis prediction [27], our features have been included in other models. Additionally, other clinical complications or adverse events should be expanded in future studies.

Conclusions

In this study, we developed an outcome prediction model for real-world applications. We considered not only performance but also the robustness of the model to temporal skewness and input delays and errors. By considering temporal skewness, physicians can more effectively intervene in disease progression. Additionally, since the models are robust to delayed input and input error, physicians can trust this model more than those that are not as robust.

Acknowledgments

This study was supported by the Severance Hospital Research Fund for Clinical Excellence (C-2019-0005), a National Health Insurance Ilsan Hospital grant (NHIMC2020-20-003). In addition, it was supported by an Institute for Information & Communication Technology Planning & Evaluation (IITP) grant (2017-0-01779, XAI), and the 2019 IT Promotion Fund (Development of AI-based Precision Medicine Emergency System) by the Korean government (Ministry of Science and Information and Communications Technology [MSIT]). Also, the study was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the MSIT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety (Project number: KMDF_PR_20200901_0089, 9991006762).

Conflicts of Interest

None declared.

Multimedia Appendix 1
The input features of the model.
[DOCX File , 14 KB - medinform_v9i11e26426_app1.docx]

Multimedia Appendix 2
Logistic regression, XGBoost, and LSTM hyperparameters. LSTM: long short-term memory.
[DOCX File , 18 KB - medinform_v9i11e26426_app2.docx]

Multimedia Appendix 3
The results of AUROC and AUPRC of each model and prediction hour. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve.
[DOCX File , 34 KB - medinform_v9i11e26426_app3.docx]

Multimedia Appendix 4
The comparison among AUROC values by each prediction hour in our models. AUROC: area under the receiver operating characteristic curve.
[DOCX File , 18 KB - medinform_v9i11e26426_app4.docx]

Multimedia Appendix 5
Mean and standard deviation of AUROC values in adding noise and input delayed experiments. AUROC: area under the receiver operating characteristic curve.
[DOCX File , 29 KB - medinform_v9i11e26426_app5.docx]

Multimedia Appendix 6
Distribution of true positive samples with event to prediction temporal gap.
[DOCX File , 62 KB - medinform_v9i11e26426_app6.docx]

Multimedia Appendix 7
Error input experiment with two fixed selected vital signs.
[DOCX File , 233 KB - medinform_v9i11e26426_app7.docx]

Multimedia Appendix 8
The AUROC values of our model and non-time-series model which inputted several representative values, such as highest, median, and lowest value of all time windows. AUROC: area under the receiver operating characteristic curve.
[DOCX File , 187 KB - medinform_v9i11e26426_app8.docx]

References

1. Nates JL, Nunnally M, Kleinpell R, Blosser S, Goldner J, Birriel B, et al. ICU admission, discharge, and triage guidelines: A framework to enhance clinical operations, development of institutional policies, and further research. Crit Care Med 2016 Aug;44(8):1553-1602. [doi: 10.1097/CCM.0000000000001856] [Medline: 27428118]
2. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med 1985 Oct;13(10):818-829. [Medline: 3928249]
3. Le Gall J. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 1993 Dec 22;270(24):2957. [doi: 10.1001/jama.1993.03510240069035]
4. Lemeshow S. Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. JAMA 1993 Nov 24;270(20):2478. [doi: 10.1001/jama.1993.03510200084037]
5. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019 Aug;572(7767):116-119 [FREE Full text] [doi: 10.1038/s41586-019-1390-1] [Medline: 31367026]
6. Rolnick JA, Weissman GE. Early warning systems: the neglected importance of timing. J Hosp Med 2019 Jul 01;14(7):445-447. [doi: 10.12788/jhm.3229] [Medline: 31251151]

7. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. J Am Med Inform Assoc 2017 Nov 01;24(6):1142-1148 [FREE Full text] [doi: 10.1093/jamia/ocx080] [Medline: 29016973]
8. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 04;380(14):1347-1358. [doi: 10.1056/NEJMra1814259] [Medline: 30943338]
9. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med 2020 Mar;26(3):364-373. [doi: 10.1038/s41591-020-0789-4] [Medline: 32152583]
10. Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sündermann SH, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med 2018 Dec;6(12):905-914. [doi: 10.1016/S2213-2600(18)30300-X] [Medline: 30274956]
11. Kim J, Park YR, Lee JH, Lee J, Kim Y, Huh JW. Development of a real-time risk prediction model for in-hospital cardiac arrest in critically ill patients using deep learning: retrospective study. JMIR Med Inform 2020 Mar 18;8(3):e16349 [FREE Full text] [doi: 10.2196/16349] [Medline: 32186517]
12. Thorsen-Meyer H, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. The Lancet Digital Health 2020 Apr;2(4):e179-e191. [doi: 10.1016/s2589-7500(20)30018-2]
13. Gutierrez G. Artificial intelligence in the intensive care unit. Crit Care 2020 Mar 24;24(1):101 [FREE Full text] [doi: 10.1186/s13054-020-2785-y] [Medline: 32204716]
14. Dall'Ora C, Griffiths P, Redfern O, Recio-Saucedo A, Meredith P, Ball J, Missed Care Study Group. Nurses' 12-hour shifts and missed or delayed vital signs observations on hospital wards: retrospective observational study. BMJ Open 2019 Feb 01;9(1):e024778 [FREE Full text] [doi: 10.1136/bmjopen-2018-024778] [Medline: 30782743]
15. Kotronis C, Routis I, Politi E, Nikolaidou M, Dimitrakopoulos G, Anagnostopoulos D, et al. Evaluating internet of medical things (IoMT)-based systems from a human-centric perspective. Internet of Things 2019 Dec;8:100125. [doi: 10.1016/j.iot.2019.100125]
16. 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] [Medline: 28903154]
17. Kellum JA, Lameire N, KDIGO AKI Guideline Work Group. Diagnosis, evaluation, and management of acute kidney injury: a KDIGO summary (Part 1). Crit Care 2013 Feb 04;17(1):204 [FREE Full text] [doi: 10.1186/cc11454] [Medline: 23394211]
18. Yuan K, Tsai L, Lee K, Cheng Y, Hsu S, Lo Y, et al. The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. Int J Med Inform 2020 Sep;141:104176 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104176] [Medline: 32485555]
19. 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]
20. Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, et al. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. Comput Biol Med 2019 Jun;109:79-84 [FREE Full text] [doi: 10.1016/j.compbiomed.2019.04.027] [Medline: 31035074]
21. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018 Apr;46(4):547-553 [FREE Full text] [doi: 10.1097/CCM.0000000000002936] [Medline: 29286945]
22. Python Software Foundation. Python 3 Reference Manual. URL: https://docs.python.org/3/reference/ [accessed 2021-10-22]
23. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015. URL: https://www.tensorflow.org/ [accessed 2021-10-22]
24. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016 Feb 23;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
25. Krogh A, Hertz J. A simple weight decay can improve generalization. In: Advances in Neural Information Processing Systems. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1992 Presented at: Proceedings of the 4th International Conference on Neural Information Processing Systems; December 2-5, 1991; Denver, CO, USA p. 950-957 URL: https://papers.nips.cc/paper/1991/hash/8eefcfdf5990e441f0fb6f3fad709e21-Abstract.html
26. Prechelt L. Early Stopping — But When? In: Neural Networks: Tricks of the Trade. Berlin, Heidelberg: Springer; 1998:53-67.
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;46(3):383-400 [FREE Full text] [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]

Abbreviations

AKI: acute kidney injury
APACHE: Acute Physiology And Chronic Health Evaluation
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
EMR: electronic medical record
GCS: Glasgow Coma Score
ICU: intensive care unit
IITP: Institute for Information & Communication Technology Planning & Evaluation
IoMT: internet of medical things
MPM: Mortality Probability Model
MSIT: Ministry of Science and Information and Communications Technology
SAPS: Simplified Acute Physiology Score
SOFA: Sequential Organ Failure Assessment
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by C Lovis; submitted 27.12.20; peer-reviewed by S Hyland; comments to author 20.01.21; revised version received 21.05.21; accepted 24.09.21; published 04.11.21.

Please cite as:
Sung M, Hahn S, Han CH, Lee JM, Lee J, Yoo J, Heo J, Kim YS, Chung KS
Event Prediction Model Considering Time and Input Error Using Electronic Medical Records in the Intensive Care Unit: Retrospective Study
JMIR Med Inform 2021;9(11):e26426
URL: https://medinform.jmir.org/2021/11/e26426
doi:10.2196/26426
PMID:34734837

Original Paper

Local Differential Privacy in the Medical Domain to Protect Sensitive Information: Algorithm Development and Real-World Validation

MinDong Sung[1*], MD; Dongchul Cha[1,2*], MD; Yu Rang Park[1], PhD

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea

[2]Department of Otorhinolaryngology, Yonsei University College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Yu Rang Park, PhD
Department of Biomedical Systems Informatics
Yonsei University College of Medicine
Yonsei-ro 50-1
Seoul, 03722
Republic of Korea
Phone: 82 2 228 2363
Fax: 82 2 227 8354
Email: yurangpark@yuhs.ac

Abstract

Background: Privacy is of increasing interest in the present big data era, particularly the privacy of medical data. Specifically, differential privacy has emerged as the standard method for preservation of privacy during data analysis and publishing.

Objective: Using machine learning techniques, we applied differential privacy to medical data with diverse parameters and checked the feasibility of our algorithms with synthetic data as well as the balance between data privacy and utility.

Methods: All data were normalized to a range between –1 and 1, and the bounded Laplacian method was applied to prevent the generation of out-of-bound values after applying the differential privacy algorithm. To preserve the cardinality of the categorical variables, we performed postprocessing via discretization. The algorithm was evaluated using both synthetic and real-world data (from the eICU Collaborative Research Database). We evaluated the difference between the original data and the perturbated data using misclassification rates and the mean squared error for categorical data and continuous data, respectively. Further, we compared the performance of classification models that predict in-hospital mortality using real-world data.

Results: The misclassification rate of categorical variables ranged between 0.49 and 0.85 when the value of ε was 0.1, and it converged to 0 as ε increased. When ε was between 10^2 and 10^3, the misclassification rate rapidly dropped to 0. Similarly, the mean squared error of the continuous variables decreased as ε increased. The performance of the model developed from perturbed data converged to that of the model developed from original data as ε increased. In particular, the accuracy of a random forest model developed from the original data was 0.801, and this value ranged from 0.757 to 0.81 when ε was 10^{-1} and 10^4, respectively.

Conclusions: We applied local differential privacy to medical domain data, which are diverse and high dimensional. Higher noise may offer enhanced privacy, but it simultaneously hinders utility. We should choose an appropriate degree of noise for data perturbation to balance privacy and utility depending on specific situations.

(JMIR Med Inform 2021;9(11):e26914) doi:10.2196/26914

KEYWORDS

privacy-preserving; differential privacy; medical informatics; medical data; privacy; electronic health record; algorithm; development; validation; big data; medical data; feasibility; machine learning; synthetic data

Introduction

Big data is a core factor in the renovation of medicine. The raw data have low utility; however, applying algorithms such as machine learning (ML) enables us to make the most of these data [1]. Unlike rule-based systems, ML algorithms are data driven and require a large amount of data. Particularly, conventional ML approaches require centralized data for learning. To obtain this substantial amount of data, it is necessary to exchange data among different organizations to develop an effective ML model.

However, the exchange of data between different parties causes privacy problems, and there are increasing concerns about privacy violations by large companies [2]. Medical data that mostly contain sensitive information should be appropriately protected when shared with third parties. The European Union’s General Data Protection Regulation [3] and the United States’ Health Insurance Portability and Accountability Act of 1996 (HIPAA) [4] recognize this problem and require users’ privacy to be strengthened. Medical data have various distinct properties in addition to their sensitive attributes. For example, serum glucose levels are continuous, whereas medical histories are usually recorded using categorical values. Medical data also contain multimodal values: some of the data may be obtained from blood tests, whereas others may originate from radiologic and physical examination tests.

Deidentification is defined as “the removal or replacement of personal identifiers so that it would be difficult to reestablish a link between the individual and his or her data [5].” Especially, in the HIPAA, data is considered as deidentified when specified data elements are removed [4]. Anonymization is defined as “the irreversible removal of the link between the individual and his or her medical record data to the degree that it would be virtually impossible to reestablish the link [5].” In such a case, the anonymized data could never be reidentified using the data in the underlying data sets. There are three primary ways to anonymize these data: suppression, generalization, and noise addition [6]. Deidentification may not necessarily be anonymized. That is, anonymization is a subset of deidentification. Following anonymization, three main measures to identify the privacy risk can be evaluated: *k*-anonymity [7], *l*-diversity [8], and *t*-closeness [9]. Deidentification tools, such as ARX [10], offer seamless privacy protection through feature generalization and the suppression of records.

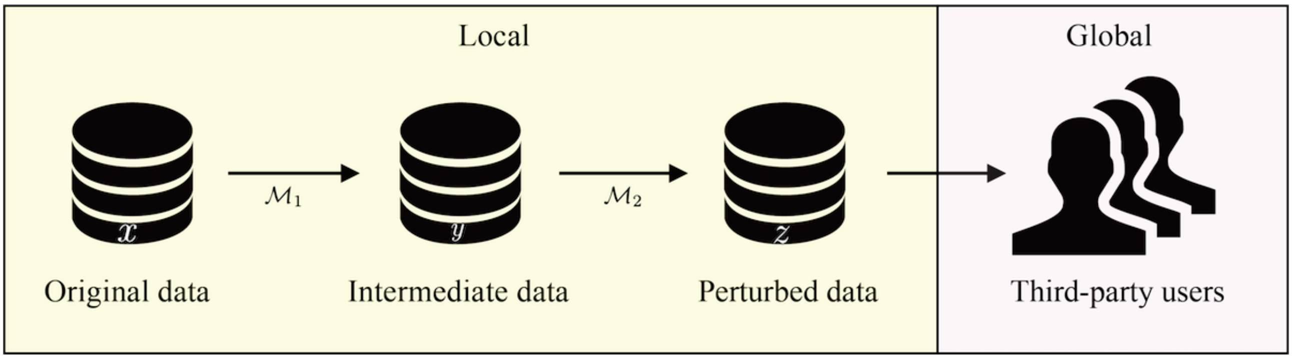
Differential privacy [11], which entails a semantic model, is another data privacy approach. Compared to syntactic anonymity, it requires less domain knowledge and is inherently robust to linkage attacks combined with domain knowledge. Moreover, differential privacy is considered to be a de facto standard for private data analysis or publishing [12,13]. Technology companies such as Apple and Google have attempted to apply differential privacy to protect the privacy of mobile data [14,15]. Moreover, the rapid development of the Internet of Things (IoT) should consider privacy risk [16]. Researchers have been actively applying differential privacy to the IoT, such as automatically driving cars [17] and sensors [16]. In ML, personal information can be leaked. Applying differential privacy to the deep learning model can overcome this threat [18,19], and the health care domain is no exception. Several studies have been performed in the health care domain. For example, Kim et al [20] introduced a local differential privacy algorithm for health data streams. Also, Suriyakumar et al [21] investigated the feasibility of differentially private stochastic gradient descent in a health care setting with the influential function. Most studies focus on a data set that has only a few features and focus on differential privacy in the deep learning model.

In this study, we focused on local differential privacy with regard to multivariate medical data. We applied differential privacy with diverse parameters and checked (1) the feasibility of training our algorithms with synthetic data and (2) the balance between data privacy and utility with regard to ML techniques.

Methods

Figure 1 presents the workflow employed to achieve differential privacy in this study. When a user requests data, we perturb the data using the bounded Laplacian method () and discretization postprocessing () to provide high-fidelity data while preserving the privacy of the original data.

Figure 1. Differential privacy upon data request from third party users. The owner perturbs the original data to preserve privacy before sending the data externally. The third-party user can be either a curator or the final user. <inline-graphic xlink:href="medinform_v9i11e26914_fig5.png" xlink:type="simple" mimetype="image"/>: bounded Laplacian method; <inline-graphic xlink:href="medinform_v9i11e26914_fig6.png" xlink:type="simple" mimetype="image"/>: discretization postprocessing.



The Value of ε for Local Differential Privacy

Dwork et al [22] defined ε-differential privacy as a randomized function. For adjacent data Y_1 and Y_2, function κ is (ε, δ)–differentially private if

$$P[\kappa(Y_1) \in S] \le \varepsilon \cdot P[\kappa(Y_2) \in S] + \delta$$

where $S \subset Range(\kappa)$. Local differential privacy is a specific case in which the random function or perturbation is applied by data owners, not by central aggregators.

Bounded Laplacian Method

Before applying local differential privacy, all variables were normalized to a range between –1 and 1. First, we applied the bounded Laplacian method. Because a conventional Laplacian distribution yields an infinite boundary, it entails some limitations when applied to clinical domains. For example, respiratory rates, which are supposed to be a positive number, may become negative after applying the conventional Laplacian method, which is illogical. There are two methods to overcome this problem: the truncation method and the bound method [23]. We focused on the latter to minimize the probability of data manipulation because changes in data in the medical domain may have a considerable impact on the desired outputs.

We used the bounded Laplacian function proposed by Holohan et al [23], assuming that the input variable is within the output domain. Given $b > 0$, $W_q: \Omega \rightarrow D$, for each $q \in D$, we defined the probability density function as:

where

We set δ=0,l (lower bound) as –1, u (upper bound) as 1, and ΔQ as 2 in our experiments and adjusted ε to measure the effect of the privacy changes.

Discretization Postprocessing for Discrete Variables

Because we applied the bounded Laplacian method to perturb the given data to a range between –1 and 1 in a continuous manner, there are infinite possibilities for a given input. Many medical domain variables are categorical (either ordinal or nominal), such as medicosurgical histories. Therefore, following the application of the bounded Laplacian method, additional postprocessing was performed for categorical variables. We distributed the intermediate output of the given data over the Bernoulli distribution, similar to the method proposed by Yang et al [17]. The perturbed data $y \in [-C,C]$ were separated into m pieces, where m is the cardinality of the original input variable (a positive integer). We first shifted the range $[-C,C]$ to $[0, m]$ by equally dividing the space, which resulted in intervals. Therefore, for given perturbed data y, we obtain the following:

After calculating k, the Bernoulli probability p was sampled such that

which is the distance between two adjacent possibilities. Finally, we discretized the perturbed data y concerning the Bernoulli probability p such that

where denotes the Bernoulli distribution function.

Data Set for Validation

We used simulated (randomly generated) data for initial validation to ensure that the bounded Laplacian method functions as expected. To simulate real-world use, we used the eICU Collaborative Research Database [24]. First, to evaluate the extent to which the proposed differential privacy algorithms effectively perturbed the given original data, we used the misclassification rate for categorical variables and mean squared error (MSE) for continuous variables when measuring the similarity between two data sets. Second, to evaluate the adverse effect of differential privacy on the utility of the data set, we compared the accuracy of predicting the mortality rate following intensive care unit admission using Acute Physiology and Chronic Health Evaluation (APACHE) [25] scoring variables under various ε values. The data set contained intubated, ventilation, dialysis, medication status (cardinality: 2), eyes (cardinality: 4), motor (cardinality: 5), and verbal status (cardinality: 6) as categorical variables. Urine output, temperature, respiratory rate, sodium, heart rate, mean blood pressure, pH, hematocrit, creatinine, albumin, oxygen pressure, CO_2 pressure, blood urea nitrogen, glucose, bilirubin, and fraction of inspired oxygen (FiO_2) values were considered continuous variables. There were initially 148,532 patients (rows) in the data set, but after the deletion of missing values, the data set contained a total of 4740 patients (3597 who were alive and 1143 who had died). The following ML methods were used for mortality prediction: decision tree, K-nearest neighbor, support vector machine, logistic regression, naïve Bayes, and random forest. The data were divided into training and test sets in a ratio of 80:20. All predictions were averaged using a 5-fold cross-validation method, and the scikit-learn [26] library was used with the Python programming language.

Results

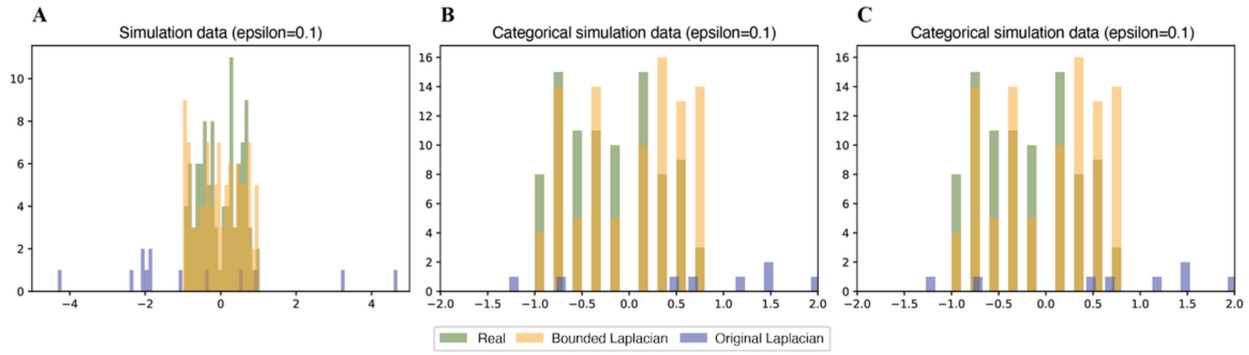
Synthetic Data for Validation of the Bounded Laplacian Function

We created an equally spaced distribution, ranging between –1 and 1, and applied the bounded Laplacian method. In contrast to the conventional Laplacian method, which has an infinite range, the bounded method entailed a range of –1 to 1.

After confirming that the bounded Laplacian method works as intended, we then created synthetic continuous data that range from –1 to 1 and applied the conventional Laplacian method and bounded Laplacian method with ε=0.1, δ=0 (Figure 2A).

The original Laplacian method had out-of-range occurrences that were not present in the bounded Laplacian method. To test the categorical data and postdiscretization processing, we created a set of 100 random integers ranging from 0 to 9, then normalized them to range from –1 to 1. The original Laplacian method had some occurrences that were out of bounds. In the categorical data, the bounded Laplacian method stayed within the data range, as in the continuous data. However, some of the categorical values were not initially present in the given data (Figure 2B), which is similar to the out-of-bounds condition. Therefore, additional postprocessing discretization was performed, and the algorithm showed that the discretization technique ensures that there are no nonexistent values in the categorical data (Figure 2C).

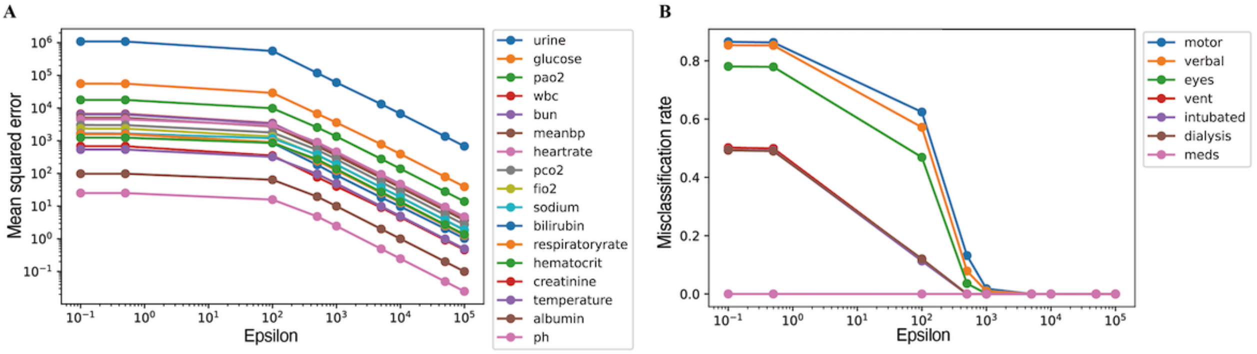
Figure 2. Comparison of conventional and bounded Laplacian methods using synthetic data. (A) Histogram of randomly generated continuous data ranging from –1 to 1. (B) Histogram of randomly generated categorical data, which originally ranged from 0 to 9 and were then normalized to range from –1 to 1. (C) Histogram obtained after application of discretization postprocessing to the data in (B). In all scenarios, the Laplacian method was applied with ε=0.1, δ=0.



Validation Using Real-World Data

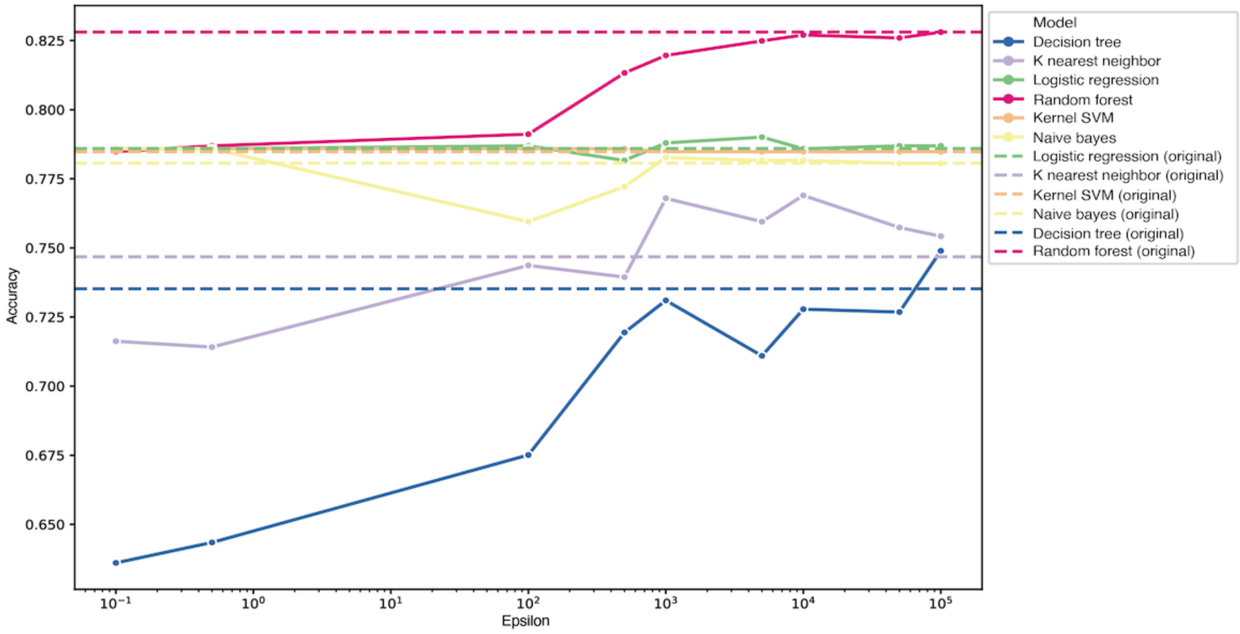
The eICU Collaborative Research Database [24] was used for validation. We used MSEs and misclassification rates as metrics for continuous and categorical variables, respectively, to calculate the differences between the original and perturbed data. Because of the variance between values in the original data, the MSE of continuous variables varies extensively in the case of eICU data. For example, pH and albumin are similar among different individuals, whereas heart rate and glucose have substantial differences (Figure 3A). Regarding the categorical variables, intubated, ventilation, and dialysis status are either 0 or 1, and the chance level is 0.5. The value for “eye” ranges from 1 to 4, that for “verbal” ranges from 1 to 5, and that for “motor” ranges from 1 to 6. Therefore, there were differences in the misclassification rates, especially when ε was small (Figure 3B). As ε increased, all perturbed values approached their original values for both continuous and categorical variables (Figures 3A and 3B).

Figure 3. ε values and degrees of data perturbation for (A) continuous variables and (B) categorical variables. bun: blood urea nitrogen; fio2: fraction of inspired oxygen; meanbp: mean blood pressure; pao2: partial pressure of oxygen, arterial; pco2: partial pressure of carbon dioxide; wbc: white blood cells.



To simulate data utility with respect to ε, we constructed a predictive classifier to predict mortality using the eICU data set. Note that 3,597 of the 4,740 patients (75.9%) were alive, yielding a chance level of 76%. A lower value of ε caused severe data perturbation, resulting in an accuracy that was near the chance level. Increasing the value of ε increased the performance of the classifiers, and the performance converged to the accuracy obtained using the original data (shown as dashed lines in Figure 4). This tendency was consistent among the different models, and the random forest model was the top performer.

Figure 4. Classification accuracies among different machine learning models with respect to ε. The performance of the models developed using original data is marked with dashed lines. SVM: support vector machine.



Discussion

Principal Findings

In this study, we developed and validated a local differential privacy method for the medical domain. We used the bounded Laplacian method to overcome the out-of-bounds problem. In addition, we used discretization postprocessing for the categorical variables to address nonexistent categorical variables following perturbation.

Various approaches and metrics are employed when publishing microdata publicly. *k*-anonymity [7] is a metric that requires each cluster (or set of persons in medical data) to have at least *k* records so that there are at least *k* – 1 individuals that are indistinguishable. However, this metric is susceptible to reidentification through linkage attacks and applications of background knowledge. *l*-diversity was introduced to overcome these limitations; it requires each equivalent block containing sensitive information to have at least *l* appropriately represented values. This method is still vulnerable to skewness and similarity attacks [9]. *t*-closeness [9] mitigates this issue by requiring an equivalence class to have a distance of less than *t* (the earth mover distance) between the distribution of a sensitive attribute and that of the overall data. However, using the earth mover distance makes it difficult to identify the closeness between *t* and the gained knowledge. In addition, in this approach, the distribution of sensitive attributes in the equivalence class must be similar to that in the entire data set.

In contrast to these privacy metrics and methods, ε-differential privacy retains the structure of the data while adding noise to prevent leakage of the original data (Figure 2). There are two main differential privacy schemas: global and local. Global differential privacy requires the database owner to trust a curator that performs data perturbation before sending the data to the requested user. Our implementation, local differential privacy, assumes the worst-case scenario by considering an untrusted curator. The leakage of a medical data set may have critical consequences because such a data set may contain sensitive information, such as disease data, medical history, and insurance status. Therefore, our method minimizes the risk of data leaks by not trusting anyone outside the network.

Medical domain data are, by nature, multidimensional and multimodal. *k*-anonymity may suffer from severe utility loss if applied to high-dimensional data [27]. ε-differential privacy also suffered from severe utility loss under a low ε, which was apparent from the low classification accuracy in predicting the mortality rate (Figure 4). Despite the fact that the given data set was multidimensional and multimodal, adjusting the value of ε affected all variables uniformly regardless of their data type.

Differential privacy usually has stronger tradeoffs between data utility, which we mainly focused on, and privacy [28,29]. There were high variances between variables with regard to the MSEs and misclassification rates when ε was low (Figure 3). As ε increased, all variables approached their actual values, enabling better utility at the cost of privacy; this is apparent from the accuracy of prediction shown in Figure 4. When publishing synthetically perturbed data with ε-differential privacy, we may consider providing the ε value along with the data. This additional information may provide users with insights into the degree of data perturbation.

According to the results, for our data set, we may heuristically choose an ε value between 10^3 and 10^4 and apply differential privacy methods to send the perturbed data upon the user's request. The optimal value of ε varies among different data sets and utility requirements, and choosing this value is beyond the scope of this study.

A limitation of this study is that we only applied our algorithms to synthetic data, and we validated the algorithms on only one data set. However, it is likely that other data sets can also be directly employed because we used a relatively small amount of prior data knowledge in our algorithm. In addition, we excluded rows that contained null values in the database. Because medical data are high-dimensional and sparse, future studies should be conducted to address null values. The distributions of data sets affect the normalization and the perturbation process. It is better to share distributions with each institute, such as the minimum and maximum values of each column. The model would be developed from perturbed data, which can be less accurate than a model based on original data. The optimal ε value, which determines the degree of perturbation, should be set to apply to the algorithm. In this study, a value of ε between 10^3 and 10^4 seemed heuristically appropriate; this depends on which data or model is used.

Conclusion

We applied local differential privacy to medical domain data, which is diverse and high-dimensional. Applying bounded Laplacian noise with discretization postprocessing ensures that no out-of-bound data are present. Higher noise may offer enhanced privacy, but it simultaneously hinders utility. Thus, choosing an appropriate degree of noise for data perturbation entails a privacy-utility tradeoff, and one should choose such parameters depending on specific situations.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project and the MD-Phd/Medical Scientist Training Program through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (KHIDIHI19C1015010020, HI21C0974).

Conflicts of Interest

None declared.

References

1. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219 [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
2. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning. ACM Trans Intell Syst Technol 2019 Feb 28;10(2):1-19. [doi: 10.1145/3298981]
3. Voigt P, von dem Bussche A. The EU General Data Protection Regulation (GDPR): A Practical Guide. Cham, Switzerland: Springer International Publishing; 2017.
4. Fact Sheet: The Health Insurance Portability and Accountability Act (HIPAA). US Department of Labor. 2004 Dec. URL: http://purl.fdlp.gov/GPO/gpo10291 [accessed 2021-09-15]
5. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. J Med Internet Res 2019 May 31;21(5):e13484-e13570 [FREE Full text] [doi: 10.2196/13484] [Medline: 31152528]
6. Anonymization. International Association of Privacy Professionals. URL: https://iapp.org/resources/article/anonymization/ [accessed 2001-09-16]
7. Sweeney L. k-Anonymity: a model for protecting privacy. Int J Unc Fuzz Knowl Based Syst 2012 May 02;10(05):557-570. [doi: 10.1142/S0218488502001648]
8. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. ACM Trans Knowl Discov Data 2007 Mar 01;1(1):3-es. [doi: 10.1145/1217299.1217302]
9. Li N, Li T, Venkatasubramanian S. t-Closeness: privacy beyond k-anonymity and l-diversity. 2007 Jun 4 Presented at: 2007 IEEE 23rd International Conference on Data Engineering; April 15-20, 2007; Istanbul, Turkey. [doi: 10.1109/icde.2007.367856]
10. Prasser F, Eicher J, Spengler H, Bild R, Kuhn KA. Flexible data anonymization using ARX—current status and challenges ahead. Softw: Pract Exper 2020 Feb 25;50(7):1277-1304. [doi: 10.1002/spe.2812]
11. Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M. Our data, ourselves: privacy via distributed noise generation. In: Advances in Cryptology - EUROCRYPT 2006. 2006 Presented at: EUROCRYPT 2006: Annual International Conference on the Theory and Applications of Cryptographic Techniques; May 28-June 1, 2006; Saint Petersburg, Russia p. 486. [doi: 10.1007/11761679_29]
12. Barthe G, Chadha R, Jagannath V, Sistla A, Viswanathan M. Deciding differential privacy for programs with finite inputs and outputs. 2020 Jul 08 Presented at: 35th Annual ACM/IEEE Symposium on Logic in Computer Science; July 8-11, 2020; Saarbrücken, Germany p. 141-154. [doi: 10.1145/3373718.3394796]
13. Li N, Lyu M, Su D, Yang W. Differential privacy: from theory to practice. Synthesis Lectures on Information Security, Privacy, and Trust. 2016 Oct 25. URL: https://www.morganclaypool.com/doi/10.2200/S00735ED1V01Y201609SPT018 [accessed 2021-09-20]
14. Differential privacy. Apple. URL: https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf [accessed 2021-09-15]
15. Google's differential privacy libraries. GitHub. URL: https://github.com/google/differential-privacy [accessed 2021-09-15]

16. Erlingsson, Pihur V, Korolova A. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In: CCS '14: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. 2014 Nov 03 Presented at: 2014 ACM SIGSAC Conference on Computer and Communications Security; November 3-7, 2014; Scottsdale, AZ p. 1054-1067. [doi: 10.1145/2660267.2660348]
17. Zhao Y, Zhao J, Yang M, Wang T, Wang N, Lyu L, et al. Local differential privacy-based federated learning for Internet of Things. IEEE Internet Things J 2021 Jun 1;8(11):8836-8853. [doi: 10.1109/jiot.2020.3037194]
18. Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. 2019 Presented at: 2019 IEEE Symposium on Security and Privacy (SP); May 19-23, 2019; San Francisco, CA p. 739. [doi: 10.1109/sp.2019.00065]
19. Ha T, Dang T, Dang T, Truong T, Nguyen M. Differential privacy in deep learning: an overview. 2019 Presented at: 2019 International Conference on Advanced Computing and Applications (ACOMP); November 26-28, 2019; Nha Trang, Vietnam. [doi: 10.1109/acomp.2019.00022]
20. Kim JW, Jang B, Yoo H. Privacy-preserving aggregation of personal health data streams. PLoS One 2018;13(11):e0207639 [FREE Full text] [doi: 10.1371/journal.pone.0207639] [Medline: 30496200]
21. Suriyakumar V, Papernot N, Goldenberg A, Ghassemi M. Chasing your long tails: differentially private prediction in health care settings. In: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021 Mar 03 Presented at: 2021 ACM Conference on Fairness, Accountability, and Transparency; March 3-10, 2021; Virtual event (Canada) p. 723-734. [doi: 10.1145/3442188.3445934]
22. Dwork C, Rothblum GN. Concentrated differential privacy. ArXiv Preprint posted online on March 6, 2016 [FREE Full text]
23. Holohan N, Antonatos S, Braghin S, Mac Aonghusa P. The bounded Laplace mechanism in differential privacy. ArXiv. Preprint posted online on August 30, 2018 2020 [FREE Full text] [doi: 10.29012/jpc.715]
24. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018 Sep 11;5:180178 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
25. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 2006 May;34(5):1297-1310. [doi: 10.1097/01.CCM.0000215112.84523.F0] [Medline: 16540951]
26. Pedregosa F, Grisel O, Weiss R, Passos A, Brucher M, Varoquax G, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
27. Rajendran K, Jayabalan M, Rana M. A study on k-anonymity, l-diversity, and t-closeness techniques of privacy preservation data publishing. Int J Innov Res Sci Eng Technol 2019;6(6):19-24 [FREE Full text]
28. Kohlmayer F, Prasser F, Kuhn KA. The cost of quality: implementing generalization and suppression for anonymizing biomedical data with minimal information loss. J Biomed Inform 2015 Dec;58:37-48 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.007] [Medline: 26385376]
29. Dankar F, El EK. Practicing differential privacy in health care: a review. Trans Data Priv 2013;6(1):35-67 [FREE Full text]

Abbreviations

APACHE: Acute Physiology and Chronic Health Evaluation
FiO2: fraction of inspired oxygen
HIPAA: Health Insurance Portability and Accountability Act
IoT: Internet of Things
ML: machine learning
MSE: mean squared error

Edited by G Eysenbach; submitted 04.01.21; peer-reviewed by J Vilaplana; comments to author 28.01.21; revised version received 10.02.21; accepted 06.09.21; published 08.11.21.

Please cite as:
Sung M, Cha D, Park YR
Local Differential Privacy in the Medical Domain to Protect Sensitive Information: Algorithm Development and Real-World Validation
JMIR Med Inform 2021;9(11):e26914
URL: https://medinform.jmir.org/2021/11/e26914
doi:10.2196/26914
PMID:34747711

Original Paper

Accurate Prediction of Stroke for Hypertensive Patients Based on Medical Big Data and Machine Learning Algorithms: Retrospective Study

Yujie Yang[1,2*], MSc; Jing Zheng[3*], PhD; Zhenzhen Du[1], MSc; Ye Li[1,4], PhD; Yunpeng Cai[1], PhD

[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

[2]University of Chinese Academy of Sciences, Beijing, China

[3]Shenzhen Health Information Center, Shenzhen, China

[4]Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen, China

[*]these authors contributed equally

Corresponding Author:
Yunpeng Cai, PhD
Shenzhen Institute of Advanced Technology
Chinese Academy of Sciences
1068 Xueyuan Blvd
Nanshan District
Shenzhen, 518055
China
Phone: 86 755 86392202
Fax: 86 755 86392299
Email: yp.cai@siat.ac.cn

Abstract

Background: Stroke risk assessment is an important means of primary prevention, but the applicability of existing stroke risk assessment scales in the Chinese population has always been controversial. A prospective study is a common method of medical research, but it is time-consuming and labor-intensive. Medical big data has been demonstrated to promote disease risk factor discovery and prognosis, attracting broad research interest.

Objective: We aimed to establish a high-precision stroke risk prediction model for hypertensive patients based on historical electronic medical record data and machine learning algorithms.

Methods: Based on the Shenzhen Health Information Big Data Platform, a total of 57,671 patients were screened from 250,788 registered patients with hypertension, of whom 9421 had stroke onset during the 3-year follow-up. In addition to baseline characteristics and historical symptoms, we constructed some trend characteristics from multitemporal medical records. Stratified sampling according to gender ratio and age stratification was implemented to balance the positive and negative cases, and the final 19,953 samples were randomly divided into a training set and test set according to a ratio of 7:3. We used 4 machine learning algorithms for modeling, and the risk prediction performance was compared with the traditional risk scales. We also analyzed the nonlinear effect of continuous characteristics on stroke onset.

Results: The tree-based integration algorithm extreme gradient boosting achieved the optimal performance with an area under the receiver operating characteristic curve of 0.9220, surpassing the other 3 traditional machine learning algorithms. Compared with 2 traditional risk scales, the Framingham stroke risk profiles and the Chinese Multiprovincial Cohort Study, our proposed model achieved better performance on the independent validation set, and the area under the receiver operating characteristic value increased by 0.17. Further nonlinear effect analysis revealed the importance of multitemporal trend characteristics in stroke risk prediction, which will benefit the standardized management of hypertensive patients.

Conclusions: A high-precision 3-year stroke risk prediction model for hypertensive patients was established, and the model's performance was verified by comparing it with the traditional risk scales. Multitemporal trend characteristics played an important role in stroke onset, and thus the model could be deployed to electronic health record systems to assist in more pervasive, preemptive stroke risk screening, enabling higher efficiency of early disease prevention and intervention.

(JMIR Med Inform 2021;9(11):e30277) doi:10.2196/30277

KEYWORDS

stroke; medical big data; electronic health records; machine learning; risk prediction; hypertension

Introduction

Stroke is the third leading cause of death globally, and China has become the country with the highest lifetime risk of stroke (39.3%) worldwide [1,2]. According to the China Stroke Report 2019, stroke has been the leading cause of death and disability among Chinese adults, and the incidence shows a younger trend. In 2018, the number of deaths from cerebrovascular diseases reached 1.57 million, accounting for 22% of the deaths of Chinese residents [3]. The damage caused by stroke is often irreversible, and stroke is prone to recur, with an annual recurrence rate of 3%-5%, and the condition aggravates with the increasing number of recurrences. However, stroke is preventable and controllable, and early intervention of modifiable risk factors can effectively reduce the occurrence and death of stroke [4].

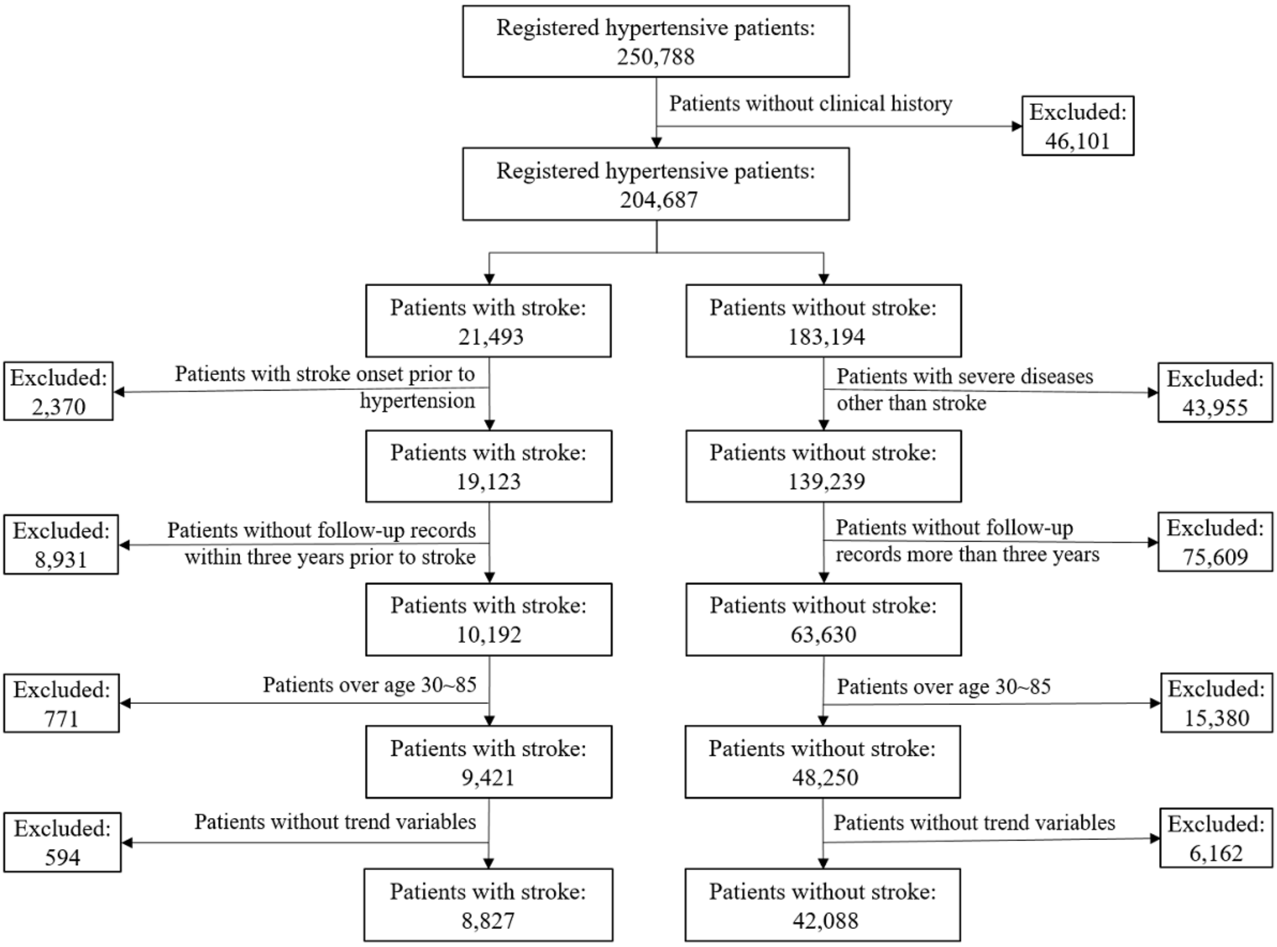
The pathogenesis of stroke is complicated and often results from the synergistic effect of various risk factors [5]. The known risk factors include gender, age, race, hypertension, diabetes, hyperlipidemia, systolic blood pressure (SBP), smoking, atrial fibrillation, etc. In recent years, studies have been discovering or proposing new risk factors of stroke, such as lipoprotein [6], triglyceride-glucose index [7], obstructive sleep apnea [8], vascular profile [9], heart failure [10], sleep disturbances [11], cerebral microbleeds [12], diet [13], imaging biomarkers [14], genetics [15], and environment [16].

Stroke risk assessment is an effective means to identify high-risk groups, and various well-known risk assessment scales have been established, such as the Framingham Stroke Risk Profile (FSRP) [17,18], SCORE-based fatal cardiovascular disease risk model [19,20], QStroke [21], pooled cohort risk equation (PCE) for atherosclerotic cardiovascular disease (ASCVD) [22,23], CHADS2 [24], CHA2DS2-VASc [25], HAS-BLED [26], and ATRIA [27]. However, the risk factors for stroke vary slightly by region and race [28,29], and these scales are mostly based on European and American populations, which tend to overestimate the risk of the Chinese population [30,31]. Some scales, such as the acute cardiovascular events risk model based on the Chinese Multiprovincial Cohort Study (CMCS) [32], the ASCVD risk model based on the China-PAR project [33,34], and a stroke risk model among adults in Taiwan [35], have also been estimated based on the Chinese population, but they have not been widely used. Moreover, these models are established based on long-term prospective studies, which are time-consuming and labor-intensive.

With the widespread application of electronic medical record (EMR) systems, a massive amount of medical data have been accumulated, which provides a fast, cost-efficient approach to collecting large-scale samples for retrospective studies. Medical big data has been demonstrated to promote medical applications such as discovering disease risk factors and prognosis, but it has also attracted extensive concerns [36-38]. Retrospective studies based on EMRs face enormous challenges, the most important of which is a large amount of missing data. How to construct effective features, especially those of medical significance, is crucial to building high-precision risk models, and the prevalence of machine learning provides interesting tools to optimize the modeling process.

In this study, we started from the substantial historical stock EMRs of registered hypertensive patients in Shenzhen and aimed to establish a high-precision stroke risk prediction model through medical big data and machine learning. A total of 250,788 registered hypertensive patients were collected, of which 21,493 developed stroke during the 3-year follow-up. After strict screening, only 57,671 samples were selected for risk modeling, as shown in Figure 1. We constructed characteristics from the multitemporal EMRs, established 3-year stroke risk prediction models based on 4 machine learning algorithms, and compared performance with well-known risk assessment scales. Finally, we analyzed the nonlinear correlation between continuous variables and the occurrence of stroke. Our study revealed the important role of multitemporal trend characteristics in improving the performance of stroke risk prediction models, which will benefit the standardized management of hypertensive patients.

Figure 1. The screening process of study population.


Registered hypertensive patients: 250,788
Patients without clinical history
Excluded: 46,101
Registered hypertensive patients: 204,687
Patients with stroke: 21,493
Patients without stroke: 183,194
Excluded: 2,370
Patients with stroke onset prior to hypertension
Patients with severe diseases other than stroke
Excluded: 43,955
Patients with stroke: 19,123
Patients without stroke: 139,239
Excluded: 8,931
Patients without follow-up records within three years prior to stroke
Patients without follow-up records more than three years
Excluded: 75,609
Patients with stroke: 10,192
Patients without stroke: 63,630
Excluded: 771
Patients over age 30~85
Patients over age 30~85
Excluded: 15,380
Patients with stroke: 9,421
Patients without stroke: 48,250
Excluded: 594
Patients without trend variables
Patients without trend variables
Excluded: 6,162
Patients with stroke: 8,827
Patients without stroke: 42,088

Methods

Data Resource and Study Population

The data used in this study are the electronic health records from the Shenzhen Health Information Big Data Platform, which has access to more than 4000 health institutions, including 85 hospitals and over 650 community health service centers. The platform covered medical service records, including disease management, outpatient service, hospitalization, laboratory test, imaging examination, and physical examination. Disease management covers patients with hypertension, diabetes, cancer, etc, who are registered and regularly followed up. At present, the platform has more than 600 million EMRs from 2010 to 2020. Medical records among different institutions of the same patient can be associated with a unique personal identification number. Since medical records were collected in routine clinical activities, patients had agreed and authorized their use during the consultation process. According to the Guidelines of the WMA Declaration of Helsinki, the study was approved by the SIAT IRB (SIAT-IRB-151115-H0084).

Hypertension is the primary risk factor for stroke. Moreover, hypertensive patients are the key population of disease management, and thus long-term physical examination results have been accumulated, which are essential data for stroke risk prediction. This study focused on registered hypertensive patients and aimed to establish a high-precision stroke risk prediction model. A total of 250,788 hypertensive patients were collected from the platform, with an average follow-up of 4.5 years. The stroke diagnosis was extracted from the main diagnosis fields of the outpatient or inpatient records according to the International Statistical Classification of Diseases and Related Health Problems 10th Revision diagnostic codes [39], including I60 (subarachnoid hemorrhage), I61 (intracerebral hemorrhage), I62 (other nontraumatic intracranial hemorrhages), I63 (cerebral infarction) and I64 (stroke, not specified as hemorrhage or infarction), and excluded I69 (sequelae of cerebrovascular disease). Finally, there were 21,493 cases of stroke onset, and the date of the first occurrence of a stroke diagnosis in the clinical records was taken as the date of stroke diagnosis.

We limited the study to patients with at least one outpatient or hospitalization record to ensure the reliability of outcomes, and thus 46,101 patients were excluded. We excluded patients with stroke (positive cases), those with stroke prior to hypertension, and those without follow-up records within 3 years before stroke onset. In addition, patients without stroke (negative cases), those with heart disease, renal failure, or tumor, and those without more than 3 years of follow-up records were also excluded. In addition, patients were limited to 30-85 years old. As a result, 57,671 hypertensive patients were included in the study, of which 9421 patients had a stroke within 3 years of follow-up. Moreover, patients were required to have trend change variables (eg, mean SBP), and thus 6756 patients were excluded. The detailed screening process of the study population is shown in Figure 1.

Feature Extraction

The medical records of 57,671 samples were extracted from the platform, including resident information, lifestyle, family history, follow-up records with registered hypertensive patients, outpatient and hospitalization records, and laboratory test results. Medical records were collected from hundreds of health institutions with slightly different medical service systems, resulting in diverse data formats, poor data quality, and even a large number of missing fields. We first performed a series of cleaning operations on the medical records, including deleting outliers or replacing them with null values, unit unification of test results, and drug classification.

Given this is a retrospective study based on real-world multitemporal medical data, the event endpoint and baseline needed to be predefined before feature extraction. For positive cases, the endpoint was the date of stroke diagnosis, and baseline was defined as the date of the first follow-up record within 3 years prior to the endpoint. For negative cases, the endpoint was the date of the last medical service record, and baseline was defined as the date of the last follow-up record 3 years before the endpoint. The physiological parameters in the follow-up record at baseline were extracted as characteristics, such as age, SBP, diastolic blood pressure (DBP), pulse pressure difference (PPD; the difference between SBP and DBP), heart rate (HR), BMI, glucose.

Secondly, trend characteristics of physiological parameters based on multitemporal follow-up records before baseline were specially constructed, such as SBP, DBP, PPD, HR, BMI, and glucose. The follow-up records were grouped by patients and sorted in ascending order of follow-up date, and the difference in the two consecutive records was calculated, which were marked with *_delta. Moreover, the maximum, minimum, mean, and derivation of physiological parameters of each patient and their differences were calculated.

Thirdly, historical symptoms were extracted from the outpatient and hospitalization records before baseline. In this study, only some symptoms that are potentially associated with stroke attack were extracted, such as diabetes, hyperlipemia, sleep disorder, etc. Moreover, demographic characteristics (ie, gender), family disease history (ie, family history of coronary heart disease or FAM_CHD), lifestyles (ie, smoking and drinking), and drug categories (ie, antihypertensive drug use) were extracted. The features were binarized based on their existence.

Finally, laboratory test results were extracted. According to statistics, less than 10% of patients had laboratory test records near the baseline. For the purpose of comparing model prediction performance with existing scales, only necessary blood lipid tests were extracted, including triglycerides, total cholesterol, low-density lipoprotein cholesterol, and high-density lipoprotein cholesterol (HDL-C).

Proper feature selection is beneficial to improve the performance of the model. First, features with missing values above 30% were removed. Then, correlation analysis and univariate trend analysis were adopted to remove redundant features, and a two-tailed *P* value <.05 was considered a significant correlation. In addition, some features of existing research were manually retained.

Prediction Modeling

An ensemble method extreme boosting gradient (XGBoost) [40] was used to establish a 3-year stroke risk prediction model for hypertensive patients and compared with the other 3 widely-used traditional machine learning algorithms, including logistic regression [41], support vector machine (SVM) [42], and random forest [43].

XGBoost is an integration algorithm based on multiple decision trees under the gradient boosting framework. Unlike traditional gradient boosting decision trees, XGBoost supports column sampling, which can reduce overfitting and calculation. In addition, XGBoost considers a sparse matrix and can automatically learn its splitting direction for samples with missing values.

Logistic regression is a classical classification algorithm widely used in epidemiology and medicine, such as risk factor discovery, disease risk prediction, and automatic disease diagnosis. Logistic regression is a generalized linear regression model that introduces the sigmoid function to normalize dependent variables, thus making it more focused on the classification boundaries and increasing its robustness.

SVM is a bicategorical algorithm, which is characterized by the ability to minimize empirical errors and maximize geometric edge regions at the same time. SVM also includes nuclear techniques, which makes it a substantial nonlinear classifier. In addition, the stability and sparsity of SVM give it good generalization capability.

Random forest is also an ensemble algorithm based on decision trees, which determines the final prediction by combining the outcome of multiple weak classifiers. In random forests, the base classifiers are trained independently, so the learning process is very fast. Moreover, random forests have the advantages of evaluating the importance of variables and resisting overfitting and supporting column sampling and missing values by default.

According to a ratio of 7:3, we randomly divided the data set into training and test sets with balanced positive and negative cases. We performed 5-fold cross-validation on the training set and validated the performance of the models on the test set. Five evaluation criteria were used to validate the models, including the area under the receiver operating characteristic curve (AUC), accuracy, recall, specificity, and F1-score. For continuous features, the missing values were filled with the mean of each feature, and the data were standardized by the mean and variance of the feature.

All the experiments were performed under the environment manager Anaconda of the Linux server in the isolated intranet, and a Python3.6.5 kernel was used for data processing and modeling. We implemented 4 algorithms using the Scikit-learn library in the Python programming environment [44].

Results

Characteristics Description

A total of 50,915 registered patients with hypertension were screened into the study cohort, and 8827 patients developed stroke within 3-year follow-up. In the study cohort, the positive/negative ratio was about 1:4.7, and the age distribution was different, as depicted in Figure 2.

In order to balance positive and negative cases, we performed a random stratified sampling of negative cases according to gender ratio and age stratification of positive cases. Age was stratified into 30 to 40, 40 to 50, 50 to 60, 60 to 70, and 70 to 85 years, and the proportion of negative to positive cases in gender and age stratification was calculated. We took the minimum proportion as the sampling rate and randomly selected the corresponding number of samples from the negative cases of each group. After stratified sampling, 11,126 negative cases and 8827 positive cases were used for modeling, and the gender and age distribution are depicted in Table 1.

Figure 2. Age distribution of stroke and nonstroke patients.

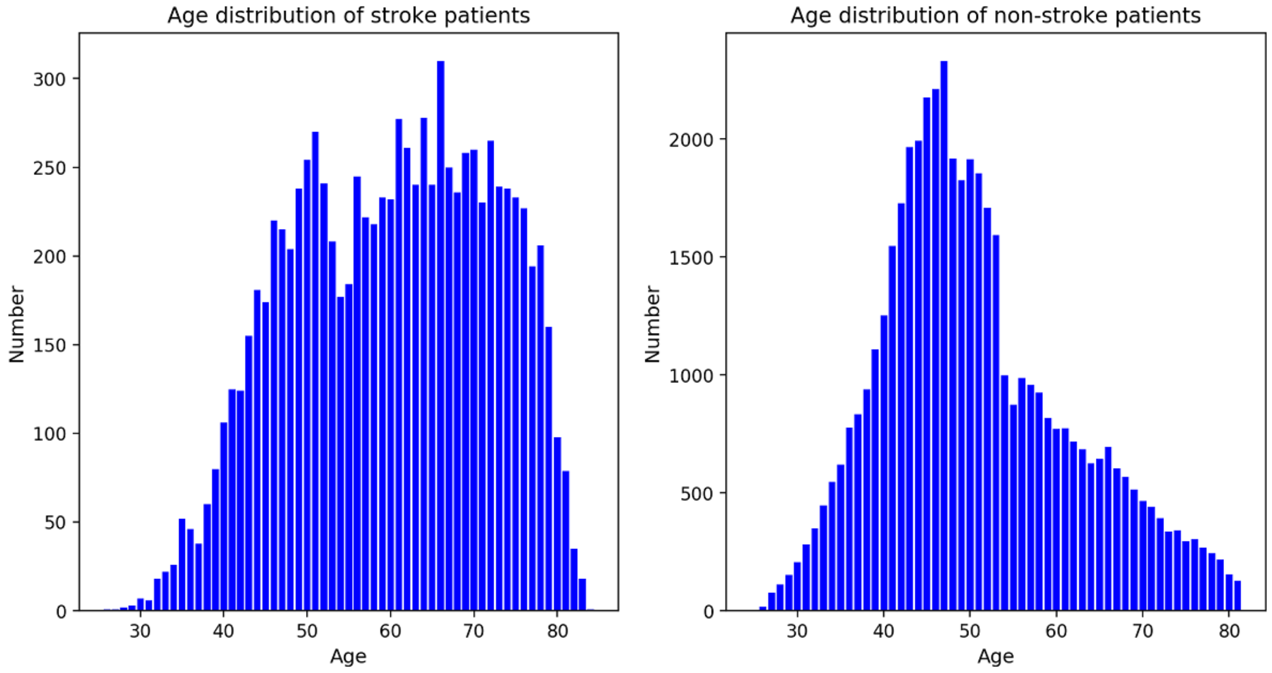


Table 1. Gender and age distribution before and after stratified sampling.

Characteristics	Positive cases (N=8,827), n (%)	Negative cases (N=42,088), n (%)	Negative cases after sampling N=11,126), n (%)
Gender, male	5251 (59.49)	25990 (61.75)	6174 (55.49)
Age, years			
30-40	414 (4.69)	5843 (13.88)	522 (4.69)
40-50	1746 (19.78)	17342 (41.20)	2204 (19.81)
50-60	2104 (23.84)	10415 (24.75)	2656 (23.87)
60-70	2462 (27.89)	5448 (12.94)	3108 (27.93)
70-85	2088 (23.65)	2636 (6.26)	2636 (23.69)

A total of 77 features were extracted from the medical records, and eventually, 49 features were used as input for the machine learning algorithms. Blood lipid test results were not included because the missing ratio was more than 80%. Table 2 shows the statistical distribution of partial features of higher correlation (*P* value less than .01).

Table 2. Distribution of the basic characteristics.

Characteristics	Positive cases (N=8,827)	Negative cases (N=11,126)	*P* value[a]
Demographics			
Gender, n (%), male	5,251 (59.49)	6174 (55.49)	<.001
Age, mean (SD), years	60.21 (11.88)	59.73 (11.94)	.005
Years_after_hypertension, mean (SD), years	6.25 (5.64)	6.78 (5.27)	<.001
Lifestyle (current or previous), n (%)			
Smoking	768 (8.70)	1233 (11.08)	<.001
Drink	1000 (11.33)	1643 (14.77)	<.001
Family history, n (%)			
FAM_hypertension	239 (2.71)	489 (4.40)	<.001
FAM_diabetes	57 (0.65)	116 (1.04)	.002
Physical examination, mean (SD)			
SBP[b], mmHg	133.76 (13.42)	131.33 (10.02)	<.001
DBP[c], mmHg	81.93 (9.56)	80.17 (7.45)	<.001
PPD[d], mmHg	52.16 (10.59)	51.15 (8.81)	<.001
Trend characteristics, mean (SD)			
N_followup_1year	4.13 (3.66)	5.89 (3.84)	<.001
SBP_max, mmHg	140.29 (14.56)	142.77 (13.46)	<.001
SBP_min, mmHg	127.17 (14.09)	122.81 (10.21)	<.001
SBP_mean mmHg	133.20 (11.58)	131.75 (7.90)	<.001
DBP_max, mmHg	86.47 (9.69)	89.10 (8.51)	<.001
DBP_min, mmHg	76.67 (10.16)	73.42 (7.36)	<.001
DBP_mean, mmHg	81.35 (8.17)	80.71 (5.80)	<.001
PPD_max, mmHg	58.41 (12.02)	61.48 (10.83)	<.001
PPD_min, mmHg	46.01 (11.05)	41.69 (8.28)	<.001
PPD_mean, mmHg	51.89 (8.75)	51.04 (6.26)	<.001
HR[e]_max, times/min	78.57 (7.08)	79.57 (7.11)	<.001
HR_min, times/min	74.29 (6.68)	72.97 (5.97)	<.001
SBP_delta_mean, mmHg	4.37 (3.53)	4.01 (3.17)	<.001
DBP_delta_mean, mmHg	3.46 (2.44)	3.24 (2.10)	<.001
PPD_delta_mean, mmHg	4.30 (3.08)	4.04 (2.68)	<.001
HR_delta_mean, times/min	1.23 (1.83)	1.08 (1.51)	<.001
Medical history, n (%)			
Prior cardiovascular diseases	176 (1.99)	11 (0.1)	<.001
Atrial fibrillation	53 (0.6)	16 (0.14)	<.001
Atherosclerosis	488 (5.53)	358 (3.22)	<.001
sleep disorder	99 (1.12)	475 (4.27)	<.001
Dizziness and headache	1094 (12.39)	1804 (16.21)	<.001
Malaise and fatigue	6 (0.07)	55 (0.49)	<.001
Giddiness	9 (0.10)	55 (0.49)	<.001
Migraine	7 (0.08)	38 (0.34)	<.001
Antihypertensive treatment	8551 (96.87)	10905 (98.01)	<.001

Characteristics	Positive cases (N=8,827)	Negative cases (N=11,126)	*P* value[a]
Lipid-lowering drug	1123 (12.72)	1046 (9.40)	<.001

[a] Pearson chi-square test was applied.

[b] SBP: systolic blood pressure.

[c] DBP: diastolic blood pressure.

[d] PPD: pulse pressure difference.

[e] HR: heart rate.

Predictive Performance Evaluation

According to the ratio of 7:3, the data set was randomly divided into a training set (N=13,967) and test set (N=5986), and the ratio of positive to negative cases was balanced (ratio=1:1.26). Table 3 shows the performance of the 4 algorithms on the test set. The tree-integration algorithm XGBoost achieved the best performance with AUC of 0.9220, followed by random forest with AUC of 0.8956. Logistic regression had the worst performance with AUC of 0.8544, as shown intuitively from the receiver operating characteristic (ROC) curve in Figure 3.

Table 3. Model performance of four different algorithms.

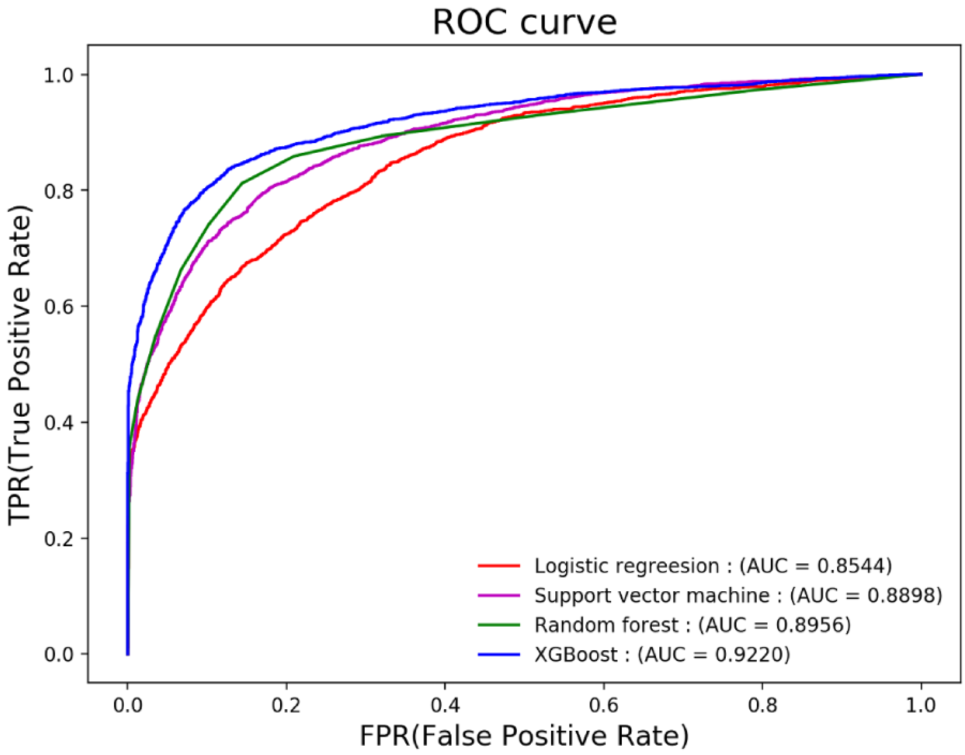
Methods	AUC[a]	Accuracy	Recall	F1-score	Specificity
Logistic regression	0.8544	0.7726	0.7141	0.7354	0.8191
SVM[b]	0.8898	0.8112	0.7844	0.7861	0.8325
Random forest	0.8956	0.8343	0.8157	0.8133	0.8490
XGBoost[c]	0.9220	0.8478	0.8512	0.8319	0.8451

[a] AUC: area under the receiver operating curve.

[b] SVM: support vector machine.

[c] XGBoost: extreme gradient boosting.

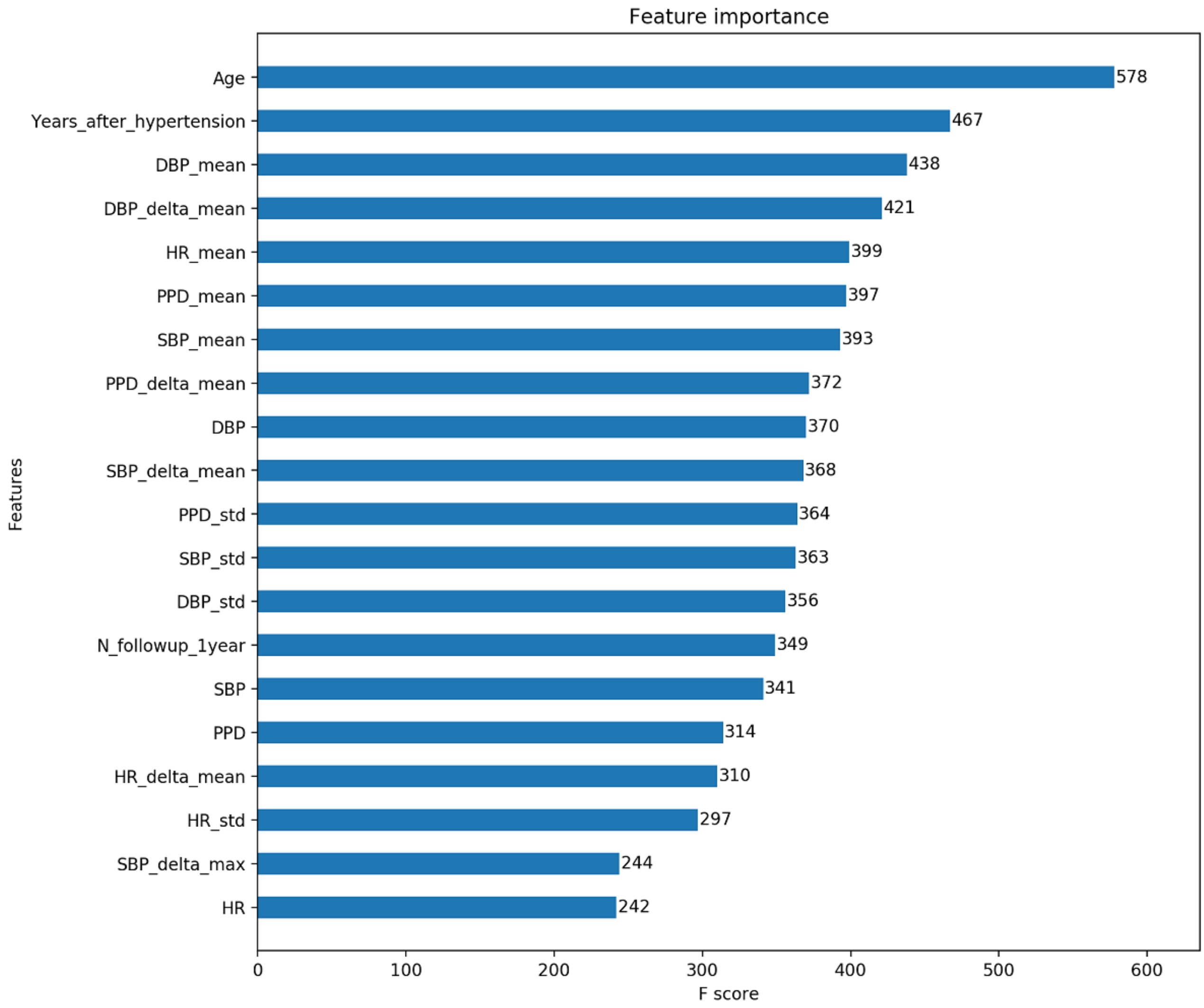
Figure 3. The receiver operating characteristic curve of the four algorithms.



Features Importance

Feature importance measures the relative contribution of the features to modeling. The top 20 features are depicted in Figure 4. In addition to the traditional risk factors contained in well-known scales, the trend characteristics of physiological parameters also played an important role in modeling, such as PPD, HR_mean, and PPD_delta_mean. The feature PPD could reflect the change of vascular elasticity, and when PPD is too large or too small, the disease's hidden danger would be indicated and should be addressed. In addition, the mean of the difference between 2 adjacent follow-up records could reflect the control level of physiological parameters, which are easily obtained in daily monitoring and promote the health management of hypertensive patients.

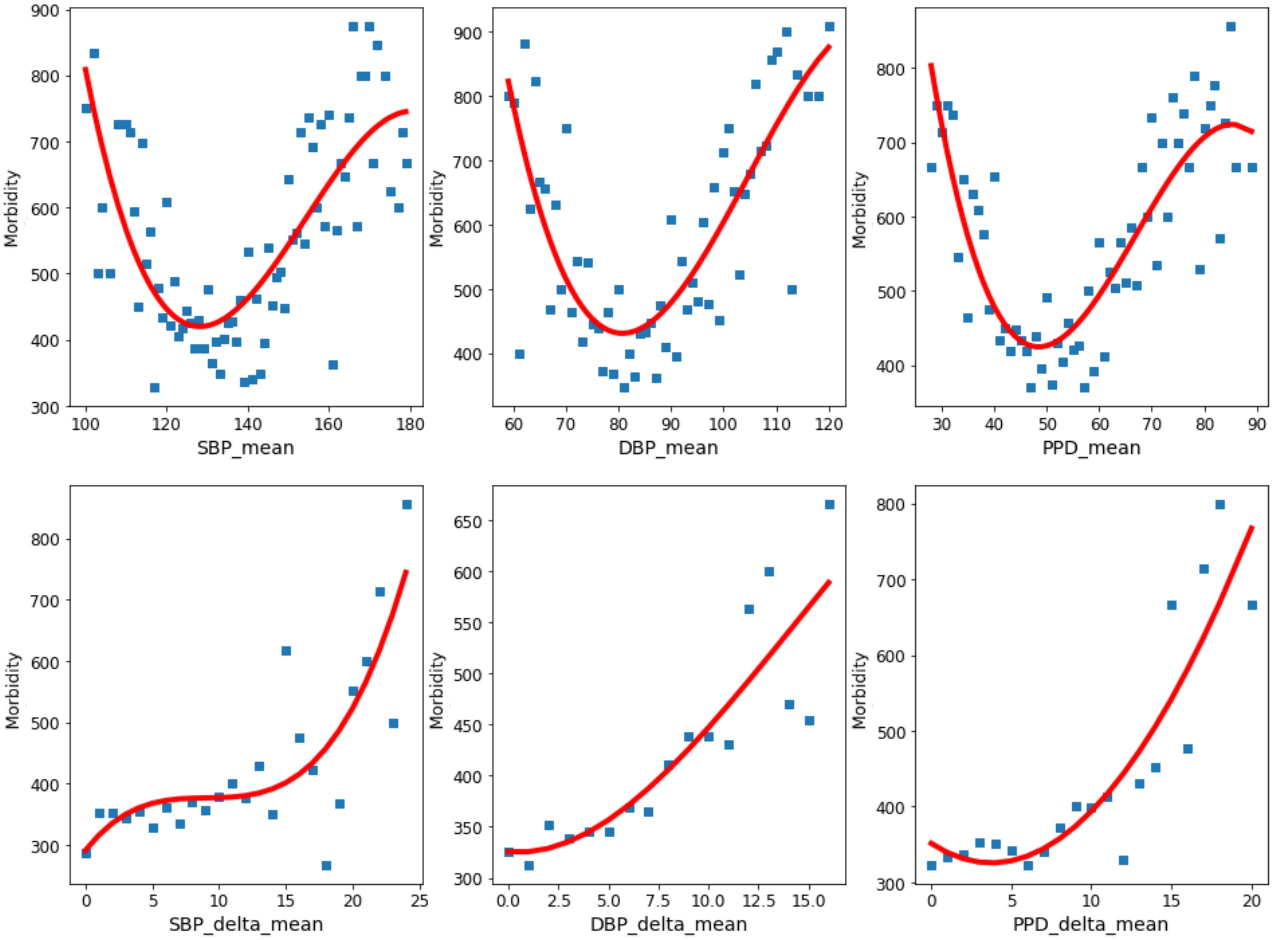
Figure 4. Features of the top 20 importance in XGBoost model. DBP: diastolic blood pressure; HR: heart rate; PPD: pulse pressure difference; SBP: systolic blood pressure; XGBoost: extreme gradient boosting.



Nonlinear Effects of Continuous Features

We performed a univariate trend analysis of continuous features based on the 3-year risk prediction data set to analyze the effect of characteristics on stroke occurrence further. Morbidity was defined as the number of stroke cases in a thousand samples under a characteristic value, and the relationship between the morbidity and characteristic values was fitted. In this study, we chose Gaussian, polynomial, and exponential functions to fit the curve, and the fitting effect was evaluated by discriminant coefficient R^2 [45]. Figure 5 showed the nonlinear effects of 6 features, which were the top modifiable risk factors in the feature importance of Figure 4. We found that the effect of some factors (eg, SBP_mean, DBP_mean, HR_mean, and PPD_mean) formed a U-shaped trend, where the marginal risk was minimized when the factor fell within a given range while increasing both when it went lower or higher. Unsurprisingly, the turn-points for the 3 factors were highly consistent with the blood pressure control targets of the latest hypertension guidelines. On the other hand, the effects of DBP_delta_mean and PPD_delta_mean formed a hinge-like sharp, which revealed the importance of stable blood pressure for stroke prevention in hypertension patients.

Figure 5. Nonlinear effect of six continuous features on the morbidity of stroke. DBP: diastolic blood pressure; PPD: pulse pressure difference; SBP: systolic blood pressure.



Discussion

Principal Findings

We had developed a high-precision risk prediction model of stroke for hypertensive patients based on large-scale electronic health records from a regional medical information platform and validated the prediction performance on an independent test set. The integrated tree-based XGBoost algorithm achieved the best prediction performance with an AUC of 0.9220 and outperformed the other 3 traditional algorithms. Besides the traditional risk factors, such as age, gender, SBP, smoking, diabetes, and antihypertensive drug use, we specially constructed several changing-trend variables from multitemporal medical records, which were confirmed to be nonlinearly correlated with stroke onset. The effect of nonlinear correlation justified the necessity of adopting sophisticated nonlinear machine learning models over traditional linear regressions. Furthermore, with nonlinear ensemble algorithms such as XGBoost used in this study, there was no need to select variables in advance even when the number of potential variables was large, which was different from most traditional clinical studies and enabled the identification of novel biomarkers with both linear and nonlinear effects during modeling process through mining large-scale population data. This was an advantage brought by big data technologies.

Comparison With Traditional Statistical Models

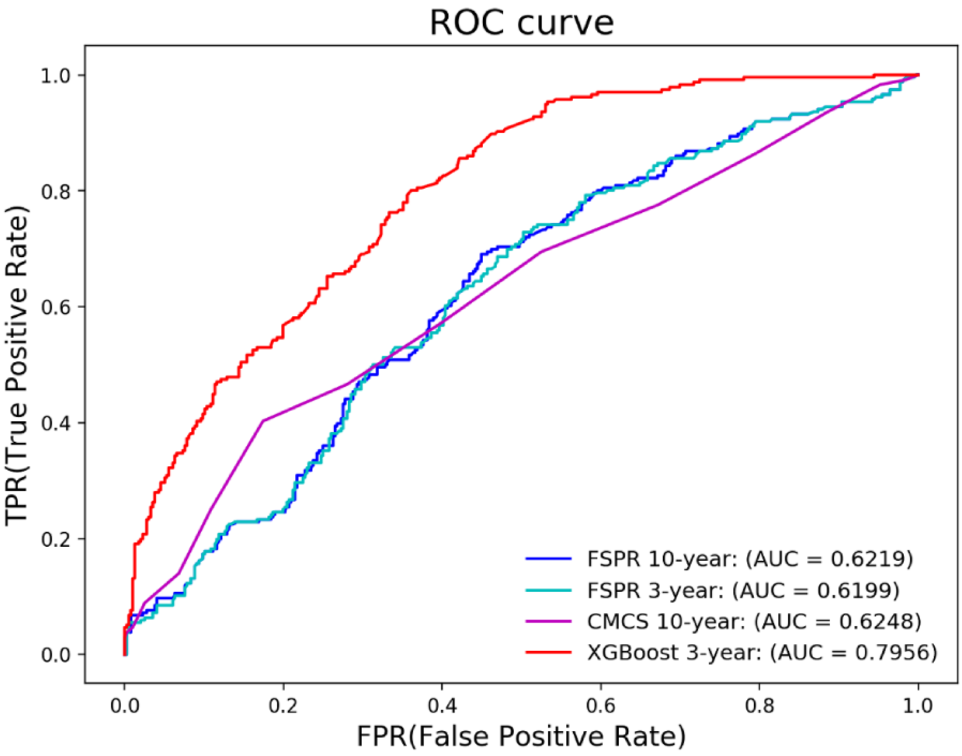
Several risk models based on long-period prospective studies have been widely used to screen high-risk populations, such as Framingham studies, QStroke, and PCE. Considering the target events and wide application of the models, we selected to compare the model's performance based on XGBoost with the revised FSRP [20] and CMCS risk scale [32].

The FSRP, originally described in 1991 [19], had been validated in other cohorts, was recommended by the American Heart Association. The study population was between 55 and 84 years old. However, the profile had been demonstrated by several studies to overestimated risk; therefore, the profile was updated in 2017. The revised FSRP better predicted current stroke risk in 3 large community samples, integrating gender, age, current smoking habits, prevalent cardiovascular disease (including myocardial infarction, angina, coronary insufficiency, intermittent claudication, and congestive heart failure), atrial fibrillation, diabetes, SBP, and antihypertensive treatment. Moreover, the profile provided a multiyear prediction model for 10 years. In this study, we selected the 3-year and 10-year models to compare the performance.

The CMCS risk scale was a 10-year risk prediction model of acute cardiovascular event (acute coronary heart disease and acute stroke) proposed in 2003. The study population was aged 35 to 64 years living in 11 provinces and cities of China. The risk factors used in the model included gender, age, diabetes, smoking, SBP, total cholesterol, and HDL-C.

We screened a subset of 632 samples from 76,494 samples that simultaneously met the FSRP and CMCS profile, of which 236 had stroke onset. These samples were assigned to the test set in the first step of our model-building process. Figure 6 depicts the ROC curve achieved by the 4 models. The developed model based on XGBoost achieved a higher performance with an AUC of 0.7956, and there was no significant difference between the other 3 scales.

Figure 6. Receiver operating characteristic curve compared with three traditional risk scales. AUROC: area under the receiver operating characteristic; CMCS: Chinese Multi-provincial Cohort Study; FSRP: Framingham Stroke Risk Profile; XGBoost: extreme gradient boosting.



Limitations and Future Research

This work was a retrospective study based on historical stock data collected at different periods. There were a large number of missing values in characteristic variables, which may affect the sample population size and the performance of the model. In addition, due to the insufficiency of laboratory test results, the established model did not include the biochemical indicators in the traditional scales, such as TC and HDL-C. However, the impact of missing information was equal for both the positive and negative cases so that no significant biases were likely to be introduced through missing data. Compared with the benefits obtained by the enlarged population and the abundance of clinical features, the data's increased noise was considered acceptable. In addition, the study cohort was imbalanced in view of the numbers of positive cases and negative cases. We performed randomly stratified sampling according to gender ratio and age stratification, which may not represent the rest of the patients accurately. We are currently accumulating longer periods of medical data as well as a larger population and trying to further validate and improve the model with recent data.

Conclusions

We established a high-precision 3-year stroke risk prediction model for hypertensive patients based on large-scale EMRs and verified that the proposed model could perform better than traditional risk scales. In addition, the features in the model are routinely accessible data, so the model could be easily implemented in EMR systems to help with a more pervasive, preemptive screening of stroke risk, enabling higher efficiency of early disease prevention and intervention.

Acknowledgments

This work was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (grant number XDB38040200) and the Shenzhen Science and Technology Program (grant numbers KQTD2019092917283566, JCYJ20180703145202065, JCYJ20180703145002040, and 20170502171625936).

Authors' Contributions

YY contributed to experimental design and method, data collection, analysis and interpretation of the data, machine learning algorithms, and draft preparation. ZJ contributed to experimental design and method, data collection, and analysis and interpretation of the data. DZ contributed to the experimental design and method and machine learning algorithms. LY contributed to the experimental design and method. CY contributed to the conception, experimental design and method, and analysis and interpretation of the data. All the authors have reviewed and agreed to the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Roth, Gregory A. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1736-1788 [FREE Full text] [doi: 10.1016/S0140-6736(18)32203-7] [Medline: 30496103]
2. 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. Lancet 2019 Sep 28;394(10204):1145-1158 [FREE Full text] [doi: 10.1016/S0140-6736(19)30427-1] [Medline: 31248666]
3. Wang L, Liu J, Yang Y, Peng B. The Prevention and Treatment of Stroke Still Face Huge Challenges-Brief Report on Stroke Prevention and Treatment in China, 2018. Chinese Circulation Journal 2019;34:105-119. [doi: 10.3969/j.issn.1000-3614.2019.02.001]
4. Yusuf S, Joseph P, Rangarajan S, Islam S, Mente A, Hystad P, et al. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. Lancet 2020 Mar 07;395(10226):795-808 [FREE Full text] [doi: 10.1016/S0140-6736(19)32008-2] [Medline: 31492503]
5. Boehme AK, Esenwa C, Elkind MSV. Stroke Risk Factors, Genetics, and Prevention. Circ Res 2017 Feb 03;120(3):472-495 [FREE Full text] [doi: 10.1161/CIRCRESAHA.116.308398] [Medline: 28154098]
6. Nave AH, Lange KS, Leonards CO, Siegerink B, Doehner W, Landmesser U, et al. Lipoprotein (a) as a risk factor for ischemic stroke: a meta-analysis. Atherosclerosis 2015 Oct;242(2):496-503. [doi: 10.1016/j.atherosclerosis.2015.08.021] [Medline: 26298741]
7. Shi W, Xing L, Jing L, Tian Y, Yan H, Sun Q, et al. Value of triglyceride-glucose index for the estimation of ischemic stroke risk: Insights from a general population. Nutr Metab Cardiovasc Dis 2020 Feb 10;30(2):245-253. [doi: 10.1016/j.numecd.2019.09.015] [Medline: 31744716]
8. King S, Cuellar N. Obstructive Sleep Apnea as an Independent Stroke Risk Factor: A Review of the Evidence, Stroke Prevention Guidelines, and Implications for Neuroscience Nursing Practice. J Neurosci Nurs 2016 Jun;48(3):133-142. [doi: 10.1097/JNN.0000000000000196] [Medline: 27136407]
9. Rutten-Jacobs LCA, Markus HS, UK Young Lacunar Stroke DNA Study. Vascular Risk Factor Profiles Differ Between Magnetic Resonance Imaging-Defined Subtypes of Younger-Onset Lacunar Stroke. Stroke 2017 Sep;48(9):2405-2411 [FREE Full text] [doi: 10.1161/STROKEAHA.117.017813] [Medline: 28765289]
10. Kim W, Kim EJ. Heart Failure as a Risk Factor for Stroke. J Stroke 2018 Jan;20(1):33-45 [FREE Full text] [doi: 10.5853/jos.2017.02810] [Medline: 29402070]
11. Koo DL, Nam H, Thomas RJ, Yun C. Sleep Disturbances as a Risk Factor for Stroke. J Stroke 2018 Jan;20(1):12-32 [FREE Full text] [doi: 10.5853/jos.2017.02887] [Medline: 29402071]
12. Wilson D, Ambler G, Lee K, Lim J, Shiozawa M, Koga M, Microbleeds International Collaborative Network. Cerebral microbleeds and stroke risk after ischaemic stroke or transient ischaemic attack: a pooled analysis of individual patient data from cohort studies. Lancet Neurol 2019 Jul;18(7):653-665 [FREE Full text] [doi: 10.1016/S1474-4422(19)30197-8] [Medline: 31130428]
13. Abdollahi AM, Virtanen HEK, Voutilainen S, Kurl S, Tuomainen T, Salonen JT, et al. Egg consumption, cholesterol intake, and risk of incident stroke in men: the Kuopio Ischaemic Heart Disease Risk Factor Study. Am J Clin Nutr 2019 Jul 01;110(1):169-176. [doi: 10.1093/ajcn/nqz066] [Medline: 31095282]
14. Saba L, Saam T, Jäger HR, Yuan C, Hatsukami TS, Saloner D, et al. Imaging biomarkers of vulnerable carotid plaques for stroke risk prediction and their potential clinical implications. Lancet Neurol 2019 Jun;18(6):559-572. [doi: 10.1016/S1474-4422(19)30035-3] [Medline: 30954372]
15. Lee Y, Chung C, Chang M, Wang S, Liao Y. cysteine-altering variant is an important risk factor for stroke in the Taiwanese population. Neurology 2020 Jan 07;94(1):e87-e96. [doi: 10.1212/WNL.0000000000008700] [Medline: 31792094]
16. Graber M, Mohr S, Baptiste L, Duloquin G, Blanc-Labarre C, Mariet AS, et al. Air pollution and stroke. A new modifiable risk factor is in the air. Rev Neurol (Paris) 2019 Dec;175(10):619-624. [doi: 10.1016/j.neurol.2019.03.003] [Medline: 31153597]

17. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham Study. Stroke 1991 Mar;22(3):312-318. [doi: 10.1161/01.str.22.3.312] [Medline: 2003301]
18. Dufouil C, Beiser A, McLure LA, Wolf PA, Tzourio C, Howard VJ, et al. Revised Framingham Stroke Risk Profile to Reflect Temporal Trends. Circulation 2017 Mar 21;135(12):1145-1159 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.021275] [Medline: 28159800]
19. Conroy RM. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. European Heart Journal 2003 Jun;24(11):987-1003. [doi: 10.1016/s0195-668x(03)00114-3] [Medline: 12788299]
20. Tillmann T, Läll K, Dukes O, Veronesi G, Pikhart H, Peasey A, et al. Development and validation of two SCORE-based cardiovascular risk prediction models for Eastern Europe: a multicohort study. Eur Heart J 2020 Sep 14;41(35):3325-3333 [FREE Full text] [doi: 10.1093/eurheartj/ehaa571] [Medline: 33011775]
21. Hippisley-Cox J, Coupland C, Brindle P. Derivation and validation of QStroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study. BMJ 2013 May 02;346:f2573 [FREE Full text] [doi: 10.1136/bmj.f2573] [Medline: 23641033]
22. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, American COCHATFOPG. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 2014 Jul 01;63(25 Pt B):2935-2959 [FREE Full text] [doi: 10.1016/j.jacc.2013.11.005] [Medline: 24239921]
23. Rana JS, Tabada GH, Solomon MD, Lo JC, Jaffe MG, Sung SH, et al. Accuracy of the Atherosclerotic Cardiovascular Risk Equation in a Large Contemporary, Multiethnic Population. J Am Coll Cardiol 2016 May 10;67(18):2118-2130 [FREE Full text] [doi: 10.1016/j.jacc.2016.02.055] [Medline: 27151343]
24. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. JAMA 2001 Jun 13;285(22):2864-2870. [doi: 10.1001/jama.285.22.2864] [Medline: 11401607]
25. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest 2010 Feb;137(2):263-272. [doi: 10.1378/chest.09-1584] [Medline: 19762550]
26. Pisters R, Lane DA, Nieuwlaat R, de Vos CB, Crijns HJGM, Lip GYH. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. Chest 2010 Nov;138(5):1093-1100. [doi: 10.1378/chest.10-0134] [Medline: 20299623]
27. Singer DE, Chang Y, Borowsky LH, Fang MC, Pomernacki NK, Udaltsova N, et al. A new risk scheme to predict ischemic stroke and other thromboembolism in atrial fibrillation: the ATRIA study stroke risk score. J Am Heart Assoc 2013 Jun 21;2(3):e000250 [FREE Full text] [doi: 10.1161/JAHA.113.000250] [Medline: 23782923]
28. Oikonomou E, Lazaros G, Georgiopoulos G, Christoforatou E, Papamikroulis GA, Vogiatzi G, et al. Environment and cardiovascular disease: rationale of the Corinthia study. Hellenic J Cardiol 2016;57(3):194-197 [FREE Full text] [doi: 10.1016/j.hjc.2016.06.001] [Medline: 27451913]
29. Bhatnagar A. Environmental Determinants of Cardiovascular Disease. Circ Res 2017 Jul 07;121(2):162-180 [FREE Full text] [doi: 10.1161/CIRCRESAHA.117.306458] [Medline: 28684622]
30. Jiang Y, Ma R, Guo H, Zhang X, Wang X, Wang K, et al. External validation of three atherosclerotic cardiovascular disease risk equations in rural areas of Xinjiang, China. BMC Public Health 2020 Sep 29;20(1):1471 [FREE Full text] [doi: 10.1186/s12889-020-09579-4] [Medline: 32993590]
31. Li J, Liu F, Yang X, Cao J, Chen S, Chen J, et al. Validating World Health Organization cardiovascular disease risk charts and optimizing risk assessment in China. Lancet Reg Health West Pac 2021 Mar;8:100096 [FREE Full text] [doi: 10.1016/j.lanwpc.2021.100096] [Medline: 34327424]
32. Wang W, Zhao D, Liu J. Prospective study on the predictive model of cardiovascular disease risk in a Chinese population aged 35-64. Chin J Cardiol 2003;31(12):902-908. [doi: 10.3760/j:issn:0253-3758.2003.12.006]
33. Yang X, Li J, Hu D, Chen J, Li Y, Huang J, et al. Predicting the 10-Year Risks of Atherosclerotic Cardiovascular Disease in Chinese Population: The China-PAR Project (Prediction for ASCVD Risk in China). Circulation 2016 Nov 08;134(19):1430-1440. [doi: 10.1161/CIRCULATIONAHA.116.022367] [Medline: 27682885]
34. Xing X, Yang X, Liu F, Li J, Chen J, Liu X, et al. Predicting 10-Year and Lifetime Stroke Risk in Chinese Population. Stroke 2019 Sep;50(9):2371-2378. [doi: 10.1161/STROKEAHA.119.025553] [Medline: 31390964]
35. Chien K, Su T, Hsu H, Chang W, Chen P, Sung F, et al. Constructing the prediction model for the risk of stroke in a Chinese population: report from a cohort study in Taiwan. Stroke 2010 Sep;41(9):1858-1864. [doi: 10.1161/STROKEAHA.110.586222] [Medline: 20671251]
36. George J, Majeed W, Mackenzie IS, Macdonald TM, Wei L. Association between cardiovascular events and sodium-containing effervescent, dispersible, and soluble drugs: nested case-control study. BMJ 2013 Nov 26;347:f6954 [FREE Full text] [doi: 10.1136/bmj.f6954] [Medline: 24284017]
37. Pike MM, Decker PA, Larson NB, St Sauver JL, Takahashi PY, Roger VL, et al. Improvement in Cardiovascular Risk Prediction with Electronic Health Records. J Cardiovasc Transl Res 2016 Jun;9(3):214-222 [FREE Full text] [doi: 10.1007/s12265-016-9687-z] [Medline: 26960568]

38. Lin H, Tang X, Shen P, Zhang D, Wu J, Zhang J, et al. Using big data to improve cardiovascular care and outcomes in China: a protocol for the CHinese Electronic health Records Research in Yinzhou (CHERRY) Study. BMJ Open 2018 Feb 12;8(2):e019698 [FREE Full text] [doi: 10.1136/bmjopen-2017-019698] [Medline: 29440217]
39. WHO. ICD-10 international statistical classification of diseases and related health problems tenth revision. World Health Organization 2010;2:2980 [FREE Full text]
40. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Xgboost: A scalable tree boosting system// Proceedings of the 22nd ACM SIGKDD international conference on knowledge discoverydata mining; 2016 Presented at: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016; San Francisco, CA p. A. [doi: 10.1145/2939672.2939785]
41. Hosmer JDW, Lemeshow S, Sturdivant R. Applied logistic regression, 3rd ed. In: Applied logistic regression, 3rd ed. Canada: John Wiley&Sons; 2013.
42. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/bf00994018]
43. Flaxman AD, Vahdatpour A, Green S, James SL, Murray CJ, Population Health Metrics Research Consortium (PHMRC). Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. Popul Health Metr 2011 Aug 04;9:29 [FREE Full text] [doi: 10.1186/1478-7954-9-29] [Medline: 21816105]
44. Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-learn. GetMobile: Mobile Comp. and Comm 2015 Jun;19(1):29-33. [doi: 10.1145/2786984.2786995]
45. Colin Cameron A, Windmeijer FA. An R-squared measure of goodness of fit for some common nonlinear regression models. Journal of Econometrics 1997 Apr;77(2):329-342. [doi: 10.1016/s0304-4076(96)01818-0]

Abbreviations

ASCVD: atherosclerotic cardiovascular disease
AUC: area under the ROC curve
CMCS: Chinese Multiprovincial Cohort Study
DBP: diastolic blood pressure
EMR: electronic medical record
FSRP: Framingham Stroke Risk Profile
HDL-C: high-density lipoprotein cholesterol
HR: heart rate
PCE: pooled cohort risk equation
PPD: pulse pressure difference
ROC: receiver operating characteristic
SBP: systolic blood pressure
SVM: support vector machine
TC: total cholesterol
XGBoost: extreme gradient boosting

Edited by G Eysenbach; submitted 08.05.21; peer-reviewed by H Ma, Z Ren; comments to author 31.05.21; revised version received 31.08.21; accepted 17.09.21; published 10.11.21.

Please cite as:
Yang Y, Zheng J, Du Z, Li Y, Cai Y
Accurate Prediction of Stroke for Hypertensive Patients Based on Medical Big Data and Machine Learning Algorithms: Retrospective Study
JMIR Med Inform 2021;9(11):e30277
URL: https://medinform.jmir.org/2021/11/e30277
doi:10.2196/30277
PMID:34757322

Original Paper

Machine Learning–Based Hospital Discharge Prediction for Patients With Cardiovascular Diseases: Development and Usability Study

Imjin Ahn[1], BSc; Hansle Gwon[1], BSc; Heejun Kang[2], MSc; Yunha Kim[1], BSc; Hyeram Seo[1], BSc; Heejung Choi[1], BSc; Ha Na Cho[2], MSc; Minkyoung Kim[2], BSc; Tae Joon Jun[3*], PhD; Young-Hak Kim[2*], MD, PhD

[1]Department of Medical Science, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[2]Division of Cardiology, Department of Internal Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[3]Big Data Research Center, Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Young-Hak Kim, MD, PhD
Division of Cardiology
Department of Internal Medicine
Asan Medical Center, University of Ulsan College of Medicine
88, Olympic-ro 43-gil
Seoul, 05505
Republic of Korea
Phone: 82 2 3010 3955
Email: mdyhkim@amc.seoul.kr

Abstract

Background: Effective resource management in hospitals can improve the quality of medical services by reducing labor-intensive burdens on staff, decreasing inpatient waiting time, and securing the optimal treatment time. The use of hospital processes requires effective bed management; a stay in the hospital that is longer than the optimal treatment time hinders bed management. Therefore, predicting a patient's hospitalization period may support the making of judicious decisions regarding bed management.

Objective: First, this study aims to develop a machine learning (ML)–based predictive model for predicting the discharge probability of inpatients with cardiovascular diseases (CVDs). Second, we aim to assess the outcome of the predictive model and explain the primary risk factors of inpatients for patient-specific care. Finally, we aim to evaluate whether our ML-based predictive model helps manage bed scheduling efficiently and detects long-term inpatients in advance to improve the use of hospital processes and enhance the quality of medical services.

Methods: We set up the cohort criteria and extracted the data from CardioNet, a manually curated database that specializes in CVDs. We processed the data to create a suitable data set by reindexing the date-index, integrating the present features with past features from the previous 3 years, and imputing missing values. Subsequently, we trained the ML-based predictive models and evaluated them to find an elaborate model. Finally, we predicted the discharge probability within 3 days and explained the outcomes of the model by identifying, quantifying, and visualizing its features.

Results: We experimented with 5 ML-based models using 5 cross-validations. Extreme gradient boosting, which was selected as the final model, accomplished an average area under the receiver operating characteristic curve score that was 0.865 higher than that of the other models (ie, logistic regression, random forest, support vector machine, and multilayer perceptron). Furthermore, we performed feature reduction, represented the feature importance, and assessed prediction outcomes. One of the outcomes, the individual explainer, provides a discharge score during hospitalization and a daily feature influence score to the medical team and patients. Finally, we visualized simulated bed management to use the outcomes.

Conclusions: In this study, we propose an individual explainer based on an ML-based predictive model, which provides the discharge probability and relative contributions of individual features. Our model can assist medical teams and patients in identifying individual and common risk factors in CVDs and can support hospital administrators in improving the management of hospital beds and other resources.

(JMIR Med Inform 2021;9(11):e32662) doi:10.2196/32662

KEYWORDS

electronic health records; cardiovascular diseases; discharge prediction; bed management; explainable artificial intelligence

Introduction

Background

The use of human and physical resources, which are both costly and scarce, is essential for the efficient operation of hospital processes. Hospitals are required to manage different kinds of resources, such as managing the schedules of the medical team and staff, bed management , and clinical pathways to improve overall management efficiency [1]. Effective resource management in hospitals can improve the quality of medical services by reducing the labor-intensive burden on staff, decreasing inpatient waiting time, and securing optimal treatment time [2].

Bed management is a form of hospital resource management. Currently, in most hospitals, clinicians manually check a patient's condition to decide whether to continue their hospitalization or discharge them [3]. On the basis of this decision, the medical team and staff identify the bed capacity available in the near future and schedule the patient's reservation. In addition, the number of patients hospitalized for a variety of chronic and acute illnesses, such as cardiovascular diseases (CVDs) [4], has been steadily increasing, and their insufficient treatment can lead to readmissions or complications. However, a stay in the hospital longer than the optimal treatment time hinders effective bed management. Thus, it is important to accurately predict the patient's hospitalization period and make judicious decisions about their discharge.

Many studies have focused on the efficiency of hospital resources, and most of them presented algorithms or models for improving bed management. Bachouch et al [5] investigated hospital bed planning and proposed the integer linear program to solve the optimization problem. They illustrated the simulated bed occupancy schedule. Troy et al [6] studied the simulation of beds for surgery patients using the Monte Carlo simulation to determine the intensive care unit (ICU) capacity. Particularly, the predicted length of stay (LOS) is one necessary piece of information for bed management, and there are many studies predicting the LOS based on electronic health records (EHRs) [7-9].

Moreover, authors have used machine learning (ML)–based models to predict the LOS [7-9], prolonged hospitalization, and unplanned readmission [10] and to find biomarkers for critical diseases [11]. Recently, there have been many studies on interpretable or explainable artificial intelligence (XAI) [12]. One XAI study [13] developed a model to predict acute illness and provide results and interpretation. Compared with EHRs, studies employing computer vision algorithms such as convolutional neural networks are more actively pursued because these models can directly visualize significant parts of an image [14,15]. Thus, we developed an ML-based predictive model to provide the daily discharge probability and *individual explainer* visualizing significant features of each patient to support bed management.

Objectives

The main contributions of this study can be summarized in the following steps: first, we developed an ML-based predictive model to predict the discharge probability daily within 3 days for each patient with CVD and to acquire the individual LOS. Patients with chronic and acute diseases, including CVDs, have high hospitalization and readmission rates and greater complications [16]. There are alternatives to transfer those who need urgent care or hospitalization to another hospital to address delays. However, it could be causing other serious problems, hospitals should continuously identify methods to reduce waiting time, and efficient bed management can be considered as one of them.

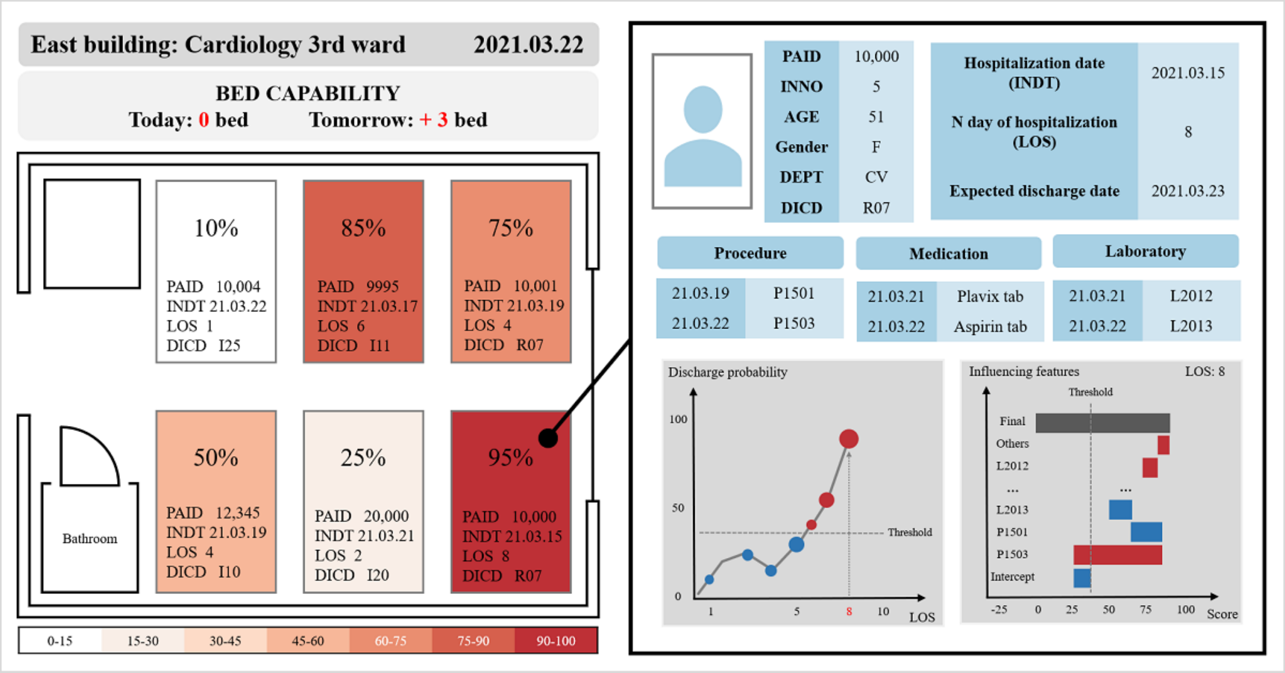
In addition, because of the diversity of diseases, it may be more advantageous to find common risk factors and implement bed management for specific departments or diseases (ie, clustered specific wards), and then expand it further to the hospital level. Therefore, we developed an ML-based model to determine the bed capacity that would be available in the near future and find risk factors by predicting the discharge of patients hospitalized with CVDs [17]. By providing persuasive discharge information such as expected individual discharge date and risk factors related to CVDs, it is possible, in practice, to assist in precise bed management, which is otherwise done manually by the medical team.

Second, we assessed the outcome of the prediction and provided the individual explainer to describe the primary risk factors of inpatients for patient-specific care. Even if patients have the same diseases and common variables represent the diseases, each patient has different characteristics, history, circumstances, and treatments. Therefore, it is also necessary to identify and monitor the unique, individual variables for each patient. In this study, our ML-based predictive model's outcomes include not only information on daily patient discharge but also the contributions of features such as feature importance. Furthermore, we visualized the day-by-day discharge probability of each patient and the features that influenced individual patients during the hospitalization. This explainer can guide the medical team and patients to produce reasonable evidence on the ML-based model's outcomes and helps them understand the conditions in detail and prepare in advance for treatment. Such individual analysis can focus on each patient, and the meaningful features identified can be used in other studies as a basis for preidentifying variables affecting hospitalization.

Third, this study could help manage bed scheduling efficiently and detect long-term inpatients in advance. Bed management refers to the process of identifying patients who are most likely to be discharged, confirming the number of available beds, and allocating beds to patients waiting for admission after reservation. As this process is complicated and usually carried out manually, we aimed to support it by providing the estimated LOS and probability of discharge returned by the model and by identifying the capacity of beds that would be available in the near future. In addition, it is possible to detect not only patients with a high probability of discharge but also patients with a consistently low probability of discharge. In other words, it helps discover and analyze the causes of long-term hospitalization of high-risk patients and provides this information to their management team.

To summarize, we developed an ML-based model to predict whether hospitalized patients with CVDs would be discharged within 3 days. On the basis of this model, we proposed an individual explainer; the simulations of bed management are depicted in Figure 1, including the probability of discharge and influenced features such as demographics, prescribed medications, and treatments. Our model can improve the efficient use of hospital resources and enhance the quality of medical services.

Figure 1. Visualized simulation of discharge prediction for machine learning–based bed management. DEPT: department; DICD: diagnostic code; INDT: the date of visitation or admission; INNO: the patient's encounter number; LOS: length of stay; PAID: the patient's identification.

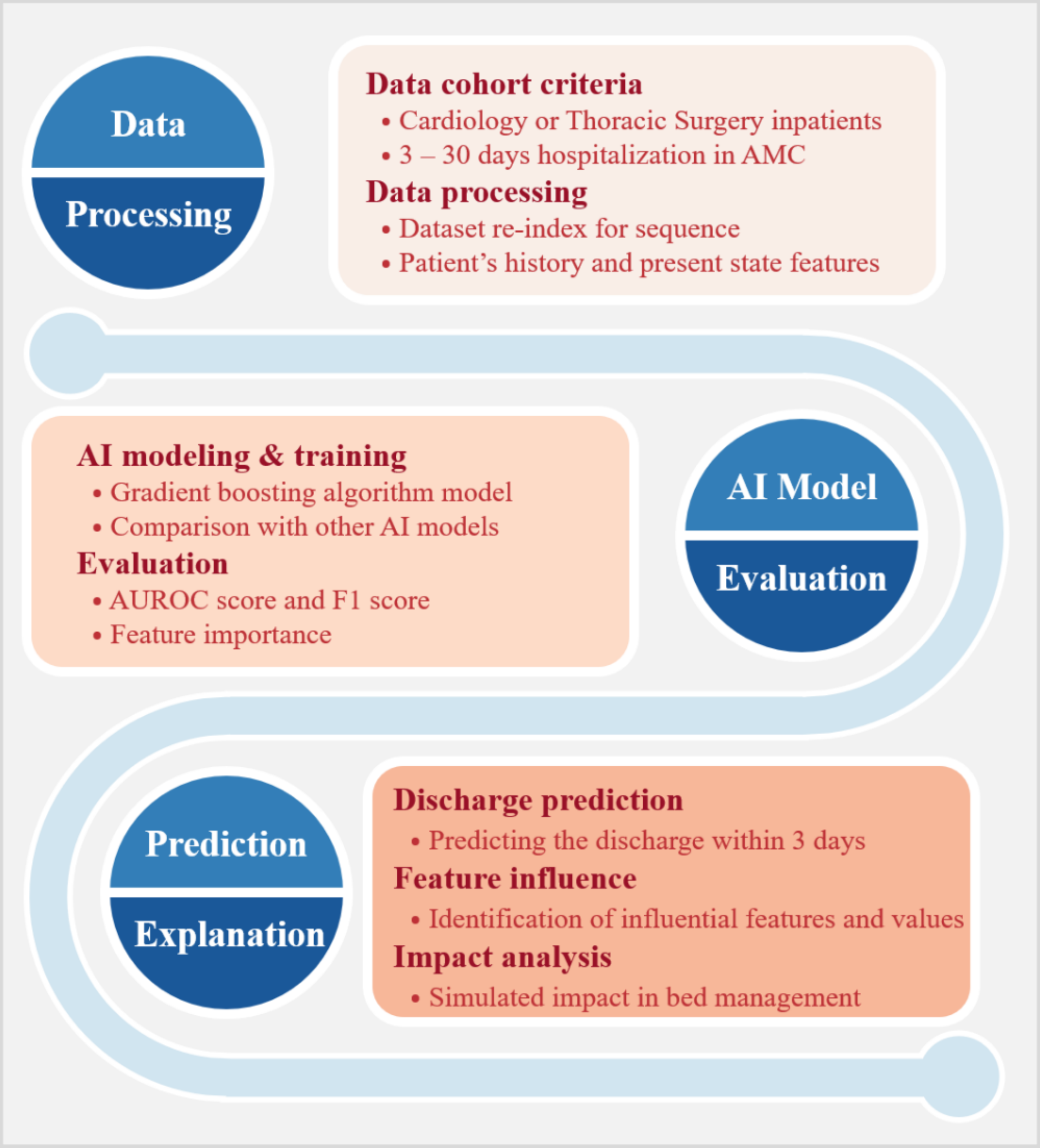


Methods

Overview

Figure 2 describes the overall flow of the prediction method employed in this study. We set up the cohort criteria and processed the data to create suitable data sets. Subsequently, we trained the ML-based predictive models and evaluated them to find an elaborate model. Finally, we predicted the discharge probability within 3 days and explained the model's outcomes by identifying, quantifying, and visualizing its features.

Figure 2. Overall flow of the prediction method for discharge within 3 days. AI: artificial intelligence; AMC: Asan Medical Center; AUROC: area under the receiver operating characteristic.


Data
Processing
Data cohort criteria
• Cardiology or Thoracic Surgery inpatients
• 3 – 30 days hospitalization in AMC
Data processing
• Dataset re-index for sequence
• Patient's history and present state features
AI modeling & training
• Gradient boosting algorithm model
• Comparison with other AI models
Evaluation
• AUROC score and F1 score
• Feature importance
AI Model
Evaluation
Prediction
Explanation
Discharge prediction
• Predicting the discharge within 3 days
Feature influence
• Identification of influential features and values
Impact analysis
• Simulated impact in bed management

Data Acquisition

Data were extracted from CardioNet [18] (Textbox 1), a manually curated EHR database specialized in CVDs. CardioNet consists of data from 572,811 patients who had visited Asan Medical Center (AMC) with CVDs between January 1, 2000, and December 31, 2016. The AMC institutional review board approved the collection of CardioNet data and waived informed consent. CardioNet contains 27 tables on topics such as visitation, demographics, diagnosis, medication, and laboratory examination. Most tables have common variables including patient identification (PAID), patient encounter number (INNO), the date of visitation or admission (INDT), and the date of discharge (OUDT). The KEY column, which concatenates the PAID and INNO columns, can connect the visitation table to other tables. Using the KEY column, we extracted the variables in each table to be analyzed.

From the 572,811 patients in CardioNet, we obtained 84,251 records of 63,261 anonymous patients hospitalized in the departments of cardiology or thoracic surgery. Furthermore, to develop a practical and usable model, we focused on predicting discharge within 3 days and detecting long-term patients. Long-term patients, defined as those hospitalized for more than 30 days, are separately managed by the AMC. Therefore, we set the LOS between 3 and 30 days.

Textbox 1. Data extracted from CardioNet.

- Visit table: patient identification, patient encounter number, KEY, date of visitation or admission, date of discharge, type of visit, medical department, and duration of stay in the intensive care unit (ICU).
  - acute care unit, coronary care unit, cardiac surgery ICU, medical ICU, neonatal ICU, neurological ICU, neurosurgical ICU, pediatric ICU, and surgical ICU.
- Diagnosis table: *International Classification of Diseases, Tenth Revision* code of diagnosis.
- Laboratory test result table: date and code of pathology examination, and the result of the examination.
- Physical information table: patient’s age, height, weight, systolic and diastolic blood pressures, respiratory rate, pulse rate, BMI, body surface area, and date of measurements.
- Medication table: date and code of prescription.
- Procedure table: date and code of order.
- Operation table: date and code of surgery or treatment.
- Picture archiving and communication system table: date and code of order.
- Transfusion order table: date and code of order.

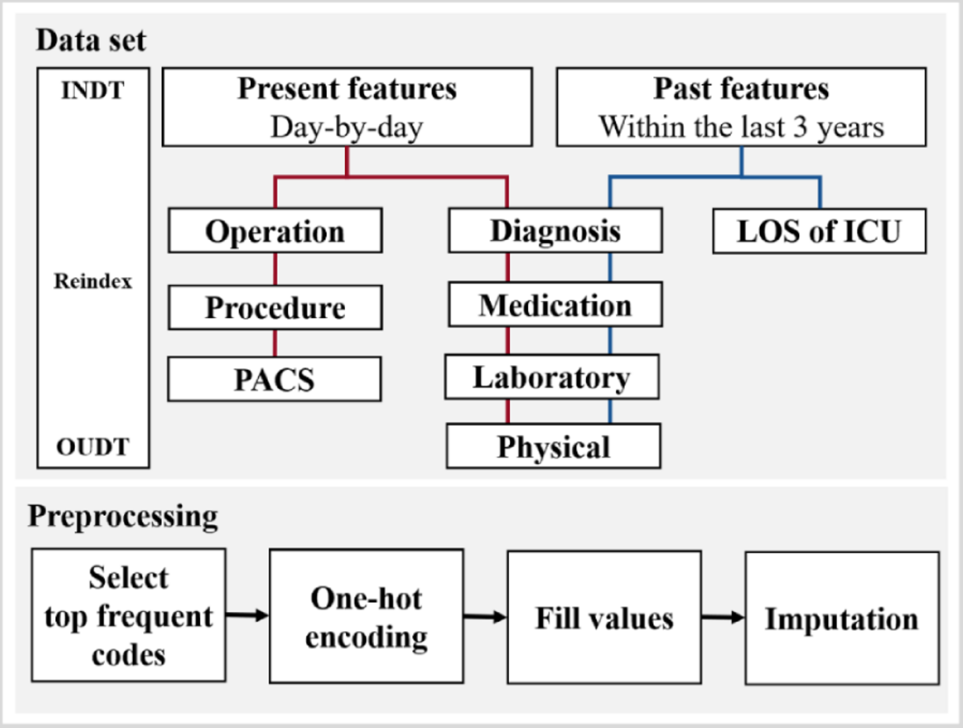
Data Preprocessing

Data Set Creation

In the visit table, which is the primary table of CardioNet, there are 4 main columns (PAID, INNO, INDT, OUDT) and visit-related variables. Each row represents a single hospitalization case for each inpatient. We reset the index to create a new data set with the duration between admission and discharge as date-index (eg, a row with an INDT of 2021.2.1 and an OUDT of 2021.2.10 has an LOS 10 of days; therefore, it was converted to 10 rows with 10 date-indexes). Finally, after preprocessing all values corresponding to PAID, INNO, and date-indexes of other tables, we merged and concatenated the tables to generate a new data set for model training.

Figure 3 shows the data set creation process. Each table of diagnosis, medication, laboratory test results, and physical information was used for both past and present features. The operation, procedure, and picture archiving and communication system (PACS) were used for the present features, and LOS in the ICU was used for the past features. The preprocessing of values for each table is discussed in the next section. The specific methods of feature handling are as follows:

Figure 3. Data set creation process for machine learning–based model training. ICU: intensive care unit; INDT: date of visitation or admission; LOS: length of stay; OUDT: date of discharge; PACS: picture archiving and communication system.



Data-Related Features

After creating the new data set, we removed the OUDT containing future information. To distinguish and recognize the time information in date by type, we created a total of 10 date-related features. INDT and date-index were sliced into integer features such as year, month, day, and weekday. Furthermore, we created a feature that denotes whether the date-index is a holiday or not and another feature that indicates the LOS at the date-index by subtracting INDT from the date-index.

Day-by-day Present Features Related to Hospitalization

As the visit table and other tables contain only one piece of information per row, it is difficult for the ML model to learn the data all at once. Therefore, we performed one-hot encoding (OHE) of clinically important orders and codes and created them as features in the new data set. Consequently, we could access aggregated records by date for each patient.

First, in the diagnosis and operation tables, we sliced all the values of the International Classification of Diseases-10th edition codes and the operation codes at the third digit to convert them into three-digit codes because the strings from the fourth digit onward represent the subhierarchy of the three-digit codes. We arranged all the frequency values in descending order and selected the first 99 codes. We transformed the remaining codes (ie, unselected codes) into the *others* feature and performed the OHE on all 100 codes. The features in the form of *Z_code*, such as *Z_DICD* and *Z_OPCD*, refer to *others* in each original table. As a result, we obtained a total of 100 codes for each table (ie, diagnosis and operation table) and filled the date-index values with *1* if there were valid prescribed or ordered data and *0* otherwise. Similarly, the values of the PACS table were converted to 100 features.

Second, similar to the diagnosis table, in the medication and procedure tables, we obtained the 99 most frequent codes and *others*, performed the OHE, and filled the corresponding data. In the case of the transfusion table, we used all 27 codes available. We filled the values with the number of prescriptions per day or at once, considering the severity of each patient's ailment.

Third, in the laboratory test result table, the 60 most frequent examination codes, examined in more than 50% of all patients, were selected. The physical information table had only 10 codes, which were all used. We performed the OHE of values and filled them with results corresponding to each examination. If a patient had been tested several times a day, the data set was populated with the average of the results.

Past Features

We considered that the patient's anamnesis (ie, medical history) should also be included in the data set, along with the day-to-day features (described in the previous paragraph) for the ML model to learn the data deeply. When the date-index in each hospitalization started from INDT, we created some past features from the principal information of hospital visit records 3 years before INDT.

For past features, OHE was performed, and values were filled in, similar to the present features. The hospitalization periods of all ICUs in the visit table were summed up. For 100 diagnostic codes, we summed up each value if there was a record of diagnosis. For 100 medication codes, the number of prescriptions per day or at once were summed up if the record existed. Finally, recent laboratory test results and physical information within 3 years were used for a total of 70 codes. In conclusion, the data set was filled with either summed up or recent values equivalent to each feature.

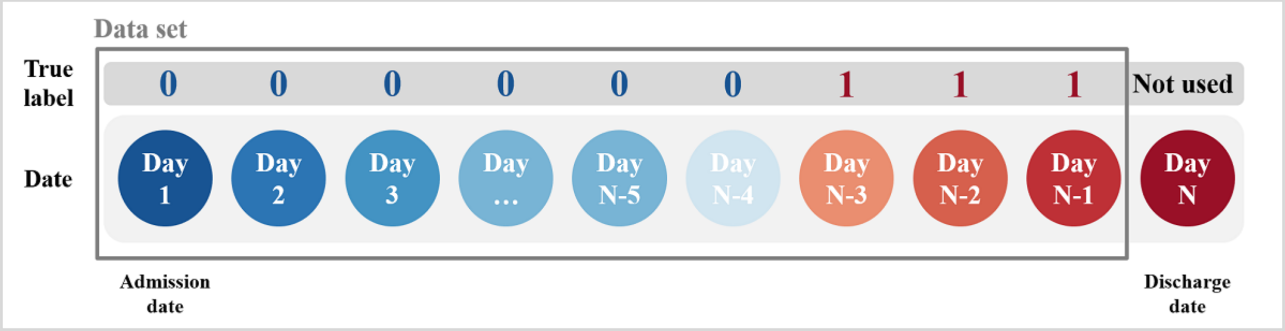
Imputation

Except for the laboratory, physical, and date-related features, we replaced all the null values with zero. The value type of most of the other features was null or integer because most were calculated by frequency. In contrast, to deal with missing values in the continuous data type of the present laboratory and physical features, we first separated the data set based on the KEY. The KEY refers to a single hospitalization case of one patient; thus, separating the data set by KEY does not mix individual hospitalizations. Therefore, we filled in null values in chronological order (ie, from past to present). Subsequently, we filled in the rest of the null values in reverse chronological order (ie, from present to past) to handle those cases where results were not measured at the beginning of the admission. Using this method, it was possible to impute the null value for each hospitalization of an individual patient. Finally, to fill the values where all the features were not ordered or measured, we filled the rest of the null values with the most frequent value for each feature.

Target Criteria

The supervised learning algorithm for classification requires the label *true* or *false* to indicate the correct answer. The target criteria for *true* labeling in this study are depicted in Figure 4.

Figure 4. Target criteria to provide the true label (ie, correct answer) to machine learning–based models.



As shown in Figure 4, *day 1* is INDT, *day N* is OUDT, and the circles represent each day of the hospitalization period. We excluded *day N* (ie, discharge date) from the data set because of information such as *discharge procedure*, which could provide the ML model with a hint. In addition, even if the accuracy of discharge prediction is higher from the discharge date to 2 days earlier, it is useful to make the prediction 3 days in advance when actually using the model. Therefore, we labeled *1* from one day before OUDT to 3 days before OUDT and labeled *0* from the INDT to 4 days before OUDT.

As a result, we transformed the diverse variables of original tables into 10 date-related features, 597 present features, and 279 past features, creating a data set of 669,667 rows with 886 features from 84,251 records of 63,261 inpatients with CVDs.

ML-Based Predictive Models

ML-Based Models

We experimented with 5 models to identify the most suitable one. We set the logistic regression [19] model as the baseline to estimate performance, and support vector machine [20,21], random forest (RF) [22], multilayer perceptron (MLP) [23], and extreme gradient boosting (XGB) [24] were selected as comparison models. We also performed hyperparameter tuning for each model through random search.

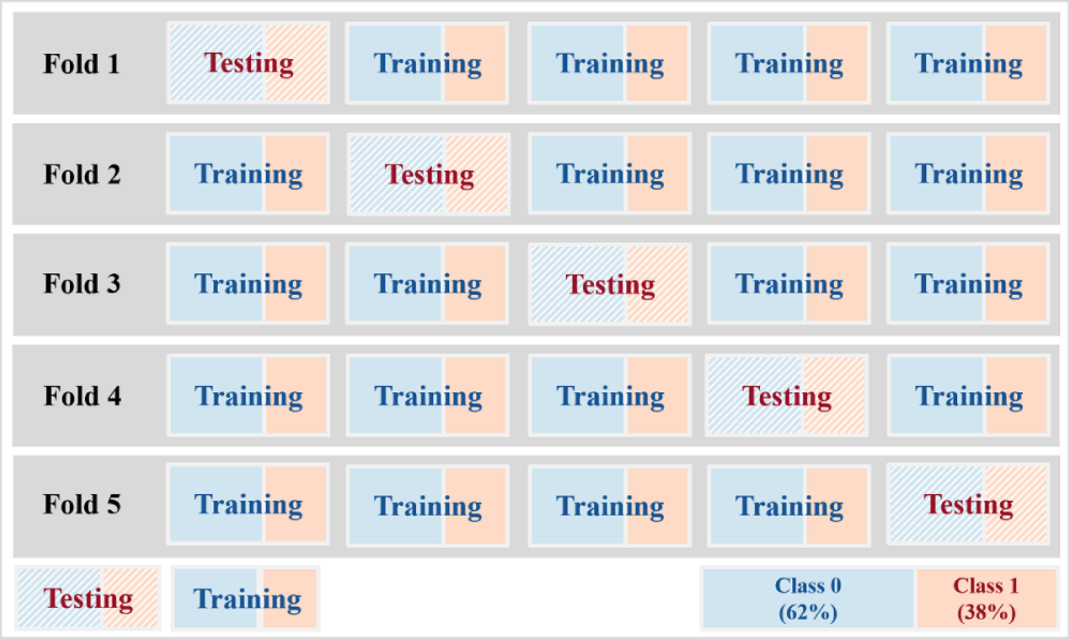
We selected XGB, which is a gradient boosting algorithm (GBM) model, as the final model. GBM is an ensemble method that combines several weak classifiers (trees). The main idea of GBM is to focus and place the weights on incorrectly predicted results. While XGB is training, one tree trains the data set and assigns weights to incorrectly predicted records with errors, and the next tree of the same model learns the weighted data set and repeats the process of assigning weights. Moreover, GBM can quantify the contribution of features to the prediction results, such as feature importance. Particularly, XGB has the advantage of regularization and performance. It can perform parallel processing, regulate to avoid overfitting, is widely used for learning structured data, and has superior prediction performance.

Evaluation

We set the positive (1) label for discharge and the negative (0) label for hospitalization. To evaluate and compare the performance of candidate models, we used metrics including accuracy, sensitivity (recall for positive), specificity, precision, positive predictive value, negative predictive value, false-positive rate, and true-positive rate. When we monitored model training and validation, we used the F1-score to reflect imbalanced targets, the receiver operating characteristic (ROC) curve to find the optimal threshold, and the area under the ROC (AUROC) score to compare models.

To prevent overfitting the ML-based models and reduce biased results, we performed stratified, 5-fold cross-validation [25] illustrated in Figure 5. First, we randomly shuffled 63,261 PAIDs and divided them into 5 groups with approximately 12,000 people because we tried not to divide the records of a single patient into training (ie, plain box in Figure 5) and testing sets (ie, diagonal hatching box in Figure 5). Second, the first PAID group becomes the testing set, and the remaining groups become the training set in fold 1. We created fold 1 to fold 5 in a similar way to ensure equal division of the imbalanced targets (ie, the data set has true labels comprising 62.4% label *0* and 37.6% label *1*) across all folds. Besides, we split 25% of the training set as the validation set to tune the hyperparameters. Consequently, in each fold, we divided the data set into approximately 133,000 rows for the testing set and 535,000 rows for the training set (including the validation set). The ML-based models trained and tested all 5 folds.

Figure 5. Stratified 5-fold cross-validation to avoid overfitting.



Individual Explainer for Outcome Assessment

Feature importance lists the features that the model considers prominent, and their contribution scores, in the process of training the data using the tree-based algorithm model. However, we considered XGB as the final model not only because of its high performance but also because of the access to the decision-making process inside the model. By approaching the trees, it is possible to describe the specific features and their influences that contribute to the prediction of each patient's daily prediction of discharge.

We demonstrate an individual explainer that can help in the interpretation of the XGB prediction results using a waterfall chart. Also called a bridge or cascade chart, it is a type of bar chart that portrays relative values and calculates the difference between adjacent values. It can show the positive or negative influence and gradual direction of the final discharge score.

To estimate values for individual explainers, we predicted the desired records with the trained XGB and obtained the contributions of all the features. The contribution refers to a feature's influence obtained by aggregating the scores that each feature contributes to all trees. Subsequently, we calculated the logistic value—$logistics\ (x) = 1 / (1 + e^{-x})$—of the feature's influence and the relative values required for the explainer. We selected the number of features to be displayed as 15, and the remaining 871 features were integrated and displayed simultaneously as *others* in the explainer.

Results

Data Characteristics

We created a data set that consisted of 669,667 records with 886 features, including diagnosis code, laboratory test results, physical information, medication, procedure, operation, PACS, and transfusion. Patients were admitted to cardiology or thoracic surgery, and their LOS ranged from 3 to 30 days. The average age of the patients was 61.03 (SD 13.42) years. The data set comprised 37.97% (254,254/669,667) women and 62.03% (415,413/669,667) men.

Performance of the ML-Based Predictive Models

Final ML-Based Model Selection

We experimented with the 5 ML-based models using 5 cross-validations. The AUROC score for each fold is listed in Table 1. The highest AUROC score for each fold is shown in italics, and the *support* column in Table 1 represents the number of each true label. Figure 6 shows the ROC curve plot; the area of the curve is represented by the AUROC and has a value between 0 and 1. The closer the AUROC score is to 1, the higher the model performance. XGB achieved the highest and a relatively stable score on all folds. Table 2 provides a comparison of the 5 ML-based models. All scores in Table 2 are the average values of the results and the SD in 5 folds, and the highest score for each metric is shown in italics. The specificity of logistic regression and support vector machine, which obtained 0.828, was the highest, but XGB achieved the highest in the rest of the metrics. Particularly, although the label of the data set was imbalanced, XGB scored 0.7 or higher for predicting label 1. Hence, we chose XGB as the final model to predict discharge probability.

Table 1. Evaluation by area under the receiver operating characteristic score of 5-fold cross-validation for each model.

	LR[a]	SVM[b]	RF[c]	MLP[d]	XGB[e]	Support (0, 1)
Fold 1	0.826	0.825	0.853	0.833	*0.866* [f]	(83,113, 50,188)
Fold 2	0.827	0.826	0.851	0.835	*0.868*	(83,538, 50,310)
Fold 3	0.824	0.824	0.850	0.821	*0.865*	(84,192, 50,585)
Fold 4	0.824	0.823	0.850	0.831	*0.864*	(83,969, 50,460)
Fold 5	0.822	0.821	0.848	0.834	*0.863*	(82,918, 50,394)
Value, mean (SD)	0.824 (0.002)	0.824 (0.002)	0.850 (0.002)	0.831 (0.005)	*0.865* (0.002)	N/A[g]

[a] LR: logistic regression.

[b] SVM: support vector machine.

[c] RF: random forest.

[d] MLP: multilayer perceptron.

[e] XGB: extreme gradient boosting.

[f] The italicized values indicate the highest score of each fold.

[g] N/A: not applicable.

Figure 6. Receiver operating characteristic curve of the machine learning–based models. LOGREG: logistic regression; MLP: multilayer perceptron; RF: random forest; SVM: support vector machine; XGB: extreme gradient boosting.

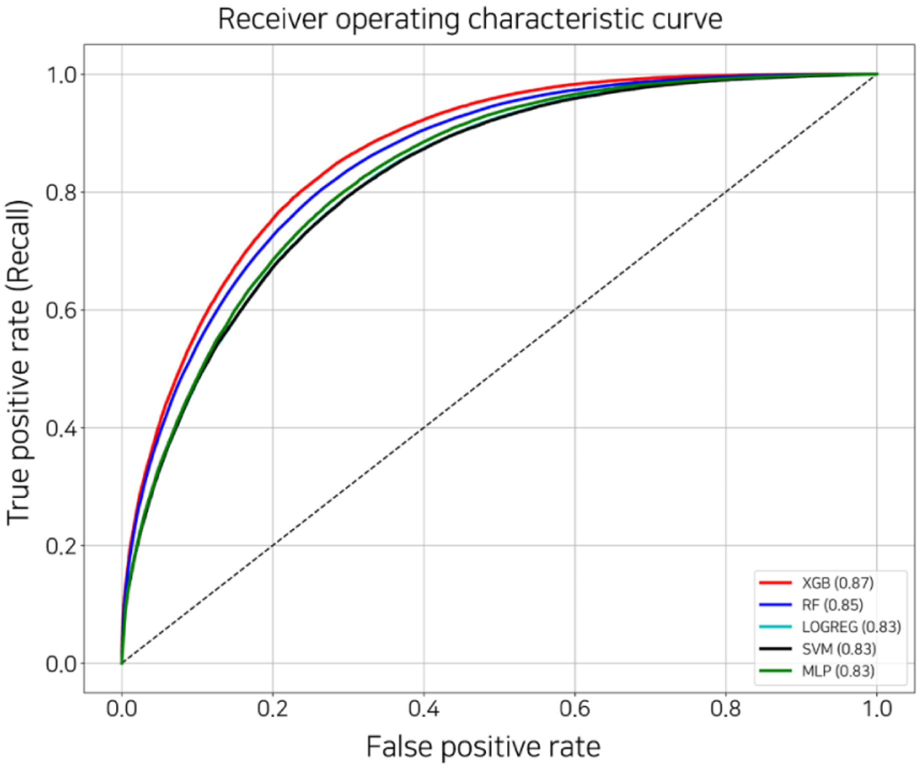


Table 2. Comparison of the 5 machine learning–based models by metric.

Model	Values, mean (SD)					
	ACC[a]	Sen[b]	Spe[c]	PPV[d]	NPV[e]	AUROC[f]
LR[g]	0.75 (0)	0.624 (0.005)	*0.828*[h] (0.004)	0.686 (0.005)	0.786 (0.005)	0.824 (0.002)
SVM[i]	0.75 (0)	0.624 (0.005)	*0.828* (0.004)	0.686 (0.005)	0.784 (0.005)	0.824 (0.002)
RF[j]	0.77 (0)	0.696 (0.005)	0.818 (0.004)	0.696 (0.005)	0.818 (0.004)	0.85 (0.002)
MLP[k]	0.758 (0.004)	0.642 (0.017)	0.822 (0.007)	0.686 (0.005)	0.792 (0.007)	0.831 (0.005)
XGB[l]	*0.782* (0.004)	*0.716* (0.005)	0.824 (0.005)	*0.71* (0)	*0.828* (0.004)	*0.865* (0.002)

[a] ACC: accuracy.

[b] Sen: sensitivity.

[c] Spe: specificity.

[d] PPV: positive predictive value.

[e] NPV: negative predictive value.

[f] AUROC: area under the receiver operating characteristic.

[g] LR: logistic regression.

[h] The italicized values refer to the highest score of each metric.

[i] SVM: support vector machine.

[j] RF: random forest.

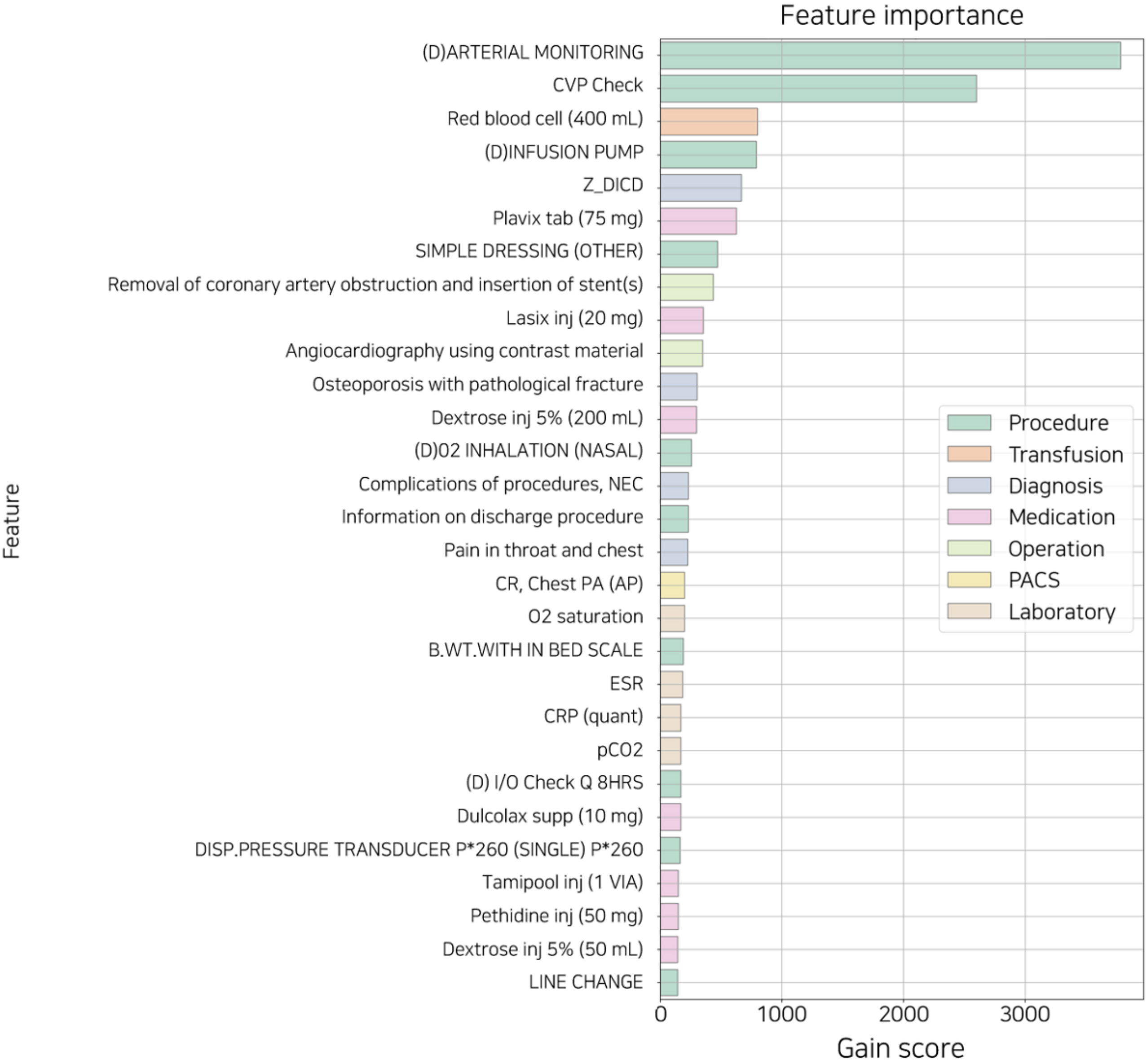
[k] MLP: multilayer perceptron.

[l] XGB: extreme gradient boosting.

Figure 7 shows the relative feature importance of XGB sorted by gain score. The gain score refers to the average gain across all splits that the feature is used in. All the features used in the model have been replaced by their names used in the AMC. Except for the date-related feature, all other features that affected the model were found in all the tables. The features in the procedure table are substantially related to clinically critical situations. For example, the terms denoted with *(D)* are likely to mean a more severe state than others. The remaining features are also associated with CVDs or include primary examination and prescriptions during hospitalization.

However, because feature importance can only explain the model but not each patient, it is insufficient for use as an individual explainer for prediction. Depending on the patient's condition, different features affect the daily probability of discharge. Therefore, we suggested an individual explainer that provides a patient-specific feature for daily prediction during hospitalization.

Figure 7. The feature importance sorted by gain score. B.WT.: body weight; CR: chest radiograph; CRP: C-reactive protein; CVP: central venous pressure; DISP: disposable; ESR: erythrocyte sedimentation rate; I/O: intake and output; supp: suppository; inj: injection; NEC: necrotizing enterocolitis; PA: posteroanterior; PACS: picture archiving and communication system; Z_DICD: all diagnostic codes not selected for one-hot encoding.



Feature Reduction

Too many features tend to reflect negatively on the model performance; therefore, it was necessary to select an appropriate number of features. We performed recursive feature elimination with cross-validation (RFECV). This algorithm aims to identify the optimal number of features by comparing model performance while eliminating the features with low feature importance one at a time. RFECV returns the ranks and names of all features; we identified approximately 150 features with a rank of 1 by applying RFECV to our final model XGB. For performance comparison, we performed 5-fold cross-validation using the same data set with the same parameters. The number of features to be compared was 886 (all), 150 selected by RFECV, and the top 50 features in the model trained with the 150 selected by RFECV.

As shown in Figure 8 and Table 3, the performance difference between the model using all the features and the models with 150 and 50 features was only approximately 1% to 2.5% based on the AUROC score. This indicates that even with 83.1% to 94.4% of feature reduction, there is only a maximum performance difference of 2.5%. Therefore, a suitable number of features should be selected considering the situation in each hospital or the data characteristic.

Figure 8. Receiver operating characteristic curve of the extreme gradient boosting models with the different number of features. FI: feature importance; RFE: recursive feature elimination; XGB: extreme gradient boosting.

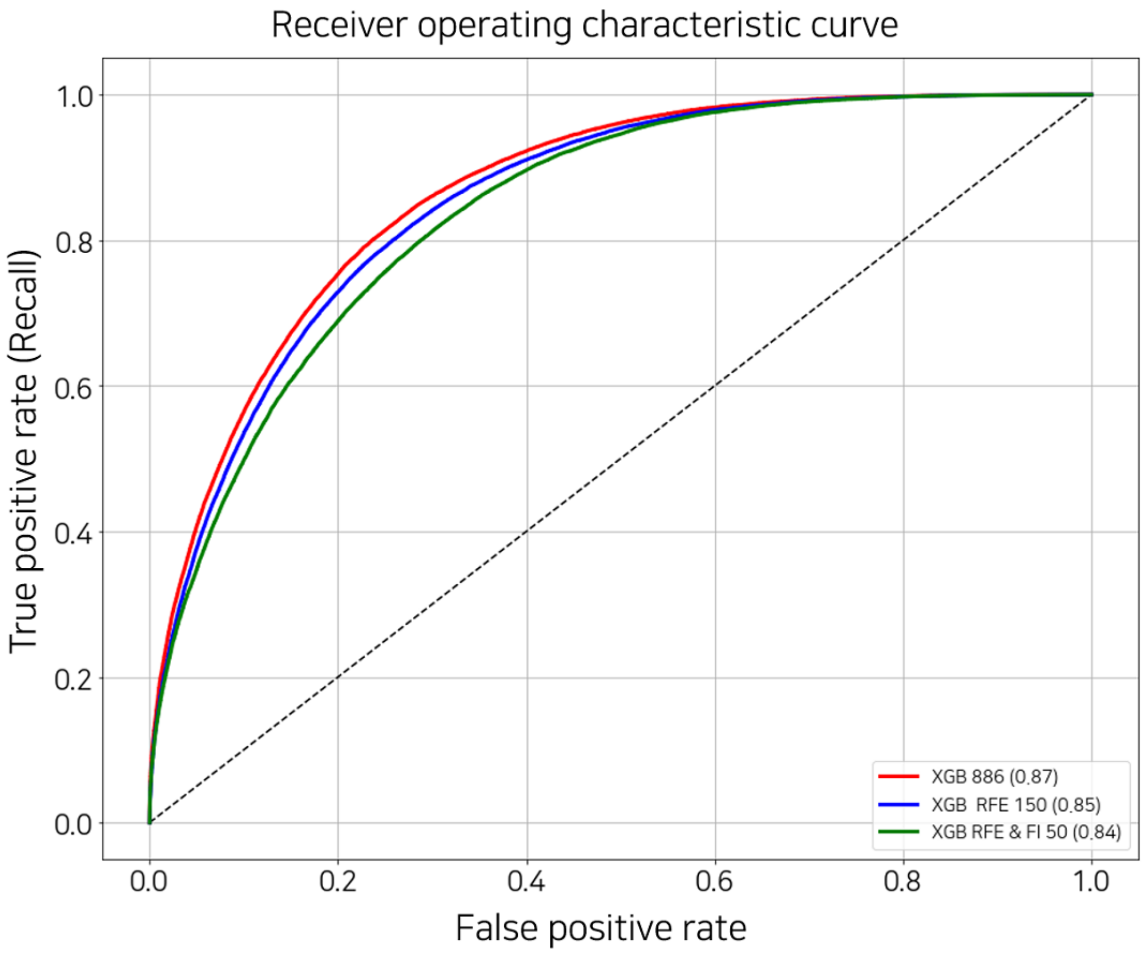


Table 3. Evaluation by area under the receiver operating characteristic (AUROC) score of 5-fold cross-validation to select features.

Number of features	Values, mean (SD)					
	ACC[a]	Sen[b]	Spe[c]	PPV[d]	NPV[e]	AUROC
886 (All)	*0.782*[f] (0.004)	*0.716* (0.005)	*0.824* (0.005)	*0.71* (0)	*0.828* (0.004)	*0.865* (0.0018)
150 (RFE[g])	0.77 (0)	0.696 (0.005)	0.814 (0.005)	0.694 (0.005)	0.818 (0.004)	0.853 (0.0018)
50 (RFE and FI[h])	0.76 (0)	0.67 (0.006)	0.812 (0.004)	0.682 (0.004)	0.802 (0.004)	0.840 (0.00096)

[a]ACC: accuracy.

[b]Sen: sensitivity.

[c]Spe: specificity.

[d]PPV: positive predictive value.

[e]NPV: negative predictive value.

[f]The italicized values refer to the highest score of each metric.

[g]RFE: recursive feature elimination.

[h]FI: feature importance.

Explainer of Individual Prediction for Outcome Assessment

Overview

The predictive model classifies the data as *0* or *1* based on a threshold. The optimal threshold is the point where the sum of sensitivity and precision can be maximized simultaneously (in the ROC curve, true-positive rate and false-positive rate are proportional to each other). However, sensitivity and precision require trade-off against each other; therefore, decreasing FN increases sensitivity, and decreasing false positive increases precision. In other words, it is necessary to adjust for the appropriate threshold to suit the decision point of the hospital operation.

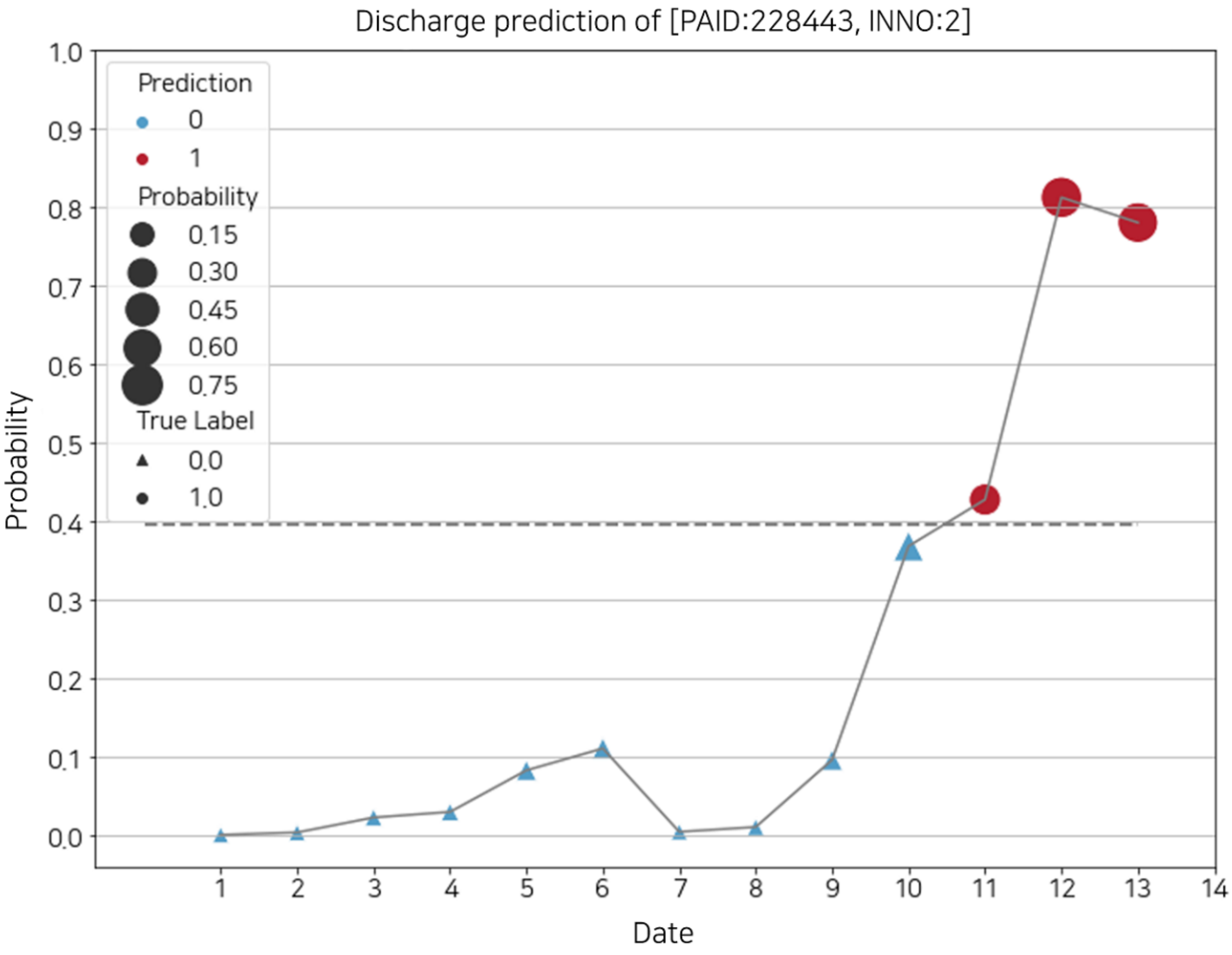
We presented the daily discharge score during hospitalization and the influence of the features by date through the explainer of individualized predictions. The following section includes a description and an example of our explainer for the sample data set, which represents one of the patients in the test set.

Discharge Score During Hospitalization

The sample data set consisted of the records of a patient with a PAID of 228,443 and an INNO of 2, hospitalized for 13 days and discharged on day 14. The patient's daily discharge score plot is depicted in Figure 9. The plot's x-axis represents the daily date excepted discharge date (ie, day 14) within the patient's hospitalization period, and the y-axis represents the probability of discharge (ie, discharge score). The model's optimal threshold was 0.39, indicated by a horizontal dotted line. The circle and the triangle represent the true labels *1* and *0*, respectively, and the size of the figure is proportional to the discharge score. The colors of the figure denote the results predicted by the model: red for positive prediction (label *1*, discharge) and blue for negative prediction (label *0*, admission).

For this sample, the model accurately predicted the discharge within 3 days. However, if the threshold is adjusted, the prediction results may change on dates 11 and 12. For example, if the current threshold rises slightly, *1* is applicable only for dates 12 and 13. This can be useful when trying to avoid false positive even if the false negative increases.

Figure 9. Daily discharge score of a patient's identification of 228,443 and patient's encounter number of 2. INNO: patient's encounter number; PAID: patient's identification.



Daily Feature Influence Score

Figures 10 and 11 describe the plot of feature influence for each day. The following is the basic description of the individual explainer: the x-axis of the plot is a score ranging from 0 to 1, and the y-axis represents the contributed features and the corresponding values that influenced the probability of discharge on that day. The threshold represented by the vertical dotted line is equal to the optimal threshold in Figure 9. The intercept, the plain blue box at the bottom of the y-axis, is a revised value reflecting that the number of each true label is imbalanced. The discharge probability, the gray box at the top of the y-axis, is the discharge score, which is the same as the probability in Figure 9. The width of each box corresponding to the feature refers to the absolute value of each score. The original score is indicated on the right side of the plot. The absolute value decreases from bottom to top, which means the contribution to the discharge score also decreases (the box of *others* is relatively wide because it is the sum of the scores of approximately 800 features, excluding the features below it). The red box with diagonal hatching represents each score of the feature that positively contributed to the discharge score and moves to the right. Conversely, the plain blue box represents negatively contributing feature scores and moves to the left.

To summarize, on the y-axis, from bottom to top, the features contributed to the prediction; the diagonal hatched red box to the right is positive, and the plain blue box to the left is negative.

Figure 10 shows the feature influence at day 7 with a low probability of discharge of 0.004, and Figure 11 shows day 12 with a high probability of 0.811. In Figure 10, *arterial monitoring=1* and *infusion pump=3* negatively affected the probability. In contrast, in Figure 11, *infusion pump=0* had a positive effect on probability. Because arterial monitoring and infusion pump are mainly prescribed for critical patients, both consist mostly of zeros in the data set. Therefore, displaying features and values together can help the medical staff interpret the plot intuitively. Moreover, each explainer may or may not have the features that appeared in the feature importance plot. This suggests that it is also necessary to identify features that contributed to individual patients rather than managing only the features of feature importance.

Figure 10. Feature influence with low probability of discharge date 7. CVP: central venous pressure; INNO: patient’s encounter number; MCH: mean corpuscular hemoglobin; MCV: mean corpuscular volume; PAID: patient’s identification; supp: suppository; ZM_ODCD: all medication codes not selected for one-hot encoding; ZP_ODCD: all procedure codes not selected for one-hot encoding.

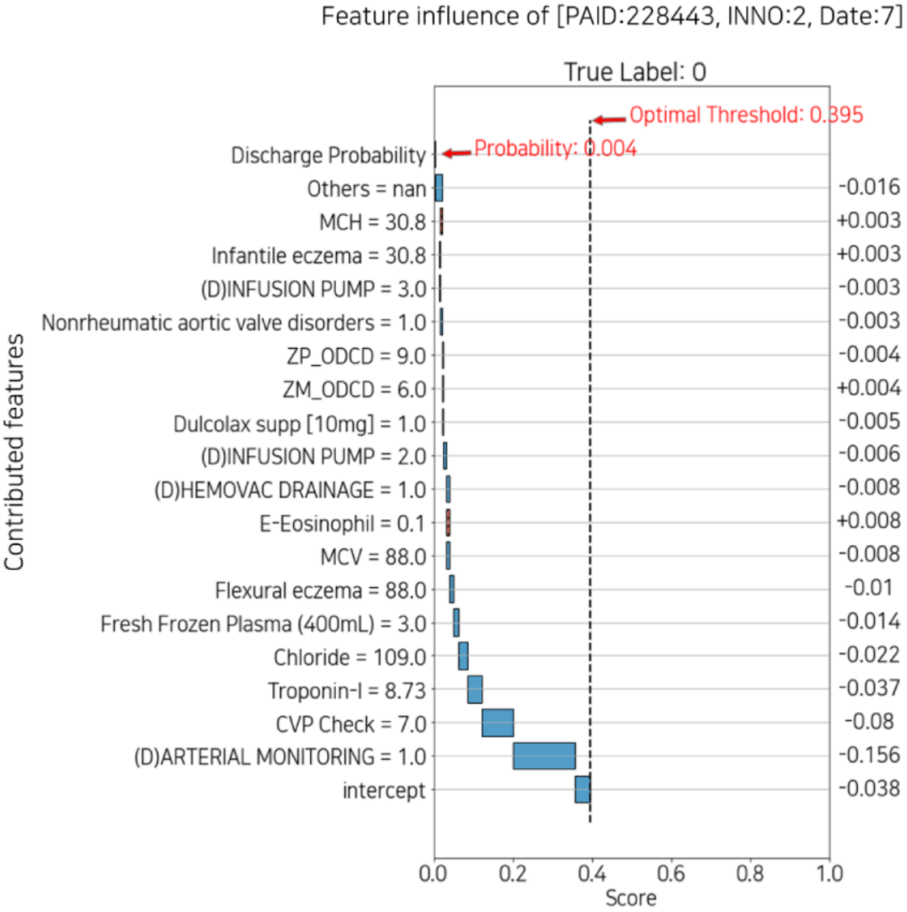
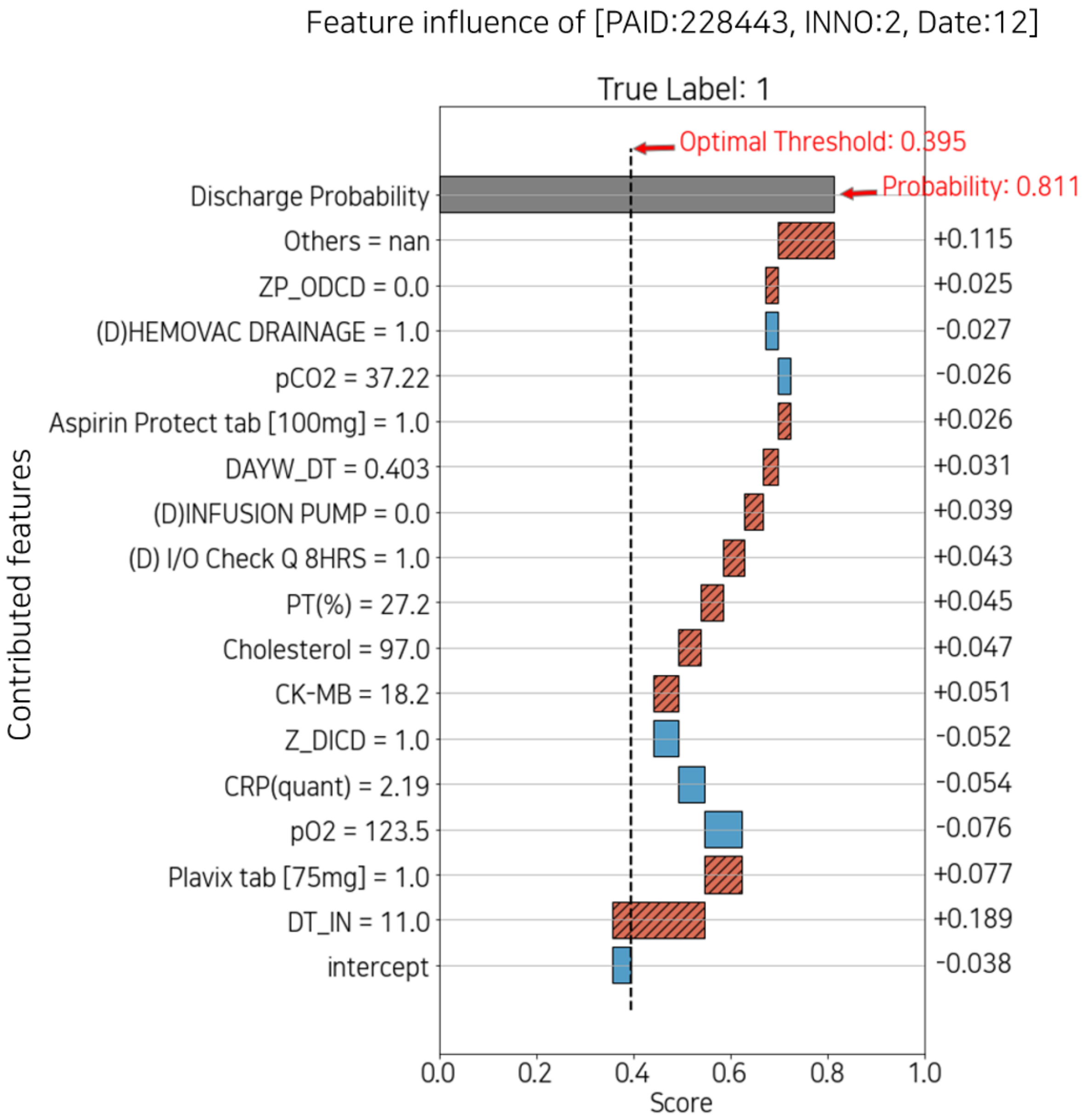


Figure 11. Feature influence with high probability of discharge on date 12. CK-MB: creatine kinase-myoglobin binding; CRP: C-reactive protein; DAYW_DT: integer feature of weekday; DT_IN: time since admission date in days; I/O: intake and output; INNO: the patient's encounter number; PAID: the patient's identification; PT: prothrombin time; Z_DICD: all diagnostic codes not selected for one-hot encoding; ZP_ODCD: all procedure codes not selected for one-hot encoding.



Outcome Assessment

Figure 1 shows the simulated impact in bed management applied with our predictive model and individual explainer. It is possible to recognize the probabilities of discharge of all patients for each ward every day. The paramount features and values that affect the discharge scores can be identified at once. It is informative for interpreting both high or low probability because the explainer implies the reasoning not only for discharge but also prolonged discharge. Similarly, it is possible to obtain information based on the expected discharge date of each patient, such as bed capacity in the near future. For the human and physical resources of the hospital to be used efficiently, future bed availability information can help reduce hospital costs through better management of beds and hospitalization reservations.

Discussion

Principal Findings

Investigations into bed management, which requires the use of hospital processes, and biomarker detection for patient-specific care, are actively pursued. In this study, we propose an ML-based predictive model to identify the discharge date for better bed management and the risk factors regarding discharge and CVDs. However, because each hospital has varying environmental variables, an algorithm that can consider them collectively was needed. Our study can contribute to improving the algorithm and supporting health care services. We have summarized the expectations of our predictive model and its explanation, along with its limitations.

First, we predicted the possibility of discharge to learn future information, but for the model to be practically applied, objective information about the current bed situation must be obtained. Currently, we are collecting bed information to combine it with the prediction results and optimize overall bed management. Consequently, our predictive model can be extended from ward-level up to hospital-level bed management. It may reduce the labor-intensive tasks for the medical team and the waiting time for patients.

Second, although our model provides adjustment of the optimal threshold according to the hospital circumstances, the ambiguity of decision-making because of results near the threshold exists, such as dates 10 and 11 in Figure 9. To solve this problem, there is a method that uses weighted average to make the result more conservative but reliable. Instead of using the probability returned by the model directly, it may be more useful to use it after weighting it for the past results, so that the target day reflects the weighted past results. It is just as necessary to produce reliable results as it is trying to explain the model and its internal features.

Finally, EHRs are longitudinal and sequential, but the sequence is different for each patient, and they do not have a regular interval. Consequently, we are preparing a preprocessing technique that can properly control the EHRs and reflect them in the model. Furthermore, compared with computer visualization, sequential data are relatively difficult to apply to XAI. Still, we are preparing explainable methods that are compatible with these data.

Conclusions

In this study, we have proposed an ML-based model to predict the daily discharge probability for each patient and demonstrated the individual explainer for any date during hospitalization, along with the reasonable contributing features. Our XGB model accomplished an AUROC of 0.865 and represented the simulated bed management based on explainable features. It could assist the medical team and patients in identifying the individual and common risk factors in CVDs and support hospital administrators in improving the management of hospital beds and other resources.

Acknowledgments

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation grant funded by the Korean government (2021-0-00982, 50%) and by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute funded by the Ministry of Health and Welfare, Republic of Korea (HR21C0198, 50%).

Conflicts of Interest

None declared.

References

1. Wei Y, Yu H, Geng J, Wu B, Guo Z, He L, et al. Hospital efficiency and utilization of high-technology medical equipment: a panel data analysis. Health Policy Technol 2018 Mar;7(1):65-72. [doi: 10.1016/j.hlpt.2018.01.001]
2. Novati R, Papalia R, Peano L, Gorraz A, Artuso L, Canta M, et al. Effectiveness of an hospital bed management model: results of four years of follow-up. Ann Ig 2017;29(3):189-196 [FREE Full text] [doi: 10.7416/ai.2017.2146] [Medline: 28383610]
3. Barnes S, Hamrock E, Toerper M, Siddiqui S, Levin S. Real-time prediction of inpatient length of stay for discharge prioritization. J Am Med Inform Assoc 2016 Apr;23(e1):2-10 [FREE Full text] [doi: 10.1093/jamia/ocv106] [Medline: 26253131]
4. Cardiovascular diseases (CVDs). World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [accessed 2021-06-11]
5. Bachouch R, Guinet A, Hajri-Gabouj S. An integer linear model for hospital bed planning. Int J Prod Econ 2012 Dec;140(2):833-843. [doi: 10.1016/j.ijpe.2012.07.023]
6. Troy PM, Rosenberg L. Using simulation to determine the need for ICU beds for surgery patients. Surgery 2009 Oct;146(4):608-617. [doi: 10.1016/j.surg.2009.05.021] [Medline: 19789019]
7. Turgeman L, May JH, Sciulli R. Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission. Expert Syst Appl 2017 Jul;78:376-385. [doi: 10.1016/j.eswa.2017.02.023]
8. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res 2017;4(1):e000234 [FREE Full text] [doi: 10.1136/bmjresp-2017-000234] [Medline: 29435343]
9. Ma F, Yu L, Ye L, Yao DD, Zhuang W. Length-of-stay prediction for pediatric patients with respiratory diseases using decision tree methods. IEEE J Biomed Health Inform 2020 Sep;24(9):2651-2662. [doi: 10.1109/jbhi.2020.2973285]
10. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 8;1(1):18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]

11. Tamarappoo BK, Lin A, Commandeur F, McElhinney PA, Cadet S, Goeller M, et al. Machine learning integration of circulating and imaging biomarkers for explainable patient-specific prediction of cardiac events: a prospective study. Atherosclerosis 2021 Feb;318:76-82. [doi: 10.1016/j.atherosclerosis.2020.11.008] [Medline: 33239189]
12. Adadi A, Berrada M. Peeking Inside the Black-Box: a survey on Explainable Artificial Intelligence (XAI). IEEE Access 2018;6:52138-52160. [doi: 10.1109/access.2018.2870052]
13. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun 2020 Jul 31;11(1):3852 [FREE Full text] [doi: 10.1038/s41467-020-17431-x] [Medline: 32737308]
14. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017 Presented at: IEEE International Conference on Computer Vision (ICCV); Oct. 22-29, 2017; Venice, Italy p. 618-626. [doi: 10.1109/iccv.2017.74]
15. Jun TJ, Eom Y, Kim D, Kim C, Park J, Nguyen H, et al. TRk-CNN: transferable ranking-CNN for image classification of glaucoma, glaucoma suspect, and normal eyes. Expert Syst Appl 2021;182:115211. [doi: 10.1016/j.eswa.2021.115211]
16. Onishi K. Total management of chronic obstructive pulmonary disease (COPD) as an independent risk factor for cardiovascular disease. J Cardiol 2017 Aug;70(2):128-134 [FREE Full text] [doi: 10.1016/j.jjcc.2017.03.001] [Medline: 28325523]
17. Levin S, Barnes S, Toerper M, Debraine A, DeAngelo A, Hamrock E, et al. Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay. BMJ Innov 2020 Dec 21;7(2):414-421. [doi: 10.1136/bmjinnov-2020-000420]
18. Ahn I, Na W, Kwon O, Yang DH, Park G, Gwon H, et al. CardioNet: a manually curated database for artificial intelligence-based research on cardiovascular diseases. BMC Med Inform Decis Mak 2021 Jan 28;21(1):29 [FREE Full text] [doi: 10.1186/s12911-021-01392-2] [Medline: 33509180]
19. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002;35(5-6):352-359 [FREE Full text] [doi: 10.1016/s1532-0464(03)00034-0] [Medline: 12968784]
20. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]
21. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46(1):389-422. [doi: 10.1023/A:1012487302797]
22. Breiman L. Random forests. Mach Learn 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
23. Yan H, Jiang Y, Zheng J, Peng C, Li Q. A multilayer perceptron-based medical decision support system for heart disease diagnosis. Expert Syst Appl 2006 Feb;30(2):272-281. [doi: 10.1016/j.eswa.2005.07.022]
24. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13 - 17, 2016; San Francisco California USA p. 785-794. [doi: 10.1145/2939672.2939785]
25. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI). 1995 Presented at: 14th International Joint Conference on Artificial Intelligence (IJCAI); August 20 - 25, 1995; Montreal Quebec Canada p. 1137-1145. [doi: 10.5555/1643031.1643047]

Abbreviations

AMC: Asan Medical Center
AUROC: area under the receiver operating characteristic
CVD: cardiovascular disease
EHR: electronic health record
GBM: gradient boosting algorithm
ICU: intensive care unit
INDT: date of visitation or admission
INNO: patient encounter number
LOS: length of stay
ML: machine learning
OHE: one-hot encoding
OUDT: date of discharge
PAID: patient identification
RFECV: recursive feature elimination with cross-validation
ROC: receiver operating characteristic
XAI: explainable artificial intelligence
XGB: extreme gradient boosting

Edited by G Eysenbach; submitted 05.08.21; peer-reviewed by A Islam; comments to author 31.08.21; revised version received 03.09.21; accepted 18.09.21; published 17.11.21.

Please cite as:
Ahn I, Gwon H, Kang H, Kim Y, Seo H, Choi H, Cho HN, Kim M, Jun TJ, Kim YH
Machine Learning–Based Hospital Discharge Prediction for Patients With Cardiovascular Diseases: Development and Usability Study
JMIR Med Inform 2021;9(11):e32662
URL: https://medinform.jmir.org/2021/11/e32662
doi:10.2196/32662
PMID:34787584

Original Paper

Deep Learning Techniques for Fatty Liver Using Multi-View Ultrasound Images Scanned by Different Scanners: Development and Validation Study

Taewoo Kim[1*], BS; Dong Hyun Lee[2*], MD, PhD; Eun-Kee Park[3], PhD; Sanghun Choi[1], PhD

[1]School of Mechanical Engineering, Kyungpook National University, Daegu, Republic of Korea

[2]Division of Gastroenterology, Department of Internal Medicine, Good Gang-An Hospital, Busan, Republic of Korea

[3]Department of Medical Humanities and Social Medicine, College of Medicine, Kosin University, Busan, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Sanghun Choi, PhD
School of Mechanical Engineering
Kyungpook National University
80 Daehak-ro
Buk-gu
Daegu, 41566
Republic of Korea
Phone: 82 53 950 5578
Fax: 82 53 950 6550
Email: s-choi@knu.ac.kr

Abstract

Background: Fat fraction values obtained from magnetic resonance imaging (MRI) can be used to obtain an accurate diagnosis of fatty liver diseases. However, MRI is expensive and cannot be performed for everyone.

Objective: In this study, we aim to develop multi-view ultrasound image–based convolutional deep learning models to detect fatty liver disease and yield fat fraction values.

Methods: We extracted 90 ultrasound images of the right intercostal view and 90 ultrasound images of the right intercostal view containing the right renal cortex from 39 cases of fatty liver (MRI–proton density fat fraction [MRI–PDFF] ≥ 5%) and 51 normal subjects (MRI–PDFF < 5%), with MRI–PDFF values obtained from Good Gang-An Hospital. We obtained combined liver and kidney-liver (CLKL) images to train the deep learning models and developed classification and regression models based on the VGG19 model to classify fatty liver disease and yield fat fraction values. We employed the data augmentation techniques such as flip and rotation to prevent the deep learning model from overfitting. We determined the deep learning model with performance metrics such as accuracy, sensitivity, specificity, and coefficient of determination (R^2).

Results: In demographic information, all metrics such as age and sex were similar between the two groups—fatty liver disease and normal subjects. In classification, the model trained on CLKL images achieved 80.1% accuracy, 86.2% precision, and 80.5% specificity to detect fatty liver disease. In regression, the predicted fat fraction values of the regression model trained on CLKL images correlated with MRI–PDFF values (R^2=0.633), indicating that the predicted fat fraction values were moderately estimated.

Conclusions: With deep learning techniques and multi-view ultrasound images, it is potentially possible to replace MRI–PDFF values with deep learning predictions for detecting fatty liver disease and estimating fat fraction values.

(JMIR Med Inform 2021;9(11):e30066) doi:10.2196/30066

KEYWORDS

fatty liver; deep learning; transfer learning; classification; regression; magnetic resonance imaging–proton density fat fraction; multi-view ultrasound images; artificial intelligence; machine imaging; imaging; informatics; fatty liver disease; detection; diagnosis

Introduction

Fatty liver disease is a disease in which fat accumulates in the liver, leading to more severe diseases, such as liver fibrosis, cirrhosis, and liver cancer [1,2]. Fatty liver disease is divided into alcoholic fatty liver disease caused by alcohol consumption and nonalcoholic fatty liver caused by metabolic diseases such as insulin resistance or abdominal obesity [3,4]. While alcoholic and nonalcoholic fatty liver have different etiologies, distinguishing them is very challenging on the basis of subjective symptoms, blood tests, imaging tests, or even histological tests; so, it usually relies on medical history based on alcohol consumption [5-7]. Recently, the prevalence of nonalcoholic fatty liver disease has reached 30% of the world's population owing to lifestyle changes, and the disease has been investigated to be highly related to cardiovascular disease and other organ cancers, attracting more attention from medical practitioners. Thus, fatty liver disease is considered a critical issue in the field of health care in today's society, whereas disease symptoms are not noticeable until the disease progresses to a critical stage. Furthermore, the disease is difficult to detect in an early stage owing to the limitation of diagnostic technology.

As of now, a liver biopsy has been regarded as the gold standard for diagnosing fatty liver disease and assessing the degree of fibrosis owing the fatty liver. However, liver biopsy is rarely performed clinically owing to its invasiveness, which can lead to serious complications. In addition, liver biopsy is limited to represent the entire liver because only a small portion of the liver is extracted. As a noninvasive method, imaging methods have been used to diagnose the fatty liver, including ultrasonography, computed tomography, and magnetic resonance imaging (MRI) of the abdomen. The MRI method consists of MRI–proton density fat fraction (MRI–PDFF) or MR spectroscopy [8-12]. The MRI–PDFF method measures fat fraction values in fatty liver, being computed by the ratio of fat protons to fat and water protons in the liver [13]. MR spectroscopy also measures the degree of fatty liver disease. Except for liver biopsy, MRI has been considered the best method in assessing fatty liver, but it is relatively expensive and cannot be carried out in hospitals without MRI equipment. On the other hand, abdominal ultrasound is the most widely used diagnostic method in clinical practice because it is relatively inexpensive and can be performed in most hospitals. However, abdominal ultrasonography has some disadvantages, such that it is highly dependent on the skill of the person conducting the examination and less sensitive to detecting early-stage fatty liver disease. Recently, several studies have been conducted to overcome the limitations of abdominal ultrasound examination and to objectify or automate fatty liver disease diagnosis through abdominal ultrasound examination [14]. Reddy et al [15] demonstrated that ultrasound images could be used to classify fatty liver diseases in computer-aided diagnosis systems, achieving 90.6% classification accuracy. In this context, we aim to develop a model that can classify fatty liver disease using B-mode ultrasound images, and to develop a regression model that can obtain fat fraction values in fatty liver based on a model architecture with the best classification performance.

Several studies have used deep learning (DL) techniques and ultrasound images to classify fatty liver disease and measure fat fraction values. Zhang et al [16] demonstrated that features of B-mode ultrasonic images can be used in a convolutional neural network (CNN)–based model, achieving 90% accuracy. They showed that unique features obtained from ultrasound images could classify fatty liver disease. Similarly, Lin et al [17] presented a novel quantitative ultrasound technique, and Han et al [18] showed a quantitative raw radiofrequency ultrasound signal method to classify fatty liver disease and measure fat fraction values. They demonstrated that preprocessed data obtained from ultrasound images may facilitate a more comprehensive characterization of fatty liver disease. However, to use preprocessed data obtained from ultrasound images, we must use a specific scanner to provide additional information, making classification using ultrasound images difficult. Therefore, we have developed a DL model that can classify fatty liver disease using liver images and kidney-liver images regardless of ultrasound scanners.

With big data sets, there were several pretrained models showing good classification performance. For example, the VGG19 model won the second prize at the 2014 imagenet large-scale visual recognition competition (ILSVRC) [19]. It had the characteristic of architectural simplicity. In addition, InceptionV3 included the batch normalization method and more layers to improve the model performance, which won the first prize at the 2014 ILSVRC [20]. However, since InceptionV3 had a more complex model architecture, people attempted numerous transfer learning methods using VGG19. Furthermore, Resnet included the skip connection method to improve classification performance using complex model architecture; so, this model won the 2015 ILSVRC [21]. Although several pretrained models including more complex model architecture showed good classification performance, they need more computational sources and time. In our previous study, VGG19 provided the best classification performance in terms of sensitivity and area under curve (AUC) scores [22]. Thus, to train our ultrasound image data set, we selected VGG19 that has a comparatively simple architecture and good classification performance.

In this study, we hypothesize that multi-view ultrasound images and DL technology can effectively classify fatty liver disease and measure fat fraction values. In addition, to validate the effectiveness of using multi-view ultrasound images for classification, we evaluated the DL model's performance on only liver images or kidney-liver images. We identified the decision-making area using a gradient class activation mapping method. Furthermore, we compared the diagnosis of a radiologist with the diagnostic predictions of the DL model using ultrasound images of fatty liver disease and normal subjects without MRI–PDFF values to demonstrate the difference in the 2 diagnoses.

Methods

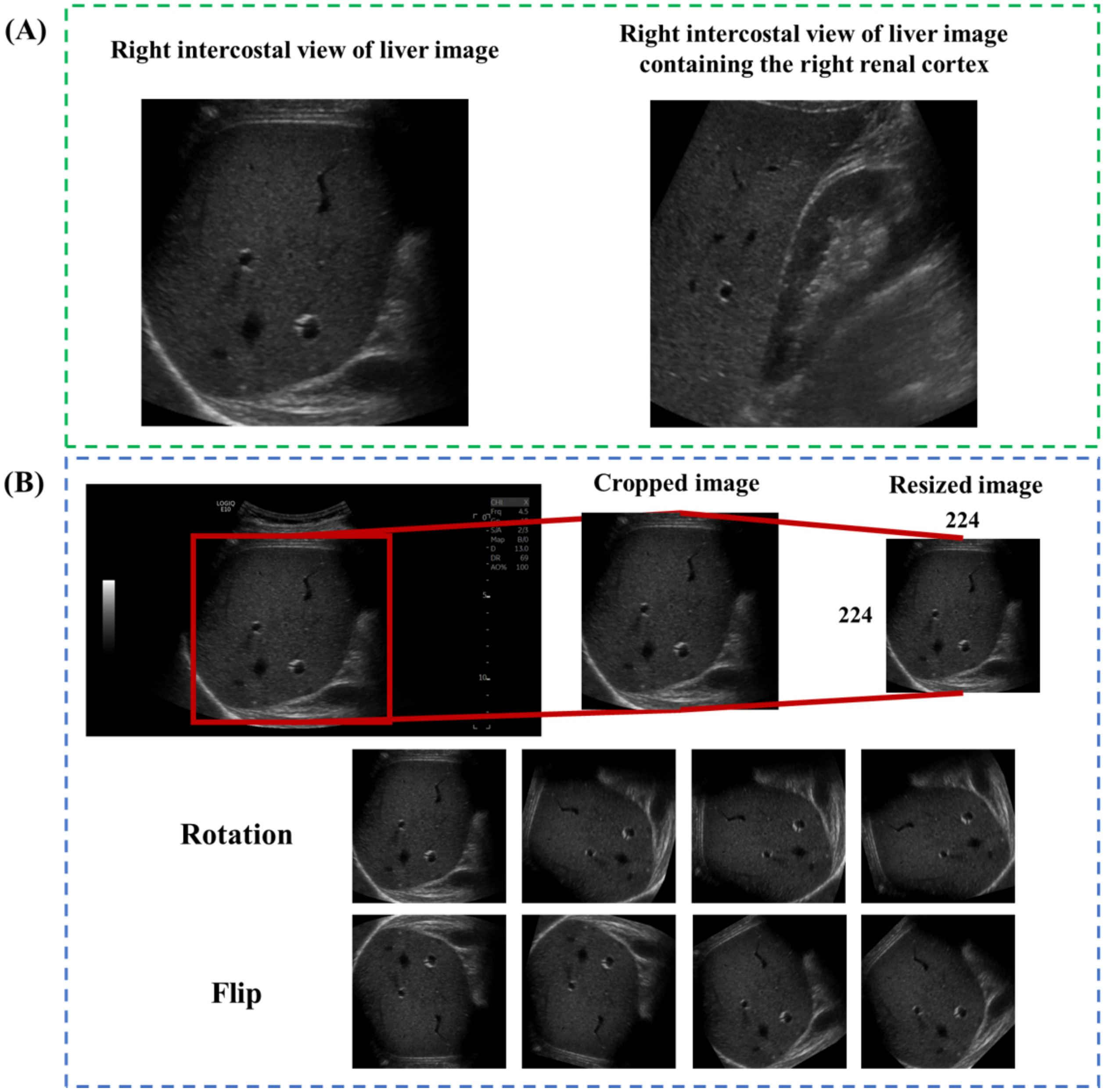
Ethics Approval

This study was approved by the institutional review board at Good Gang-An Hospital (GGAH 2020-06).

Study Population

To classify fatty liver disease, we obtained ultrasound images from 90 subjects with assigned MRI–PDFF values from Good Gang-An Hospital, Busan, Republic of Korea. The subjects comprised 39 individuals with fatty liver disease and 51 normal subjects. The criterion of a 5% MRI–PDFF was used to differentiate subjects with fatty liver from normal subjects [23,24]. From their ultrasound images, we extracted the right intercostal view of the liver (liver image), and the right intercostal view of the liver containing the right renal cortex (kidney-liver image) [14] (Figure 1A). For the DL analysis, we employed 90 liver images and liver-kidney images with MRI–PDFF values, respectively. We further used images of 50 additional subjects without MRI–PDFF values to compare the DL model's classification performance with the diagnosis of a competent radiologist. Ultrasound images were obtained using either PHILIPS or GE scanners (C5-1/ABD, PHILIPS; LOGIQ E10, GE). In addition, MRI–PDFF values were obtained using either GE or Siemens MR scanners (SIGNA Creator, GE; Skyra, Siemens). The ultrasound images were obtained using 0.5-1 MHz (PHILIPS) and 1-6 MHz (GE) multifrequency transducer. Since PDFF values were obtained in accordance with regions of interest (ROI), we used the average value of PDFF values with ROI. All subjects had ultrasound and MR scans on different days, which varied by an average of 45.1 days. Since clinical test results were collected on the date of recording of MR or ultrasound images, we collected clinical and demographic information obtained on the date of ultrasound imaging. Otherwise, we selected clinical and demographic information recorded as close as possible to the ultrasound imaging date. The metrics of clinical tests included hemoglobin, hematocrit, platelet count, aspartate aminotransferase (AST), alanine aminotransferase (ALT), total bilirubin, albumin, glucose, total cholesterol, high-density lipoprotein (HDL), and low-density lipoprotein (LDL).

Figure 1. Representative ultrasound images (A) and detailed preprocessing (B).


(A)
Right intercostal view of liver image
Right intercostal view of liver image containing the right renal cortex
(B)
Cropped image
Resized image
224
224
Rotation
Flip

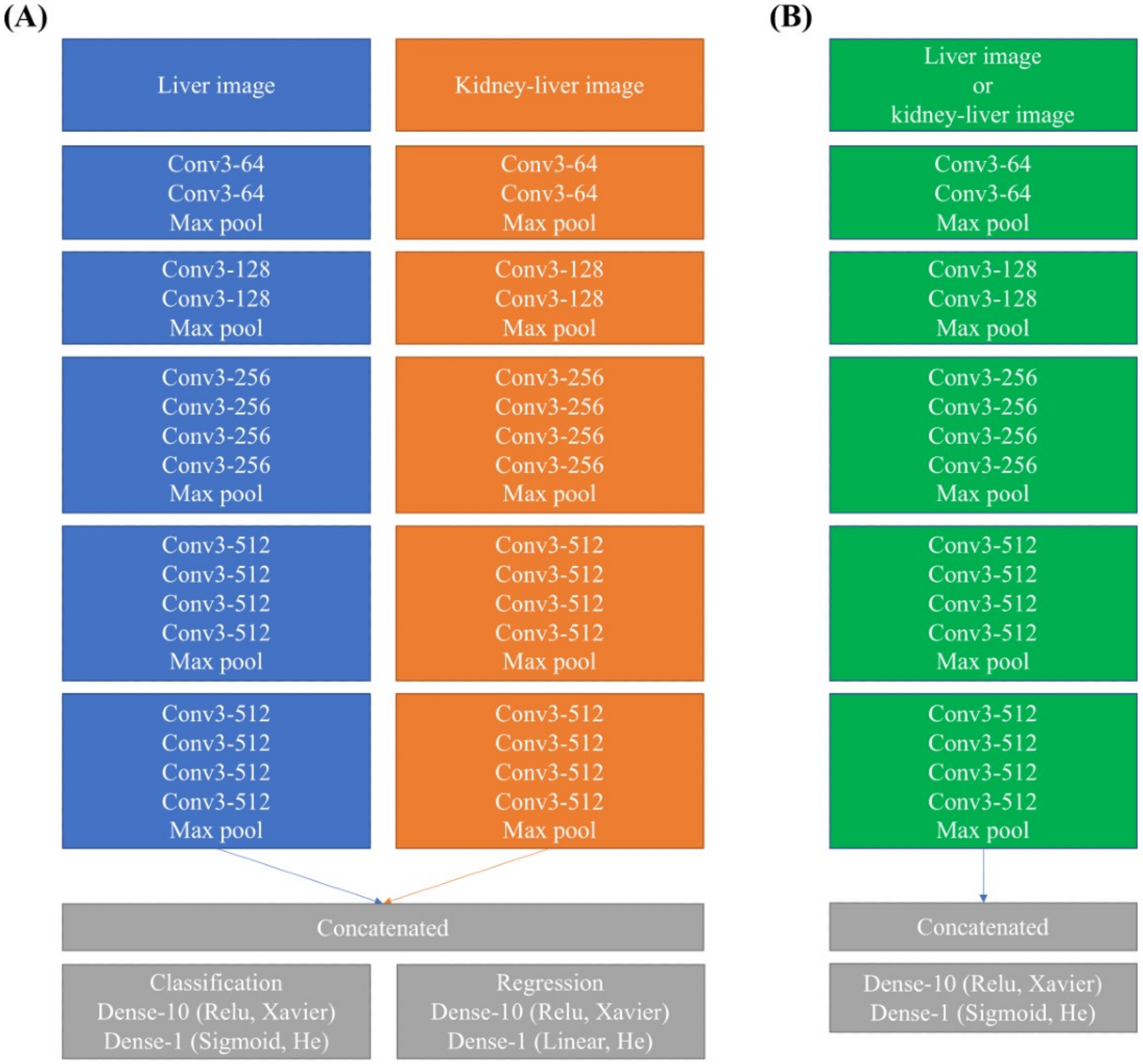
Data Processing

For image preprocessing, the red box region from an ultrasound image was cropped, and the cropped image was resized to a fixed size of 224 × 224 pixels (Figure 1B). To increase the number of ultrasound images, we used data augmentation techniques such as rotation and flipping [25]. We increased the rotation angle by 15° from 0° to 180° to obtain a total of 13 images. In addition, each image was flipped along the x- and y-axes to obtain 39 images from a single image. These image transformation techniques were essential to ensure robustness of the DL model in case samples were not enough. The technique was only applied on training images. To reduce the confounding effect of scanners, the same portions of image samples from respective PHILIPS and GE scanners were provided into the training and testing sets, respectively. In addition, to validate the deep learning model, we employed the 5-fold cross-validation method for all data, which included 17-19 ultrasound images in each fold. Thus, the training images comprised approximately 2800 images for the augmentation method. In addition, we used MinMax scaler to normalize the data to prevent the model from overfitting [26]. The test data were also transformed using this scaler.

CNN-Based Classification and Regression

Our model architectures of combined liver and kidney-liver (CLKL) images model and only liver or kidney-liver images model are shown in Figure 2. We applied a pretrained DL model of VGG19 on the preprocessed ultrasound images. This model was developed by University of Oxford, being typically used for image classification and localization. We extracted the weights of each node and architecture of the existing model of VGG19. To train the CLKL images, we concatenated their weights at the last layer of each VGG19 model (Figure 2A). In the combined VGG19 model, we constructed 2 layers as the classifier, which were composed of 10 and 1 nodes. We used the Xavier initialization method [27] and the He initialization method [28] in the classifier to improve classification performance. In addition, we used the stochastic gradient descent method with the Nesterov momentum. The momentum parameter is generally used to avoid a local minimum issue because it uses both past and current gradients to update weights in the deep learning model. The momentum parameter has been set to 0.9 in the original VGG19 model; so, we selected the same momentum parameter value [29]. Besides, we selected a learning rate of 10^{-4} because a study [30] using the VGG19 model demonstrated that the learning rate of 10^{-4} showed the best classification accuracy, compared to other learning rates. We also used the sigmoid activation function in the classifier. The regression model was similar to the classification model, but we used the linear activation function in the regressor instead of sigmoid function. To train only liver or kidney-liver images, we used the VGG19 model and same classifier in combined VGG19 model (Figure 2B). In the training phase, we use 64 batch sizes and 1000 epochs for training. We used the gradient-weighted class activation mapping (Grad-CAM) method to visualize the CNN learning process, generating a 2D spatial heatmap of input images that indicate the important regions of CNN predictions [31]. Furthermore, we employed the SHapley Additive exPlanations (SHAP) method to explain the decision evidence of our model [32].

Figure 2. Our model architectures of the combined liver and kidney-liver image model based on 2 VGG19 models (A) and only the liver or kidney-liver image model based on a VGG19 model (B).



Optimizer : the stochastic gradient descent method with Nesterove momentum
Learning rate = 0.0001, momentum parameter = 0.9
Batch size = 64, Total epoch = 1000 (Classification), 500 (Regression)

To confirm the classification performance of the combined pretrained model without fine-tuning, we set control models: the combined pretrained model with fine-tuning and without data augmentation. The pretrained model including convolutional layers and fully connected layers is updated by new data set, which is called the fine-tuning process. Without the fine-tuning process, fully connected layers of the pretrained model were only updated by the new data set.

Performance Evaluation Methods and Statistical Analysis

We evaluated the pretrained DL model's performance in 5 different cases, using the preprocessed ultrasound images. The pretrained DL model was tested on only liver images, only kidney-liver images, and both liver and kidney-liver images with or without augmentation and fine-tuning. We used the same data set and ultrasound images in each case to compare the classification performance. We used 6 performance metrics to evaluate the classification performance of the model in each case: accuracy, precision, recall (sensitivity), specificity, and F1 score. These metrics were obtained from a confusion matrix, which consists of true positive, true negative, false negative, and false positive. We used the R^2 score to compare the regression model's performance in this study with that of other studies [33]. For demographic data, we respectively used the Kruskal-Wallis and Fisher exact tests to compare continuous and categorical data between subjects with fatty liver disease and normal subjects in their history of drinking or the lack thereof (Table 1). Keras library (Keras version 2.2.4) were employed to construct deep learning models in the Python framework (version 3.6.5). In addition, statistical analyses were conducted using R software (version 3.6.1).

Table 1. Comparisons of demographic metrics between normal subjects and those with fatty liver with regard to their history of drinking or lack thereof.

Characteristics	No history of drinking (n=74)			History of drinking (n=16)		
	Normal subjects (n=42)	Subjects with fatty liver (n=32)	*P* value	Normal subjects (n=9)	Subjects with fatty liver (n=7)	*P* value
Age (years), mean (SD)	57.29 (11.74)	52.47 (13.55)	.19	53.6 (11.8)	61.3 (2.9)	.14
Females, n (%)	22 (52.4)	14 (43.8)	.49	2 (22.2)	2 (28.6)	>.99
Magnetic resonance imaging–proton density fat fraction (%)	2.96 (0.90)	11.82 (8.74)	<.001	3.11 (1.08)	11.49 (5.49)	<.001
Weight (kg), mean (SD)	63.6 (9.0)	70.5 (10.6)	<.05	73.5 (12.8)	73.1 (9.0)	.71
Height (cm), mean (SD)	165.3 (8.1)	164.6 (9.7)	>.99	170.6 (8.0)	169.2 (5.6)	.60
Hemoglobin (g/dL), mean (SD)	13.8 (1.6)	14.7 (1.9)	<.05	13.6 (2.4)	13.3 (1.5)	.60
Hematocrit (%)	41.1 (4.0)	43.3 (5.4)	<.05	40.2 (6.2)	39.4 (5.0)	.96
Platelet count (10^3/uL), mean (SD)	170.7 (65.0)	204.4 (82.8)	.16	173.6 (75.9)	141.3 (51.7)	.32
Aspartate transaminase (U/L), mean (SD)	43.7 (37.4)	61.0 (74.3)	.24	40.0 (19.7)	82.4 (39.5)	<.05
Alanine transaminase (U/L), mean (SD)	39.7 (54.1)	58.9 (74.0)	<.05	25.3 (16.5)	42.4 (24.6)	.11
Total bilirubin (mg/dL), mean (SD)	1.07 (1.03)	0.83 (0.40)	.44	1.08 (0.54)	2.02 (2.31)	.13
Albumin (g/dL), mean (SD)	4.09 (0.48)	4.30 (0.44)	<.05	4.17 (0.55)	3.47 (1.01)	.19
Glucose (mg/dL), mean (SD)	118.4 (31.8)	132.5 (67.2)	.78	124.6 (26.5)	137.5 (65.8)	.85
Total cholesterol (mg/dL), mean (SD)	172.6 (61.8)	183.5 (57.8)	.63	162.7 (37.3)	135.0 (48.8)	.31
High-density lipoprotein cholesterol (mg/dL), mean (SD)	52.3 (11.8)	49.1 (16.9)	.10	50.2 (16.6)	41.7 (19.0)	.66
Low-density lipoprotein cholesterol (mg/dL), mean (SD)	104.1 (30.0)	113.9 (39.0)	.36	103.0 (47.1)	93.5 (43.1)	.81

Results

Demographic Information

Table 1 shows the comparison between subjects with fatty liver disease and normal subjects with respect to their history of drinking or the lack thereof. Regarding both history of drinking and the no history of drinking groups, age, weight, height, and gender were not significantly different between normal and fatty liver groups. Regarding clinical metrics, hemoglobin, hematocrit, ALT, and albumin values were different between the two groups in no history of drinking group, whereas the AST levels of only subjects with fatty liver were higher than those of control subjects in history of drinking group.

CNN-Based Classification

Figure 3 shows the accuracy, precision, recall, F1 score, and specificity of the pretrained models along with the types of input image (Figure 3A) and along with transfer learning with or without fine-tuning and without augmentation (Figure 3B). Compared to other models, the CLKL image–trained model had the highest accuracy, precision, and F1 score (Figure 3A). In particular, the precision of the combined model was 86.2%, which is 14.2% higher than that of the other models. The kidney-liver image–trained model had the lowest classification performance with regard to accuracy, precision, and F1 score, whereas the liver image–trained model had the lowest specificity, compared to that of other models. With fine-tuning (Figure 3B), the fine-tuned CLKL image–trained model also had lower accuracy, precision, recall, and F1 scores, compared to those of the transfer learning model without fine-tuning. However, the CLKL image–trained model without fine-tuning had the highest classification performance than that of other models with fine-tuning and without the augmentation method. Figure 4 shows the confusion matrix and the receiver operating characteristic (ROC) curve of the CLKL image–trained model. The CLKL image–trained model had 1 false positive and 2 false negative value in the average-confusion matrix, and the average AUC score was 0.87, indicating that the DL model had good classification performance. To validate the performance of DL model–based predictions, we applied Grad-CAM to the trained CNN model. Figure 5 shows the focal region of CNN predictions of the combined liver and kidney-liver image-trained model using the Grad-CAM method. Figure 5A shows a normal subject's image, with an MRI–PDFF lower than 5%, and Figure 5B shows an image of a subject with fatty liver, with an MRI–PDFF higher than 5%. The heatmaps highlighted the general liver region in liver images and both the central region of the kidney and liver region in kidney-liver images. Figure 6 shows the focal region of CNN predictions of the CLKL image–trained model using the SHAP method. The SHA*P* values of fatty liver images were positively higher in the hepatic portal and kidney regions. On the other hand, the SHA*P* values of normal images were negatively higher in the liver and kidney regions.

Figure 3. The classification performance of the transfer learning model along with the input ultrasound image view (A) and comparison of classification performance between the transfer learning model and transfer learning with fine-tuning or without augmentation (B), including accuracy, precision, recall, F1 score, and specificity.

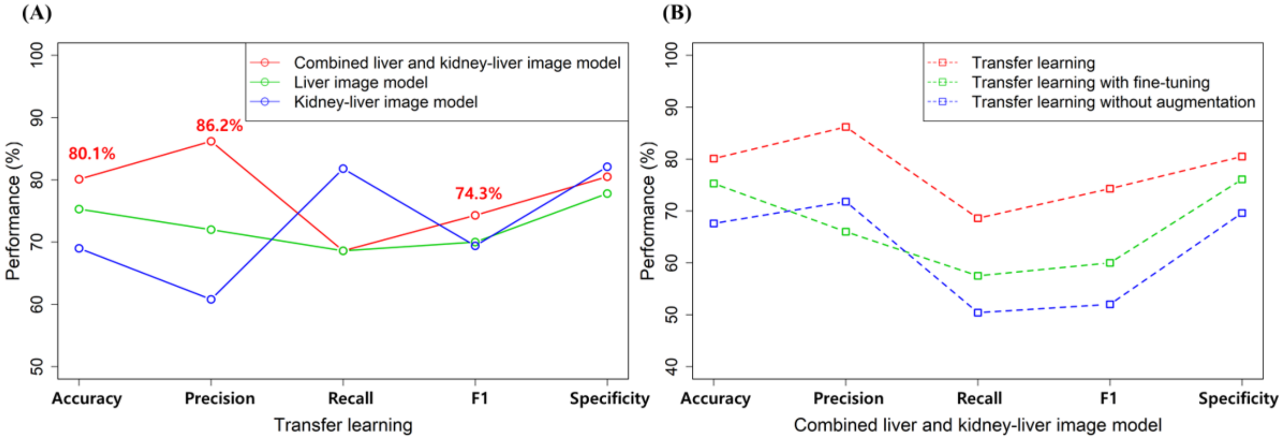


Figure 4. The average-confusion matrix (A) and average-ROC curve (B) using transfer learning model. AUC: area under the curve; ROC: receiver operating characteristic.

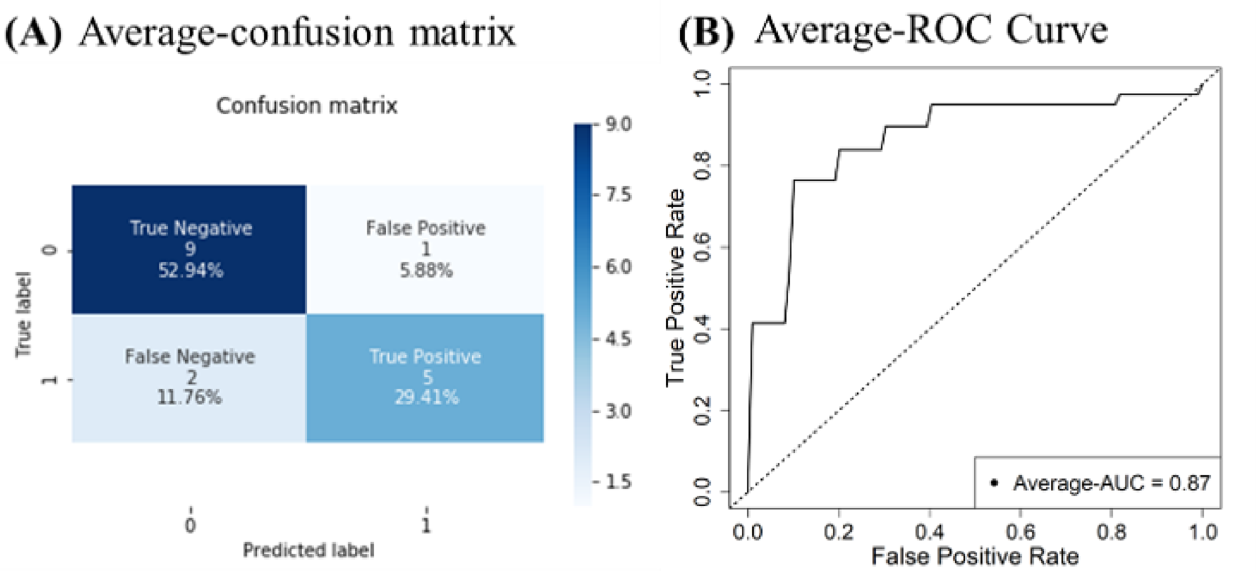


Figure 5. The focal region of CNN predictions of the combined liver and kidney-liver image–trained model. CNN: convolutional neural network.

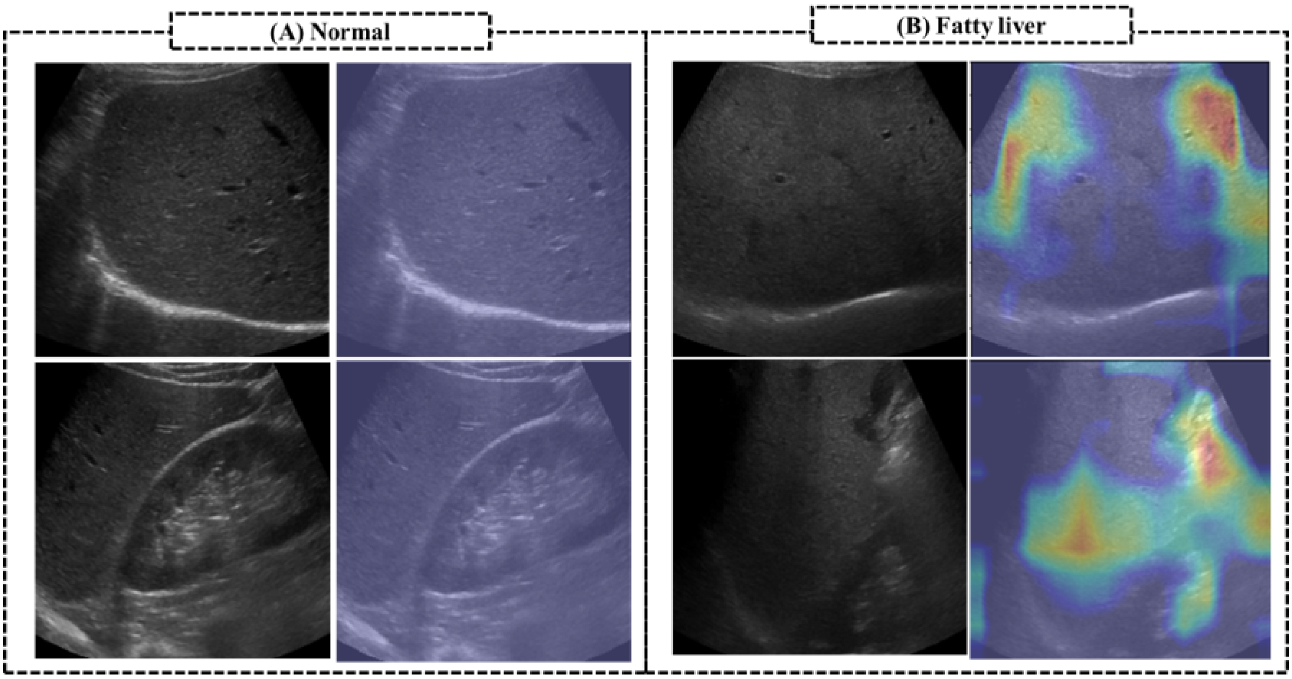
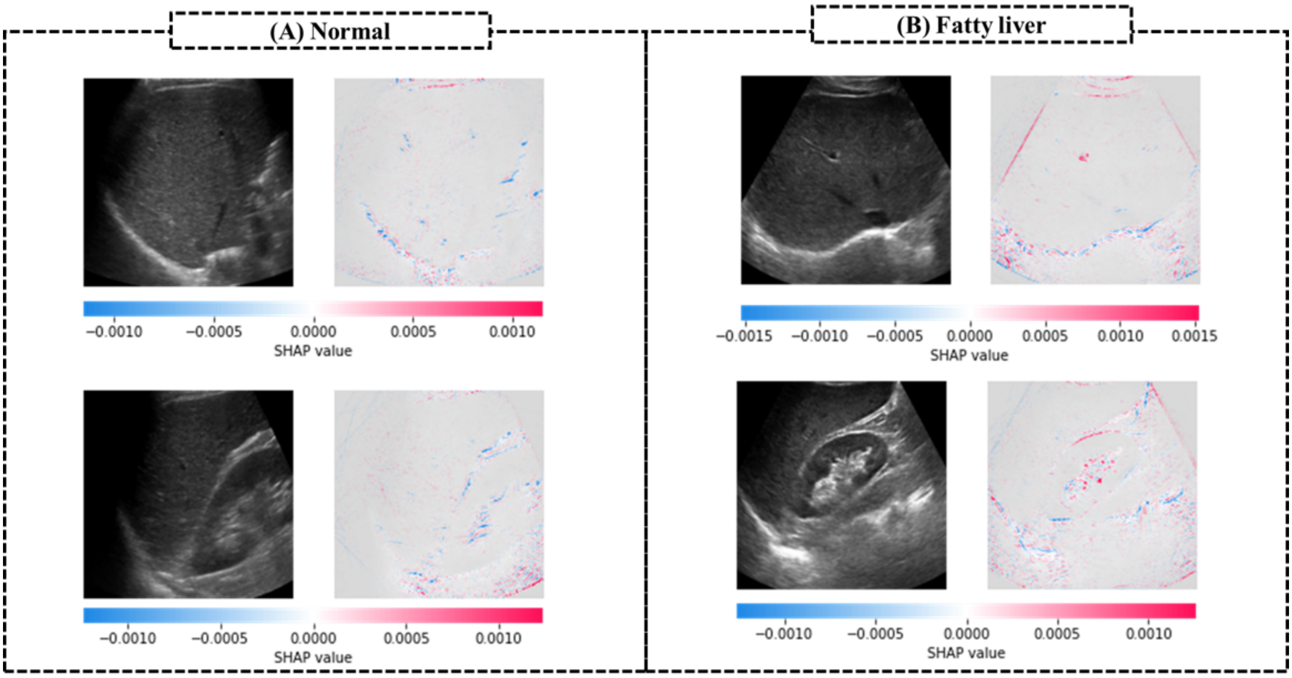


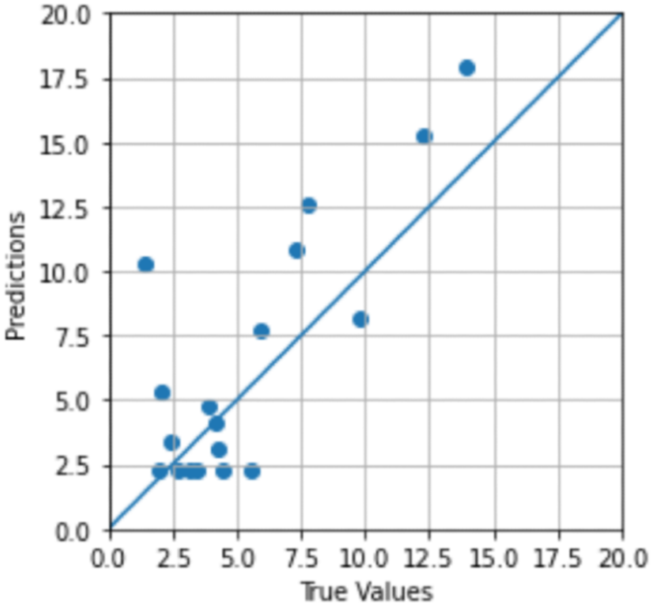
Figure 6. The focal region of CNN predictions of the combined liver and kidney-liver image–trained model using the SHAP method. CNN: convolutional neural network, SHAP: SHapley Additive exPlanations.



Regression Model Derived From the Classification Model of the Best Performance

Using the architecture of the CLKL image–trained model, which achieved the best classification performance, we developed the regression DL model using the CLKL images and MRI–PDFF values for 1 among 5 folds. Figure 7 shows the predicted fat fraction values correlated with the MRI–PDFF values, using transfer learning. When training the pretrained DL regression model using the CLKL images, the R^2 score was approximately 0.633, indicating that the predicted fat fraction values were moderately estimated. However, when using 5 folds, the regression models were not trained owing to overfitting problems.

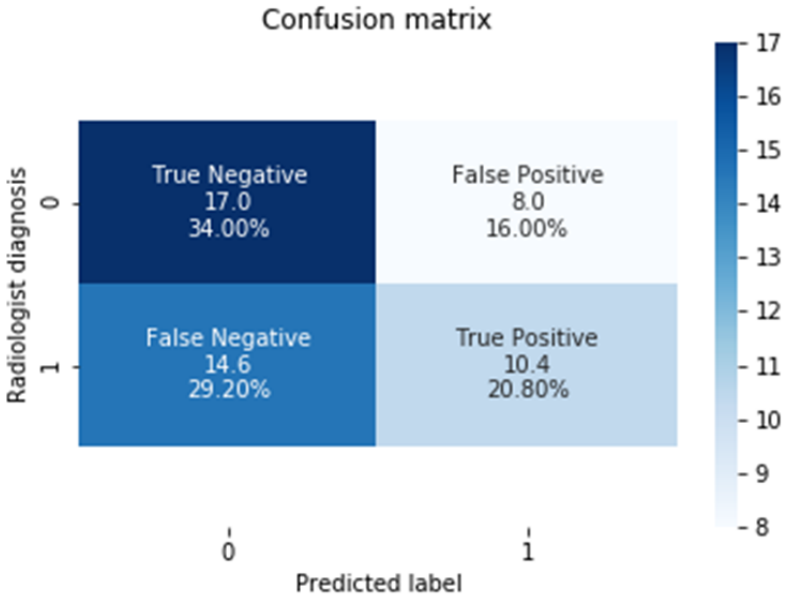
Figure 7. The correlation map of the predicted fat fraction values with MRI–PDFF values. MRI–PDFF: magnetic resonance imaging–proton density fat fraction.



Comparison Between Radiologist Diagnosis and the CNN-Based DL Model's Prediction

Using ultrasound images of the additional subjects without MRI–PDFF values, we estimated the predicted classes for the pretrained DL model with the best classification. In addition, we obtained the radiologist's diagnosis of fatty liver disease for the additional subjects' ultrasound images and compared it with the model's prediction. Figure 8 shows the confusion matrix between the classification model and the radiologist's diagnosis. The accuracy of the pretrained model was 54.8%, which indicates that predictions of the pretrained model were different from the radiologist's diagnosis.

Figure 8. The confusion matrix for the additional subjects without MRI–PDFF values between the pretrained model's prediction and the radiologist's diagnosis. MRI–PDFF: magnetic resonance imaging–proton density fat fraction.



Discussion

Principal Findings

Using the ultrasound images and pretrained DL model, we have demonstrated that multi-view ultrasound images and DL technology could effectively classify fatty liver disease and measure fat fraction values as well, regardless of the disease type, alcoholic or nonalcoholic fatty liver disease. Not only was the classification model's accuracy 80.1%, but also the R^2 value of the predicted fat fraction values obtained using the regression model was also approximately 0.633. Despite using different scanners to obtain the ultrasound images, the performance of the classification was similar to that of MRI–PDFF values. In addition, to diagnose fatty liver, radiologists often used multi-view ultrasound images including the right intercostal image of the liver and the right intercostal image of the liver including the right renal cortex [14]. We confirmed that the deep learning model could also use those ultrasound images to classify the subjects with fatty liver.

When using a different data set rather than a pretrained data set, transfer learning with fine-tuning originally had better performance than transfer learning without fine-tuning. However, in our study, transfer learning without fine-tuning had better classification performance than that with fine-tuning. This is most likely because the size of our data set for updating the weights of all layers was small. Moreover, for this reason, the regression model also showed poor performance in the 5-fold cross-validation models. On comparing the classification model's prediction and the radiologist's diagnosis, the diagnosis and the model prediction for many subjects were inconsistent. However, although the MRI–PDFF values for additional subjects should be confirmed, the possibility that it could be applied clinically in the future could be confirmed by matching for half of the radiologist's diagnosis (Figure 8).

Limitations

There were several limitations in this study. First, a radiologist scanned ultrasound images with 2 scanners. As the confounding effect of different scanners may affect DL models, DL models should be developed using ultrasound images scanned from a single scanner. Second, in this study, we collected liver images of subjects with fatty liver disease, including alcoholic fatty liver and nonalcoholic fatty liver disease. However, regardless of the type of fatty liver disease, our model could estimate the predicted fat fraction values and classify fatty liver classes, so this is not a fatal flaw in our study. Third, ultrasound and MR imaging were performed on different dates, which may have a confounding effect on our results. Although our models showed good classification and regression performance, this study has been retrospectively designed using ultrasound images and MRI–PDFF determined on different dates. Thus, ultrasound images and MRI–PDFF obtained on the same date should be used in future studies. Finally, the location of the liver or kidney in the ultrasound images was different; so, this may have a confounding effect on the DL models' performance. We used a data augmentation technique, including rotation and flip, to reduce the confounding effect of the liver location. Thus, we may be able to free our models from this confounding effect of the liver location.

Comparison With Prior Work

Several previous studies have used ultrasound images and DL techniques in this context (Table 2). Reddy et al [15] proposed a novel computer-aided diagnosis framework for fatty liver disease. They scanned and collected 86 normal liver images and 76 fatty liver images using the same scanner and used the pretrained DL model with transfer learning and fine-tuning. They obtained 90.6% accuracy, 95% sensitivity, and 85% specificity. Byra et al [34] also proposed a similar DL framework and obtained 96.3% accuracy, 100% sensitivity, and 88.2% specificity using transfer learning and B-mode images scanned using the same scanner. In addition, Han et al [18] proposed a noninvasive diagnosis system of nonalcoholic fatty liver disease and a quantification system of the liver fat fraction values using features extracted from ultrasound images. They collected ultrasound images and MRI–PDFF values of 204 prospectively enrolled participants with nonalcoholic fatty liver disease and participants without fatty liver disease. They used raw radiofrequency ultrasound signals obtained from the ultrasound image scanner and obtained 96% accuracy, 97% sensitivity, 94% specificity, and an R^2 value of 0.79 using DL techniques. Although the classification performance of our model was inferior to that reported in previous studies, it is inadequate to compare our study with previous studies using the same scanner. It is impossible to generalize the DL model using ultrasound images obtained from the same scanner. Thus, we believe that our study is more generalized than other studies because our study used ultrasound images obtained using 2 different scanners.

Table 2. Previously published classification results of the fatty liver versus the normal data sets.

Related work	Data	Methods	Accuracy
Reddy et al [15]	86 normal liver images and 76 fatty liver images using the same scanner	Transfer learning	90.6
Byra et al [34]	B-mode ultrasound images	Transfer learning	96.3
Han et al [18]	Raw radiofrequency ultrasound signals	Convolutional neural network algorithm	96.0
This study	The combined liver and kidney-liver images scanned by 2 scanners (n=90)	Transfer learning	80.1

Conclusions

In conclusion, using the pretrained DL model and ultrasound images, we demonstrated that transfer learning had the best classification (80.1% accuracy), using multi-view ultrasound images including liver and kidney-liver images. Furthermore, our study demonstrated that the predictions of fatty liver disease using the classification DL models could be implemented in the clinical field without complying with MRI–PDFF values, the gold standard, in the future. A prospective future study is required to develop DL techniques using more ultrasound images with MRI–PDFF values to confirm this study's results. Future studies can prove that ultrasound images can be used as assistant components in the clinical field, achieving more robust classification and regression performance.

Acknowledgments

This work was supported by the Korea Ministry of Environment as “The Environmental Health Action Program” (2018001360004), and the National Research Foundation of Korea Grant funded by the Korea government (MSIT) (2018R1D1A1B07040886, NRF-2021R1F1A1060436).

Authors' Contributions

TK, SC, DHL, and EKP designed the experiments and interpreted the results. DHL, EKP, and TK collected experimental data. SC and TK performed the experiments. TK, EKP, and SC performed the analyses and wrote the manuscript. EKP and SC served as co-corresponding authors. All authors provided feedback on the manuscript.

Conflicts of Interest

None declared.

References

1. Zhang YN, Fowler KJ, Hamilton G, Cui JY, Sy EZ, Balanay M, et al. Liver fat imaging-a clinical overview of ultrasound, CT, and MR imaging. Br J Radiol 2018 Sep;91(1089):20170959 [FREE Full text] [doi: 10.1259/bjr.20170959] [Medline: 29722568]
2. Dowman JK, Tomlinson J, Newsome P. Pathogenesis of non-alcoholic fatty liver disease. QJM 2010 Feb;103(2):71-83 [FREE Full text] [doi: 10.1093/qjmed/hcp158] [Medline: 19914930]
3. Stefan N, Kantartzis K, Häring HU. Causes and metabolic consequences of Fatty liver. Endocr Rev 2008 Dec;29(7):939-960. [doi: 10.1210/er.2008-0009] [Medline: 18723451]
4. Kotronen A, Yki-Järvinen H, Männistö S, Saarikoski L, Korpi-Hyövälti E, Oksa H, et al. Non-alcoholic and alcoholic fatty liver disease - two diseases of affluence associated with the metabolic syndrome and type 2 diabetes: the FIN-D2D survey. BMC Public Health 2010 May 10;10:237 [FREE Full text] [doi: 10.1186/1471-2458-10-237] [Medline: 20459722]
5. Obika M, Noguchi H. Diagnosis and evaluation of nonalcoholic fatty liver disease. Exp Diabetes Res 2012;2012:145754 [FREE Full text] [doi: 10.1155/2012/145754] [Medline: 22110476]
6. Toshikuni N, Tsutsumi M, Arisawa T. Clinical differences between alcoholic liver disease and nonalcoholic fatty liver disease. World J Gastroenterol 2014 Jul 14;20(26):8393-8406 [FREE Full text] [doi: 10.3748/wjg.v20.i26.8393] [Medline: 25024597]
7. Wong VW, Wong GL, Choi PC, Chan AW, Li MK, Chan H, et al. Disease progression of non-alcoholic fatty liver disease: a prospective study with paired liver biopsies at 3 years. Gut 2010 Jul;59(7):969-974. [doi: 10.1136/gut.2009.205088] [Medline: 20581244]
8. Cassidy FH, Yokoo T, Aganovic L, Hanna RF, Bydder M, Middleton MS, et al. Fatty liver disease: MR imaging techniques for the detection and quantification of liver steatosis. Radiographics 2009;29(1):231-260. [doi: 10.1148/rg.291075123] [Medline: 19168847]
9. Dulai PS, Sirlin CB, Loomba R. MRI and MRE for non-invasive quantitative assessment of hepatic steatosis and fibrosis in NAFLD and NASH: Clinical trials to clinical practice. J Hepatol 2016 Nov;65(5):1006-1016 [FREE Full text] [doi: 10.1016/j.jhep.2016.06.005] [Medline: 27312947]
10. Reeder SB, Sirlin CB. Quantification of liver fat with magnetic resonance imaging. Magn Reson Imaging Clin N Am 2010 Aug;18(3):337-357, ix [FREE Full text] [doi: 10.1016/j.mric.2010.08.013] [Medline: 21094444]
11. Noh H, Song X, Heo SH, Kim JW, Shin SS, Ahn KY, et al. Comparative Study of Ultrasonography, Computed Tomography, Magnetic Resonance Imaging, and Magnetic Resonance Spectroscopy for the Diagnosis of Fatty Liver in a Rat Model. J Korean Soc Radiol 2017;76(1):14. [doi: 10.3348/jksr.2017.76.1.14]
12. Kim JW, Lee Y, Park YS, Kim B, Lee SY, Yeon JE, et al. Multiparametric MR Index for the Diagnosis of Non-Alcoholic Steatohepatitis in Patients with Non-Alcoholic Fatty Liver Disease. Sci Rep 2020 Feb 14;10(1):2671 [FREE Full text] [doi: 10.1038/s41598-020-59601-3] [Medline: 32060386]
13. Reeder SB, Hu HH, Sirlin CB. Proton density fat-fraction: a standardized MR-based biomarker of tissue fat concentration. J Magn Reson Imaging 2012 Nov;36(5):1011-1014 [FREE Full text] [doi: 10.1002/jmri.23741] [Medline: 22777847]
14. Kim M, Kang B, Jun DW. Comparison of conventional sonographic signs and magnetic resonance imaging proton density fat fraction for assessment of hepatic steatosis. Sci Rep 2018 May 17;8(1):7759 [FREE Full text] [doi: 10.1038/s41598-018-26019-x] [Medline: 29773823]
15. Reddy D, Bharath R, Rajalakshmi P. A Novel Computer-Aided Diagnosis Framework Using Deep Learning for Classification of Fatty Liver Disease in Ultrasound Imaging. 2018 Presented at: 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom); September 17-20, 2018; Ostrava. [doi: 10.1109/HealthCom.2018.8531118]

16. Zhang L, Zhu H, Yang T. Deep Neural Networks for fatty liver ultrasound images classification. 2019 Presented at: 2019 Chinese Control And Decision Conference (CCDC); June 3-5, 2019; Nanchang. [doi: 10.1109/ccdc.2019.8833364]
17. Lin SC, Heba E, Wolfson T, Ang B, Gamst A, Han A, et al. Noninvasive Diagnosis of Nonalcoholic Fatty Liver Disease and Quantification of Liver Fat Using a New Quantitative Ultrasound Technique. Clin Gastroenterol Hepatol 2015 Jul;13(7):1337-1345.e6 [FREE Full text] [doi: 10.1016/j.cgh.2014.11.027] [Medline: 25478922]
18. Han A, Byra M, Heba E, Andre MP, Erdman JW, Loomba R, et al. Noninvasive Diagnosis of Nonalcoholic Fatty Liver Disease and Quantification of Liver Fat with Radiofrequency Ultrasound Data Using One-dimensional Convolutional Neural Networks. Radiology 2020 May;295(2):342-350 [FREE Full text] [doi: 10.1148/radiol.2020191160] [Medline: 32096706]
19. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv. Preprint posted online on September 4, 2014. [FREE Full text]
20. Christian S, Vincent V, Sergey I, Jonathon S, Zbigniew W. Rethinking the Inception Architecture for Computer Vision. 2015 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV. [doi: 10.1109/CVPR.2016.308]
21. Kaiming H, Xiangyu Z, Shaoqing R, Jian S. Deep Residual Learning for Image Recognition. 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV. [doi: 10.1109/CVPR.2016.90]
22. Ho TT, Kim T, Kim WJ, Lee CH, Chae KJ, Bak SH, et al. A 3D-CNN model with CT-based parametric response mapping for classifying COPD subjects. Sci Rep 2021 Jan 08;11(1):34 [FREE Full text] [doi: 10.1038/s41598-020-79336-5] [Medline: 33420092]
23. Caussy C, Alquiraish MH, Nguyen P, Hernandez C, Cepin S, Fortney LE, et al. Optimal threshold of controlled attenuation parameter with MRI-PDFF as the gold standard for the detection of hepatic steatosis. Hepatology 2018 Apr;67(4):1348-1359 [FREE Full text] [doi: 10.1002/hep.29639] [Medline: 29108123]
24. Middleton MS, Heba ER, Hooker CA, Bashir MR, Fowler KJ, Sandrasegaran K, NASH Clinical Research Network. Agreement Between Magnetic Resonance Imaging Proton Density Fat Fraction Measurements and Pathologist-Assigned Steatosis Grades of Liver Biopsies From Adults With Nonalcoholic Steatohepatitis. Gastroenterology 2017 Sep;153(3):753-761 [FREE Full text] [doi: 10.1053/j.gastro.2017.06.005] [Medline: 28624576]
25. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. J Big Data 2019 Jul 6;6(1). [doi: 10.1186/s40537-019-0197-0]
26. Jayalakshmi T, Santhakumaran A. Statistical Normalization and Back Propagationfor Classification. IJCTE 2011:89-93. [doi: 10.7763/ijcte.2011.v3.288]
27. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. 2010 Presented at: 13th International Conference on Artificial Intelligence and Statistics (AISTATS); 2010; Sardinia URL: https://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf [doi: 10.1049/pbpo161e_ch6]
28. He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. arXiv. Preprint posted online February 6, 2015 2015. [doi: 10.1109/iccv.2015.123]
29. Liu C, Belkin M. Accelerating SGD with momentum for over-parameterized learning. arXiv. Preprint posted online on October 31, 2018. [FREE Full text]
30. Anusha C, Avadhani PS. Optimal Accuracy Zone Identification in Object Detection Technique - A Learning Rate Methodology. IJEAT 2019 Oct 30;9(1):6470-6476. [doi: 10.35940/ijeat.a2258.109119]
31. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 Presented at: 2017 IEEE International Conference on Computer Vision (ICCV); October 22-29, 2017; Venice. [doi: 10.1109/iccv.2017.74]
32. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018 Oct;2(10):749-760 [FREE Full text] [doi: 10.1038/s41551-018-0304-0] [Medline: 31001455]
33. Healy MJR. The Use of R 2 as a Measure of Goodness of Fit. J R Stat Soc 1984;147(4):608. [doi: 10.2307/2981848]
34. Byra M, Styczynski G, Szmigielski C, Kalinowski P, Michałowski ?, Paluszkiewicz R, et al. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. Int J Comput Assist Radiol Surg 2018 Dec;13(12):1895-1903 [FREE Full text] [doi: 10.1007/s11548-018-1843-2] [Medline: 30094778]

Abbreviations

ALT: alanine aminotransferase
AST: aspartate aminotransferase
AUC: area under the curve
CLKL: combined liver and kidney-liver
CNN: convolutional neural network
DL: deep learning

Grad-CAM: gradient-weighted class activation mapping
HDL: high-density lipoprotein
ILSVRC: imagenet large-scale visual recognition competition
LDL: low-density lipoprotein
MRI: magnetic resonance imaging
MRI–PDFF: magnetic resonance imaging–proton density fat fraction
ROC: receiver operating characteristic
ROI: regions of interest
SHAP: SHapley Additive exPlanations

Edited by C Lovis; submitted 30.04.21; peer-reviewed by JA Benítez-Andrades, S Nagavally, T Lee; comments to author 10.07.21; revised version received 04.08.21; accepted 30.09.21; published 18.11.21.

Please cite as:
Kim T, Lee DH, Park EK, Choi S
Deep Learning Techniques for Fatty Liver Using Multi-View Ultrasound Images Scanned by Different Scanners: Development and Validation Study
JMIR Med Inform 2021;9(11):e30066
URL: https://medinform.jmir.org/2021/11/e30066
doi:10.2196/30066
PMID:34792476

Original Paper

Prediction Model of Osteonecrosis of the Femoral Head After Femoral Neck Fracture: Machine Learning–Based Development and Validation Study

Huan Wang[1], MPH; Wei Wu[2], BM; Chunxia Han[1], BSc; Jiaqi Zheng[1], MPH; Xinyu Cai[3], BM; Shimin Chang[4], PhD; Junlong Shi[5], MPA; Nan Xu[6], MD; Zisheng Ai[1], PhD

[1]Department of Medical Statistics, Tongji University School of Medicine, Shanghai, China

[2]Department of Spinal Surgery, Shanghai East Hospital, Shanghai, China

[3]Department of Orthopedics, Shanghai Tenth People's Hospital, Shanghai, China

[4]Department of Orthopedic Surgery, Yangpu Hospital, Tongji University School of Medicine, Shanghai, China

[5]Medical Record Department, Shanghai Ninth People’s Hospital, Shanghai, China

[6]Department of Radiology, Shanghai East Hospital, Shanghai, China

Corresponding Author:
Zisheng Ai, PhD
Department of Medical Statistics
Tongji University School of Medicine
No. 1239 Singping Road
Shanghai, 200092
China
Phone: 86 1 377 438 0743
Fax: 86 021 65986270
Email: azs1966@126.com

Abstract

Background: The absolute number of femoral neck fractures (FNFs) is increasing; however, the prediction of traumatic femoral head necrosis remains difficult. Machine learning algorithms have the potential to be superior to traditional prediction methods for the prediction of traumatic femoral head necrosis.

Objective: The aim of this study is to use machine learning to construct a model for the analysis of risk factors and prediction of osteonecrosis of the femoral head (ONFH) in patients with FNF after internal fixation.

Methods: We retrospectively collected preoperative, intraoperative, and postoperative clinical data of patients with FNF in 4 hospitals in Shanghai and followed up the patients for more than 2.5 years. A total of 259 patients with 43 variables were included in the study. The data were randomly divided into a training set (181/259, 69.8%) and a validation set (78/259, 30.1%). External data (n=376) were obtained from a retrospective cohort study of patients with FNF in 3 other hospitals. Least absolute shrinkage and selection operator regression and the support vector machine algorithm were used for variable selection. Logistic regression, random forest, support vector machine, and eXtreme Gradient Boosting (XGBoost) were used to develop the model on the training set. The validation set was used to tune the model hyperparameters to determine the final prediction model, and the external data were used to compare and evaluate the model performance. We compared the accuracy, discrimination, and calibration of the models to identify the best machine learning algorithm for predicting ONFH. Shapley additive explanations and local interpretable model-agnostic explanations were used to determine the interpretability of the black box model.

Results: A total of 11 variables were selected for the models. The XGBoost model performed best on the validation set and external data. The accuracy, sensitivity, and area under the receiver operating characteristic curve of the model on the validation set were 0.987, 0.929, and 0.992, respectively. The accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve of the model on the external data were 0.907, 0.807, 0.935, and 0.933, respectively, and the log-loss was 0.279. The calibration curve demonstrated good agreement between the predicted probability and actual risk. The interpretability of the features and individual predictions were realized using the Shapley additive explanations and local interpretable model-agnostic explanations algorithms. In addition, the XGBoost model was translated into a self-made web-based risk calculator to estimate an individual’s probability of ONFH.

Conclusions: Machine learning performs well in predicting ONFH after internal fixation of FNF. The 6-variable XGBoost model predicted the risk of ONFH well and had good generalization ability on the external data, which can be used for the clinical prediction of ONFH after internal fixation of FNF.

(JMIR Med Inform 2021;9(11):e30079) doi:10.2196/30079

KEYWORDS

femoral neck fracture; osteonecrosis of the femoral head; machine learning; interpretability

Introduction

Background

The incidence of hip fractures is changing worldwide. In most Western and Northern European countries, the incidence is decreasing, as well as in Singapore [1-5]. The incidence in China and America is stabilizing [6,7], whereas that in Germany, Japan, and Korea is still increasing [8-10]. Although the age-adjusted incidence of hip fractures is declining or stabilizing in some countries, the absolute number of hip fractures and the costs of associated medical care are still increasing. Femoral neck fractures (FNFs) account for approximately 48.22% to 52.26% of hip fractures [9,11,12], and 23% of young patients have osteonecrosis of the femoral head (ONFH) after internal fixation [13]. Early stages of ONFH allow hip joint preservation surgery, such as free fibula transplantation and osteotomy, before collapse of the femoral head occurs [14]. A previous study demonstrated that the short-term and medium-term success rates of early hip-preserving treatment of ONFH were between 55% and 87% [15,16].

Published papers regarding ONFH prediction are primarily based on changes in the blood circulation of the femoral head by radiological investigations, such as single-photon emission computed tomography [17]/single-photon emission computed tomography-computed tomography [18], positron emission tomography [19]/positron emission tomography-computed tomography [20], magnetic resonance imaging [21]/dynamic contrast-enhanced-magnetic resonance imaging [22], and digital subtraction angiography [23]. The sample sizes of most studies were not large enough, and their prediction results were not confirmed in subsequent prospective studies. Cui et al [24] first applied machine learning to predict small samples of ONFH. However, the accuracy, sensitivity, and area under the receiver operating characteristic curve (AUC) of the model based on naive Bayes were all lower than 80%. Zheng et al [25] and Zhu et al [26] developed a nomogram for the risk assessment of femoral head necrosis based on a traditional regression analysis. The AUCs of the validation cohort were 0.94 and 0.95. However, these studies lacked external validation.

Prediction research using machine learning involves learning models from sample data and making predictions and decisions on the new data. The support vector machine (SVM) algorithm exhibits good prediction performance and generalization ability when dealing with small sample binary classification problems [27]. Random forest (RF) and eXtreme Gradient Boosting (XGBoost) are ensemble learning algorithms that establish multiple models using the data and then integrate the modeling results from all models. Currently, they are the most popular models in the industry.

In machine learning, the black box describes models that cannot be understood by examining their parameters (eg, neural network and XGBoost) [28]. Interpretability is defined as the ability to explain or provide meaning in understandable terms to a human. The pursuit of interpretability of the black box model helps to improve users' trust in the machine learning model and provides support for human decision-making. Arrieta et al [29] summarized and distinguished between transparent models and those that can be interpreted by post hoc explainability techniques. Transparent models convey some degree of interpretability by themselves, such as logistic regression (LR) and decision trees. Post hoc explainability techniques are model-agnostic methods, including local explanations, explanations by simplification, and feature relevance explanation techniques. Further exploration based on these methods could help overcome the difficulties related to explainability and make the machine learning models more persuasive and dependable.

Objectives

This study aims to explore and compare the application value of different machine learning algorithms for the prediction of ONFH after internal fixation of FNF and to develop a prediction model of ONFH based on a machine learning algorithm. In this study, the development and validation of the prediction model was guided by the Prediction Model Risk of Bias Assessment Tool (PROBAST) [30] and adhered to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) [31] statement for reporting. To provide appropriate models to the clinic, the design, implementation, and report also considered the 20 key issues raised by Vollmer et al [32] regarding transparency, reproducibility, ethics, and effectiveness of the study.

Methods

Study Population

This multicenter retrospective follow-up study was performed at least 30 months after follow-up in patients undergoing internal fixation of FNFs. The study population comprised patients with FNF with internal fixation who were discharged from Shanghai Ninth People's Hospital, Dongfang Hospital, and Yangpu District Central Hospital from January 1, 2015, to May 1, 2018, and the Tenth People's Hospital from January 1, 2017, to May 1, 2018. By searching the inpatient electronic medical record system, medical imaging information system, laboratory information system, manual reading of cases, and follow-up, we collected 47 clinical features from 316 patients with FNF. After excluding patients who were lost to follow-up or who died, 259 patients with FNF and associated 43 variables were included in this study. The external data (n=376) were obtained from our previous retrospective cohort study [25], which collected data from a cohort of patients with FNFs from May 2013 to January 2017 at Shanghai Sixth Hospital, Tenth Hospital, and Tongji Hospital. Two patients aged >75 years were excluded.

The inclusion criteria were as follows: (1) patients with FNF treated with internal fixation aged between 18 and 75 years, (2) patients with FNF with complete baseline data, (3) follow-up time ≥30 months, (4) The American Society of Anesthesiologists anesthesia risk of approximately grade I-III, and (5) no moderate or severe pain or limitation of movement of the injured hip side that occurred before the fracture. Exclusion criteria included the following: (1) patients with FNF with pathological fractures or old fractures admitted for >2 weeks after injury; (2) patients with failed internal fixation operation; (3) patients with a history of malignant tumors, nontraumatic fractures, ipsilateral lower limb, or other fractures; (4) patients with a history of long-term diving, alcohol abuse, and fluoroquinolone, antiplatelet drug, or hormone use; (5) patients with multiple fractures at the same site, injuries on the opposite side, or fracture of both lower limbs in the past 6 months; (6) patients who experienced acute myocardial infarction, cerebrovascular accident, severe trauma, or major operation within half a year; (7) patients with vascular transplantation or free fibula transplantation during internal fixation; and (8) patients with poor compliance. The diagnosis of ONFH was based on the updated version of the Association Research Circulation Osseous grading system [14], which was displayed and approved at the Association Research Circulation Osseous conference in Dalian, China, in 2019, and was used at the same time as the ONFH Chinese grading system developed in China in 2015 [33].

Ethics Statement

The protocol for this research project was approved by the Medical and Life Science Ethics Committee of Tongji University (2019tjdx285; date June 18, 2019). Given the retrospective nature of this study, the requirement for informed consent was waived.

Independent Variables (Features)

A total of 47 clinical features were collected, and features with missing values >20% were excluded. Overall, 43 candidate variables were included in the following categories: (1) demographic information: age, sex, smoking, drinking, and age-adjusted Charlson Comorbidity Index [34]; (2) fracture related: injury cause [35], injured side, fracture position, impaction, preoperative displacement, vertical axis of the neck angle [36], and Garden classification [37]; (3) preoperative biochemical characteristics: total protein, albumin, albumin/globulin, total bilirubin, alanine aminotransferase, aspartate aminotransferase, creatinine, uric acid, and urea nitrogen; (4) preoperative routine blood parameters: red blood cells, hemoglobin, white blood cells, platelets, and hematocrit; (5) preoperative coagulation parameters: prothrombin time, fibrinogen, activated partial thromboplastin time, and international normalized ratio; (6) surgery-related parameters: American Society of Anesthesiologists grade, time to surgery, type of anesthesia, surgical treatment, and surgical method; (7) postoperative related characteristics: reduction quality [38], Lowell curve, Gotfried reduction [39], and femoral neck shortening [40]; and (8) follow-up information: interval to part weightbearing, interval to weightbearing, implant removal, and visual analog scale (VAS) score [41]. The values and definitions of the variables are presented in Multimedia Appendix 1.

Data Preprocessing

Outlier detection was performed on the raw data. Each source of outlier was checked by looking through the medical history twice, so that we knew whether the value of the outlier was true. The errors caused by incorrect manual collection were rectified. In this study, the proportion of missing variables was <5% and was substituted with the mean value. The original data, such as blood biochemical indices, were continuous variables, which were converted into low, normal, and high categorical variables according to clinical significance. According to the modeling requirements, categorical variables were transformed into dummy variables. Standardization of continuous variables was not a necessary step in preprocessing. Although there were few continuous variables included in this study, we compared the effects of standardization with nonstandardization during the modeling process. Finally, the processed data were randomly divided into a training set and validation set at a ratio of 7:3.

Development of Prediction Models

Data Balance

The ratio of ONFH to non-ONFH was 1:5, which is unbalanced. When unbalanced data are used to fit the model, the classification interface will be biased toward the minority, resulting in low sensitivity and high specificity [42]. To address this issue, we used the synthetic minority oversampling technique (SMOTE) algorithm to balance the training set. The steps of the SMOTE algorithm [43] are as follows: (1) for each sample X in the minority sample set, use the Euclidean distance as the standard to calculate the distance from all samples in the minority sample set to obtain its K nearest neighbors; and (2) set the sampling ratio according to the sample imbalance ratio and determine the sampling magnification N. For each minority sample X, randomly select several samples from its K nearest neighbors, assuming that the selected nearest neighbor is xn; (3) for each randomly selected neighbor xn, combine the original sample to construct a new sample according to the following formula: where X_j is the sample in a few classes used to synthesize new samples, is the nearest neighbor. Although the features of adjacent points in the *feature space* are similar, the newly synthesized sample set will not affect the spatial boundary of the original minority samples.

Variable Selection

A large amount of collected clinical data will inevitably contain redundant features and noise data, which will lead to overfitting in modeling and cannot be effectively classified. Variable selection is a process that can remove irrelevant and redundant features and reduce the impact of noise data on classifier performance to a certain extent [44]. We used a combination of least absolute shrinkage and selection operator (LASSO) regression and the SVM algorithm for variable selection.

Modeling and Parameter Adjustment

Four classification algorithms, LR, RF, SVM, and XGBoost, were used to establish the models. The parameter learning curve, grid search, and cross-validation methods were used to adjust the parameters of the model, and the best parameter combination was determined by checking the accuracy, sensitivity, and AUC of the model on the validation set. The parameter learning curve is a curve with different parameter values as the abscissa, and the model score under different parameter values as the ordinate. We can see that the change in trend of the model evaluation index under different parameter values, initially obtain a small parameter search interval, or select the value of the best point of the model performance as the optimal parameter value. Grid search refers to the selection of all candidate parameters through loop traversal. The system tries every possibility, and the best-performing parameter is the final result, which is the process of training and comparison. Cross-validation refers to randomly dividing the data set into K parts without replacement, k-1 parts are used to train the model, and the remaining part is used for performance evaluation. This process was repeated K times to obtain k models and performance evaluation results.

Evaluation of Model Fitting Effect

The aim of parameter tuning is to minimize the generalization error of the model. The generalization error was used to measure the accuracy of the model with unknown data in machine learning. A model that is too simple or too complex will cause high generalization errors. If the model is too complex, it will overfit; if the model is too simple, it will underfit. By comparing the sample learning curves of the training with validation sets, we can observe the fitting effect of the model. The sample learning curve is drawn with the number of different training samples as the abscissa, and the accuracy of the training or validation sets under the number of samples as the ordinate. When the errors of the training and validation sets converge but the accuracy is low, it indicates a high bias. When the deviation of the upper left corner of the curve is very large and the accuracy of the training and validation sets is very low, the model is underfitted. When the errors of the training and validation sets are large, there is high variance; the variance in the upper right corner of the curve is high, the accuracy of the training and validation sets are too different, and the model is overfitted. If one of the biases and variances is large, this indicates that the generalization error is large.

Model Evaluation and Comparison

Confusion matrix indicates the count of the true outcome and prediction under different labels (ONFH or non-ONFH). A series of indicators can be calculated using the confusion matrix. The accuracy of the model is a key indicator for measuring the quality of the model. According to PROBAST [30], reviewers should evaluate the model performance, including discrimination and calibration ability. The receiver operating characteristic (ROC) curve reflects the dynamic relationship between the false-positive rate (FPR) and sensitivity (also known as recall). Different FPRs and sensitivities were obtained by classifying different predicted values corresponding to each point on the ROC curve. The area under the ROC curve is known as AUC, which is also considered as the possibility of the fact that *sensitivity is larger than FPR*. In addition, the precision-recall (PR) curve reflects the dynamic relationship between the precision and sensitivity. The area under the PR curve equals the average precision (AP), which is the AP at all thresholds. A larger AP usually indicates better discrimination ability, so does AUC. In particular, PR curves are available for unbalanced data [45]. By evaluating the fraction of true positives among positive predictions and actual positives (real ONFH patients), we illustrated the discrimination ability more properly and specifically. In this study, sensitivity, specificity, F1 score, ROC curve, and PR curve were used to evaluate discrimination ability.

Calibration includes log-loss and the calibration curve. Log-loss is the negative logarithm of the probability of the real probability for a given probability classifier under the condition of prediction probability. A smaller value of the log-likelihood function indicates a more accurate prediction. In this study, all samples were reordered according to the predicted probability and divided into 10 equal groups. The calibration curve showed the distance between the predicted probabilities and the true incidence of ONFH in each group. A curve closer to the ideal line (y=x) shows a better calibration ability of the model.

Model Interpretability

The black box model is explained through both global and local explanations. Shapley additive explanations (SHAP) is based on the theoretically optimal Shapley values [28]. The Shapley value explains the prediction of instance x by computing the contribution of each feature to the prediction. For each prediction sample, the model produces a prediction value, and the sum or average of the Shapley absolute value of each feature of all individuals is the overall feature importance. Features with large absolute Shapley values are very important; therefore, the importance of features can be ranked from a global perspective according to the absolute value of Shapley. The local interpretable model-agnostic explanations (LIME) algorithm obtains the probability value of each category by selecting specific samples in the data set and explains the reason for the distribution probability. LIME decomposes the sample space into parts and attempts to use simple models (such as linear models) that are easy to explain to fit complex models that are not easy to explain. LIME focuses on training local surrogate models to explain individual predictions [28].

Statistical Analysis

Qualitative variables are expressed as ratios or constituent ratios. The Kolmogorov-Smirnov test was used to test the normality of the quantitative variables. Variables that fit the normal distribution were expressed as the mean (SD), and variables that did not fit the normal distribution were expressed as the median (25th percentile [P_{25}] and 75th percentile [P_{75}]). The Kendall correlation coefficient and Spearman correlation coefficient were used to describe the correlation between the qualitative and quantitative variables, respectively. A coefficient greater than 0.6 indicates that there is a correlation between the 2 variables. LASSO regression was used to eliminate multicollinearity. Statistical analysis was performed using Python 3.7.4 (Anaconda 4.9.2). The main Python library and version information used for modeling are listed in Table 1.

The flowchart of the study is shown in Figure 1.

Table 1. Python library and function.

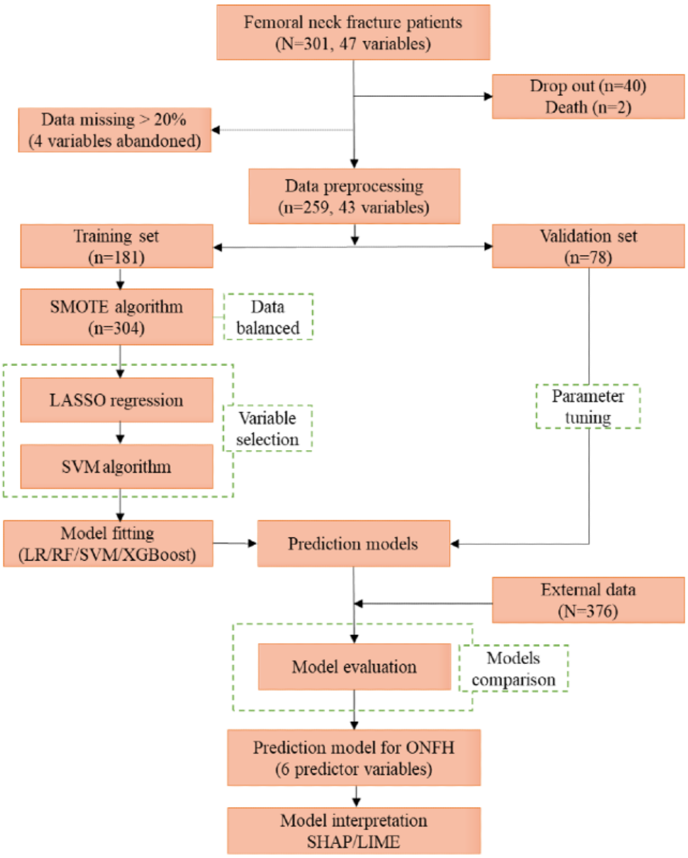
Library	Version	Function
scikit-learn	0.24.1	Machine learning
NumPy	1.16.5	Scientific computing
pandas	0.25.1	Data analysis
Matplotlib	3.3.4	Visualization
imblearn	0.0	Imbalanced data set
statsmodels	0.12.2	Statistical computations
XGBoost[a]	1.3.3	Gradient boosting framework
SHAP[b]	0.39.0	Explain the output of machine learning model
LIME[c]	0.2.0.1	Explain the output of machine learning model
Flask	1.1.1	Web development
Gunicorn	20.1.0	HTTP server

[a] XGBoost: eXtreme Gradient Boosting.

[b] SHAP: Shapley additive explanations.

[c] LIME: local interpretable model-agnostic explanations.

Figure 1. Flowchart of the study. LASSO: least absolute shrinkage and selection operator; LIME: local interpretable model-agnostic explanations; LR: logistic regression; ONFH: osteonecrosis of the femoral head; RF: random forest; SHAP: Shapley additive explanations; SMOTE: synthetic minority oversampling technique; SVM: support vector machine; XGBoost: eXtreme Gradient Boosting.



Results

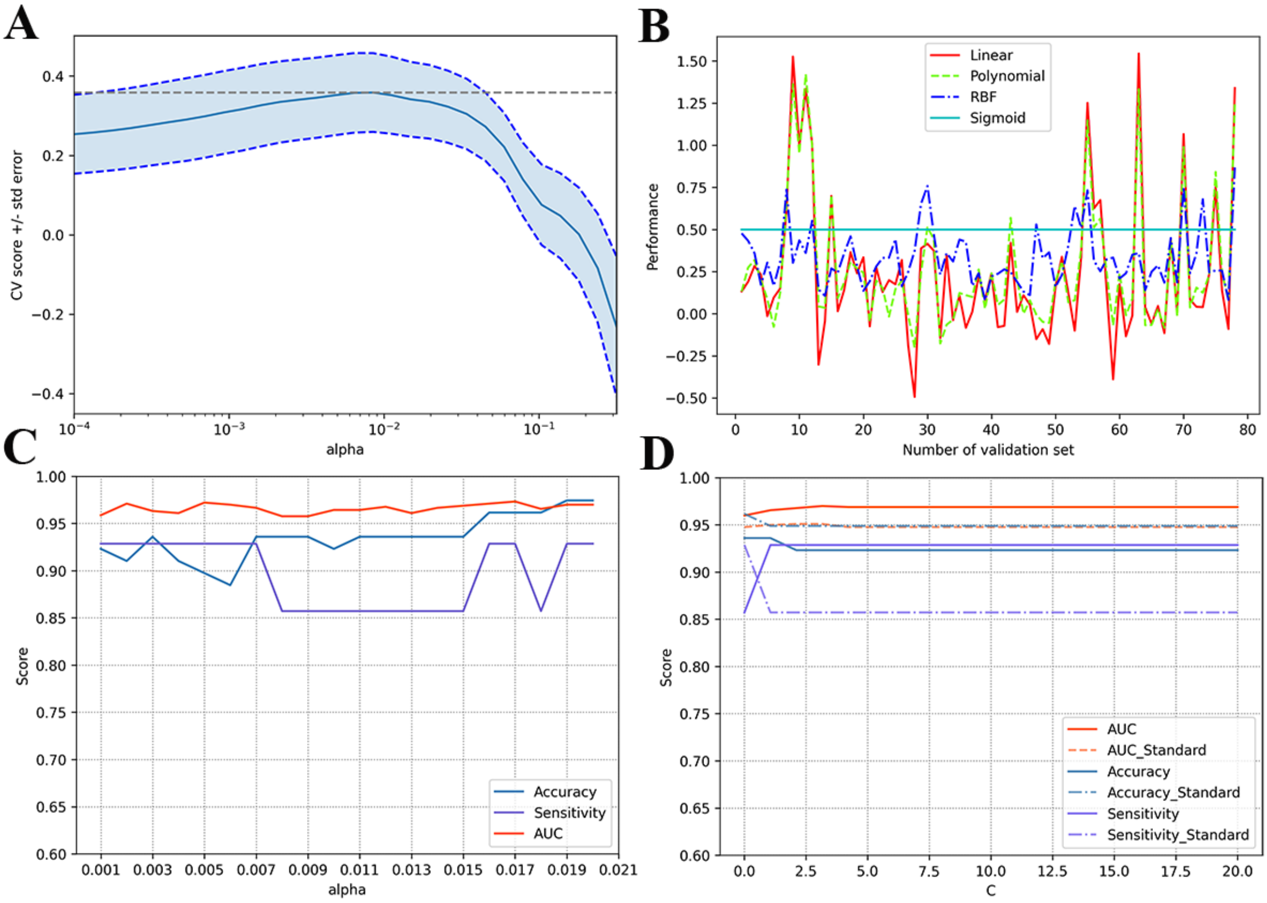
Patient Characteristics

A total of 259 patients with FNF were included in this study, comprising 124 (47.8%) men and 135 (52.1%) women, and the median (P_{25}, P_{75}) age was 57 (49, 62) years. A total of 43 patients experienced ONFH after internal fixation surgery, with an incidence of ONFH of 16.6%. All data were randomly divided into a training set (181/259, 69.8%) and a validation set (78/259, 30.1%) at a ratio of 7:3 (randomstate=420). There were 29 patients with ONFH and 152 patients without ONFH in the training set. After using the SMOTE algorithm to oversample the femoral head necrosis group in the training set, the number of ONFH and non-ONFH groups reached a balance (152 cases in each group). There were 14 patients with ONFH and 64 patients without ONFH in the validation set. Patient characteristics in the 3 data sets are presented in Multimedia Appendix 2. Overall, the composition of the patients' variables in the 3 data sets was the same. The results of the feature correlation analysis are presented in Multimedia Appendix 3.

Variable Selection

First, we used the grid search and 10-fold cross-validation estimators (GridSearchCV) to explore the LASSO regression regularization parameter α (Figure 2A). The results of the cross-validation showed that the optimal α was 0.0016. Second, we used the LassoCV object that sets its α parameter automatically from the data by internal cross-validation and used external cross-validation to evaluate the reliability of the selection of α. After 3 cross validations, we obtained 3 alphas for different subsets of the data, which were 0.00646, 0.00281, and 0.00281. However, the scores for these alphas differed substantially, with 0.49697, 0.76142, and 0, respectively. It can be seen that the reliability of LASSO for selecting variables is not very high.

Figure 2. The process of exploring the optimal feature subset. (A) Grid search and 10-fold cross-validation estimators of least absolute shrinkage and selection operator regression regularization parameter α. The y-axis represents the average and SD of 10 cross validations. (B) Prediction results of support vector machine (SVM) with different validation samples under 4 kernel functions of linear, polynomial, radial basis function (RBF) and sigmoid. (C) Best α that makes the SVM model have the best accuracy, sensitivity, and area under the receiver operating characteristic curve performance on the validation set. (D) Comparison of standardized with nonstandardized results of continuous variables under different support vector machine parameters C (kernel=linear). AUC: area under the receiver operating characteristic curve; CV: cross validation.



Therefore, to obtain a reliable α and identify the optimal feature subset, the feature subset under the 10-fold cross-validation was first introduced into the SVM classifier for modeling. The performance of the SVM classifier on the validation set using different kernel functions was determined (other parameters used default values). Figure 2B shows the prediction results of SVM with different validation samples under 4 kernel functions: linear, polynomial, radial basis function, and sigmoid. Figure 2B shows that the linear kernel performs best. Next, we identified an α between 0.001 and 0.02, which made the SVM model have the best accuracy, sensitivity, and AUC performance on the validation set. Figure 2C shows that when the α was

0.017, the total accuracy, sensitivity, and AUC were the best. The optimal feature subset included age, sex, time to surgery, injury cause, low energy, fracture position, subcapital, fracture position, neck-to-head, Garden classification Ⅳ, reduction quality, interval to weightbearing, femoral neck shortening, and VAS score.

In addition, we compared the results of standardization with nonstandardization on the accuracy, sensitivity, and AUC of the verification set under different SVM parameters C (kernel=*linear*). In Figure 2D, the solid line is the result of no standardized treatment for continuous variables, and the dotted line is the result of standardization. The figure shows that the performance of the model decreased after the standardization of continuous variables. Therefore, continuous variables were not standardized.

Modeling and Parameter Tuning

After confirming the optimal feature subset, LR, RF, SVM, and XGBoost algorithms were selected to fit the models on the balanced training set. Table 2 presents the comparison of accuracy, sensitivity, and AUC on the validation set before and after tuning the parameters. The LR and SVM models did not significantly improve. However, the accuracy, sensitivity, and AUC of the RF model were increased by 0.012, 0.072, and 0.006, respectively, and those of the XGBoost model were increased by 0.025, 0.072, and 0.003, respectively. The hyperparameters tuned in each of the 4 classifiers are listed in Table 3.

Table 2. Comparison of model performance on the validation set.

Model	Before tuning			After tuning		
	Accuracy	Sensitivity	AUC[a]	Accuracy	Sensitivity	AUC
LR[b]	0.962	0.929	0.982	0.962	0.929	0.984
RF[c]	0.962	0.857	0.985	0.974	0.929	0.991
SVM[d]	0.962	0.929	0.973	0.962	0.929	0.979
XGBoost[e]	0.962	0.857	0.989	0.987	0.929	0.992

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

[b] LR: logistic regression.

[c] RF: random forest.

[d] SVM: support vector machine.

[e] XGBoost: eXtreme Gradient Boosting.

Table 3. Hyperparameter configuration for algorithms.

Algorithm and parameter name	Initial value	Adjustment range	Result
LR[a]			
Penalty	L1	(L1, L2)	L2
C	0.3	(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 1)	0.5
RF[b]			
n_estimators	100	Range (0, 200, 10)	11
max_depth	8	(1, 3, 5, 6, 7, 8, 9, 10, 15, 20)	8
max_features	3	(2, 3, 4, 5, 6, 7, 8)	5
min_samples_leaf	1	(1, 2, 3, 4)	2
SVM[c]			
Kernel	Rbf	(Linear, polynomial, rbf, sigmoid)	Linear
C	1	Range (0.01, 20, 20)	7.37
XGBoost[d]			
n_estimators	100	Range (0, 200, 10)	51
max_depth	6	Range (1, 20, 1)	8
min_child_weight	1	(2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20)	7
learning_rate	0.3	(0.3, 0.31, 0.32, 0.33, 0.335, 0.36, 0.38, 0.4)	0.335
gamma	0	(0, 1, 2, 3, 4)	1

[a] LR: logistic regression.

[b] RF: random forest.

[c] SVM: support vector machine.

[d] XGBoost: eXtreme Gradient Boosting.

Figure 3 shows the ROC curves of the 4 models for the training and validation sets. The AUC values of each algorithm were similar, and the XGBoost model had the highest AUC value in the validation set. Figure 4 shows the learning processes of the 4 models. Except for the RF model, which exhibited slight overfitting when the number of training samples was less than 250, all other models fit well. The parameter configuration of the machine learning models is shown in Multimedia Appendix 4.

Figure 3. Receiver operating characteristic curves of the logistic regression, random forest, support vector machine (SVM) and eXtreme Gradient Boosting (XGBoost) prediction models on the training set and the validation set, which indicate discrimination ability. The more convex the upper left corner of the curve, the better. AUC: area under the receiver operating characteristic curve.

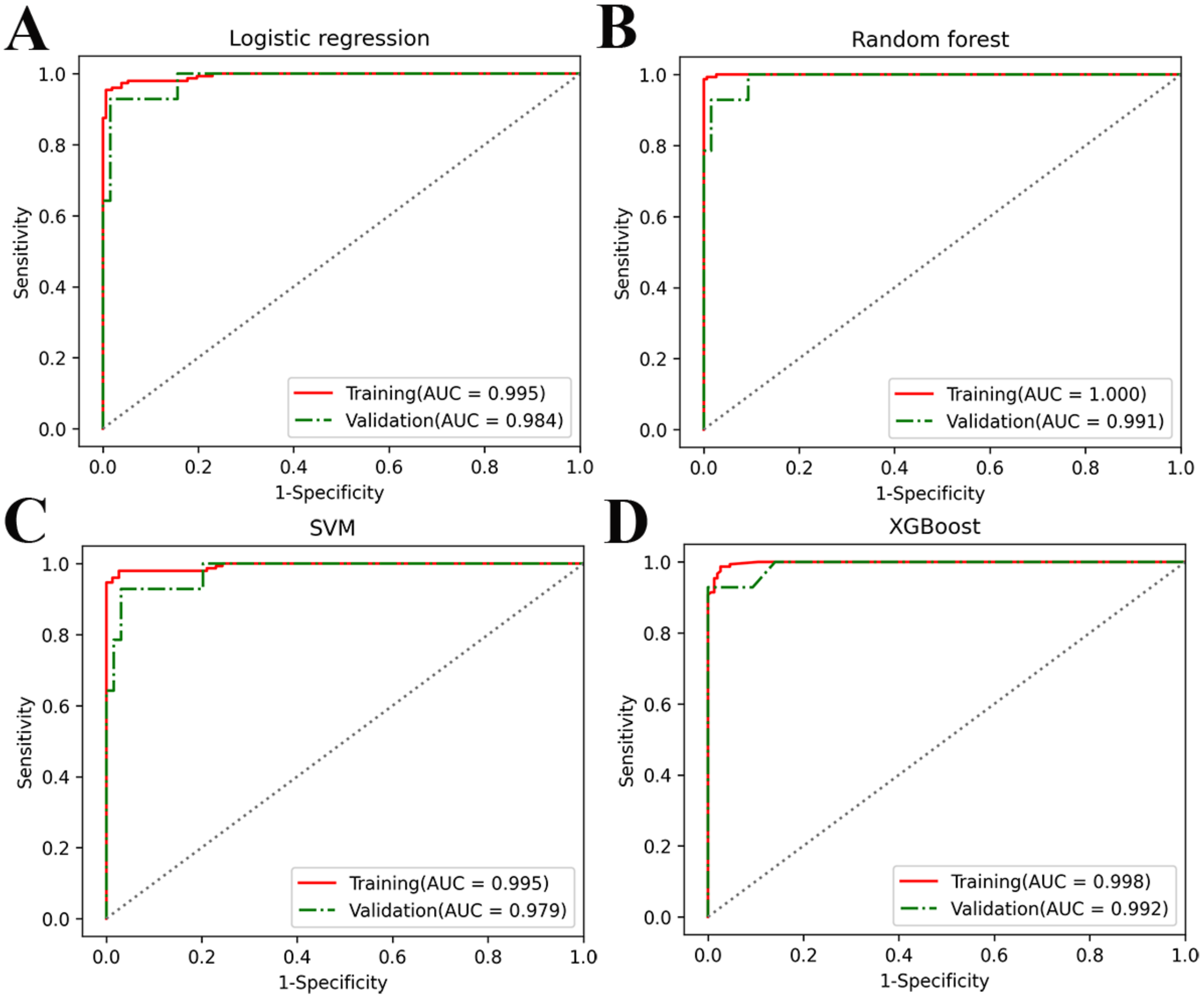
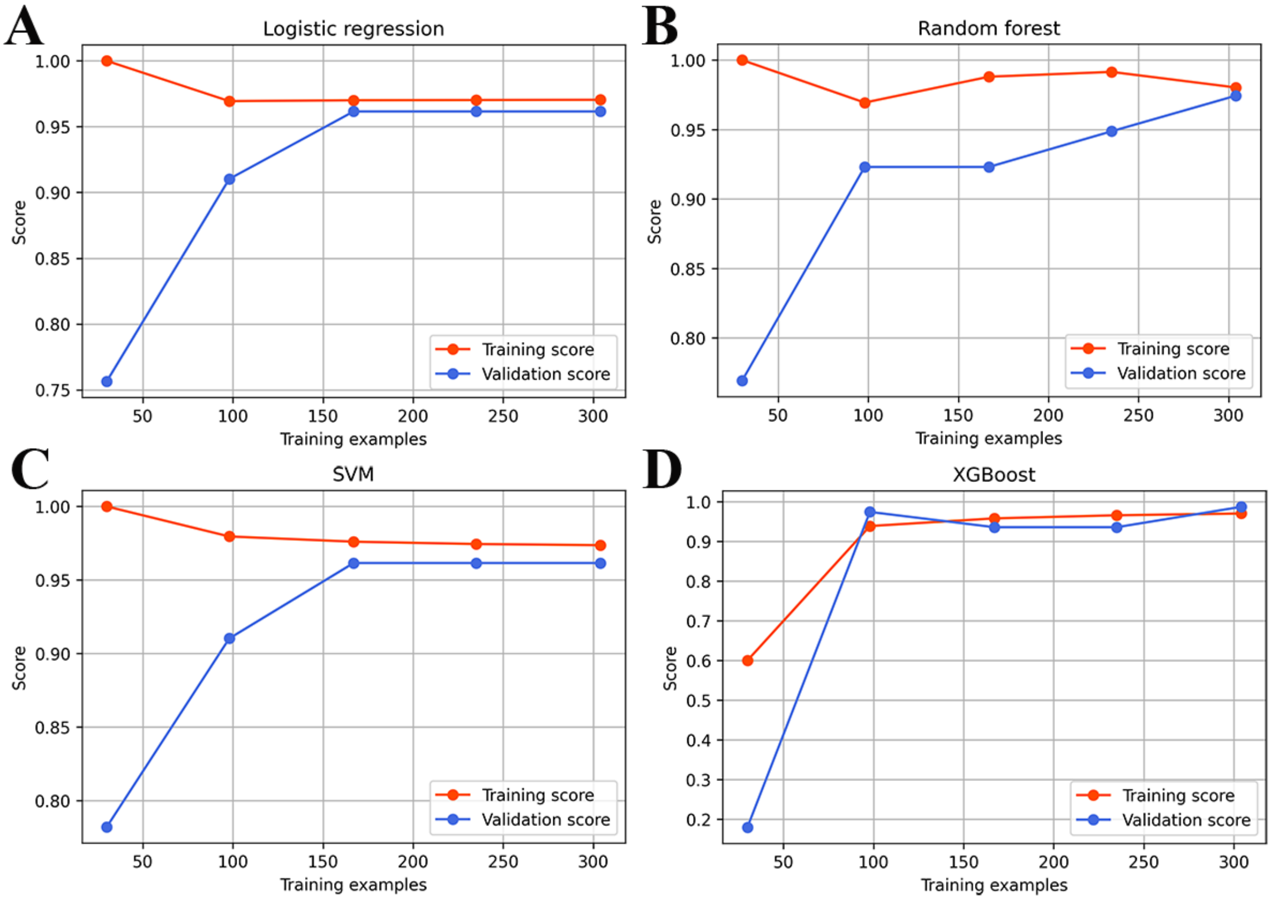


Figure 4. The learning curves of the logistic regression, random forest, support vector machine (SVM) and eXtreme Gradient Boosting (XGBoost) prediction models on the training and validation sets. The 2 curves of logistic regression, support vector machine and XGBoost model are consistent at a higher accuracy level, which indicates that the model is well fitted for training. The 2 curves of random forest are not well merged, indicating that they are slightly overfitted.



Model Evaluation and Comparison

Table 4 presents the confusion matrix of the classification results for the 4 models on the external data. XGBoost generated the maximum number of ONFH (67), and SVM generated the maximum number of non-OFNH (278). Table 5 presents the evaluation results of the 4 models for the external validation data. The XGBoost model exhibited the highest accuracy (0.907). With the exception of the specificity of the XGBoost model being lower than SVM (0.949), the sensitivity (0.807), AUC (0.933), and F1 score (0.793) were all higher than those of the other models. In addition, the XGBoost model presented the smallest log-loss (0.279). A comparison of the ROC curves of the 4 models is shown in Figure 5A. The AUCs of the 4 models were above 0.9, and the XGBoost model achieved the largest AUC. The shapes of the 4 ROC curves were similar. A comparison of the PR curves of the 4 models is shown in Figure 5B. When the sensitivity of the prediction model was greater than 0.7, the XGBoost model had the highest prediction precision. In addition, the LR model achieved the largest AP. The calibration curve is shown in Figure 6, and the curve of the XGBoost model was closest to the ideal calibrated line (y=x).

Table 4. Confusion matrices of the prediction models of ONFH[a].

Model and actual	Predictive	
	ONFH	Non-ONFH
LR[b]		
ONFH	64	19
Non-ONFH	21	272
RF[c]		
ONFH	64	19
Non-ONFH	24	269
SVM[d]		
ONFH	61	22
Non-ONFH	15	278
XGBoost[e]		
ONFH	67	16
Non-ONFH	19	274

[a] ONFH: osteonecrosis of the femoral head.

[b] LR: logistic regression.

[c] RF: random forest.

[d] SVM: support vector machine.

[e] XGBoost: eXtreme Gradient Boosting.

Table 5. Performance comparison on external data.

Model	Accuracy	Discrimination				Calibration
		Sensitivity	Specificity	AUC[a]	F1 score	Log-loss
LR[b]	0.894	0.771	0.928	0.927	0.762	0.288
RF[c]	0.886	0.771	0.918	0.910	0.749	0.775
SVM[d]	0.901	0.735	0.949	0.904	0.767	0.327
XGBoost[e]	0.907	0.807	0.935	0.933	0.793	0.279

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

[b] LR: logistic regression.

[c] RF: random forest.

[d] SVM: support vector machine.

[e] XGBoost: eXtreme Gradient Boosting.

Figure 5. (A) Comparison of receiver operating characteristic curves of the 4 models on external data. The curve closer to the upper left corner showed better overall discrimination ability. (B) Comparison of precision-recall curves of the 4 models on external data. The curve closer to the upper right corner also showed the ability to combine precision with sensitivity. AP: average precision; AUC: area under the receiver operating characteristic curve; ROC: receiver operating characteristic; SVM: support vector machine; XGBoost: eXtreme Gradient Boosting.

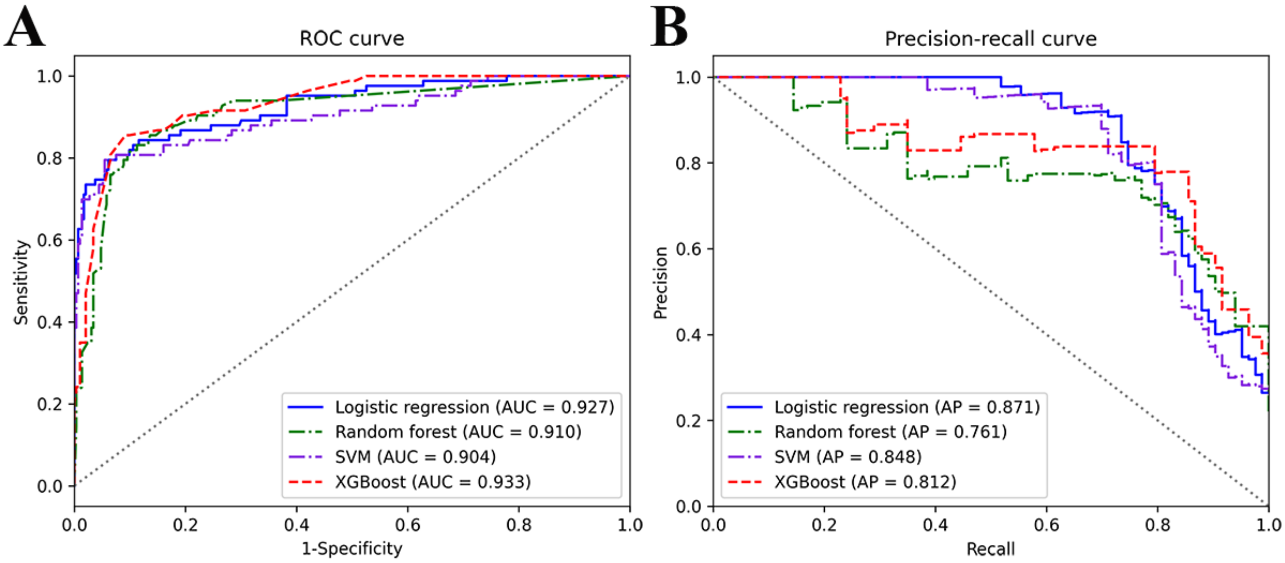
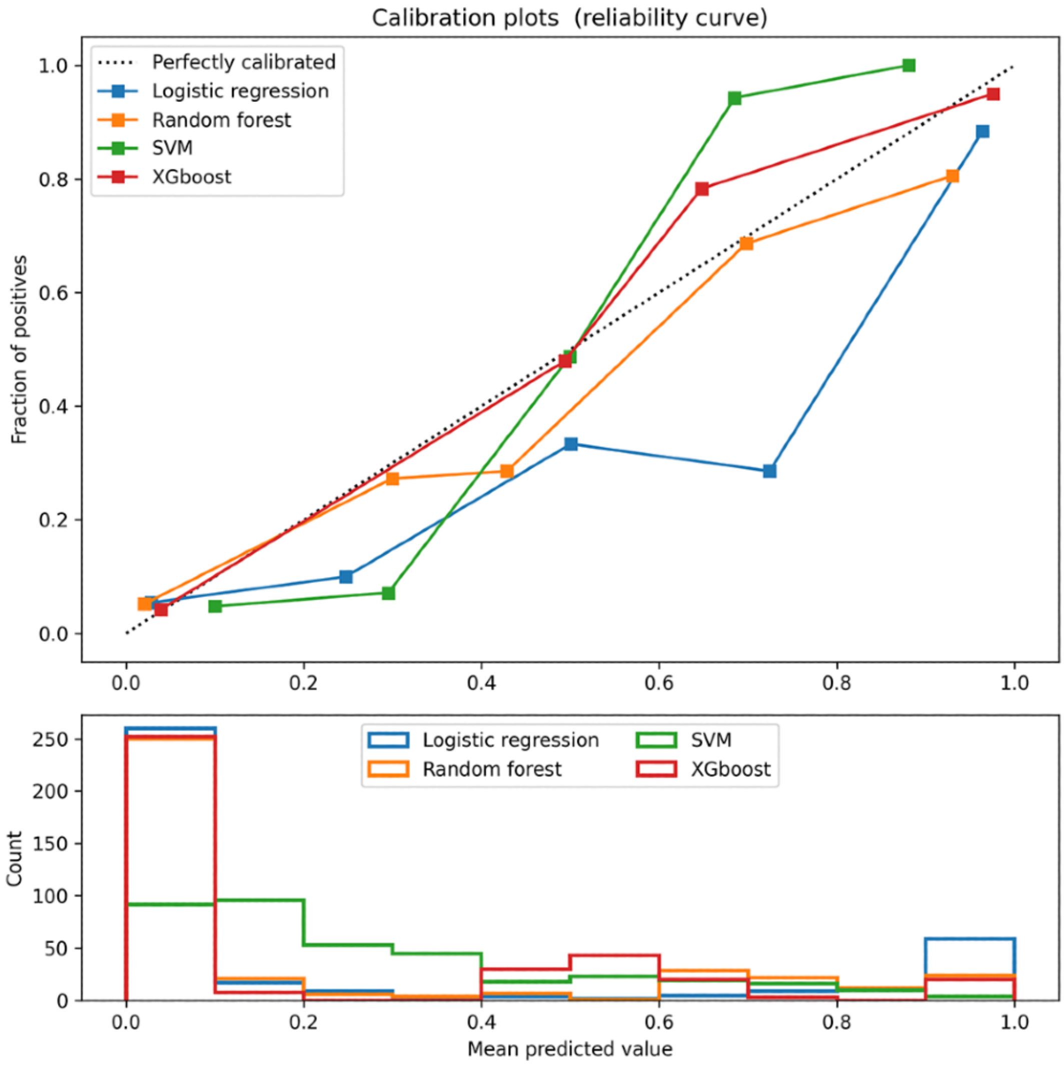


Figure 6. Comparison of calibration curves of the 4 models on external data. The calibration curve of the model is consistent with the ideal calibrated line (y=x), indicating that the predicted value of the model is close to the actual probability of the outcome. SVM: support vector machine; XGBoost: eXtreme Gradient Boosting.



Interpretability of the Prediction Model

On the basis of the above comparisons, we determined that the XGBoost model was the best predictive model for ONFH. Take the average of the absolute value of the SHAP of each feature as the importance of the feature. The predictor variables of the XGBoost model and their importance ranking are as follows: reduction quality (1.759), VAS score (1.483), Garden classification (0.299), time to surgery (0.247), cause of injury (0.127), and fracture position (0.090). Figure 7 shows a summary of the SHAP values of each feature in each sample. The color represents the feature value, and the redder the color, the greater the feature value. Therefore, we can see that the VAS score, Garden classification IV, time to surgery, and fracture position_subcapital are all risk factors for ONFH. Reduction quality_good and injury cause_low energy are protective factors for femoral head necrosis.

Figure 7. Global explanations of the eXtreme Gradient Boosting model based on Shapley additive explanations (SHAP) values. Summary of the SHAP values of each feature in each sample. The abscissa is the SHAP value (the impact on the model output), the ordinate is the different features, a point represents a sample, and the color represents the feature value. The larger the feature value is, the redder the color, and the smaller the feature is, the bluer the color. VAS: visual analog scale.

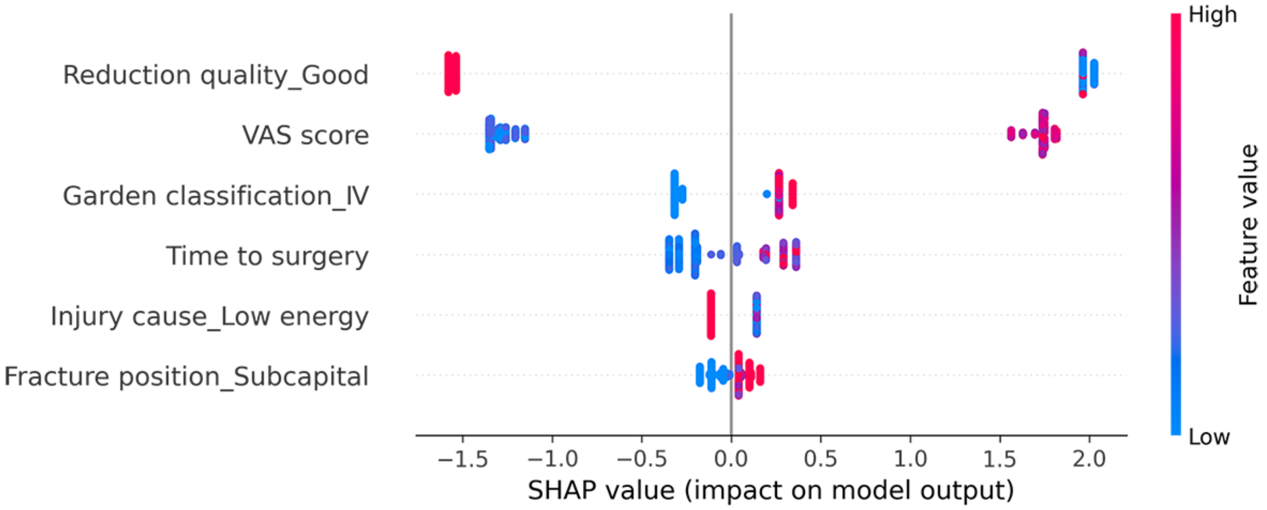
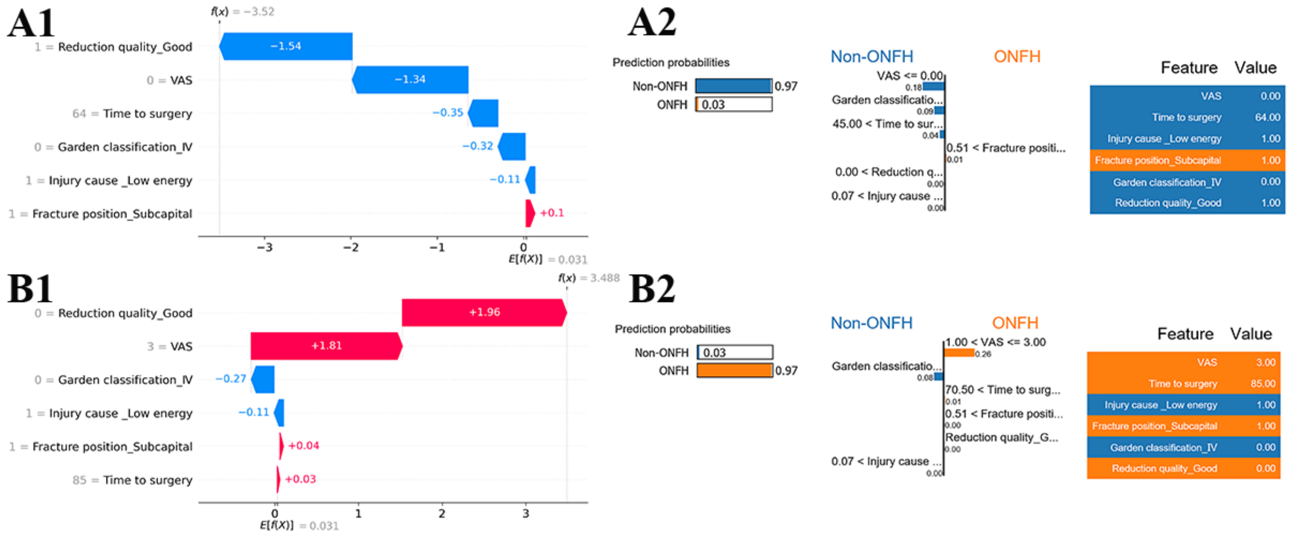


Figure 8 shows the decision process for the single-sample prediction. These are local explanations of XGBoost model based on SHAP and LIME. The true outcome of the first sample is non-ONFH, and the predicted outcome is non-ONFH, as shown in Figure 8A. The true outcome of the second sample is ONFH, and the predicted outcome is ONFH, as shown in Figure 8B. The 2 figures on the left and the 2 figures on the right are the results of local explanations by the SHAP and LIME algorithms, respectively. As can be seen from A1 and A2, both SHAP and LIME show that the features determined the outcome of non-ONFH, including reduction quality_good (1), VAS score (0), time to surgery (64), Garden classification_IV (0), and injury cause_Low energy (1). The difference is that LIME can provide a predicted probability of non-ONFH of 0.97. Similarly, B1 and B2 show that the features determined the outcome of ONFH, including reduction quality_good (0), VAS score (3), fracture position_subcapital (1), and time to surgery (85). LIME also shows that the prediction probability of the ONFH is 0.97.

Figure 8. Local explanations of the XGBoost model. (A) The true outcome is nonosteonecrosis of the femoral head (ONFH), and the predicted outcome is non-ONFH. (B) The true outcome is ONFH, and the predicted outcome is ONFH. A1 and B1 are local explanations realized by Shapley additive explanations. Variables in blue decided the sample to be classified into category non-ONFH, and variables in red decided the sample to be classified into category ONFH. A2 and B2 are local explanations realized by local interpretable model-agnostic explanation (LIME). LIME can obtain the probability value of each category, and showing which variables determine the sample to be classified into category non-ONFH (blue) and which variables determine the samples to be classified into category ONFH (orange), specifically listing the numerical size of the samples in these features. VAS: visual analog scale.



Discussion

Principal Findings

In this study, we compared the application of different machine learning algorithms in the prediction of femoral head necrosis after internal fixation of FNFs and obtained a 6-variable XGBoost model that could be used for the clinical prediction of traumatic ONFH. This model was translated into a self-made web-based risk calculator to estimate an individual's probability of ONFH. The predictors included reduction quality, VAS score, Garden classification, time to surgery, cause of injury, and fracture position. This prediction model exhibited good discrimination and calibration and showed good generalization performance on external data. Performance on the internal validation set yielded an accuracy of 0.987, sensitivity of 0.929, and AUC of 0.992. Performance on external data revealed an accuracy of 0.907, sensitivity of 0.807, specificity of 0.935, AUC of 0.933, F1 score of 0.793, and log-loss of 0.279. The web-based risk calculator can be found on the Herokuapp website [46].

While constructing the predictive model, we also conducted an excavation on the predictive variables of ONFH after internal surgery for FNF. In the design stage of the study, we made our best effort to collect relevant injury and clinical information throughout the clinical course, such as preoperative coagulation indicators, preoperative routine blood tests, and other indicators that have not been analyzed in previous studies. However, these indicators did not pass variable selection. A new British study [47] revealed that poor nutritional status was correlated with mortality and worse postoperative outcomes in patients with FNF. We did not find any indices related to femoral head necrosis on preoperative biochemical examination. Huang et al [48] reported that compared with open reduction in pursuit of anatomical reduction, which may cause vascular injury, positive support can provide reasonable reduction support and reduce the occurrence of vascular necrosis. When Gotfried reduction had a positive buttress pattern, the medial cortex of the distal end of the fracture straddled the *medial femoral neck support bridge* due to sliding compression of the femoral head. The special stress transfer effect of the arch structure can effectively resist the longitudinal shear force between the fracture pieces and stabilize the fracture. We used Gotfried reduction as a predictive variable to participate in the model. However, the overall performance of the model decreased. This is similar to the results reported by Zhao et al [49]. Other controversial risk factors that might be related to femoral head necrosis, such as early weight bearing, removal of internal fixation implants, and reduction methods, were not selected by the classification model.

Among the 6 predictors in the XGBoost model, poor reduction, severe fracture displacement, and delay in operation time were clear risk factors for ONFH after internal fixation of FNF. The VAS pain score is widely used in clinical prognosis research and has high reliability and validity. After internal fixation, patients generally experience slight soreness when they get up and sit down and when the temperature suddenly drops. When osteocytes of the hip joint change histologically, patients may experience pain. Through finite element analysis based on biomechanics, Li et al [50] reported that when necrosis occurs, the increase in mechanical load on the hip joint in patient's daily life will increase the area of necrotic lesions, especially lesions in the anterior and lateral areas of the femoral head, which are more likely to accelerate expansion and collapse in advance. The causes and mechanisms of postoperative hip pain have not been fully explored and require further investigation. A subcapital fracture indicates that the fracture line is completely located at the bottom of the femoral head. When a subcapital fracture of the femoral neck occurs, it is usually accompanied by a rupture of the medial and lateral femoral circumflex artery, in which the epiphyseal artery from the medial femoral circumflex artery supplies most of the blood to the femoral head. Subcapital fractures are usually accompanied by rupture of the medial and lateral femoral circumflex arteries. However, the epiphyseal artery from the medial circumflex femoral artery supplies most of the blood to the femoral head. The lateral epiphyseal artery is the primary blood vessel for the femoral head. Injury causes most of the blood supply to the femoral head to be interrupted. Only the internal artery of the round ligament supplies the blood. Its blood vessels are thin and supply the femoral head over a small range, but it cannot provide the necessary blood volume to the entire femoral head. Therefore, the necrosis rate of subcapital fractures is higher [51]. Different trauma mechanisms cause different fracture injuries. In FNFs caused by high-energy traumas, such as road traffic accidents and falling from a height, the displacement of the broken end is usually large, often tearing the ascending cervical branches that stem off the arterial ring supply formed by the circumflex arteries, destroying the blood supply to the femoral head and causing nonunion of the fracture or complications, such as femoral head necrosis [52]. The cause of injury has also been considered a risk factor for femoral head necrosis [53].

It is worth noting that before the categorical variables entered the machine learning classifier to fit the model, they were converted into dummy variables according to the category. The Garden classification is a 4-category variable. After XGBoost modeling, only Garden classification IV became a predictor variable. At this time, Garden classification is no longer a 4-category variable but becomes a 2-category variable of Garden classification_IV.

Obtaining a sufficient number of training samples is difficult and time-consuming for the prediction of femoral head necrosis after FNF. LR, RF, SVM, and XGBoost can learn effectively from a limited training set. As a strongly integrated algorithm, the performance of XGBoost was not only better than that of SVM and RF but also more accurate and reliable than traditional LR in our study.

In addition, we opened the black box of machine learning with the help of post hoc interpretability techniques of the machine learning model. Through the global interpretation based on SHAP, we can understand the relationship between predictors and outcomes in the XGBoot model. The variables of reduction quality_good and injury cause_low energy correlated negatively with the outcomes and were protective factors; VAS score, Garden classification_IV, time to surgery, and fracture position_subcapital correlated positively with the outcomes and were risk factors. Both SHAP and LIME can provide local explanations for a single sample. The explanatory plot produced by the SHAP is close to the one generated by LIME in that it shows the variables' names and contributions that are used in the explanation [54]. The advantage of the LIME method is that the explanation is based on the local regression model, which allows physicians to make statements about changes in explanations for changes in the features of the patient to be explained. The disadvantage is that the instability of the explanations is insufficient. For a single sample, if you get the explanation twice, you may have 2 different explanations. Comparably, the principle of the SHAP method is strictly improved from the classical Shapley value estimation method [55], so the interpretation results of the SHAP algorithm have variable consistency and model stability.

Limitations

This study has several limitations. First, the number of ONFH cases was insufficient. According to the requirements of PROBAST for the number of participants in clinical events, the ratio of participants in clinical events to the number of candidate predictors was at least 10. There were 6 predictors in the model, and only 43 patients had ONFH. However, we used the SMOTE algorithm to balance the training set and increase the number of ONFH to 152 cases. Second, the sensitivity and F1 score on the external data were approximately 0.8, which is low compared with other indicators. When using the LIME algorithm to explain individual predictions, we discovered that most samples used only 4 variables for prediction. Therefore, the reasons for the low sensitivity and F1 score may include the following: (1) the number of ONFH is insufficient and (2) there are still risk factors related to ONFH that have not been identified. In the future, we will conduct prospective validation based on this model, continue to explore important risk factors for ONFH, and modify the model to further improve the accuracy of the XGBoost prediction model.

Comparison With Prior Work

The patients with FNF in this study were from 6 hospitals in Shanghai, which are more representative. We included a wider range of candidate variables. Instead of using traditional single-variable analysis for variable selection, LASSO was integrated into the SVM as a new variable selection method. The performance of our model on the validation set was better than that of the naive Bayesian prediction model proposed by Cui et al [24], whose accuracy, sensitivity, and AUC were 0.744, 0.742, and 0.746, respectively. The AUC of our model on the validation set was higher than that of the hybrid nomogram based on LR developed by Zhu et al (0.948) [26] and the nomogram based on Cox regression developed by Zheng et al (0.97) [25]. It also exhibited satisfactory generalization ability on external data, with accuracy, specificity, AUC, and log-loss values of 0.907, 0.935, 0.933, and 0.279, respectively.

Conclusions

Machine learning performs well in predicting ONFH after internal fixation of FNF. The 6-variable XGBoost model predicts the risk of ONFH well and has good generalization ability in external data, which can be used for the clinical prediction of ONFH after internal fixation of FNF.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (grant number 81872718), Shang Hai Municipal Health and Family Planning Commission (grant number 201840041), The Outstanding Clinical Discipline Project of Shang Hai Pu Dong (grant number PWYgy2018-10), and Key Undergraduate Course Project of Shang Hai Education Commission (201965).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Variables and definitions.
[XLSX File (Microsoft Excel File), 14 KB - medinform_v9i11e30079_app1.xlsx]

Multimedia Appendix 2
Characteristics of 3 groups of patients with femoral neck fracture.
[DOCX File , 20 KB - medinform_v9i11e30079_app2.docx]

Multimedia Appendix 3
Correlation coefficient matrix heat maps.
[DOCX File , 577 KB - medinform_v9i11e30079_app3.docx]

Multimedia Appendix 4
Parameters of machine learning models.
[DOCX File , 16 KB - medinform_v9i11e30079_app4.docx]

References

1. Garofoli R, Maravic M, Ostertag A, Cohen-Solal M. Secular trends of hip fractures in France: impact of changing characteristics of the background population. Osteoporos Int 2019 Feb;30(2):355-362. [doi: 10.1007/s00198-018-4666-7] [Medline: 30215115]
2. Kannus P, Niemi S, Parkkari J, Sievänen H. Continuously declining incidence of hip fracture in Finland: analysis of nationwide database in 1970-2016. Arch Gerontol Geriatr 2018;77:64-67. [doi: 10.1016/j.archger.2018.04.008] [Medline: 29684740]
3. Søgaard AJ, Holvik K, Meyer HE, Tell GS, Gjesdal CG, Emaus N, et al. Continued decline in hip fracture incidence in Norway: a NOREPOS study. Osteoporos Int 2016 Jul;27(7):2217-2222. [doi: 10.1007/s00198-016-3516-8] [Medline: 26902091]
4. Requena G, Abbing-Karahagopian V, Huerta C, De Bruin ML, Alvarez Y, Miret M, et al. Incidence rates and trends of hip/femur fractures in five European countries: comparison using e-healthcare records databases. Calcif Tissue Int 2014 Jun;94(6):580-589. [doi: 10.1007/s00223-014-9850-y] [Medline: 24687523]
5. Yong EL, Ganesan G, Kramer M, Logan S, Lau T, Cauley J, et al. Hip fractures in Singapore: ethnic differences and temporal trends in the new millennium. Osteoporos Int 2019 Apr;30(4):879-886. [doi: 10.1007/s00198-019-04839-5] [Medline: 30671610]
6. Lehtonen EJ, Stibolt RD, Smith W, Wills B, Pinto MC, McGwin G, et al. Trends in surgical treatment of femoral neck fractures in the elderly. Einstein (Sao Paulo) 2018 Sep 06;16(3):eAO4351 [FREE Full text] [doi: 10.1590/S1679-45082018AO4351] [Medline: 30208153]
7. Zhang C, Feng J, Wang S, Gao P, Xu L, Zhu J, et al. Incidence of and trends in hip fracture among adults in urban China: a nationwide retrospective cohort study. PLoS Med 2020 Aug 6;17(8):e1003180 [FREE Full text] [doi: 10.1371/journal.pmed.1003180] [Medline: 32760065]
8. Hagino H, Osaki M, Okuda R, Enokida S, Nagashima H. Recent trends in the incidence of hip fracture in Tottori Prefecture, Japan: changes over 32 years. Arch Osteoporos 2020 Oct 02;15(1):152 [FREE Full text] [doi: 10.1007/s11657-020-00823-3] [Medline: 33006016]
9. Kim B, Lim J, Ha Y. Recent epidemiology of hip fractures in South Korea. Hip Pelvis 2020 Sep;32(3):119-124 [FREE Full text] [doi: 10.5371/hp.2020.32.3.119] [Medline: 32953703]
10. Muhm M, Amann M, Hofmann A, Ruffing T. [Changes in the patient population with proximal femur fractures over the last decade : incidence, age, comorbidities, and length of stay]. Unfallchirurg 2018 Aug;121(8):649-656. [doi: 10.1007/s00113-017-0425-z] [Medline: 29058020]
11. Tian F, Sun X, Liu J, Liu Z, Liang C, Zhang L. Unparallel gender-specific changes in the incidence of hip fractures in Tangshan, China. Arch Osteoporos 2017 Dec;12(1):18. [doi: 10.1007/s11657-017-0313-8] [Medline: 28190173]

12. Zhang Y. [Selection strategy and progress on the treatment of femoral neck fractures]. Zhongguo Gu Shang 2015 Sep;28(9):781-783. [Medline: 26647555]
13. Damany DS, Parker MJ, Chojnowski A. Complications after intracapsular hip fractures in young adults. A meta-analysis of 18 published studies involving 564 fractures. Injury 2005 Jan;36(1):131-141. [doi: 10.1016/j.injury.2004.05.023] [Medline: 15589931]
14. Zhao D, Zhang F, Wang B, Liu B, Li L, Kim S, et al. Guidelines for clinical diagnosis and treatment of osteonecrosis of the femoral head in adults (2019 version). J Orthop Translat 2020 Mar;21:100-110 [FREE Full text] [doi: 10.1016/j.jot.2019.12.004] [Medline: 32309135]
15. Sultan AA, Khlopas A, Surace P, Samuel LT, Faour M, Sodhi N, et al. The use of non-vascularized bone grafts to treat osteonecrosis of the femoral head: indications, techniques, and outcomes. Int Orthop 2019 Jun;43(6):1315-1320. [doi: 10.1007/s00264-018-4056-y] [Medline: 30039197]
16. Xie H, Wang B, Tian S, Liu B, Qin K, Zhao D. Retrospective long-term follow-up survival analysis of the management of osteonecrosis of the femoral head with pedicled vascularized iliac bone graft transfer. J Arthroplasty 2019 Aug;34(8):1585-1592. [doi: 10.1016/j.arth.2019.03.069] [Medline: 31031157]
17. Noh JH, Lee JY, Hwang S, Lee KH. Prediction of femoral head avascular necrosis following femoral neck fracture: "pin-tract sign" of Tc-HDP pinhole bone scan after metallic fixation. Hip Int 2020 Sep;30(5):641-648 [FREE Full text] [doi: 10.1177/1120700019860492] [Medline: 31280602]
18. Yuan H, Shen F, Zhang J, Shi H, Gu Y, Yan Z. Predictive value of single photon emission computerized tomography and computerized tomography in osteonecrosis after femoral neck fracture: a prospective study. Int Orthop 2015 Jul;39(7):1417-1422. [doi: 10.1007/s00264-015-2709-7] [Medline: 25711398]
19. Kubota S, Inaba Y, Kobayashi N, Tateishi U, Ike H, Inoue T, et al. Prediction of femoral head collapse in osteonecrosis using 18F-fluoride positron emission tomography. Nucl Med Commun 2015 Jun;36(6):596-603. [doi: 10.1097/MNM.0000000000000284] [Medline: 25714808]
20. Kumar MN, Belehalli P, Ramachandra P. PET/CT study of temporal variations in blood flow to the femoral head following low-energy fracture of the femoral neck. Orthopedics 2014 Jun;37(6):e563-e570. [doi: 10.3928/01477447-20140528-57] [Medline: 24972438]
21. Kamano M, Narita S, Honda Y, Fukushima K, Yamano Y. Contrast enhanced magnetic resonance imaging for femoral neck fracture. Clin Orthop Relat Res 1998 May(350):179-186. [Medline: 9602818]
22. Cionca D, Alexa O, Leka V. [Early contrast-enhanced MR imaging assessment of femoral head viability after femoral neck fracture]. Rev Med Chir Soc Med Nat Iasi 2007;111(4):959-964. [Medline: 18389787]
23. Zhao D, Xiaobing Y, Wang T, Wang B, Liu B, Fengde T, et al. Digital subtraction angiography in selection of the vascularized greater trochanter bone grafting for treatment of osteonecrosis of femoral head. Microsurgery 2013 Nov;33(8):656-659. [doi: 10.1002/micr.22179] [Medline: 24115327]
24. Cui S, Zhao L, Wang Y, Dong Q, Ma J, Wang Y, et al. Using Naive Bayes Classifier to predict osteonecrosis of the femoral head with cannulated screw fixation. Injury 2018 Oct;49(10):1865-1870. [doi: 10.1016/j.injury.2018.07.025] [Medline: 30097310]
25. Zheng J, Wang H, Gao Y, Ai Z. A study on the evaluation of a risk score of osteonecrosis of the femoral head based on survival analysis. J Arthroplasty 2021 Jan;36(1):62-71. [doi: 10.1016/j.arth.2020.07.046] [Medline: 32800435]
26. Zhu W, Zhang X, Fang S, Wang B, Zhu C. Deep learning improves osteonecrosis prediction of femoral head after internal fixation using hybrid patient and radiograph variables. Front Med (Lausanne) 2020 Oct 7;7:573522 [FREE Full text] [doi: 10.3389/fmed.2020.573522] [Medline: 33117834]
27. Belousov A, Verzakov S, von Frese J. A flexible classification approach with optimal generalisation performance: support vector machines. Chemometr Intell Lab Syst 2002 Oct 28;64(1):15-25. [doi: 10.1016/s0169-7439(02)00046-1]
28. Molnar C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. USA: Lulu.com; 2020.
29. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020 Jun;58:82-115. [doi: 10.1016/j.inffus.2019.12.012]
30. Moons KG, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 01;170(1):W1-33 [FREE Full text] [doi: 10.7326/M18-1377] [Medline: 30596876]
31. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015 Jan 06;162(1):W1-73 [FREE Full text] [doi: 10.7326/M14-0698] [Medline: 25560730]
32. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020 Mar 20;368:l6927 [FREE Full text] [doi: 10.1136/bmj.l6927] [Medline: 32198138]
33. Joint Surgery Group of the Orthopaedic Branch of the Chinese Medical Association. Guideline for diagnostic and treatment of osteonecrosis of the femoral head. Orthop Surg 2015 Aug;7(3):200-207 [FREE Full text] [doi: 10.1111/os.12193] [Medline: 26311093]

34. Koppie TM, Serio AM, Vickers AJ, Vora K, Dalbagni G, Donat SM, et al. Age-adjusted Charlson comorbidity score is associated with treatment decisions and clinical outcomes for patients undergoing radical cystectomy for bladder cancer. Cancer 2008 Jun;112(11):2384-2392 [FREE Full text] [doi: 10.1002/cncr.23462] [Medline: 18404699]
35. Zhu Y, Chen W, Xin X, Yin Y, Hu J, Lv H, et al. Epidemiologic characteristics of traumatic fractures in elderly patients during the outbreak of coronavirus disease 2019 in China. Int Orthop 2020 Aug;44(8):1565-1570 [FREE Full text] [doi: 10.1007/s00264-020-04575-0] [Medline: 32350584]
36. Zhang Y, Zhang W, Zhang C. A new angle and its relationship with early fixation failure of femoral neck fractures treated with three cannulated compression screws. Orthop Traumatol Surg Res 2017 Apr;103(2):229-234 [FREE Full text] [doi: 10.1016/j.otsr.2016.11.019] [Medline: 28093376]
37. Collin PG, D'Antoni AV, Loukas M, Oskouian RJ, Tubbs RS. Hip fractures in the elderly—a clinical anatomy review. Clin Anat 2017 Jan;30(1):89-97. [doi: 10.1002/ca.22779] [Medline: 27576301]
38. Garden RS. Malreduction and avascular necrosis in subcapital fractures of the femur. J Bone Joint Surg Br 1971 May;53(2):183-197. [Medline: 5578215]
39. Gotfried Y, Kovalenko S, Fuchs D. Nonanatomical reduction of displaced subcapital femoral fractures (Gotfried reduction). J Orthop Trauma 2013 Nov;27(11):e254-e259. [doi: 10.1097/BOT.0b013e31828f8ffc] [Medline: 23481921]
40. Zlowodzki M, Ayeni O, Ayieni O, Petrisor BA, Bhandari M. Femoral neck shortening after fracture fixation with multiple cancellous screws: incidence and effect on function. J Trauma 2008 Jan;64(1):163-169. [doi: 10.1097/01.ta.0000241143.71274.63] [Medline: 18188116]
41. Le May S, Ballard A, Khadra C, Gouin S, Plint AC, Villeneuve E, et al. Comparison of the psychometric properties of 3 pain scales used in the pediatric emergency department: Visual Analogue Scale, Faces Pain Scale-Revised, and Colour Analogue Scale. Pain 2018 Aug;159(8):1508-1517. [doi: 10.1097/j.pain.0000000000001236] [Medline: 29608509]
42. Bauder RA, Khoshgoftaar TM. The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. Health Inf Sci Syst 2018 Sep 3;6(1):9 [FREE Full text] [doi: 10.1007/s13755-018-0051-3] [Medline: 30186595]
43. Fernandez A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Art Intell Res 2018 Jan;61(1):863-905. [doi: 10.1613/jair.1.11192]
44. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157-1182.
45. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015 Mar 4;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
46. Calculator of osteonecrosis.: Tongji Univeriity URL: http://calculator-of-osteonecrosis.herokuapp.com/ [accessed 2021-11-01]
47. O'Leary L, Jayatilaka L, Leader R, Fountain J. Poor nutritional status correlates with mortality and worse postoperative outcomes in patients with femoral neck fractures. Bone Joint J 2021 Jan;103-B(1):164-169. [doi: 10.1302/0301-620X.103B1.BJJ-2020-0991.R1] [Medline: 33380184]
48. Huang K, Fang X, Li G, Yue J. Assessing the effect of Gotfried reduction with positive buttress pattern in the young femoral neck fracture. J Orthop Surg Res 2020 Nov 07;15(1):511 [FREE Full text] [doi: 10.1186/s13018-020-02039-0] [Medline: 33160395]
49. Zhao G, Liu C, Chen K, Lyu J, Chen J, Shi J, et al. Nonanatomical reduction of femoral neck fractures in young patients (≤65 Years Old) with internal fixation using three parallel cannulated screws. Biomed Res Int 2021 Jan 4;2021:3069129 [FREE Full text] [doi: 10.1155/2021/3069129] [Medline: 33490267]
50. Li H, Li F, Liu N, Li P. Risk prediction of femoral head necrosis: a finite element analysis based on fracture mechanics. Int J Comput Methods 2019 Apr 04;17(06):1950019. [doi: 10.1142/s0219876219500191]
51. Barney J, Piuzzi N, Akhondi H. Femoral Head Avascular Necrosis. Treasure Island: StatPearls Publishing; Jan 2021.
52. Kazley J, Bagchi K. Femoral Neck Fractures. Treasure Island: StatPearls Publishing; 2021.
53. Liu BC, Sun C, Xing Y, Zhou F, Tian Y, Ji HQ, et al. [Analysis of risk factors for necrosis of femoral head after internal fixation surgery in young and mid-aged patients with femoral neck fracture]. Beijing Da Xue Xue Bao Yi Xue Ban 2020 Apr 18;52(2):290-297 [FREE Full text] [Medline: 32306013]
54. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak 2019 Jul 29;19(1):146 [FREE Full text] [doi: 10.1186/s12911-019-0874-0] [Medline: 31357998]
55. Meng Y, Yang N, Qian Z, Zhang G. What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values. J Theor Appl Electron Res 2020 Nov 20;16(3):466-490. [doi: 10.3390/jtaer16030029]

Abbreviations

AP: average precision
AUC: area under the receiver operating characteristic curve
FNF: femoral neck fracture
FPR: false-positive rate

LASSO: least absolute shrinkage and selection operator
LIME: local interpretable model-agnostic explanation
LR: logistic regression
ONFH: osteonecrosis of the femoral head
PR: precision-recall
PROBAST: Prediction Model Risk of Bias Assessment Tool
RF: random forest
ROC: receiver operating characteristic
SHAP: Shapley additive explanations
SMOTE: synthetic minority oversampling technique
SVM: support vector machine
VAS: visual analog scale
XGBoost: eXtreme Gradient Boosting

Edited by R Kukafka; submitted 02.05.21; peer-reviewed by JL Domínguez-Olmedo, A Lamer; comments to author 28.06.21; revised version received 01.09.21; accepted 28.09.21; published 19.11.21.

Please cite as:
Wang H, Wu W, Han C, Zheng J, Cai X, Chang S, Shi J, Xu N, Ai Z
Prediction Model of Osteonecrosis of the Femoral Head After Femoral Neck Fracture: Machine Learning–Based Development and Validation Study
JMIR Med Inform 2021;9(11):e30079
URL: https://medinform.jmir.org/2021/11/e30079
doi:10.2196/30079
PMID:34806984

Corrigenda and Addenda

Correction: Use of Deep Learning to Predict Acute Kidney Injury After Intravenous Contrast Media Administration: Prediction Model Development Study

Donghwan Yun[1,2], MD; Semin Cho[2], MD; Yong Chul Kim[2], MD, PhD; Dong Ki Kim[2], MD, PhD; Kook-Hwan Oh[2], MD, PhD; Kwon Wook Joo[2], MD, PhD; Yon Su Kim[1,2], MD, PhD; Seung Seok Han[1,2], MD, PhD

[1]Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, Republic of Korea

[2]Department of Internal Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea

Corresponding Author:
Seung Seok Han, MD, PhD
Department of Biomedical Sciences
Seoul National University College of Medicine
103 Daehakro, Jongno-gu
Seoul, 03080
Republic of Korea
Phone: 82 027408093
Email: hansway80@gmail.com

Related Article:

Correction of: https://medinform.jmir.org/2021/10/e27177

(JMIR Med Inform 2021;9(11):e34411) doi:10.2196/34411

In "Use of Deep Learning to Predict Acute Kidney Injury After Intravenous Contrast Media Administration: Prediction Model Development Study" (JMIR Med Inform 2021;9(10):e27177), one error was noted.

In the originally published paper, names of 5 authors (Yong Chul Kim, Dong Ki Kim, Kwon Wook Joo, Yon Su Kim, and Seung Seok Han) were inadvertently formatted with middle initials instead of the full author names.

The full authorship list was listed as follows in the originally published paper.

> *Donghwan Yun, Semin Cho, Yong C Kim, Dong K Kim, Kook-Hwan Oh, Kwon W Joo, Yon S Kim, Seung S Han*

This has been corrected to:

> *Donghwan Yun, Semin Cho, Yong Chul Kim, Dong Ki Kim, Kook-Hwan Oh, Kwon Wook Joo, Yon Su Kim, Seung Seok Han*

The correction will appear in the online version of the paper on the JMIR Publications website on November 1, 2021, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 21.10.21; this is a non–peer-reviewed article;accepted 26.10.21; published 02.11.21.

Please cite as:
Yun D, Cho S, Kim YC, Kim DK, Oh KH, Joo KW, Kim YS, Han SS
Correction: Use of Deep Learning to Predict Acute Kidney Injury After Intravenous Contrast Media Administration: Prediction Model Development Study
JMIR Med Inform 2021;9(11):e34411
URL: https://medinform.jmir.org/2021/11/e34411
doi:10.2196/34411
PMID:34727045

Corrigenda and Addenda

Figure Correction: Antibiotic Prescription Rates After eVisits Versus Office Visits in Primary Care: Observational Study

Artin Entezarjou[1], MD; Susanna Calling[1], MD, PhD; Tapomita Bhattacharyya[1], MD; Veronica Milos Nymberg[1], MD, PhD; Lina Vigren[2], MD, PhD; Ashkan Labaf[3], MD, PhD; Ulf Jakobsson[1], PhD; Patrik Midlöv[1], MD, PhD

[1]Center for Primary Health Care Research, Department of Clinical Sciences in Malmö/Family Medicine, Lund University, Malmö, Sweden

[2]Capio Go AB, Gothenburg, Sweden

[3]Department of Clinical Sciences in Lund, Lund University, Lund, Sweden

Corresponding Author:
Artin Entezarjou, MD
Center for Primary Health Care Research
Department of Clinical Sciences in Malmö/Family Medicine
Lund University
Box 50332
Malmö
Sweden
Phone: 46 40 391400
Email: artin.entezarjou@med.lu.se

Related Article:

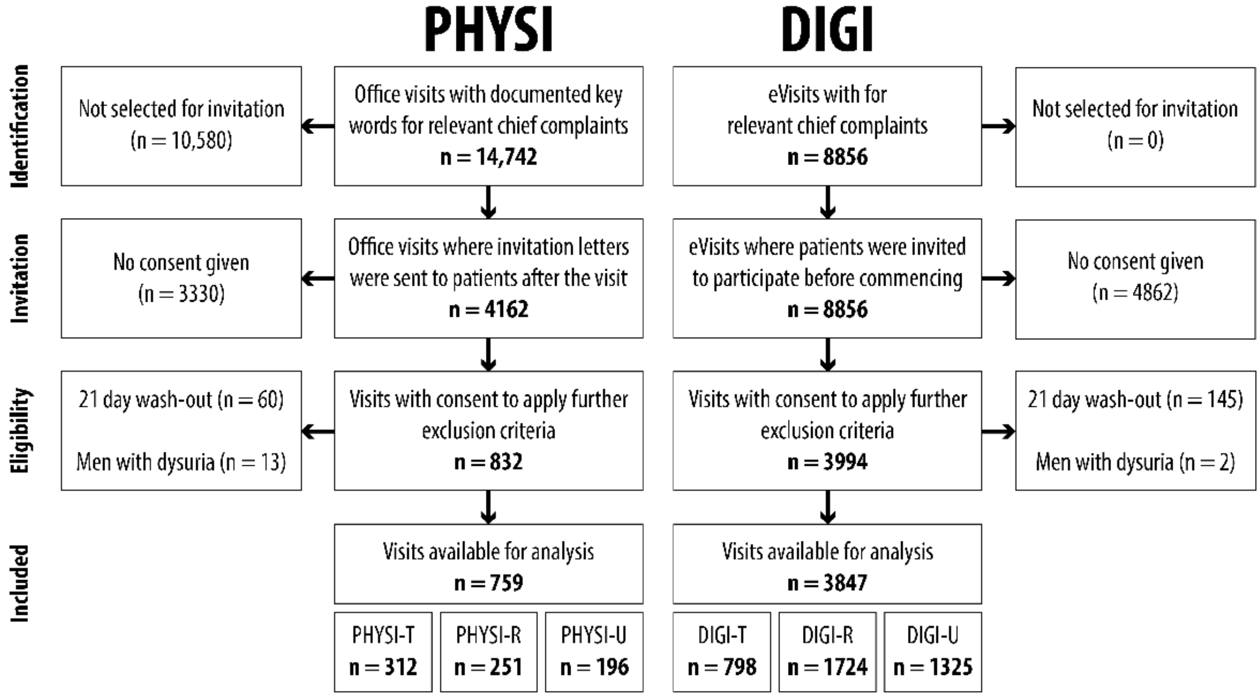
Correction of: https://medinform.jmir.org/2021/3/e25473/

(JMIR Med Inform 2021;9(11):e34529) doi:10.2196/34529

In “Antibiotic Prescription Rates After eVisits Versus Office Visits in Primary Care: Observational Study” (JMIR Med Inform 2021;9(3):e25473) the authors noted one error.

In the originally published paper, the flowchart of patient recruitment (Figure 1) contained incorrect exclusion criteria which are irrelevant for the current publication and thus had incorrect numbers. The figure also did not include the acronyms PHYSI, DIGI, PHYSI-T, PHYSI-R, PHYSI-U, DIGI-T, DIGI-R, and DIGI-U as specified in the figure caption. The figure has been replaced with the correct version, which includes all previously stated acronyms. The headings “Office visits” and “eVisits” have been replaced with “PHYSI” and “DIGI”, respectively. “Patients available for analysis” has been replaced with “Visits available for analysis.” Exclusion criteria now correctly include only “21 day wash-out” and “Men with dysuria.” The updated version of Figure 1 that will appear in the corrected manuscript is displayed below. The originally published version of Figure 1 can be found in Multimedia Appendix 1.

Figure 1. Flowchart of patient recruitment. PHYSI: primary care office visits; DIGI: eVisits; PHYSI-T: office visits with a chief complaint of sore throat; PHYSI-R: office visits with a chief complaint of common cold/influenza or cough; PHYSI-U: office visits with a chief complaint of dysuria; DIGI-T: eVisits with a chief complaint of sore throat; DIGI-R: eVisits with a chief complaint of common cold/influenza or cough; DIGI-U: eVisits with a chief complaint of dysuria.


PHYSI
DIGI
Identification
Invitation
Eligibility
Included
Not selected for invitation (n = 10,580)
Office visits with documented key words for relevant chief complaints n = 14,742
eVisits with for relevant chief complaints n = 8856
Not selected for invitation (n = 0)
No consent given (n = 3330)
Office visits where invitation letters were sent to patients after the visit n = 4162
eVisits where patients were invited to participate before commencing n = 8856
No consent given (n = 4862)
21 day wash-out (n = 60)
Men with dysuria (n = 13)
Visits with consent to apply further exclusion criteria n = 832
Visits with consent to apply further exclusion criteria n = 3994
21 day wash-out (n = 145)
Men with dysuria (n = 2)
Visits available for analysis n = 759
Visits available for analysis n = 3847
PHYSI-T n = 312
PHYSI-R n = 251
PHYSI-U n = 196
DIGI-T n = 798
DIGI-R n = 1724
DIGI-U n = 1325

The correction will appear in the online version of the paper on the JMIR Publications website on November 26, 2021, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Multimedia Appendix 1
Originally published Figure 1.
[PNG File , 158 KB - medinform_v9i11e34529_app1.png]

Submitted 27.10.21; this is a non–peer-reviewed article;accepted 27.10.21; published 26.11.21.

Please cite as:
Entezarjou A, Calling S, Bhattacharyya T, Nymberg VM, Vigren L, Labaf A, Jakobsson U, Midlöv P
Figure Correction: Antibiotic Prescription Rates After eVisits Versus Office Visits in Primary Care: Observational Study
JMIR Med Inform 2021;9(11):e34529
URL: https://medinform.jmir.org/2021/11/e34529
doi:10.2196/34529
PMID:34889754

Original Paper

A Blockchain-Based Dynamic Consent Architecture to Support Clinical Genomic Data Sharing (ConsentChain): Proof-of-Concept Study

Faisal Albalwy[1,2,3], BSc, MS; Andrew Brass[1,3], BSc, PhD; Angela Davies[3], BSc, PhD

[1]Department of Computer Science, University of Manchester, Manchester, United Kingdom

[2]Department of Computer Science, College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia

[3]Division of Informatics, Imaging and Data Sciences, University of Manchester, Manchester, United Kingdom

Corresponding Author:
Faisal Albalwy, BSc, MS
Department of Computer Science
University of Manchester
Oxford Road
Manchester, M13 9PL
United Kingdom
Phone: 44 161 306 6000
Email: faisal.albalwy@manchester.ac.uk

Abstract

Background: In clinical genomics, sharing of rare genetic disease information between genetic databases and laboratories is essential to determine the pathogenic significance of variants to enable the diagnosis of rare genetic diseases. Significant concerns regarding data governance and security have reduced this sharing in practice. Blockchain could provide a secure method for sharing genomic data between involved parties and thus help overcome some of these issues.

Objective: This study aims to contribute to the growing knowledge of the potential role of blockchain technology in supporting the sharing of clinical genomic data by describing blockchain-based dynamic consent architecture to support clinical genomic data sharing and provide a proof-of-concept implementation, called ConsentChain, for the architecture to explore its performance.

Methods: The ConsentChain requirements were captured from a patient forum to identify security and consent concerns. The ConsentChain was developed on the Ethereum platform, in which smart contracts were used to model the actions of patients, who may provide or withdraw consent to share their data; the data creator, who collects and stores patient data; and the data requester, who needs to query and access the patient data. A detailed analysis was undertaken of the ConsentChain performance as a function of the number of transactions processed by the system.

Results: We describe ConsentChain, a blockchain-based system that provides a web portal interface to support clinical genomic sharing. ConsentChain allows patients to grant or withdraw data requester access and allows data requesters to query and submit access to data stored in a secure off-chain database. We also developed an ontology model to represent patient consent elements into machine-readable codes to automate the consent and data access processes.

Conclusions: Blockchains and smart contracts can provide an efficient and scalable mechanism to support dynamic consent functionality and address some of the barriers that inhibit genomic data sharing. However, they are not a complete answer, and a number of issues still need to be addressed before such systems can be deployed in practice, particularly in relation to verifying user credentials.

(JMIR Med Inform 2021;9(11):e27816) doi:10.2196/27816

KEYWORDS

blockchain; smart contracts; dynamic consent; clinical genomics; data sharing

Introduction

Overview

With the advent of fast and effective next-generation sequencing technologies, unlinked and dispersed genomic data have emerged as a major challenge in diagnosing rare diseases. The molecular diagnosis of a rare disease involves comparing a patient’s genetic variant data with the variants of others with similar diseases in a large population. Therefore, sharing of data between genetic databases and laboratories is essential to identify overlapping results and for determining the pathogenic significance of variants to enable the diagnosis of rare genetic diseases.

One of the most common challenges to be overcome is that genomic data are often kept in centralized restricted-access repositories because of privacy and security concerns [1-7]; therefore, the data are difficult to locate or unavailable outside of the laboratories that own them. An in-depth qualitative study has revealed that current approaches to genomic data access and sharing through restricted-access repositories are time consuming and difficult and emphasized that the availability, discoverability, and accessibility of genomic data are bottlenecks to facilitating genomic data sharing [8]. There are also further challenges that hinder the large-scale sharing of genomic data, including a lack of time and the resources required to obtain consent to share [9], insufficient resources and infrastructure to track and recontact patients [10,11], lack of interoperability [1,2,12,13], and ethical issues [1,13-15].

Some of the above-mentioned challenges are the result of adopting centralized architectures for storing, sharing, and accessing genomic data. In such architectures, the data are stored in centralized databases and accessed through controlled access mechanisms. Although this approach to the gathering and management of genomic data has proven successful in the past, studies have revealed that such centralized architectures fail to properly address the growing demand for accessing genomic data [16,17]. This is concerning because the discoverability, availability, and accessibility of genomic data are essential for enabling the diagnosis of rare genetic diseases [8,18].

Various solutions to the challenges associated with the centralized storage of genomic data have been proposed. For example, federated data storage systems have been proposed to support genomic data sharing. The GA4GH Beacon Project [19] and i2b2 Data Sharing Network [20] are examples of such systems. Both use a federated network to connect institutions’ genomic databases, which enables them to process queries concerning the presence of genetic variants and traits. This also reduces the cost of genomic data transfers and allows institutions to maintain data control [21]. However, such systems have some drawbacks, including their failure to support complex queries, limitations to research institutions and hospitals, nonallowance of patient engagement in contributing or controlling their genomic data, and lack of decentralized governance [21,22].

Decentralized and distributed technologies have been suggested as a potential solution to promote genomic data sharing [23,24]. One emerging example of such a technology is blockchain technology. As decentralized and distributed technology, blockchain technology has many appealing properties, such as data integrity and accountability, that could be used to improve the integrity, discoverability, and accessibility of genomic data, thereby moving toward a new trusted infrastructure to support the promotion of genomic data sharing. This paper proposes blockchain-based dynamic consent architecture to support genomic data sharing. We present some design considerations and describe a proof-of-concept implementation for the proposed architecture called ConsentChain. The source code is available on Mendeley data [25] under the MIT license.

Background

Blockchain

Overview

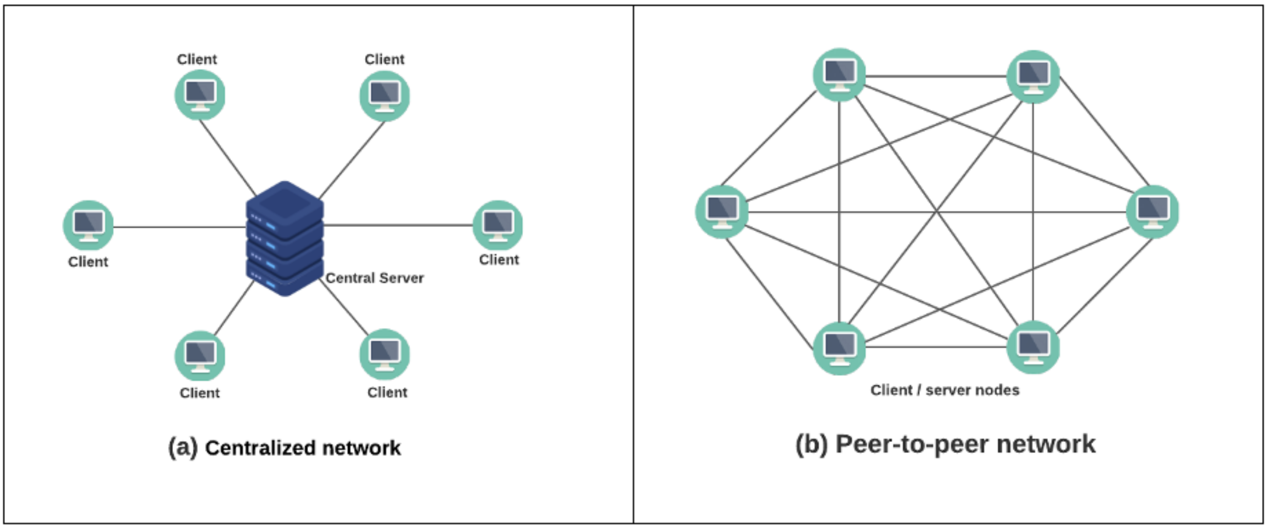
A blockchain is a protocol that enables a network of computers, known as nodes, to maintain a shared database called a ledger, without the need for complete trust between the network’s nodes [26]. It was originally developed as the underlying infrastructure for the peer-to-peer electronic cash system Bitcoin in 2009 [27]. Other blockchain platforms, including Ethereum [28] and Hyperledger Fabric [29], have emerged as the next generation of blockchain technology and implemented the concept of smart contracts, which was first introduced by Nick Szabo in the 1990s to build a digital relationship between 2 parties over computer networks [30]. In blockchain, a smart contract is a computer program that is stored, executed, and verified in the blockchain according to predefined conditions without the need for any trusted-third party [31]. The result of smart contract execution is a transaction recorded on a blockchain [28]. Ethereum smart contracts are written using high-level programming languages, such as Solidity and Vyper; therefore, they are vulnerable to coding bugs and malicious flaws [32].

Blockchain Architecture

A blockchain consists of 2 main components: a peer-to-peer network and a distributed ledger.

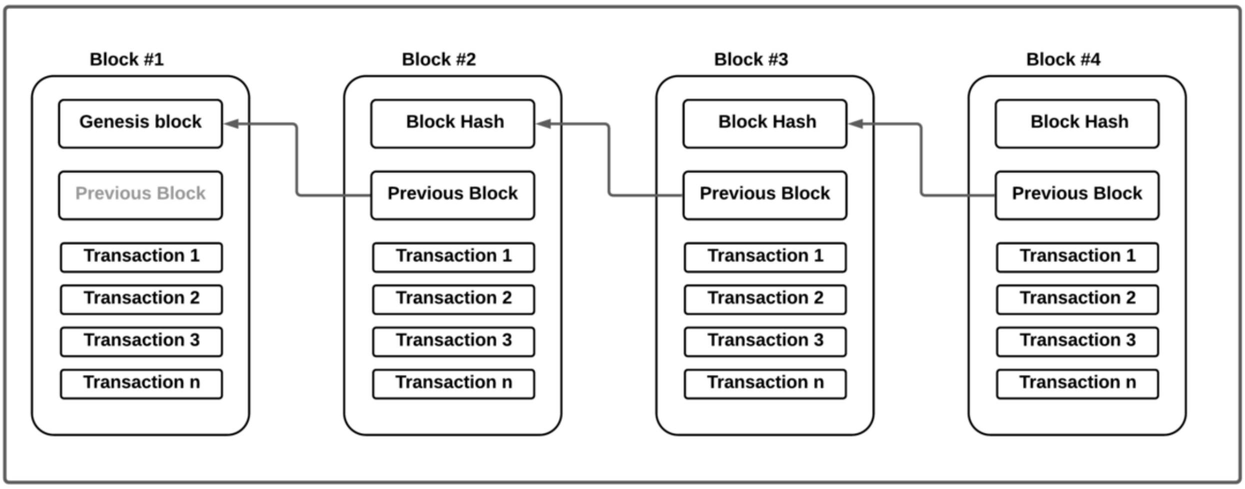
- Peer-to-peer network: understanding peer-to-peer networks is essential for understanding blockchains because, at its core, a blockchain is a peer-to-peer network. As stated, a peer-to-peer network consists of numerous connected computers called nodes. Each node in the network has a direct or indirect connection with the other network nodes. Each node makes a portion of its computational resources (ie, processing power or storage capacity) available directly to other nodes, without the need for central coordination by servers [33]. Unlike centralized networks, peer-to-peer networks have no central control, and each network node is equal to all others. Furthermore, all nodes function as both servers and clients. Figure 1 illustrates the architecture of the centralized and peer-to-peer networks.

Figure 1. The architectures of centralized and peer-to-peer networks.



- Distributed ledger: all transactions in the network are stored in a shared ledger. This consists of a chain of blocks, with each block containing a set of transactions. Each block is timestamped and linked to the blocks immediately preceding it. Each node maintains an identical copy of the shared ledger. To add a new transaction, the network nodes use a consensus protocol to evaluate and verify the new transaction. This protocol guarantees that a transaction is appended to the shared ledger only if most nodes validate the transaction. Once the transaction is appended to the shared ledger, it cannot be changed or reverted, and because all nodes have an identical copy of the shared ledger, no node has the power to change the data. This ensures the integrity of the shared ledger. However, recent research has proven that altering the shared ledger is feasible with 51% attacks where an adversary can control more than half of the total nodes in the blockchain network to alter the shared ledger [34]. Figure 2 illustrates a simplified blockchain concept.

Figure 2. Simplified blockchain concept.



Types of Blockchains

In terms of access to data and the role of nodes participating in the network, blockchain is classified into 4 types [35].

1. Public permissionless. Anyone can participate in the network and read or write data from the blockchain. Bitcoin and Ethereum are examples of a public permissionless blockchain.
2. Public permissioned. Anyone can participate in the network and read data from the blockchain, but a limited set of participants can write data in the blockchain. Ripple [36] and EOSIO blockchain [37] are examples of public permissioned blockchains.
3. Private permissionless. A limited set of participants can participate in a network in which all participate can read or write data from or in the blockchain. Holochain [38] is an example of a private permissionless blockchain.
4. Private permissioned. A limited set of participants can participate in the network and read data from the blockchain, but a subset of them can write data in the blockchain. Hyperledger Fabric [39] and Hyperledger Besu [40] are examples of privately permissioned blockchains.

Dynamic Consent and Blockchain

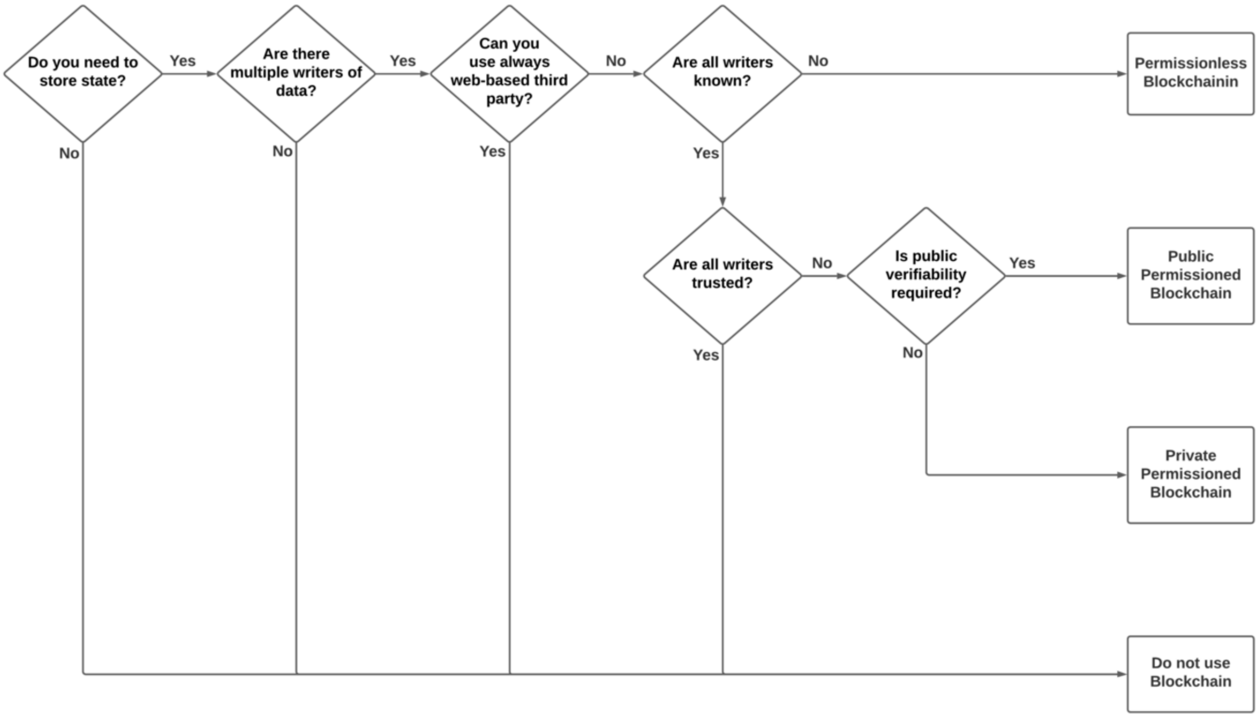
Dynamic consent is a two-way communication method that enables individuals to specify what data they are willing to share with various health care providers by setting and modifying their consent preferences. It enables individuals to control their data by granting and revoking access to their data, tracking their data, and updating their consent preferences. Despite these benefits, the implementation of dynamic consent in clinical genetics is limited because of ethical, legal, and data security concerns. The lack of patient trust [41,42], confidentiality data and misuse [42,43], and the lack of traceability and transparency mechanisms [44-47] are among the greatest concerns. Blockchain technology has many appealing properties, such as immutability, transparency, and accountability, that can address some of the barriers that inhibit the implementation of dynamic consent. Blockchain can support dynamic consent, as follows: data transparency and accountability through an immutable ledger, data security and privacy using cryptography mechanisms, and an efficient management system through smart contracts.

Methods

Blockchain Potential in Genomic Data Sharing

Determining whether blockchain is applicable to a particular scenario is not an easy task. Although no general formula or rule exists for the applicability of blockchain, several decision schemes have been proposed to determine whether a blockchain should be used depending on situational requirements [48-50]. Wüst and Gervais [48] proposed a decision tree to identify the scenario-based applicability of blockchain, as shown in Figure 3. This decision tree consists of 6 questions. Next, we answer these questions by considering our genomic data-sharing scenario.

Figure 3. Decision tree to determine the use of blockchain [48].


Do you need to store state?
Yes
Are there multiple writers of data?
Yes
Can you use always web-based third party?
No
Are all writers known?
No
Permissionless Blockchainin
No
No
Yes
Yes
Are all writers trusted?
No
Is public verifiability required?
Yes
Public Permissioned Blockchain
Yes
No
Private Permissioned Blockchain
Do not use Blockchain

1. Do you need to store state? The answer to this question is yes. Diagnosing a patient with a rare genetic disease is a complex and time-consuming task, as it involves gathering data from multiple sources [51]. For instance, to answer a simple question of whether a mutation in a patient associated with a particular disease has been previously reported with the same or similar disorders in another individual requires accessing preexisting genetic and phenotypic data from multiple databases relevant to the clinical case [51,52]. Therefore, uniform access to preexisting genotype and phenotypic data using blockchain could improve the discovery and diagnosis of rare diseases. Moreover, accessing such databases involves legal and ethical obligations, including patient consent. For example, patients must control their own data and keep track of who has access to their data at any given time. Therefore, the storage and collection of patient consent as well as the administration of consent and data traceability will be guaranteed by using blockchain.
2. Are there multiple writers of data? In clinical genomics, multiple parties are involved in the patient treatment pathway, such as clinicians, scientists, and clinical laboratory technicians [51]. Therefore, a single source of truth is required for the patient data. Owing to the immutability of blockchain, the existence of patient data as well as the ownership and integrity of the data can be guaranteed. Therefore, considering that multiple parties would produce and deliver patient data, this question can be answered with yes.
3. Can you use an always web-based trusted third party? Trust and consent are important factors in the successful advancement of genome medicine and research. Patients

should feel confident that their data are handled safely and are only used with their consent. A recent Genome UK report [53] showed that patients and the public are optimistic about the potential of genome medicine, but they have concerns related to the security and use of their data. It is reasonable to mention that patients trust health care providers more than any third party with their data. However, because of the high profile of patient data breaches [54,55] by health care providers, this trust has been broken. Blockchain can eliminate the need for a trusted party by establishing trust between system actors through its robust technical infrastructure and cryptography mechanisms. Therefore, the answer to this question is probably no.

4. Are all writers known? To produce, manage, and store patient data, health care providers must identify themselves. Moreover, patients need to identify themselves to connect with health care providers. Therefore, a clear answer to this question is yes.
5. Are all writers trusted? Although a minimum level of trust is required between patients and health care providers, health care providers might use patient data for research purposes without obtaining explicit consent from patients [56-58]. Blockchain enables accountability and transparency in the system by providing an audit trail and traceability of the stored data, which in turn reinforces patients' trust in health care providers. Therefore, the answer to this question is probably no.
6. Is public verifiability required? Even though patient data are not stored in the blockchain directly (off-chain storage), access to the system should be private and permissioned. Thus, the answer to this question is no.

On the basis of the answers to these 6 questions, it is clear that the use of blockchain for the proposed genomic data sharing scenario is justifiable.

Design Requirements

Overview

To identify the design requirements for ConsentChain, we analyzed a recent deliberative focus group study with National Health Service (NHS) Genomic Medicine Service patients regarding public opinion on sharing genomic data (National Research Ethics Committees ethical approval reference 18/NW/0510) [59]. We used the user stories method [60] to capture the main system design requirements. We used card sorting to collect data from the manuscript. We used our interpretation to represent the statements made by the study participants in simple user stories. We then discussed these user stories with a focus group study team to refine them. We emphasize that the findings from the focus group study are partially applicable to the scenario of our blockchain use case. Finally, 6 design requirements were identified.

Requirement 1: Data Discovery

User Stories

> *As a patient, I want my data to be available for sharing to facilitate my diagnosis and treatment.*
>
> *As a patient, I want my unidentifiable data to be available for wider sharing to help others' treatment and facilitate extensive research.*
>
> *As a patient, I want my data to be available for different healthcare providers, so I won't have to repeat myself every time I visit a new healthcare provider.*

Context

The study participants allowed the sharing of their genomic data to support the diagnosis and treatment of their conditions across multiple health care providers. They also agreed to use their genomic data to benefit other patients with similar genetic conditions and for future research.

Implications for System Design

The system should allow information about a genomic data set of interest stored in an individual genetic laboratory to be discoverable and accessible by health care professionals and researchers.

Requirement 2: Data Security

User Stories

> *As a patient, I want best practices in data security to be implemented to protect my data so that it can be safeguarded against hacking and loss.*
>
> *As a patient, I want to have different levels of purpose to access my data, so they can be used for authorised purposes.*

Context

There was consensus among the participants that genomic data should be stored and shared securely without unauthorized alteration while making them available for authorized purposes.

Implications for System Design

Security techniques, such as data encryption and access control, should be used to protect sensitive data. Owing to the open and transparent nature of blockchains, sensitive data (either encrypted or not) should not be stored in the chain.

Requirement 3: Data Privacy

User Stories

> *As a patient, I want my genetic data to be shared without my identifiable information (eg, my name), so my identity will not be compromised.*

Context

The participants emphasized that sharing genomic data outside of the patient's direct care should be anonymized to protect their identity.

Implications for System Design

The system should allow the flow of patient data among involved parties while minimizing the risk of patient identity disclosure.

Requirement 4: Patient Control Over Data and Requirement 5: Traceability

User Stories

> *As a patient, I want to give my consent to share my data for certain purposes that are clearly outlined so that no further consent is required for these purposes.*
>
> *As a patient, I want to be told whether the purpose of sharing my data is changed so I'll have the option of giving explicit permission for the new changes.*
>
> *As a patient, I want to have the option to update/withdraw my consent in a straightforward and easy way so I can change my mind later.*
>
> *As a patient, I want to be able to track my shared data so that I know when and with whom my data are being shared.*

Context

The participants thought that they should be asked for permission to share their data and be informed about how their data would be used and for what purpose. Moreover, some believed that they would exercise their right to opt out.

Implications for System Design

The system should enable patients to update their permissions dynamically and track data that are being shared with different parties.

Requirement 6: Minimum Data Disclosure

User Stories

> *As a patient, I want to have different levels of role requesters designated to access my data so only authorised parties can gain access.*
>
> *As a patient, I want to have a time limit for my shared data, so they cannot be used for other purposes in the future.*

Context

Some participants were concerned about unauthorized disclosure of their data to third parties, including family members, employers, and law enforcement agencies, whereas others were concerned with restricting access to their data by commercial entities.

Implications for System Design

The system should be designed in a way that allows the sharing of patient data for a given time frame and specific purpose.

Consent Elements

Inspired by the Global Alliance for Genomics and Health (GA4GH) data use ontology effort to model genomic data use restrictions and data access requests [61,62], we developed an ontology model to represent patient consent elements into machine-readable codes. The model includes consent elements describing the data type, purpose, and role of the data requester (DR). Tables 1-3 show an abstract view of the consent elements and their codes. We also introduced an access policy tree representing a Boolean formula that defines a combination of consent elements. Any data access request that satisfies the tree can obtain access to patient data. Figure 4 shows an example of an access policy tree that allows patient genotype data to be accessed by a clinician for treatment.

Table 1. Code representing the data type in consent element.

Data type	Code
Genotype	GNE
Phenotype	PHE
Metadata	MEA

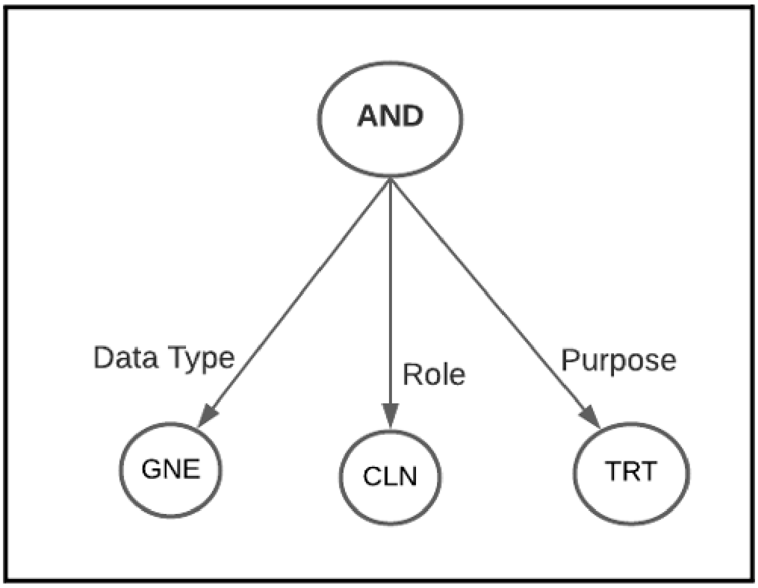
Table 2. Code representing the role in consent element.

Purpose	Code
Treatment	TRT
Research	REH
Clinical	CLL

Table 3. Code representing the purpose in consent element.

Role	Code
Clinician	CLN
Researcher	REE
Bioinformatician	BIN

Figure 4. Example of an access policy tree where patient genotype data to be accessed by a clinician for treatment. CLN: clinician; GNE: patient genotype data; TRT: treatment.

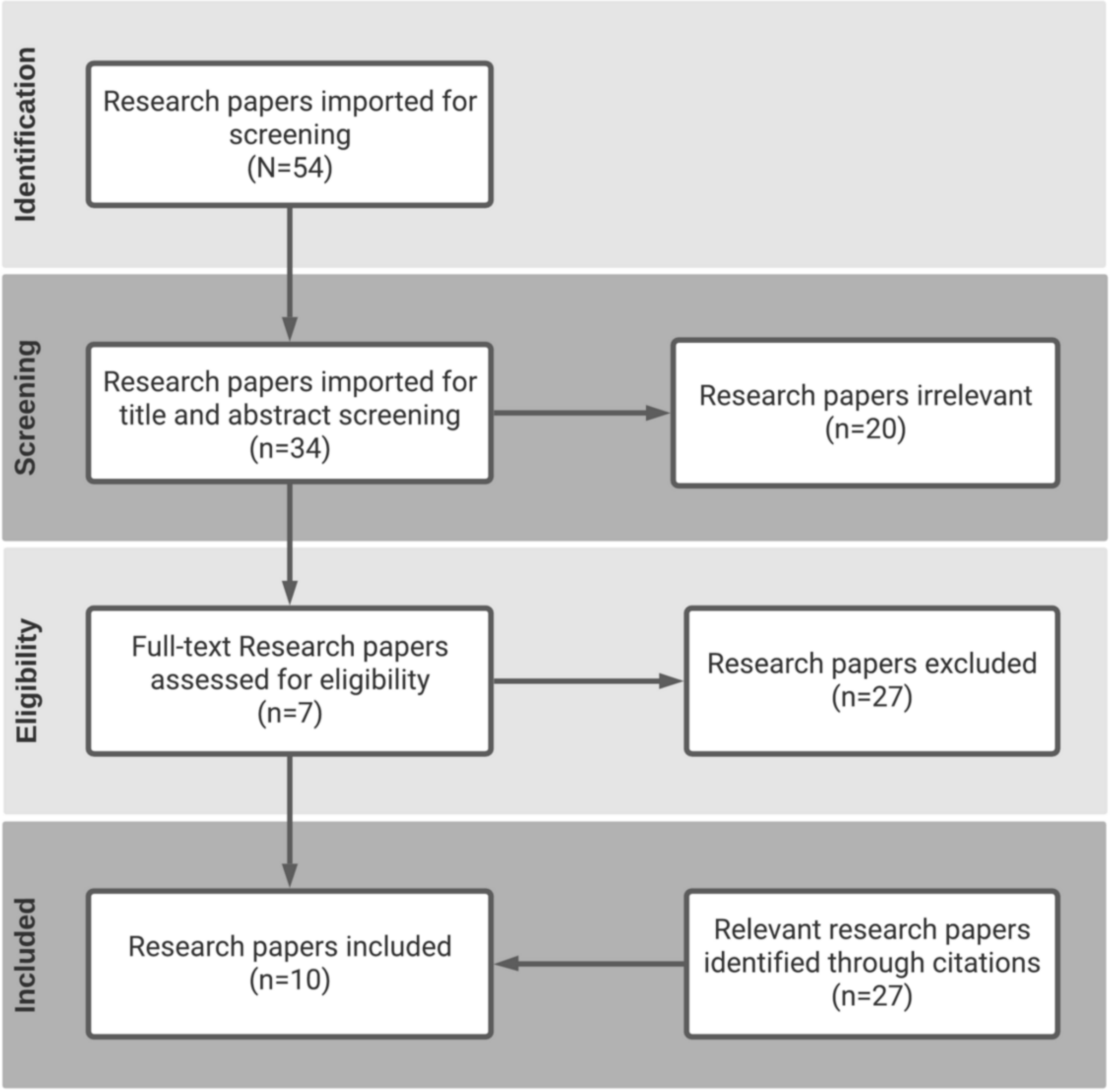


Related Work

We used PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to conduct a systematic review to analyze the existing literature on blockchain-based consent data used in health care management systems. The PRISMA flowchart for this systematic review is shown in Figure 5. For the purposes of this review, a reputable database (PubMed) was searched using the search query shown in Textbox 1. The resulting research papers (N=54) were imported into Covidence, a web-based app tool used to manage systematic reviews. In the next step, research papers were screened against titles and abstracts, and research papers unrelated to consent management systems were excluded (n=20). Then, the remaining research papers (n=34) were assessed for full-text eligibility, with the following exclusion criteria:

- No consent management explained (n=13)
- No implementation provided (n=2)
- No access to the full text (n=2)
- Reviews and ideas (n=6)

Figure 5. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow for this review.


Identification
Research papers imported for screening (N=54)
Screening
Research papers imported for title and abstract screening (n=34)
Research papers irrelevant (n=20)
Eligibility
Full-text Research papers assessed for eligibility (n=7)
Research papers excluded (n=27)
Included
Research papers included (n=10)
Relevant research papers identified through citations (n=27)

Textbox 1. Research query.

((blockchain[Title/Abstract]) OR (Smart contracts [Title/Abstract]) OR (blockchain-based[Title/Abstract]) OR (Smart contracts-based[Title/Abstract])) AND ((Consent*[Title/Abstract]) OR (permission*[Title/Abstract]) OR (access control[Title/Abstract])) AND ((healthcare[Title/Abstract]) OR (EMR[Title/Abstract]) OR (genomic[Title/Abstract]) OR (Genetic [Title/Abstract]) OR (electronic health records[Title/Abstract]) OR (EHR[Title/Abstract]) OR (electronic Medical Records [Title/Abstract]) OR (Medical[Title/Abstract]) OR (Clinical Trial[Title/Abstract]) OR (Patient*[Title/Abstract]))

Additional relevant research papers were identified through citations (n=3). The remaining research papers and the identified relevant research papers (n=10) were analyzed thoroughly. The final findings are summarized in Multimedia Appendix 1 [63-72].

Chenthara et al [63] proposed a blockchain-based privacy-preserving framework called Healthchain to support electronic health record (EHR) access control and management. The framework was implemented using the Hyperldger Fabric InterPlanetary File System (IPFS). To achieve the immutability of EHRs, they were stored off-chain in an IPFS, with only the hash values of the EHRs being stored in the blockchain. Smart contracts were used to model the logic of EHR transactions in the system, including data exchange, access management, and EHR management. Azaria et al [64] proposed a decentralized management system called MedRec, which was built using Ethereum smart contracts to facilitate the management of EHRs between health care providers. MedRec enables patients to have full control over their data by granting or revoking access to their data. To keep patients anonymous, their identification strings are mapped to their blockchain addresses. Smart contracts are used to define how data are managed and accessed. MedRec provides an immutable access history summary that improves accountability and transparency in the system. It can be integrated with current providers' existing databases, and other medical stakeholders can participate.

Cryan [65] proposed a blockchain-based architecture capable of enabling patient data sharing across hospital systems. The proposed architecture was implemented using Ethereum smart contracts and IPFS to protect sensitive patient data and enable patients to own and share their data with designated clinicians and revoke that permission later. Choudbhury et al [66] developed a decentralized system using Hyperledger Fabric for informed consent management and secondary data sharing. The system enhances compliance in human subject regulations for institutional review board regulations by leveraging smart contracts to enable a quick and efficient recording of consent and enforce the guidelines of a clinical trial protocol. Mamo et al [67] presented a well-designed system called Dwarna that harnesses blockchain technology to enable dynamic consent in biobanking. This system aims to increase transparency by storing the research participants' consent changes on the blockchain and presents a solution to overcome the blockchain incompatibility with Article 17 of the European Union's General Data Protection Regulation (GDPR), known as the right to erasure, by using a different representation of research participants in both off-chain databases and blockchain. The proposed system was implemented using a Hyperledger Fabric blockchain.

Tith et al [68] proposed a blockchain-based consent management model to support the sharing of EHRs. The model was implemented using Hyperledger Fabric and where smart contracts were used to manage patient consent. Patient consent preferences, metadata of patient records, and data access logs are stored immutably on the blockchain, enabling transparency and traceability of patient data and consent. Dubovitskaya et al [69] proposed a secure blockchain-based record management system that facilitates the secure sharing and aggregation of EHR data. The system is patient-centric and allows patients to manage their own EHRs across multiple hospitals. It uses proxy re-encryption algorithms and a fine-grained access control mechanism to ensure patient privacy and confidentiality. Dubovitskaya et al [70] proposed a framework on a permissioned blockchain for sharing EHRs for care of patients with cancer. The proposed framework is implemented with the Hyperledger Fabric blockchain and uses a membership service to authenticate registered users using username or password credentials. To create patient identity, personally identifying information, such as name, social security number, and date of birth, are hashed and encrypted for security. Medical data were stored off-chain in secure cloud storage, where access management is managed by smart contract logic.

Rajput et al [71] presented a blockchain-based access control framework that maintains patient data privacy under emergency conditions. The framework was implemented on the permissioned blockchain Hyperledger Fabric, and smart contracts were used to enable patients to manage the access rules for their data. The system keeps the history-of-transactions logs while patients are in an emergency, enabling auditing at any time point. Zhuang et al [72] presented a generalized blockchain-based architecture that provides generic functions and methods for a wide spectrum of health care apps. These functions and methods include requesting patient data, data access permission granting or revoking, and data tracking. The presented architecture was implemented on the Ethereum blockchain in 2 relevant health app domains: health information exchange and subject recruitment for clinical trials.

Compared with existing relevant literature, the proposed system is dynamic and supports minimum data disclosure. To the best of our knowledge, no relevant literature has reported on the 6 design requirements and provides a detailed analysis of the system performance. Multimedia Appendix 1 [63-72] summarizes the literature for blockchain-based consent management systems.

System Architecture

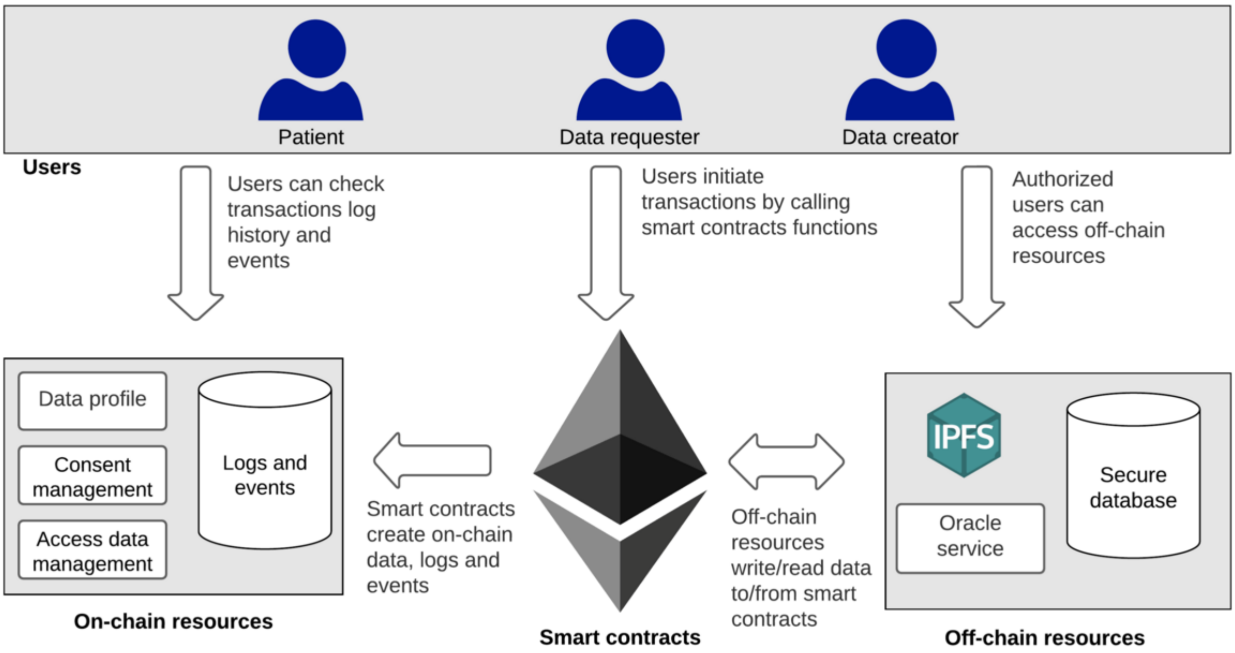
In this section, we describe the proposed blockchain-based dynamic consent architecture for supporting clinical genomic data sharing. This generic architecture can be customized and used in different use cases where dynamic consent is required. As illustrated in Figure 6, the components of the proposed architecture are as follows:

1. Users
   - A data creator (DC): an organizational entity, such as a genetic testing laboratory, where patient data are collected and stored in secure databases.
   - Patient: an individual whose data are stored off-chain in a secure database managed by the DC; a patient can provide consent to the system using the
     consent elements
     code.
   - DR: a domain expert or organizational entity that wishes to discover and request access to patient data for a specific purpose, including research and health care.
2. Smart contracts, which are used to provide system functionalities, such as registering new users, managing patient consent, and processing access requests to patient data. In addition, smart contracts create transaction logs and events that enable the tracing and auditing of all system data and actions.
3. On-chain resources
   - Logs and events: smart contracts create logs and events for all system transactions. These logs and events are stored on-chain, and they are an important resource for tracing and auditing all system actions, thus making all system users accountable for their actions.
   - Data profile (DP): This is a description of preexisting genomic data for a specific patient that is stored off-chain in a genetic laboratory database. A patient DP contains information including the location of the patient data, patient condition, and gene name, and it does not reveal any sensitive and identifiable information. Storing patient DPs on-chain helps the DR to discover and identify a genomic data set of interest stored in several genetic laboratory databases.

- Consent management: This is used to handle patient consent operations, such as adding, updating, and deleting consent.
- Access data management: This is used to handle access to patient data procedures, including validating access requests and providing secure access to off-chain data.

4. Off-chain resources
   - Secure database: a private database managed by a DC in which all information related to the required DP is stored.
   - Oracle service: by design, blockchain and smart contracts cannot access and read off-chain data; therefore, oracle services are used. An oracle service is a trusted data feed service that provides off-chain data to the blockchain. In the proposed system, an oracle service is used to enable smart contracts to communicate with a secure database.
   - IPFS: This is a decentralized file storage system that stores and shares various types of files permanently. Each stored file is given a unique hash value based on its content. This hash value is then used to retrieve the file from the system. In the context of this study, we leverage IPFS as a key management service to store users' public key (PU). We believe that IPFS is the best candidate for users' PU because of its high availability and low cost.

Figure 6. The components of the proposed architecture. IPFS: InterPlanetary File System.


Patient
Data requester
Data creator
Users
Users can check transactions log history and events
Users initiate transactions by calling smart contracts functions
Authorized users can access off-chain resources
Data profile
Consent management
Access data management
Logs and events
On-chain resources
Smart contracts create on-chain data, logs and events
Smart contracts
Off-chain resources write/read data to/from smart contracts
IPFS
Oracle service
Secure database
Off-chain resources

Results

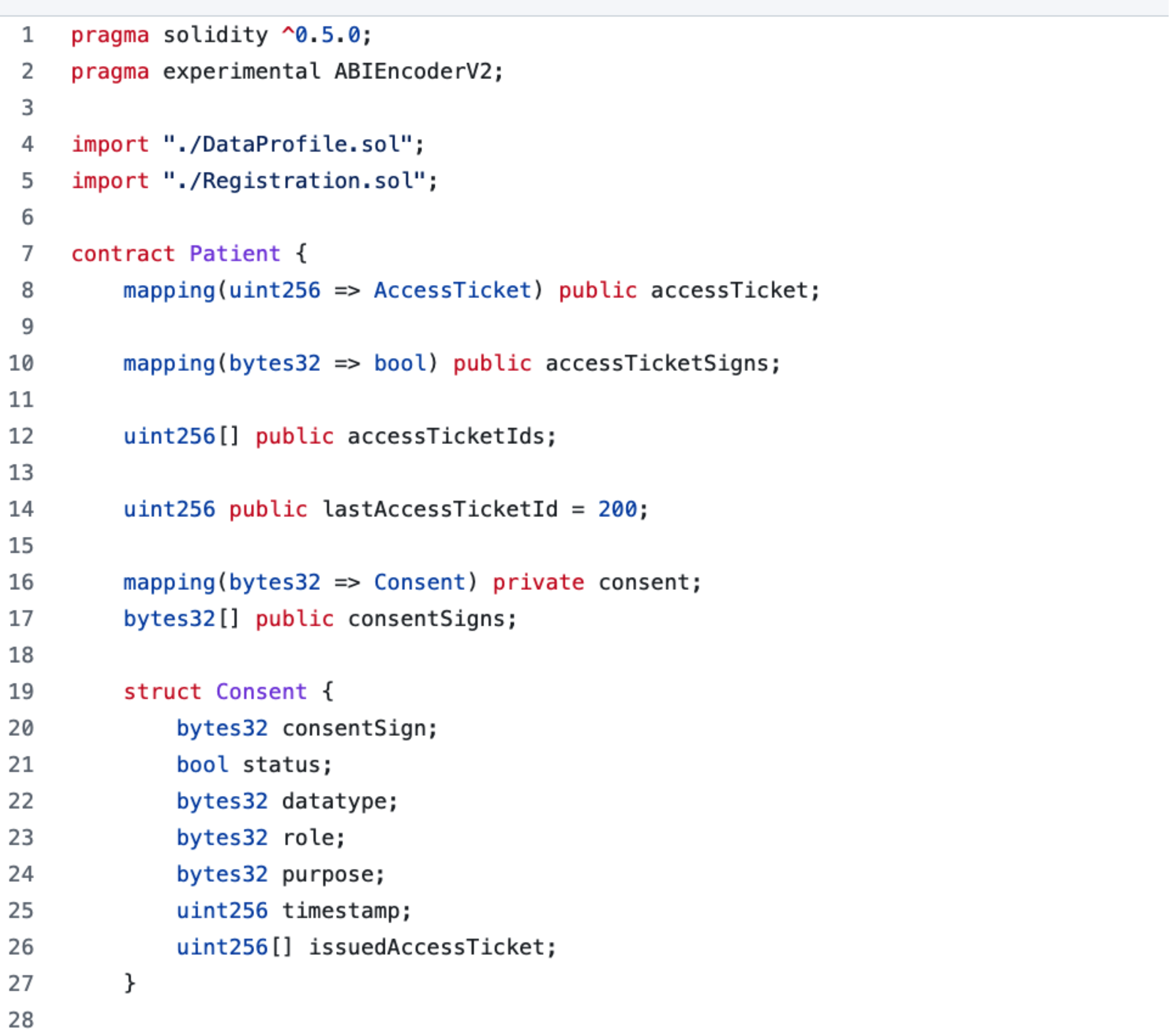
Implementation

Overview

We implemented our proof-of-concept on a privately permissioned blockchain to demonstrate the feasibility of our blockchain-based architecture. At the infrastructure level, Hyperledger Besu [40], an open-source Ethereum client that provides permissioned private blockchain networks, was used to build a private blockchain. The Solidity programming language was used to write the system smart contracts and truffle framework, a development tool for developing and testing Ethereum smart contracts, to test, compile, and deploy system smart contracts. Figure 7 shows a portion of the patient's smart contract code. Finally, we used Provable [73] as an oracle service and MongoDB to create an off-chain database.

Six smart contracts are written to manage on-chain transactions: registration smart contract (RSC), patient smart contract (PSC), data profile smart contract (DPSC), data creator smart contract (DCSC), data requester smart contract (DRSC), and oracle service smart contract (OSSC). These smart contracts provide 8 main system functions: *createNewDataRequestorContract*, *createNewPatientContract*, *CreateNewDataCreatorContract*, *setConsent*, *cancelConsent*, *checkConsent*, *setupDataProfile*, *requestAccessTicket*, and *requestAccessToken*. We used smart contract modifiers to restrict the calling of these functions to authorized users. Any unauthorized function call results in stopping the execution of the function and reverting all changes to the original state. The remainder of this section explains the implementation of the main system functionalities using smart contract functions.

Figure 7. An illustrative example of patient smart contract code.



```
pragma solidity ^0.5.0;
pragma experimental ABIEncoderV2;

import "./DataProfile.sol";
import "./Registration.sol";

contract Patient {
    mapping(uint256 => AccessTicket) public accessTicket;

    mapping(bytes32 => bool) public accessTicketSigns;

    uint256[] public accessTicketIds;

    uint256 public lastAccessTicketId = 200;

    mapping(bytes32 => Consent) private consent;
    bytes32[] public consentSigns;

    struct Consent {
        bytes32 consentSign;
        bool status;
        bytes32 datatype;
        bytes32 role;
        bytes32 purpose;
        uint256 timestamp;
        uint256[] issuedAccessTicket;
    }

```

Registration

Each system participant interacts with the system via his or her smart contract, which includes all the required information to interact with the system. Therefore, the participant should be registered in a system in which a smart contract is created. All users' identities and professional registrations should be verified by a system admin, who is responsible for setting up the system and inviting the authorities to join the system, such as the NHS, before proceeding with the process of system registration. Textboxes 2-4 describe the user registration process for the patient, DC, and DR, respectively. The system admin executes a specific smart contract function for each user, which creates a new smart contract and assigns the user as the owner of the contract. This is done by using modifiers to restrict the calling of the user smart contract functions to the user's Ethereum address.

Textbox 2. Pseudocode of registering new patient.

```
Algorithm 1:createNewPatientContracter
Input:caller, patientWalletAddress
Output: smartContractAddress
If caller=admin$\wedge$patientWalletAddress$\neq$null then
    Create newPatientSmartContract
    Set newPatientSmartContract owner to patientWalletAddress
    Output newPatientSmartContract address
Else
    Revert smart contract state and show an error message
```

Textbox 3. Pseudocode of registering new data creator.

```
Algorithm 2:createNewDataCreatorContract
Input: caller, dataCreatorWalletAddress
Output: smartContractAddress
If caller = admin$\wedge$dataCreatorWalletAddress $\neq$ null then
    Create new DataCreatorSmartContract
    set newDataCreatorSmartContract owner to
    dataCreatorWalletAddress
    Output newDataCreatorSmartContract address
Else
    revert smart contract state and show an error message
```

Textbox 4. Pseudocode of registering new data requester.

```
Algorithm 3: createNewDataRequestorContract
Input: caller, dataRequesterWalletAddress, dataRequesterPUK
Output: smartContractAddress
If caller=admin$\wedge$dataCreatorWalletAddress$\neq$null$\wedge$
dataRequesterPUK$\neq$null then
    Create newDataRequesterSmartContract
    Set newDataRequesterSmartContract owner to
    dataRequesterWalletAddress
    set newDataRequesterSmartContract’s public key to dataRequesterPUK
    output newDataRequesterSmartContract address
Else
    revert smart contract state and show an error message
```

Consent Management

Textbox 5 describes the process of creating and storing patient consent by submitting the elements of the access policy tree, which represents the patient’s consent, to the patient’s smart contract". The tree elements are then hashed to create a consent signature, which is then stored in the patient’s smart contract. A mapping data structure, a data structure type that consists of key types and corresponding value type pairs, is used to store the consent signature, which is used as a key associated with a Boolean value to indicate its status (eg, the value is true for valid consent and false for invalid consent). Hashing and storing the consent tree in a mapping data structure would enable efficient consent status retrieval and validation. As shown in Textbox 6, if the patient wants to cancel his or her consent, the associated value with the consent signature would be set to false. Textbox 7 describes the process of checking a patient’s consent status by returning the associated value with the consent signature.

Textbox 5. Pseudocode of storing patient consent

```
Algorithm 4: setConsent
Input: caller, dataType, role, purpose
Output: status
$CONSENT \leftarrow mapping$
If $caller = contractOwner \wedge dataType \neq null \wedge role \neq null \wedge purpose \neq null$, then
    $h \leftarrow hash(dataType, role, purpose)$
    if CONSENT.contain(h,true) then
        revert smart contract state and show an error message
    else
        CONSENT.insert(h,true)
        Output true
Else
    Revert smart contract state and show an error message
```

Textbox 6. Pseudocode of cancelling patient consent.

```
Algorithm 5: cancelConsent
Input: caller, dataType, role, purpose
Output:status
$CONSENT \leftarrow mapping$
If $caller = contractOwner \wedge dataType \neq null \wedge role \neq null \wedge purpose \neq null$, then
    $h \leftarrow hash(dataType, role, purpose)$
    if CONSENT.contain(h,false) then
        revert smart contract state and show an error message
    Else
        CONSENT.insert(h,false)
        output true
Else
    Revert smart contract state and show an error message
```

Textbox 7. Pseudocode of checking patient consent.

```
Algorithm 6: checkConsent
Input: dataType, role, purpose
Output: status
$CONSENT \leftarrow mapping$
If $dataType \neq null \wedge role \neq null \wedge purpose \neq null$, then
    $h \leftarrow hash(dataType, role, purpose)$
    $r \leftarrow CONSENT.return(h)$
    output r
Else
    revert smart contract state and show an error message
```

Patient Data

Textbox 8 describes the process of submitting the patient data to the system. After collecting and storing patient data in a secure, off-chain database (eg, a genomic laboratory database), the DC submits the patient metadata, a description of the patient data that does not reveal sensitive and identifiable information, such as the hash of the stored data, conditions, data type, and gene name, to the system. The patient metadata are then stored in a data structure, where the hash of the stored data is used as a key and the remaining patient data are the value.

Textbox 8. Pseudocode of creating patient data profile.

```
Algorithm 7: setupDataProfile
Input: caller, patientSmartContract, dataHash, condition, dataType,gene
Output: id
DATAPROFILE←mapping
i←counter
if caller = dataCreatorSmartContract ∧ patientSmartContract ≠ null ∧ datatHash ≠ null ∧ condition ≠ null ∧ dataType ≠ null ∧ gene ≠ null
then
    i++
    DATAPROFILE.insert(i,[patientSmartContract, dataHash, condition, dataType, gene, dataCreatorSmartContract])
    output i
Else
    revert smart contract state and show an error message
```

Access Management

To access patient data, the DR needs to obtain an access ticket (ATi) and access token (ATo). The ATi is used to control access to patient data, whereas the ATo is used to minimize access to the requested data to the lowest level. Textbox 9 describes the process of requesting an ATi for the patient data. After identifying a potential patient's data, the DR must submit an ATi request to the system to provide the hash of the requested data, his role, and the purpose of accessing the data. Then, the request is verified by the patient's smart contract in which the patient's consent is stored. If there is valid consent that matches a DR request, an ATi is created automatically for the DR.

Textbox 9. Pseudocode for requesting access tickets to access off-chain patient data.

```
Algorithm 8: requestAccessTicket
Input: caller, dataProfileId, role, purpose
Output:ticketId
DATAPROFILE←mapping
If caller=contractOwner ∧ dataProfileId ≠ null ∧ role≠ null ∧ purpose ≠ null, then
    d←DATAPROFILE.return(dataProfileId)
    patient←d.patientSmartContract
    dataType←d.dataType
    h←hash(dataType, role, purpose)
    if patient.CONSENT.return(h)=true then
        ticket←patient.CreateAccessTicket(caller, dataProfileId)
        ticket.status=true
        output ticket.id
    Else
        revert smart contract state and show an error message
Else
    revert smart contract state and show an error message
```

To obtain an ATo, the DR must submit a valid ATi to the system. Textbox 10 describes the process of requesting an ATo.

If the ATi is still valid and patient consent has not been updated or cancelled, an ATo is generated automatically by the DC for the DR. The ATo includes a secure one-time URL that can be used to gain access to the patient data stored off-chain.

Textbox 10. Requesting an access token to retrieve off-chain patient data.

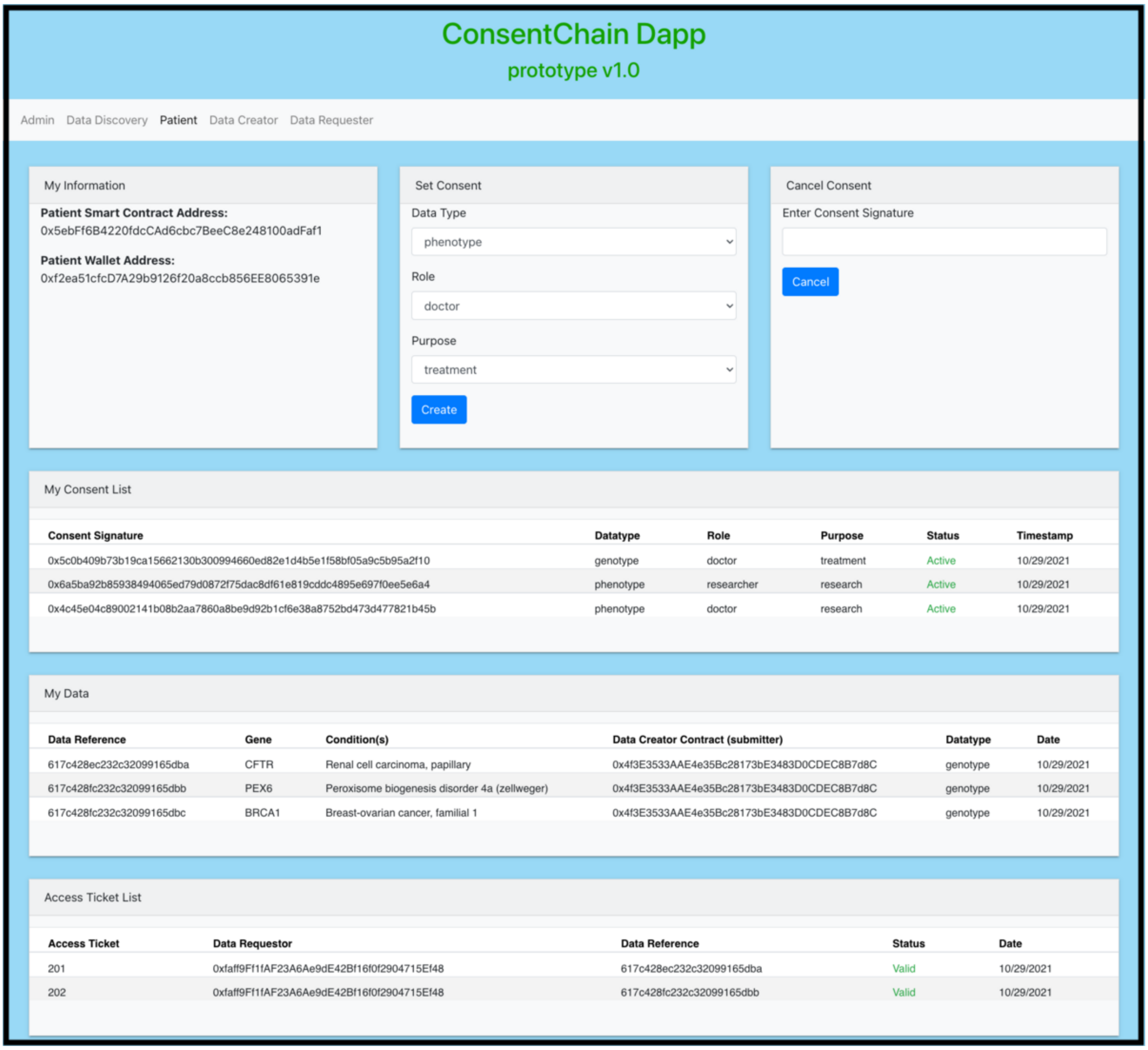
```
Algorithm 9: requestAccessToken
Input:caller, dataProfileId, ticketId
Output: tokenId
DATAPROFILE⟵mapping
If caller=contractOwner ∧ dataProfileId ≠ null ∧ ticketId ≠ null, then
d⟵DATAPROFILE.return(dataProfileId)
dataCreator⟵d.dataCreatorSmartContract
patient⟵d.patientSmartContract
if patient.ticket[ticketId].status=true then
token
⟵dataCreator.createAccessToken(caller, dataProfileId)
Token.status=true
Output token.id
Else
revert smart contract state and show an error message
Else
revert smart contract state and show an error message
```

A Proof-of-Concept (ConsentChain)

This section presents ConsentChain, a proof-of-concept implementation of the proposed architecture, to explore the efficacy of applying blockchain technology to support clinical genomic data sharing. The ConsentChain provides a web portal for patients, DCs, and DRs to interact with the system. It enables patients to provide or withdraw their consent regarding the sharing of their data and DCs to collect and store patient data and DRs to query and access patient data. Figure 8 shows the patient interface provided by the ConsentChain. The high-level structure and workflow of ConsentChain is shown in Figure 9, and the corresponding description of each step is as follows:

1. During registration, DR generates a pair of keys: a PU and a private key (PR). DR then uploads PU to the IPFS and records its location returned by the IPFS.
2. DR sends a blockchain transaction to store the PU's location returned by the IPFS in the RSC.
3. Patient sends a blockchain transaction to store their consent elements (data type, role, and purpose) in PSC.
4. DC collects patient's data and stores it in a secure, off-chain database. The DC also records patient's data reference (DRef) returned by the database.
5. DC creates a DP that includes DRef, a PSC address, and other information related to patient's data that do not reveal any sensitive and identifiable information. Then, the DC sends a blockchain transaction to store the DP in the DPSC.
6. DR queries DPSC to discover a specific DP of interest and reads transaction information related to that DP.
7. DR obtains the PSC address from the DP and sends a blockchain transaction to the PSC to request an ATi to access patient's data stored in the off-chain database. The request is accepted or rejected automatically, based on patient consent stored in the PSC. On acceptance, ATi is generated and stored in PSC, and DR receives the transaction ID related to ATi.
8. DR sends a blockchain transaction including ATi to DCSC to request an ATo to retrieve patient's data stored in the off-chain database. The request is accepted or rejected automatically based on ATi validation. On acceptance of the request, the ATo is stored in the DCSC, and DR receives the transaction ID related to the ATo.
9. DR sends a blockchain transaction including ATo to the oracle service smart contract to retrieve patient's data stored in the off-chain database. The request is accepted or rejected automatically based on the ATo validation.
10. On acceptance of the request, the request is forwarded to the Oracle Service Server (OSS).
11. OSS retrieves the DR's PU location on the IPFS from the RSC.
12. OSS downloads the PU of the DR from the IPFS.
13. OSS fetches patient's data from the database and creates a temporary JSON file that contains patient's data. This JSON file can be accessed via HTTPS requests and is available for one-time access.
14. The OSS encrypts the URL for a JSON file using the PU of the DR. Then, the OSS sends a blockchain transaction to store the encrypted URL in the DRSC.
15. DR retrieves encrypted URL from DRSC and decrypts it using the corresponding PR to access the JSON file.

Figure 8. Patient interface.



ConsentChain Dapp

prototype v1.0

My Information

Patient Smart Contract Address:
0x5ebFf6B4220fdcCAd6cbc7BeeC8e248100adFaf1

Patient Wallet Address:
0xf2ea51cfcD7A29b9126f20a8ccb856EE8065391e

Set Consent

Data Type: phenotype

Role: doctor

Purpose: treatment

Cancel Consent

Enter Consent Signature:

My Consent List

Consent Signature	Datatype	Role	Purpose	Status	Timestamp
0x5c0b409b73b19ca15662130b300994660ed82e1d4b5e1f58bf05a9c5b95a2f10	genotype	doctor	treatment	Active	10/29/2021
0x6a5ba92b85938494065ed79d0872f75dac8df61e819cddc4895e697f0ee5e6a4	phenotype	researcher	research	Active	10/29/2021
0x4c45e04c89002141b08b2aa7860a8be9d92b1cf6e38a8752bd473d477821b45b	phenotype	doctor	research	Active	10/29/2021

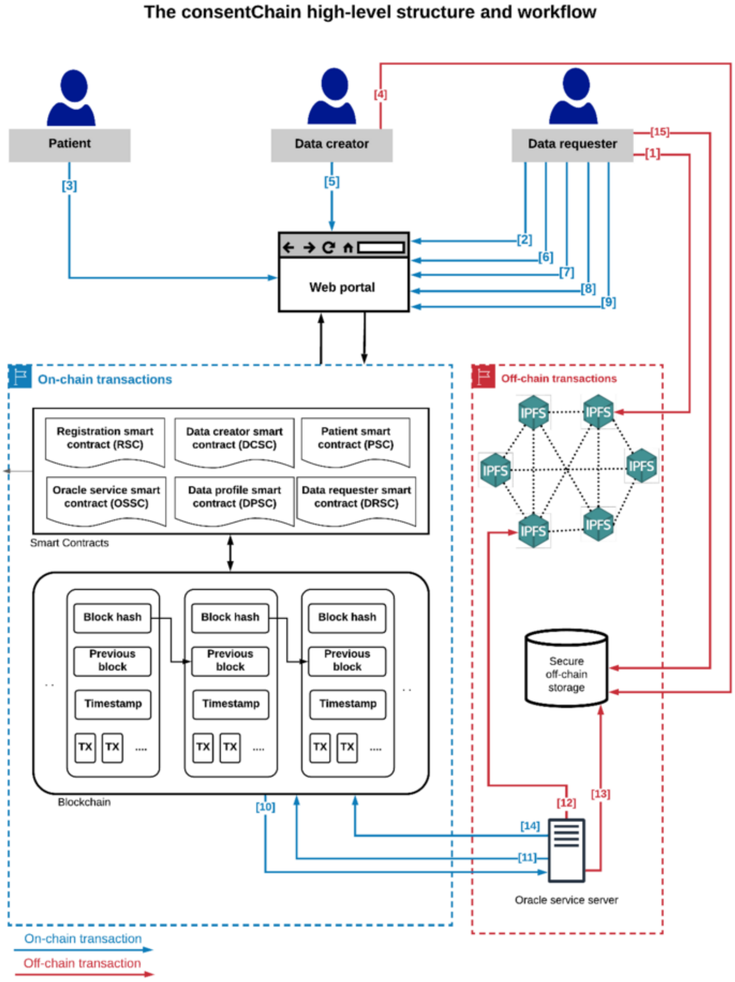
My Data

Data Reference	Gene	Condition(s)	Data Creator Contract (submitter)	Datatype	Date
617c428ec232c32099165dba	CFTR	Renal cell carcinoma, papillary	0x4f3E3533AAE4e35Bc28173bE3483D0CDEC8B7d8C	genotype	10/29/2021
617c428fc232c32099165dbb	PEX6	Peroxisome biogenesis disorder 4a (zellweger)	0x4f3E3533AAE4e35Bc28173bE3483D0CDEC8B7d8C	genotype	10/29/2021
617c428fc232c32099165dbc	BRCA1	Breast-ovarian cancer, familial 1	0x4f3E3533AAE4e35Bc28173bE3483D0CDEC8B7d8C	genotype	10/29/2021

Access Ticket List

Access Ticket	Data Requestor	Data Reference	Status	Date
201	0xfaff9Ff1fAF23A6Ae9dE42Bf16f0f2904715Ef48	617c428ec232c32099165dba	Valid	10/29/2021
202	0xfaff9Ff1fAF23A6Ae9dE42Bf16f0f2904715Ef48	617c428fc232c32099165dbb	Valid	10/29/2021

Figure 9. The high-level structure and workflow of ConsentChain. Ati: access ticket; DR: data requester; IPFS: InterPlanetary File System; OSS: Oracle Service Server; OSSC: oracle service smart contract; P: patient; PSC: patient smart contract; RSC: registration smart contract.



[1] DR generates a pair of keys: public key (PU) and private key (PR) and uploads PU to IPFS

[2] DR store the PU's location RSC

[3] P stores their consent elements in PSC

[4] DC collects P's data and stores it in a secure, off-chain database.

[5] DC creates a data profile (DP) that includes DRef, a PSC address related to P's and stores it in DPSC.

[6] DR queries DPSC for a specific DP of interest

[7] DR obtains PSC address from DP and sends to PSC to request an access ticket (ATi)

[8] DR sends ATi to DCSC to request an Access Token (ATo)

[9-10] DR sends ATo to OSSC to retrieve P's data stored in the off-chain database.

[11] OSS retrieves the DR's PU location on IPFS from RSC.

[12] OSS downloads DR's PU from IPFS.

[13] OSS fetches P's data from the database and creates a JSON file that contains P's data.

[14] OSS encrypts the URL for JSON file using DR's PU and stores the encrypted URL (eURL) in DRSC.

[15] DR retrieves eURL from DRSC and decrypts it using the corresponding PR in order to access the JSON file.

Discussion

Principal Findings

In this section, we discuss how our proof-of-concept, ConsentChain, meets the requirements captured from the patient forum, and we provide a detailed analysis of its performance.

Addressing Requirement

Requirement 1: Data Security

In ConsentChain, we used a hybrid data storage model that included on-chain or off-chain storage. Sensitive patient data are stored securely off-chain, whereas metadata for patient data are stored on-chain along with a reference pointer to the data source. This reference pointer is constrained by a short time frame and is encrypted. Only an authorized DR can decrypt it within the given time frame to access patient data. Moreover, implementing ConsentChain on a private or consortium blockchain adds a security layer in which all users are verified before joining the network.

Requirement 2: User Control Over Data

Smart contracts act as autonomous actors whose behavior is predictable [74]. However, because of blockchain immutability, once a smart contract is deployed, it cannot be modified; hence, bugs and security vulnerabilities found in the deployed smart contract are difficult to resolve. Therefore, smart contract security audits and testing are essential for developing smart contracts to minimize the risk of mismatches between a smart contract intended behavior and the actual behavior [75]. Using a smart contract to manage consent would enable patients to dynamically grant and revoke access to their data. In ConsentChain, patients record consent preferences in their smart contract, and they can amend or delete these preferences at any time. These changes were reflected in the system in real time.

Requirement 3: Data Privacy

By leveraging blockchain authenticity and verifiability features, ConsentChain maintains privacy by using permissioned blockchain and anonymized accounts. Only authorized users can access the blockchain via their anonymized accounts, enabling patients to provide their consent without revealing their real identities.

Requirements 4 and 5: Data Discovery and Minimum Data Disclosure

In the health care context, balancing the maximization of data discovery while minimizing data disclosure risk is a challenging task [76-78]. Inspired by the one-time password scheme, we proposed a one-time-access-token mechanism to minimize the data disclosure risk in ConsentChain. In this mechanism, an ATo is automatically generated for an authorized access request. The token is valid for one-time use, and it contains an encrypted reference pointer to the data source along with a digital signature on the shared data to ensure data integrity against tampering. Only an authorized DR can decrypt the reference pointer to access the data within a given time frame. If the DR needs to access data in the future, the generation of a new ATo is required. Through the implementation of a one-time access-based token and public-key cryptography, a compromised reference pointer to patient data will not lead to data leakage. This is because of the limited access and time restrictions given to access patient data, further increasing the security of ConsentChain and decreasing the likelihood of data leakage.

To maximize data discovery, we leveraged the blockchain features. One of these is the replication of data stored on-chain across the network; a consensus mechanism ensures that each node obtains a local identical copy of the data. Using their local copy of the on-chain data, a DR can identify potential patient data instead of individually querying each off-chain storage. Therefore, storing patients' metadata on the chain would provide DRs with a broader vision of similar patient data, which are stored off-chain across different laboratories.

Requirement 6: Traceability

By leveraging the blockchain's immutability, our system maintains an immutable log of all system transactions. As the process of sharing patient data is managed by smart contracts, all involved transactions are recorded permanently on the blockchain. This would enable patients to inspect the blockchain for all information and transactions related to their data, including where data are stored off-chain and who have access to them and for what purpose. This feature is a significant upgrade toward patient-centric approaches to advance data sharing. It would also enable regulators to investigate claims in the event of disputes among involved parties, thereby increasing confidence in ConsentChain.

Security Analysis

This section provides a security analysis of ConsentChain in terms of patient privacy preservation, data storage, data sharing, and tamper-proofing.

Patient Privacy Preservation

Genomic data are highly sensitive and should not be disclosed without proper permission. In ConsentChain, genomic data are stored in an off-chain private secure storage with an access control mechanism, thereby reducing the risk of patient data leakage. Moreover, to ensure participant anonymity, a randomly generated unique account was generated for the participants who were associated with a PU. This account is used to send transactions to the blockchain; these transactions are anonymous and cannot be linked to the real identity of participants. In addition, multiple accounts can be created for one participant; hence, transactions sent to the blockchain by the same participant cannot be inferred by an adversary.

Data Storage

In ConsentChain, genomic data are stored in an off-chain private secure storage system. The security of this storage is beyond the scope of this paper, and we assume that it is secured by its owner (the DC). Only the metadata, hash, and reference of the off-chain stored data are shared on the blockchain. The off-chain DRef stored in the blockchain is tamper-proof.

Data Sharing

Only authorized users can request access to off-chain data through permissions that are preset in smart contracts. After receiving a valid request, the DC creates a JSON file that contains the requested data and stores it in the temporary access off-chain storage from where it can be accessed via HTTPS. Access to the JSON file is restricted by a one-time visit and a short time frame. The DC then retrieves the PU of the user who requested the data from the IPFS and encrypts the URL that allows access to the JSON file and then stores it in the blockchain. The user requesting the data can then obtain the URL from the blockchain and decrypt it using their PR and access the JSON file. Once the JSON file is accessed, it is removed from the temporary access off-chain storage, making the URL stored in the blockchain useless; therefore, if the adversary compromises the PR of the user requesting the data to decrypt the URL, the URL would lead to nothing. Further, if the JSON file is not accessed within the specified time frame, it is removed from the temporary access off-chain storage, reducing the risk of unauthorized access to the data.

Tamper-Proofing

In ConsentChain, data access activities are recorded in the blockchain and can be audited and tracked. In addition, the data stored in the blockchain are immutable and cannot be arbitrarily modified owing to the consensus mechanisms used in the blockchain, which guarantees that the added blocks cannot be modified unless an adversary can launch a 51% attack. It is worth noting that the mechanism of launching a 51% attack differs depending on the type of consensus mechanism used in the blockchain. For instance, public blockchains such as Ethereum and Bitcoin use the proof-of-work consensus mechanism, which requires high computational power to generate new blocks, whereas in a private permissioned blockchain, the proof-of-authority consensus mechanism can be used to generate new blocks [79-82]. To launch a 51% attack on a blockchain that uses the proof-of-work consensus mechanism, an adversary needs to obtain 51% of the network's computational power. In contrast, when the proof-of-authority consensus mechanism is used, a 51% attack can only be launched by controlling over 51% of the network nodes, which is much more difficult than obtaining 51% of the network computational power [80]. Therefore, in ConsentChain, the proof-of-authority consensus mechanism is used to reduce the risk of a 51% attack.

Performance Evaluation

To test and validate ConsentChain, we built a real production environment for the deployment and hosting of ConsentChain. A detailed performance analysis of ConsentChain is provided in Multimedia Appendix 2. In summary, the analysis of the performance of the *Transaction* and *Read* operations of ConsentChain indicated an average *Transaction Throughput* of 13.59 tps and an average *Read Throughput* of 135.78 tps. The *Transaction Latency* was 2.76 seconds, whereas the average *Read Latency* was 0.288 seconds. In addition, the system performance analysis shows that a large number of read operations (reading a state from blockchain), that is, 10,000 transactions, could be handled by the system at very low latency, whereas transaction operations are processed with higher latency owing to the complexity involved (reading or writing a state from or to blockchain).

Conclusions

Genomic data are useful when shared within the clinical genomics community and compared with other patient data, indicating that clinicians might need to share data to efficiently treat patients. However, many challenges hinder large-scale genomic data sharing, such as the availability, discoverability, and accessibility of genomic data [8,51,52], preventing clinicians and researchers from generating an integrated view of rare genetic diseases. In this study, we proposed a blockchain-based dynamic consent architecture to support genomic data sharing and implemented a proof-of-concept for the architecture. We also developed an ontology model to represent patient consent elements into machine-readable codes to automate the consent and data access processes. The proof-of-concept has been implemented on a private Ethereum blockchain, and it shows that the proposed architecture can achieve a large-scale sharing of genomic data among the parties involved. The evaluation showed that patients achieved greater control over their data using this system. Performance analysis showed that the system was efficient and scalable.

Nonetheless, several limitations of this study need to be addressed. Owing to the openness and distributed nature of blockchain technology, verifying user identity is challenging. Our system operates under the assumption that the system is implemented on a private blockchain, and all users are invited to join the system. User identity verification is performed before one can join the system, and each user is given a pseudonymous identifier to represent them on the system. A more reliable and practical solution to overcome this issue might be linking patient identity with an external trusted source of information, such as GOV.UK Verify and NHS Identity. In addition, DR and DC identity verification could be achieved by linking to their professional registration.

Another issue is blockchain’s GDPR compliance, which needs to be considered [83-85]. Although blockchains can help dynamic consent systems comply with some GDPR objectives, including the rights to be informed and to withdraw, blockchains’ immutability seems to conflict with the GDPR, which encourages data minimization and gives data owners the right to erasure. A study conducted by the European Parliamentary Research Service concluded that although private and permissioned blockchains could easily comply with GDPR requirements, it is difficult to determine whether blockchains are, as a whole, either completely compliant or incompliant with GDPR [86]. However, since the GDPR came into effect, several studies have taken initial steps toward designing and building GDPR-compliant blockchain-based use cases [44,87-91]. Therefore, GDPR compliance should be considered during the design of blockchain-based systems [92,93].

The objective of this work was not to design a system that could be used in practice in health care environments, but to show that blockchain technology has the potential to address several genomic data sharing challenges. We found that facilitating genomic data sharing through blockchain technology and smart contracts is promising. However, they are not the complete answer, and a number of issues still need to be addressed before such systems can be deployed in practice, particularly in relation to verifying user credentials.

Conflicts of Interest

None declared.

Multimedia Appendix 1

System design requirements in existing blockchain solutions in health care.

[DOCX File , 19 KB - medinform_v9i11e27816_app1.docx]

Multimedia Appendix 2

Detailed performance analysis of the proposed model.

[DOCX File , 268 KB - medinform_v9i11e27816_app2.docx]

References

1. Borry P, Bentzen HB, Budin-Ljøsne I, Cornel MC, Howard HC, Feeney O, et al. The challenges of the expanded availability of genomic information: an agenda-setting paper. J Community Genet 2018 Apr;9(2):103-116 [FREE Full text] [doi: 10.1007/s12687-017-0331-7] [Medline: 28952070]
2. Agarwala V, Khozin S, Singal G, O'Connell C, Kuk D, Li G, et al. Real-world evidence in support of precision medicine: clinico-genomic cancer data as a case study. Health Aff (Millwood) 2018 May;37(5):765-772. [doi: 10.1377/hlthaff.2017.1579] [Medline: 29733723]
3. Shabani M, Borry P. Challenges of web-based personal genomic data sharing. Life Sci Soc Policy 2015;11:3 [FREE Full text] [doi: 10.1186/s40504-014-0022-7] [Medline: 26085313]

4. Rehm HL. Evolving health care through personal genomics. Nat Rev Genet 2017 Apr;18(4):259-267 [FREE Full text] [doi: 10.1038/nrg.2016.162] [Medline: 28138143]
5. Wang S, Jiang X, Singh S, Marmor R, Bonomi L, Fox D, et al. Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. Ann N Y Acad Sci 2017 Jan;1387(1):73-83 [FREE Full text] [doi: 10.1111/nyas.13259] [Medline: 27681358]
6. Tabor HK, Berkman BE, Hull SC, Bamshad MJ. Genomics really gets personal: how exome and whole genome sequencing challenge the ethical framework of human genetics research. Am J Med Genet A 2011 Dec;155A(12):2916-2924 [FREE Full text] [doi: 10.1002/ajmg.a.34357] [Medline: 22038764]
7. Kaye J. The tension between data sharing and the protection of privacy in genomics research. Annu Rev Genomics Hum Genet 2012;13:415-431 [FREE Full text] [doi: 10.1146/annurev-genom-082410-101454] [Medline: 22404490]
8. van Schaik TA, Kovalevskaya NV, Protopapas E, Wahid H, Nielsen FG. The need to redefine genomic data sharing: a focus on data accessibility. Appl Transl Genom 2014 Sep 28;3(4):100-104 [FREE Full text] [doi: 10.1016/j.atg.2014.09.013] [Medline: 27294022]
9. Riggs ER, Azzariti DR, Niehaus A, Goehringer SR, Ramos EM, Rodriguez LL, Clinical Genome Resource Education Working Group. Development of a consent resource for genomic data sharing in the clinical setting. Genet Med 2019 Jan;21(1):81-88 [FREE Full text] [doi: 10.1038/s41436-018-0017-5] [Medline: 29899502]
10. Dheensa S, Carrieri D, Kelly S, Clarke A, Doheny S, Turnpenny P, et al. A 'joint venture' model of recontacting in clinical genomics: challenges for responsible implementation. Eur J Med Genet 2017 Jul;60(7):403-409 [FREE Full text] [doi: 10.1016/j.ejmg.2017.05.001] [Medline: 28501562]
11. Carrieri D, Dheensa S, Doheny S, Clarke AJ, Turnpenny PD, Lucassen AM, et al. Recontacting in clinical practice: an investigation of the views of healthcare professionals and clinical scientists in the United Kingdom. Eur J Hum Genet 2017 Feb;25(3):275-279 [FREE Full text] [doi: 10.1038/ejhg.2016.188] [Medline: 28051074]
12. Lawler M, Siu LL, Rehm HL, Chanock SJ, Alterovitz G, Burn J, Clinical Working Group of the Global Alliance for GenomicsHealth (GA4GH). All the world's a stage: facilitating discovery science and improved cancer care through the global alliance for genomics and health. Cancer Discov 2015 Nov;5(11):1133-1136 [FREE Full text] [doi: 10.1158/2159-8290.CD-15-0821] [Medline: 26526696]
13. Siu LL, Lawler M, Haussler D, Knoppers BM, Lewin J, Vis DJ, et al. Facilitating a culture of responsible and effective sharing of cancer genome data. Nat Med 2016 May 05;22(5):464-471 [FREE Full text] [doi: 10.1038/nm.4089] [Medline: 27149219]
14. Vis DJ, Lewin J, Liao RG, Mao M, Andre F, Ward RL, Clinical Working Group of the Global Alliance for GenomicsHealth. Towards a global cancer knowledge network: dissecting the current international cancer genomic sequencing landscape. Ann Oncol 2017 May 01;28(5):1145-1151 [FREE Full text] [doi: 10.1093/annonc/mdx037] [Medline: 28453708]
15. McDonald SA, Mardis ER, Ota D, Watson MA, Pfeifer JD, Green JM. Comprehensive genomic studies: emerging regulatory, strategic, and quality assurance challenges for biorepositories. Am J Clin Pathol 2012 Jul;138(1):31-41 [FREE Full text] [doi: 10.1309/AJCPXBA69LNSCVMH] [Medline: 22706855]
16. Chaterji S, Koo J, Li N, Meyer F, Grama A, Bagchi S. Federation in genomics pipelines: techniques and challenges. Brief Bioinform 2019 Jan 18;20(1):235-244 [FREE Full text] [doi: 10.1093/bib/bbx102] [Medline: 28968781]
17. Thorisson G, Muilu J, Brookes A. Genotype-phenotype databases: challenges and solutions for the post-genomic era. Nat Rev Genet 2009 Jan;10(1):9-18. [doi: 10.1038/nrg2483] [Medline: 19065136]
18. Acmg Board Of Directors. Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics and Genomics. Genet Med 2017 Jul;19(7):721-722. [doi: 10.1038/gim.2016.196] [Medline: 28055021]
19. Global Alliance for GenomicsHealth. GENOMICS. A federated ecosystem for sharing genomic, clinical data. Science 2016 Jun 10;352(6291):1278-1280. [doi: 10.1126/science.aaf6162] [Medline: 27284183]
20. Weber G, Murphy S, McMurry A, Macfadden D, Nigrin D, Churchill S, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc 2009;16(5):624-630 [FREE Full text] [doi: 10.1197/jamia.M3191] [Medline: 19567788]
21. Grishin D, Obbad K, Estep P, Quinn K, Zaranek SW, Zaranek AW, et al. Accelerating genomic data generation and facilitating genomic data access using decentralization, privacy-preserving technologies and equitable compensation. Blockchain Healthc Today 2018 Dec 19;1:1-23. [doi: 10.30953/bhty.v1.34]
22. Dyke SO, Philippakis AA, Rambla De Argila J, Paltoo DN, Luetkemeier ES, Knoppers BM, et al. Consent codes: upholding standard data use conditions. PLoS Genet 2016 Jan 21;12(1):e1005772 [FREE Full text] [doi: 10.1371/journal.pgen.1005772] [Medline: 26796797]
23. Shabani M. Blockchain-based platforms for genomic data sharing: a de-centralized approach in response to the governance problems? J Am Med Inform Assoc 2019 Jan 01;26(1):76-80 [FREE Full text] [doi: 10.1093/jamia/ocy149] [Medline: 30496430]
24. Ozercan HI, Ileri AM, Ayday E, Alkan C. Realizing the potential of blockchain technologies in genomics. Genome Res 2018 Sep;28(9):1255-1263 [FREE Full text] [doi: 10.1101/gr.207464.116] [Medline: 30076130]
25. Consent-chain-project. Mendeley Data. 2021. URL: https://data.mendeley.com/datasets/vwy3hj5h8n/1 [accessed 2021-09-17]

26. Bashir I. Mastering Blockchain. Birmingham: Packt Publishing; 2017.
27. Bitcoin: a peer-to-peer electronic cash system. bitcoin.org. URL: https://bitcoin.org/bitcoin.pdf? [accessed 2021-01-15]
28. Buterin V. Ethereum white paper. etherium.org. URL: https://blockchainlab.com/pdf/Ethereum_white_paper-a_next_generation_smart_contract_and_decentralized_application_platform-vitalik-buterin.pdf [accessed 2021-01-14]
29. Androulaki E, Barger A, Bortnikov V, Muralidharan S, Cachin C, Christidis K, et al. Hyperledger fabric: a distributed operating system for permissioned blockchains. In: Proceedings of the Thirteenth EuroSys Conference. 2018 Presented at: Proceedings of the Thirteenth EuroSys Conference; Apr 23-26, 2018; Porto Portugal. [doi: 10.1145/3190508.3190538]
30. Szabo N. Formalizing and securing relationships on public networks. First Monday 1997 Sep 1;2(9):-. [doi: 10.5210/fm.v2i9.548]
31. Cannarsa M. Interpretation of contracts and smart contracts: smart interpretation or interpretation of smart contracts? Eur Rev Priv Law 2018;26(6):773-785.
32. Delmolino K, Arnett M, Kosba A, Miller A, Shi E. Step by step towards creating a safe smart contract: lessons and insights from a cryptocurrency lab. IACR. 2015. URL: https://eprint.iacr.org/2015/460.pdf [accessed 2021-04-01]
33. Schollmeier R. A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications. In: Proceedings First International Conference on Peer-to-Peer Computing. 2001 Presented at: First International Conference on Peer-to-Peer Computing; Aug 27-29, 2001; Sweden. [doi: 10.1109/p2p.2001.990434]
34. Sayeed S, Marco-Gisbert H. Assessing blockchain consensus and security mechanisms against the 51% attack. Appl Sci 2019 Apr 29;9(9):1788. [doi: 10.3390/app9091788]
35. Oliveira M, Carrara G, Fernandes N, Albuquerque C, Carrano R, Medeiros DV. Towards a performance evaluation of private blockchain frameworks using a realistic workload. In: Proceedings of the 2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN). 2019 Presented at: 2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN); Feb 19-21, 2019; Paris, France. [doi: 10.1109/icin.2019.8685888]
36. Schwartz D, Youngs N, Britto A. The Ripple protocol consensus algorithm. Ripple Consensus. 2018. URL: http://www.naation.com/ripple-consensus-whitepaper.pdf [accessed 2021-05-05]
37. EOS.IO technical white paper. steemit. URL: https://steemit.com/eos/@eosio/eos-io-technical-white-paper [accessed 2021-05-05]
38. Brock A, Braden D, Day J. Holochain—a framework for distributed applications. Google Patents. 2021. URL: https://patents.google.com/patent/US20200389521A1/en [accessed 2021-05-05]
39. Hyperledger Fabric. Hyperledger. URL: https://www.hyperledger.org/projects/fabric [accessed 2021-05-05]
40. Dawson R, Baxter M. Announcing Hyperledger Besu. Hyperledger. URL: https://www.hyperledger.org/blog/2019/08/29/announcing-hyperledger-besu [accessed 2021-05-05]
41. Dankar FK, Gergely M, Malin B, Badji R, Dankar SK, Shuaib K. Dynamic-informed consent: a potential solution for ethical dilemmas in population sequencing initiatives. Comput Struct Biotechnol J 2020 Apr 2;18:913-921 [FREE Full text] [doi: 10.1016/j.csbj.2020.03.027] [Medline: 32346464]
42. Karlson E, Boutin N, Hoffnagle A, Allen N. Building the partners healthcare biobank at partners personalized medicine: informed consent, return of research results, recruitment lessons and operational considerations. J Pers Med 2016 Jan 14;6(1):2 [FREE Full text] [doi: 10.3390/jpm6010002] [Medline: 26784234]
43. Chen C, Lee P, Pain KJ, Delgado D, Cole CL, Campion TR. Replacing paper informed consent with electronic informed consent for research in academic medical centers: a scoping review. AMIA Jt Summits Transl Sci Proc 2020 May 30;2020:80-88 [FREE Full text] [Medline: 32477626]
44. Mamo N, Martin GM, Desira M, Ellul B, Ebejer JP. Dwarna: a blockchain solution for dynamic consent in biobanking. Eur J Hum Genet 2020 May;28(5):609-626 [FREE Full text] [doi: 10.1038/s41431-019-0560-9] [Medline: 31844175]
45. Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melham K. Dynamic consent: a patient interface for twenty-first century research networks. Eur J Hum Genet 2015 Feb;23(2):141-146 [FREE Full text] [doi: 10.1038/ejhg.2014.71] [Medline: 24801761]
46. Andrews SM, Raspa M, Edwards A, Moultrie R, Turner-Brown L, Wagner L, et al. "Just tell me what's going on": the views of parents of children with genetic conditions regarding the research use of their child's electronic health record. J Am Med Inform Assoc 2020 Mar 01;27(3):429-436 [FREE Full text] [doi: 10.1093/jamia/ocz208] [Medline: 31913479]
47. Samuel GN, Dheensa S, Farsides B, Fenwick A, Lucassen A. Healthcare professionals' and patients' perspectives on consent to clinical genetic testing: moving towards a more relational approach. BMC Med Ethics 2017 Aug 08;18(1):47 [FREE Full text] [doi: 10.1186/s12910-017-0207-8] [Medline: 28789658]
48. Wust K, Gervais A. Do you need a Blockchain? In: Proceedings of the 2018 Crypto Valley Conference on Blockchain Technology (CVCBT). 2018 Presented at: 2018 Crypto Valley Conference on Blockchain Technology (CVCBT); Jun 20-22, 2018; Zug, Switzerland. [doi: 10.1109/cvcbt.2018.00011]
49. Koens T, Poll E. What blockchain alternative do you need? In: Data Privacy Management, Cryptocurrencies and Blockchain Technology. Cham: Springer; 2018.
50. Yaga D, Mell P, Roby N, Scarfone K. Blockchain technology overview. National institute of Standards and Technology. 2018. URL: https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8202.pdf [accessed 2021-05-05]

51. Data sharing to support UK clinical genetics and genomics services. PHG Foundation. 2015. URL: https://www.phgfoundation.org/media/79/download/Data%20sharing%20to%20support%20UK%20clinical%20genetics%20and%20genomics%20services.pdf?v=1&inline=1 [accessed 2021-05-25]
52. MacArthur DG. Challenges in clinical genomics. Genome Med 2012 May 12;4:43. [doi: 10.1186/gm342]
53. Genome UK: the future of healthcare. gov.uk. 2020. URL: https://www.gov.uk/government/publications/genome-uk-the-future-of-healthcare [accessed 2021-05-25]
54. Alice G, Briggs B. NHS patient data breached 1395 times in the last two years. The Ferret. 2021. URL: https://theferret.scot/nhs-patient-data-breached-1395-times-in-two-years/ [accessed 2021-05-29]
55. Martin G, Ghafur S, Kinross J, Hankin C, Darzi A. WannaCry-a year on. BMJ 2018 Jun 04;361:k2381. [doi: 10.1136/bmj.k2381] [Medline: 29866711]
56. Parker L. Using human tissue: when do we need consent? J Med Ethics 2011 Dec;37(12):759-761. [doi: 10.1136/medethics-2011-100043] [Medline: 21873308]
57. Cardinal RN. Clinical records anonymisation and text extraction (CRATE): an open-source software system. BMC Med Inform Decis Mak 2017 Apr 26;17(1):50 [FREE Full text] [doi: 10.1186/s12911-017-0437-1] [Medline: 28441940]
58. Brown I, Brown L, Korff D. Using NHS patient data for research without consent. Law Innov Technol 2015 May 07;2(2):219-258. [doi: 10.5235/175799610794046186]
59. 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] [Medline: 32664786]
60. Hebig R, Bendraou R. On the need to study the impact of model driven engineering on software processes. In: Proceedings of the 2014 International Conference on Software and System Process. 2014 Presented at: 2014 International Conference on Software and System Process; May 26-28, 2014; Nanjing China. [doi: 10.1145/2600821.2600846]
61. Woolley JP, Kirby E, Leslie J, Jeanson F, Cabili MN, Rushton G, et al. Responsible sharing of biomedical data and biospecimens via the "Automatable Discovery and Access Matrix" (ADA-M). NPJ Genom Med 2018 Jul 23;3:17 [FREE Full text] [doi: 10.1038/s41525-018-0057-4] [Medline: 30062047]
62. Dyke SOM, Philippakis AA, Rambla De Argila J, Paltoo DN, Luetkemeier ES, Knoppers BM, et al. Consent Codes: Upholding Standard Data Use Conditions. PLoS Genet 2016 Jan 21;12(1):e1005772. [doi: 10.1371/journal.pgen.1005772]
63. Chenthara S, Ahmed K, Wang H, Whittaker F, Chen Z. Healthchain: a novel framework on privacy preservation of electronic health records using blockchain technology. PLoS One 2020 Dec 9;15(12):e0243043 [FREE Full text] [doi: 10.1371/journal.pone.0243043] [Medline: 33296379]
64. Azaria A, Ekblaw A, Vieira T, Lippman A. MedRec: Using blockchain for medical data access and permission management. In: Proceedings of the 2016 2nd International Conference on Open and Big Data (OBD). 2016 Presented at: 2016 2nd International Conference on Open and Big Data (OBD); Aug 22-24, 2016; Vienna, Austria. [doi: 10.1109/obd.2016.11]
65. Cyran M. Blockchain as a foundation for sharing healthcare data. Blockchain Healthc Today 2018 Mar 23;1:1-6 [FREE Full text] [doi: 10.30953/bhty.v1.13]
66. Choudhury O, Sarker H, Rudolph N, Foreman M, Fay N, Dhuliawala M, et al. Enforcing human subject regulations using blockchain and smart contracts. Blockchain Healthc Today 2018 Mar 23;1:1-14 [FREE Full text] [doi: 10.30953/bhty.v1.10]
67. Mamo N, Martin GM, Desira M, Ellul B, Ebejer JP. Dwarna: a blockchain solution for dynamic consent in biobanking. Eur J Hum Genet 2020 May;28(5):609-626 [FREE Full text] [doi: 10.1038/s41431-019-0560-9] [Medline: 31844175]
68. Tith D, Lee J, Suzuki H, Wijesundara WM, Taira N, Obi T, et al. Patient consent management by a purpose-based consent model for electronic health record based on blockchain technology. Healthc Inform Res 2020 Oct;26(4):265-273 [FREE Full text] [doi: 10.4258/hir.2020.26.4.265] [Medline: 33190460]
69. Dubovitskaya A, Baig F, Xu Z, Shukla R, Zambani PS, Swaminathan A, et al. ACTION-EHR: patient-centric blockchain-based electronic health record data management for cancer care. J Med Internet Res 2020 Aug 21;22(8):e13598 [FREE Full text] [doi: 10.2196/13598] [Medline: 32821064]
70. Dubovitskaya A, Xu Z, Ryu S, Schumacher M, Wang F. Secure and trustable electronic medical records sharing using blockchain. arXiv.org. 2017. URL: http://arxiv.org/abs/1709.06528 [accessed 2021-05-27]
71. Rajput AR, Li Q, Ahvanooey MT. A blockchain-based secret-data sharing framework for personal health records in emergency condition. Healthcare (Basel) 2021 Feb 14;9(2):206 [FREE Full text] [doi: 10.3390/healthcare9020206] [Medline: 33672991]
72. Zhuang Y, Chen Y, Shae Z, Shyu C. Generalizable layered blockchain architecture for health care applications: development, case studies, and evaluation. J Med Internet Res 2020 Jul 27;22(7):e19029 [FREE Full text] [doi: 10.2196/19029] [Medline: 32716300]
73. The ProvableTM blockchain oracle for modern DApps. Provable. URL: https://provable.xyz/ [accessed 2021-01-14]
74. Christidis K, Devetsikiotis M. Blockchains and smart contracts for the internet of things. IEEE Access 2016 May 10;4:2292-2303. [doi: 10.1109/ACCESS.2016.2566339]
75. Atzei N, Bartoletti M, Cimoli T. A survey of attacks on Ethereum Smart Contracts (SoK). In: Principles of Security and Trust. Berlin, Heidelberg: Springer; 2017.

76. Oliver JM, Slashinski MJ, Wang T, Kelly PA, Hilsenbeck SG, McGuire AL. Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. Public Health Genomics 2012;15(2):106-114 [FREE Full text] [doi: 10.1159/000334718] [Medline: 22213783]
77. Slavkovic A, Yu F. O privacy, where art thou?: genomics and privacy. CHANCE 2015 Apr 27;28(2):37-39. [doi: 10.1080/09332480.2015.1042736]
78. Bacchus A. Towards secure and privacy preserving e-health data exchanges through consent based access control internet. ProQuest. 2017. URL: https://www.proquest.com/openview/4c24433193f4293fca2bcdfccda1cef5/1?pq-origsite=gscholar&cbl=18750 [accessed 2021-05-17]
79. Cash M, Bassiouni M. Two-tier permission-ed and permission-less blockchain for secure data sharing. In: Proceedings of the 2018 IEEE International Conference on Smart Cloud (SmartCloud). 2018 Presented at: 2018 IEEE International Conference on Smart Cloud (SmartCloud); Sep 21-23, 2018; New York, NY, USA. [doi: 10.1109/smartcloud.2018.00031]
80. Proof-of-Authority consensus. GitHub. URL: https://apla.readthedocs.io/en/latest/concepts/consensus.html#attack [accessed 2021-06-01]
81. Angelis SD, Aniello L, Baldoni R, Lombardi F, Margheri A, Sassone V. PBFT vs proof-of-authority: applying the CAP theorem to permissioned blockchain. In: Proceedings of the Italian Conference on Cybersecurity. 2017 Presented at: Italian Conference on Cybersecurity; Feb 6, 2018; Milan, Italy.
82. Proof of authority. GitHub. URL: https://github.com/openethereum/parity-ethereum [accessed 2021-06-01]
83. Van Humbeeck A. The blockchain-GDPR paradox. J Data Prot Priv 2019;2(3):208-212 [FREE Full text]
84. Berberich M, Steiner M. Practitioner's corner · blockchain technology and the GDPR – how to reconcile privacy and distributed ledgers? Eur Data Prot Law Rev 2016;2(3):422-426. [doi: 10.21552/edpl/2016/3/21]
85. Blockchain and GDPR: how blockchain could address five areas associated with GDPR compliance. IBM Security. 2018. URL: https://iapp.org/media/pdf/resource_center/blockchain_and_gdpr.pdf [accessed 2021-05-21]
86. Finck M. Blockchains and the General Data Protection Regulation. Brussels: European Union; 2019.
87. Zheng X, Mukkamala R, Vatrapu R, Ordieres-Mere J. Blockchain-based personal health data sharing system using cloud storage. In: Proceedings of the 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom). 2018 Presented at: 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom); Sep 17-20, 2018; Ostrava, Czech Republic. [doi: 10.1109/healthcom.2018.8531125]
88. Rantos K, Drosatos G, Kritsas A, Ilioudis C, Papanikolaou A, Filippidis AP. A blockchain-based platform for consent management of personal data processing in the IoT ecosystem. Sec Commun Netw 2019;2019:1-15. [doi: 10.1155/2019/1431578]
89. Wirth C, Kolain M. Privacy by BlockChain design: a blockchain-enabled GDPR-compliant approach for handling personal data. In: Proceedings of 1st ERCIM Blockchain Workshop 2018. 2018 Presented at: Proceedings of 1st ERCIM Blockchain Workshop 2018; May 8-9, 2018; Amsterdam, Netherlands. [doi: 10.18420/blockchain2018_03]
90. Camilo J. Blockchain-based consent manager for GDPR compliance. Open Identity Summit. 2019. URL: https://dl.gi.de/bitstream/handle/20.500.12116/20985/proceedings-14.pdf?isAllowed=y&sequence=1 [accessed 2021-05-27]
91. Farshid S, Reitz A, Roßbach P. Design of a forgetting blockchain: a possible way to accomplish GDPR compatibility. In: Proceedings of the 52nd Hawaii International Conference on System Sciences. 2019 Presented at: 52nd Hawaii International Conference on System Sciences; Jan 8-11, 2019; Maui, Hawaii, USA. [doi: 10.24251/hicss.2019.850]
92. Eichler N, Jongerius S, McMullen G, Naegele O, Steininger L, Wagner K. Blockchain, data protection, and the GDPR. Blockchain Bundesverband. 2021. URL: https://www.crowdfundinsider.com/wp-content/uploads/2018/06/GDPR_Position_Paper_v1.0.pdf [accessed 2021-05-21]
93. Rose A. GDPR challenges for blockchain technology. Interact Enterain Law Rev 2019 Jun;2(1):35-41. [doi: 10.4337/ielr.2019.01.03]

Abbreviations

ATi: access ticket
ATo: access token
DC: data creator
DCSC: data creator smart contract
DP: data profile
DPSC: data profile smart contract
DR: data requester
DRef: data reference
DRSC: data requester smart contract
GDPR: General Data Protection Regulation
IPFS: InterPlanetary File System
NHS: National Health Service
OSS: Oracle Service Server

PR: private key
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSC: patient smart contract
PU: public key

Edited by C Lovis; submitted 08.02.21; peer-reviewed by M Platt, M Doerr; comments to author 21.03.21; revised version received 15.06.21; accepted 25.07.21; published 03.11.21.

Please cite as:
Albalwy F, Brass A, Davies A
A Blockchain-Based Dynamic Consent Architecture to Support Clinical Genomic Data Sharing (ConsentChain): Proof-of-Concept Study
JMIR Med Inform 2021;9(11):e27816
URL: https://medinform.jmir.org/2021/11/e27816
doi:10.2196/27816
PMID:34730538

Original Paper

An Open-Source, Standard-Compliant, and Mobile Electronic Data Capture System for Medical Research (OpenEDC): Design and Evaluation Study

Leonard Greulich[1], MSc; Stefan Hegselmann[1], MSc; Martin Dugas[2], Prof Dr

[1]Institute of Medical Informatics, University of Münster, Münster, Germany

[2]Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany

Corresponding Author:
Leonard Greulich, MSc
Institute of Medical Informatics
University of Münster
Albert-Schweitzer-Campus 1, Building A11
Münster, 48149
Germany
Phone: 49 15905368729
Email: leonard.greulich@uni-muenster.de

Abstract

Background: Medical research and machine learning for health care depend on high-quality data. Electronic data capture (EDC) systems have been widely adopted for metadata-driven digital data collection. However, many systems use proprietary and incompatible formats that inhibit clinical data exchange and metadata reuse. In addition, the configuration and financial requirements of typical EDC systems frequently prevent small-scale studies from benefiting from their inherent advantages.

Objective: The aim of this study is to develop and publish an open-source EDC system that addresses these issues. We aim to plan a system that is applicable to a wide range of research projects.

Methods: We conducted a literature-based requirements analysis to identify the academic and regulatory demands for digital data collection. After designing and implementing OpenEDC, we performed a usability evaluation to obtain feedback from users.

Results: We identified 20 frequently stated requirements for EDC. According to the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 25010 norm, we categorized the requirements into functional suitability, availability, compatibility, usability, and security. We developed OpenEDC based on the regulatory-compliant Clinical Data Interchange Standards Consortium Operational Data Model (CDISC ODM) standard. Mobile device support enables the collection of patient-reported outcomes. OpenEDC is publicly available and released under the MIT open-source license.

Conclusions: Adopting an established standard without modifications supports metadata reuse and clinical data exchange, but it limits item layouts. OpenEDC is a stand-alone web app that can be used without a setup or configuration. This should foster compatibility between medical research and open science. OpenEDC is targeted at observational and translational research studies by clinicians.

(JMIR Med Inform 2021;9(11):e29176) doi:10.2196/29176

KEYWORDS

electronic data capture; open science; data interoperability; metadata reuse; mobile health; data standard; mobile phone

Introduction

High-quality data are crucial for obtaining medical research results [1] and successful machine learning applications [2]. To collect and manage structured data in digital format, researchers can use computer programs called electronic data capture (EDC) systems [3]. There is a consensus that EDC leads to improved data quality as well as cost and time efficiency compared with paper-based methods. Direct data entry at an investigator site reduces the probability of transcription errors [4,5]. By detecting missing fields and data type and range violations, EDC systems offer data validation at the time of entry instead of days or even weeks later [6]. Finally, real-time access allows information managers to continuously monitor the collection process [7], review and analyze data in real time [8], and improve the feedback loop with local investigators [9].

Data exchange and data compatibility are two of the most important areas in medical research. However, proprietary or customized data formats used by EDC vendors render this endeavor a major point of concern. As a result, incompatible electronic case report form (eCRF) data structures impede data integration and analysis from different sources and, hence, the full potential of captured information [10]. A system that fosters compatible data structures through standardization could pave the way toward more open science to improve scientific understanding and enhance patient care [11]. In addition, EDC remains underused despite its benefits [6]. The configuration, maintenance, and financial requirements of typical EDC systems are common obstacles for dissemination and may prevent small-scale studies to profit from digital data collection [12]. Owing to these shortcomings of professional EDC systems, practitioners frequently resort to inappropriate software for data collection, such as general-purpose spreadsheet applications [13]. However, these are considered inflexible, insecure, and complicate data compatibility [14,15]. In addition, they do not provide an audit trail to trace data changes.

In this study, we describe the development process of OpenEDC to address the aforementioned issues. OpenEDC is an EDC system based on the results of a systematic requirements analysis. To the best of our knowledge, this study makes two unique contributions. First, OpenEDC is entirely based on the regulatory-compliant and internationally accepted Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM) standard [16]. It is used without modifications to allow for fully standardized metadata and research data import and export. This facilitates metadata reuse and clinical data exchange, whereas most EDC systems use custom or highly modified formats. Second, a client-based web approach allows researchers worldwide to use OpenEDC without installation or configuration needs. Therefore, it is a valuable alternative to spreadsheet applications. An optional server enables distributed data capture and access whenever necessary. In addition, OpenEDC focuses on cross-platform support for desktop computers and mobile devices to allow the collection of increasingly important patient-reported outcomes. We made OpenEDC publicly available [17] and released it under the MIT open-source license [18].

The remainder of this paper is structured as follows: the Methods section outlines the requirements analysis and evaluation process of OpenEDC. The Results section gives an overview of the identified requirements, the resulting software, and its evaluation outcomes. The contributions, limitations, and future work are discussed in the Discussion section.

Methods

Requirements Analysis

OpenEDC was developed within the context of a large-scale medical register project for chronic diseases. For an intended period of more than 10 years, most German university hospitals were to collect patient-reported outcomes and medical routine data with tablet and desktop computers. During the system selection process, however, the shortcomings of the present EDC systems became apparent. On the basis of the register's long-lasting nature, an ideal system was open source so that it could be maintained in the future without manufacturer dependency or insecure licensing conditions. Being open source would also reduce the risk of unaffordable expenses once the funding of the register might have expired. In addition, standardized metadata import was requested as we had the most eCRFs in the standardized CDISC ODM format. This would allow us to use these methods without time-consuming and error-prone manual transmission. A standardized system would also allow us to export metadata or captured clinical data in a reusable, interoperable, and nonproprietary format in the future. Finally, an easy-to-use and network-independent support for mobile devices was necessary for data collection at the participating sites.

In addition to the project-specific demands, we performed a literature-based requirements analysis to ensure the applicability of OpenEDC in a wide range of research projects. This analysis included the following three steps: first, a literature search revealed the EDC requirements stated by both academics and public bodies. Keywords for searching in the academic repositories PubMed and ScienceDirect were *electronic data capture*, *EDC system*, *digital data collection*, *data management*, and *electronic case report form*. For ScienceDirect, we also added the keywords *clinical trial*, *health study*, and *medical research*. We scanned the top 60 search results for each query. The selection criteria were thematization of EDC-related functionality, low to moderate resource settings, and generalizability (ie, very specific use cases were excluded). After identifying appropriate titles, reading abstracts, and recursively evaluating references, 18 publications were chosen for in-depth analysis (Table 1). The identified publications can be categorized into review studies that evaluated EDC implementation and use (n=8), original reports of trials that used EDC (n=6), and descriptions of EDC system development (n=4). Second, 2 team members with experience in several EDC projects prioritized the identified requirements. This prioritization happened amid the aforementioned internal register requirements and therefore influenced prioritization (see the Discussion section). Third, the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 25010 norm and its software quality model were used to categorize the prioritized requirements [19]. The resulting requirements are listed in Table 1.

Table 1. Main requirements and subrequirements of OpenEDC. Subrequirements are based on commonly stated electronic data capture requirements in the literature. The main categorizing requirements and their definitions originate from the ISO/IEC 25010 norm [19].

Requirement	Definition	Subrequirements
Functional suitability	Product or system provides functions that meet stated and implied needs when used under specified conditions	• Design [13,20] and reuse [14,21] of metadata • Capture and store clinical data [15,20-23] • Form completion tracking [3,24,25] • Field validations (edit checks) [8,13-15,20,21,23,25,26] • Conditional fields (skip patterns) [8,13,15,23,24,26] • Multicentric (multisite) studies [5,13,14,21] • Longitudinal studies (with defined events) [13,24,26] • Multilingual forms [9,24]
Availability	System, product, or component is operational and accessible when required for use	• Open source [4,8,12-14,20,22,23] • Minimal setup and configuration [4,6,12] • Distributed (near) real-time access [5,8,14,21,22,24] • Cross-platform (mobile device support) [9,13,15,26,27] • Offline-capable [15,26,27]
Compatibility	Product, system, or component can exchange information with other products, systems, or components	• Standard-compliant import and export of metadata and clinical data [4,6,13,14] • Semantic annotation (medical coding) of items [12,14,20]
Usability	Product or system can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use	• Ease of use (user-friendly) [4-6,13-15,21] • Medical staff and patient accessibility [4,5]
Security	Product or system protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization	• Authentication and authorization (user rights and roles) [13,20,21] • Encrypted data storage and transmission [14,15,22,26] • Audit trail [3,8,12,14,20-22]

An iterative waterfall model was used to implement the identified requirements [28]. Fundamental and technological requirements were used to design the software architecture. The architecture determines the overall structure, programming language, and supported computer platforms. After architecture design, the most basic and EDC-inherent functions were implemented first, while specific functions were subsequently added. For example, as EDC systems typically follow a metadata-driven approach, that is, allow data capture within the constraints of previously designed eCRFs, the metadata design module was implemented first, followed by modules for data capture and data export. After implementing the key requirements and internal testing, we evaluated OpenEDC to receive the first feedback from potential users. This linear sequence of requirements analysis, design, implementation, and evaluation will be further used to iteratively add secondary functions during the system life cycle.

Evaluation

OpenEDC was evaluated against the identified requirements. However, whereas most of these requirements can be evaluated qualitatively in an absolute sense, that is, achieved or not achieved, usability is perceived subjectively and difficult to generalize [29]. To perform a generalizable and comparable assessment of usability, OpenEDC was evaluated using the system usability scale (SUS) [29]. The SUS is a “10-item scale giving a global view of subjective assessments of usability” that has been used in numerous research projects [30]. After defining a system-related task, 16 participants completed the task and answered the 10-question survey. The participants were recruited via an institution-wide mailing list. At the time of evaluation, none of them were involved in the OpenEDC project. All survey responses resulted in an average score ranging from 0 to 100, estimating the usability of the system. A score above 70 denotes a user-friendly system [31].

Before the actual task, a video was shown to the users to explain the main functionalities of OpenEDC. This video is openly available and can be consulted by prospective users as well [32]. Subsequently, the users were asked to accomplish the following typical EDC tasks using the desktop version of OpenEDC. First, they were asked to reuse and modify case report forms using the metadata design module. For the designed forms, they were prompted to capture data for multiple simulated patients. In the third step, they were instructed to export the collected data while verifying that all data were correctly included in the export file. Finally, the users were invited to answer the SUS survey and 2 open questions about what they liked or disliked in particular. OpenEDC was used to capture all participants’ answers remotely.

Results

Requirements Analysis

Functional Suitability

A fundamental requirement for EDC systems is the support for metadata design [13,20]. The ability to reuse defined metadata elements can facilitate the design process [14,21] (also see the *Compatibility* section). Data element types and range checks enable an EDC system to evaluate data entries in real time and report violations. This is called real-time field validation [14] (plausibility [15] or edit checks [25]), and it increases the data quality significantly. Conditional fields [15] (or skip patterns [26]) can reduce form completion time and improve data quality by showing or hiding items based on prior inputs. After data entry, form completion tracking enables investigators to recognize which form has been completed for which subject [25]. When research is conducted at geographically dispersed sites, that is, research or clinical centers, it is important to keep track of which subject has been included at which site [5]. Typically, local investigators can access only subject data from one site, in contrast to central information managers [14]. To support longitudinal studies, the system needs to allow the definition of events or visits [24]. Finally, medical research studies may include patients with different demographic backgrounds that require multilingual forms to capture patient-reported outcomes [9].

Availability

Availability is frequently stated as an important property or, if absent, the reason for the limited dissemination of EDC systems [4,6,12]. However, it was noted that “open-source EDCs have the potential of increasing and improving public health research activities and raising academic standards because of their availability” [22]. A community that forms around open-source software can collaboratively enhance software quality and increase the probability of its long-term existence. However, even when a system is open source, implementation and maintenance requirements can hamper the availability of software [12]. A web server has been stated as an essential resource for supporting an EDC system [14], together with challenging installation, customization, and configuration requirements [22]. A system that can be used without assistance from potentially expensive information technology specialists could prevent practitioners from resorting to inappropriate spreadsheet applications [14,15]. Once an EDC system is established, distributed access usually allows remote data entry and real-time monitoring, which is particularly important for multicentric studies [22]. Moreover, multiple computer platforms may be used within one research study, such as tablets for data capture and desktop computers for data management [13]. A related subrequirement is offline capability. An active internet connection cannot always be guaranteed, such as in rural research settings or owing to hospital walls that shield mobile network signals [26].

Compatibility

Standardized data formats and coding of data elements can foster data compatibility [33]. Standardized data formats result in syntactic compatibility and facilitate integration and interpretation of clinical research data without the need to apply a proprietary data format [34]. Moreover, it allows the reuse of metadata elements from previous subject-related studies, leading to a simplified metadata development process and data compatibility at the design stage [10]. We agreed to support the regulatory-compliant and well-established CDISC ODM standard, which is “a vendor-neutral, platform-independent format for exchanging and archiving clinical and translational research data, along with their associated metadata, administrative data, reference data, and audit information” [16]. It is the fundamental part of Define-XML [35], which is included in the United States Food and Drug Administration Data Standards Catalog [10,36]. See the Discussion section for comparison with other medical data standards. Finally, the annotation or medical coding of data elements facilitates semantic compatibility. A unique semantic code from a terminology such as the unified medical language system assigned to an item allows unambiguous mapping independent of language or wording [37].

Usability

It was reported that “the lack of a simple, intuitive, and user-friendly EDC system is noteworthy” [14]. Moreover, Franklin et al [13] stated that in “a 2-year qualitative evaluation we found that the importance of ease of use and training materials outweighed number of features and functionality” of EDC systems. As ease of use is also considered to positively impact adoption, data quality, and overall success of EDC initiatives [14], it was added to the list of requirements. In addition, patient-reported outcomes have recently become more relevant to medical research [1]. Data entry for both investigators and patients is regarded as a desirable characteristic of an EDC system [4]. As a result, medical staff and patient accessibility are formulated requirements to enable investigators to capture both routine data and patient-reported outcomes.

Security

Regulatory bodies frequently address the data protection and privacy measures of computerized systems in clinical trials. The General Data Protection Regulation (GDPR) of the European Union, for example, became enforceable in all European Union member states in May 2018 [38]. It covers the personal data of all European Union residents [39] and is considered a driving force for international data protection standards [40]. GDPR demands that personal data, and health data as a special category of personal data in particular, is “processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing” and suggests the “encryption of personal data” [38]. Authentication and authorization, as well as encrypted storage and transmission of data, were added to the system requirements. In addition to the GDPR, national regulations specifically targeting clinical EDC systems exist. Title 21, Part 11 of the Code of Federal Regulations, for example, issued by the Food and Drug Administration, requires EDC systems to maintain a continuous audit trail [41]. In fact, an audit trail is an often stated and essential accountability requirement for every EDC system [12]. It allows investigators, sponsors, and public bodies to seamlessly trace any changes made to electronic records, including time and author.

Design and Implementation

CDISC ODM Data Schema

OpenEDC [17] is based on the CDISC ODM standard. Although initially targeted for allowing the reuse of metadata, we capitalized on the standard's data schema to achieve several other identified requirements. This was possible because this standard not only provided guidance in defining metadata but also in storing associated clinical and administrative data. As a result, OpenEDC can be seen as an editor or user interface for CDISC ODM documents and facilitates the application of this standard.

The CDISC ODM provides the groundwork for achieving the following system requirements. From the metadata perspective, events are at the highest hierarchical level with subordinate forms to allow the representation of longitudinal studies. Descriptive or interrogative texts can be defined as multiple translations for multilingual projects. Most frequently, these texts are assigned to data items for which data are to be collected. Data items have data types and may also have specified value ranges to enable real-time field validation. Moreover, items can be dynamically hidden to support conditional fields. Item definitions can be further referenced and reused in other locations. To complement the data schema for the remaining *functional suitability* requirements, the CDISC ODM specifies structures for subject-related data storage, including references to form completion states and site information.

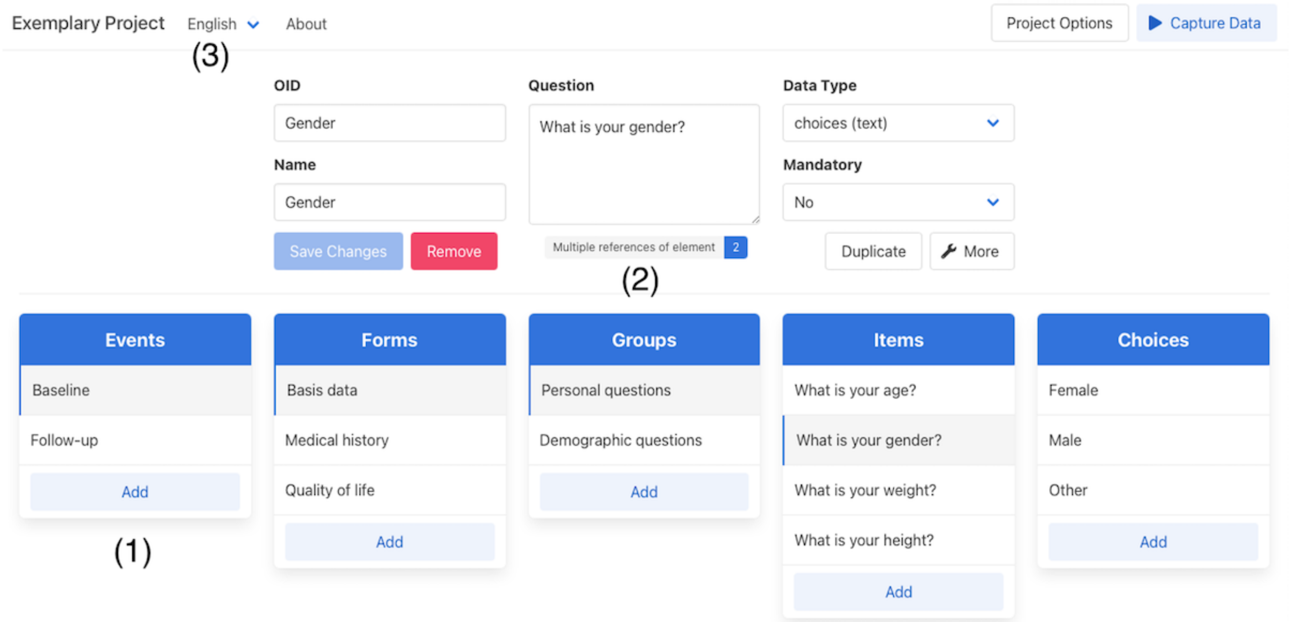
All of the specifications were implemented and internally used by OpenEDC. This results in fully standard-compliant imports and exports of both metadata and clinical research data. In addition, the CDISC ODM enables the annotation of data items with an arbitrary number of semantic codes. These features constitute the *compatibility* requirements. Finally, the CDISC ODM provides guidance for implementing an audit trail, including author information and timestamps for data modifications. This specification was used to partially address *security* requirements.

Client-Based Web App

The *availability* requirements are addressed using a client-based web app. Client-based refers to a static web app that includes all business logic and persistence. It allows researchers to design or reuse eCRFs and capture clinical data without the assistance of information technology specialists or the need for a web server. Moreover, web technology supports the development of cross-platform apps running on all devices using a web browser. Users are not required to install or configure any external software that is important in a clinical setting, where users do not have administrative rights on a standard computer. OpenEDC is a progressive web app that can be installed as a stand-alone system on desktop computers, tablets, and smartphones [42]. In addition, a service worker enables offline data capture that is needed in regions without consistent and reliable internet connections [43]. We decided not to use a third-party JavaScript framework to reduce long-term functional dependencies and developed OpenEDC by using modern browser technologies such as web components [44] and modules [45].

We implemented a simple user interface to address the *usability* requirements. The interface is structured into 2 modes, from which one is used for metadata design and the other for clinical data capture. Modes can be switched at any time to see the rendered form previews during the metadata design phase. We further integrated known concepts such as drag-and-drop, keyboard navigation, and a hierarchical, column-based file structure. A special mode was implemented to support the collection of patient-reported outcomes. Once activated, unnecessary user interface elements become hidden and access to data from other patients is prevented. The user interface of both modes on a desktop computer is shown in Figure 1 (metadata design) and Figure 2 (clinical data capture).

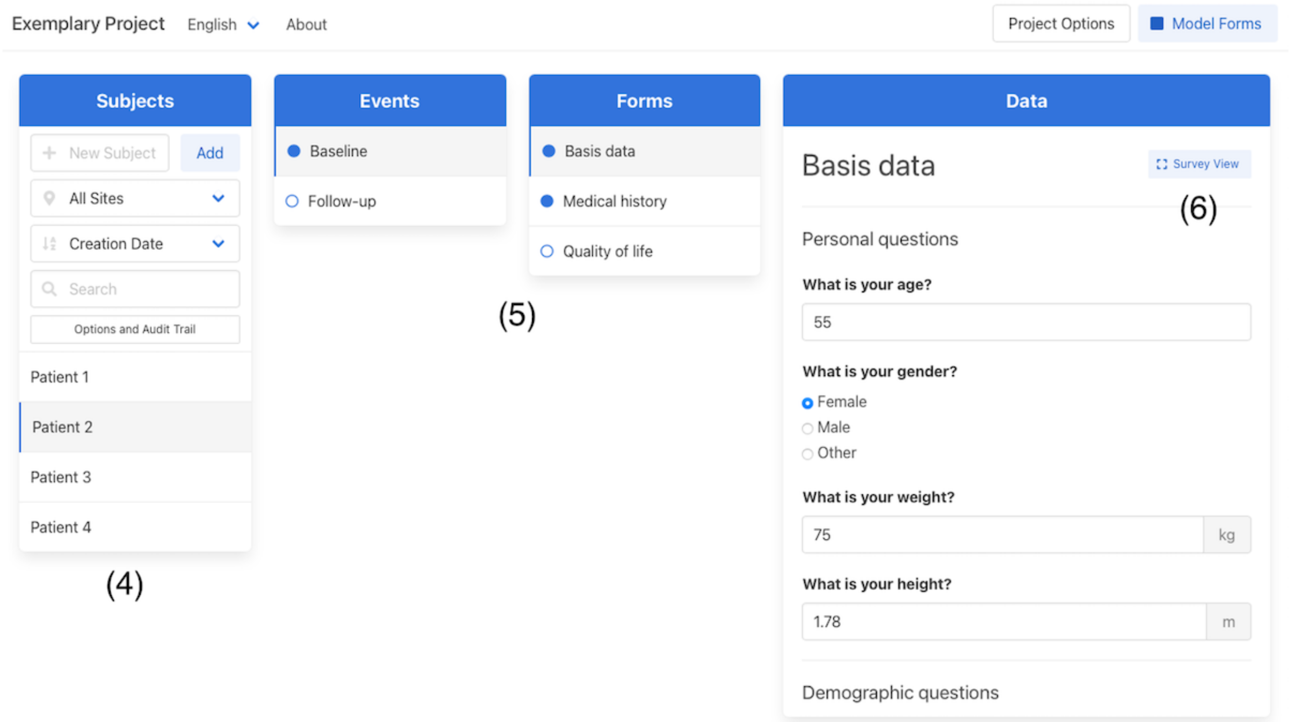
Figure 1. User interface of the metadata design mode. The hierarchical order of metadata elements is represented by the centered column view (1). By means of a referencing system, electronic case report forms (eCRFs) can be reused entirely or partially (2). The language of eCRFs can be changed with the drop-down at the top left (3).



OID: Gender
Question: What is your gender?
Data Type: choices (text)
Name: Gender
Mandatory: No

Events	Forms	Groups	Items	Choices
Baseline	Basis data	Personal questions	What is your age?	Female
Follow-up	Medical history	Demographic questions	What is your gender?	Male
	Quality of life		What is your weight?	Other
			What is your height?	

Figure 2. User interface of the clinical data capture mode. Subjects can be managed with the left column where an audit trail can be accessed as well (4). Filled or empty circles in the 2 center columns indicate whether an event or form has been completed (5). A survey view button within the right electronic case report form column switches to a mode for patient-reported outcomes (6).



Subjects: Patient 1, Patient 2, Patient 3, Patient 4

Events: ● Baseline, ○ Follow-up

Forms: ● Basis data, ● Medical history, ○ Quality of life

Basis data

Personal questions

What is your age? 55

What is your gender?
● Female
○ Male
○ Other

What is your weight? 75 kg

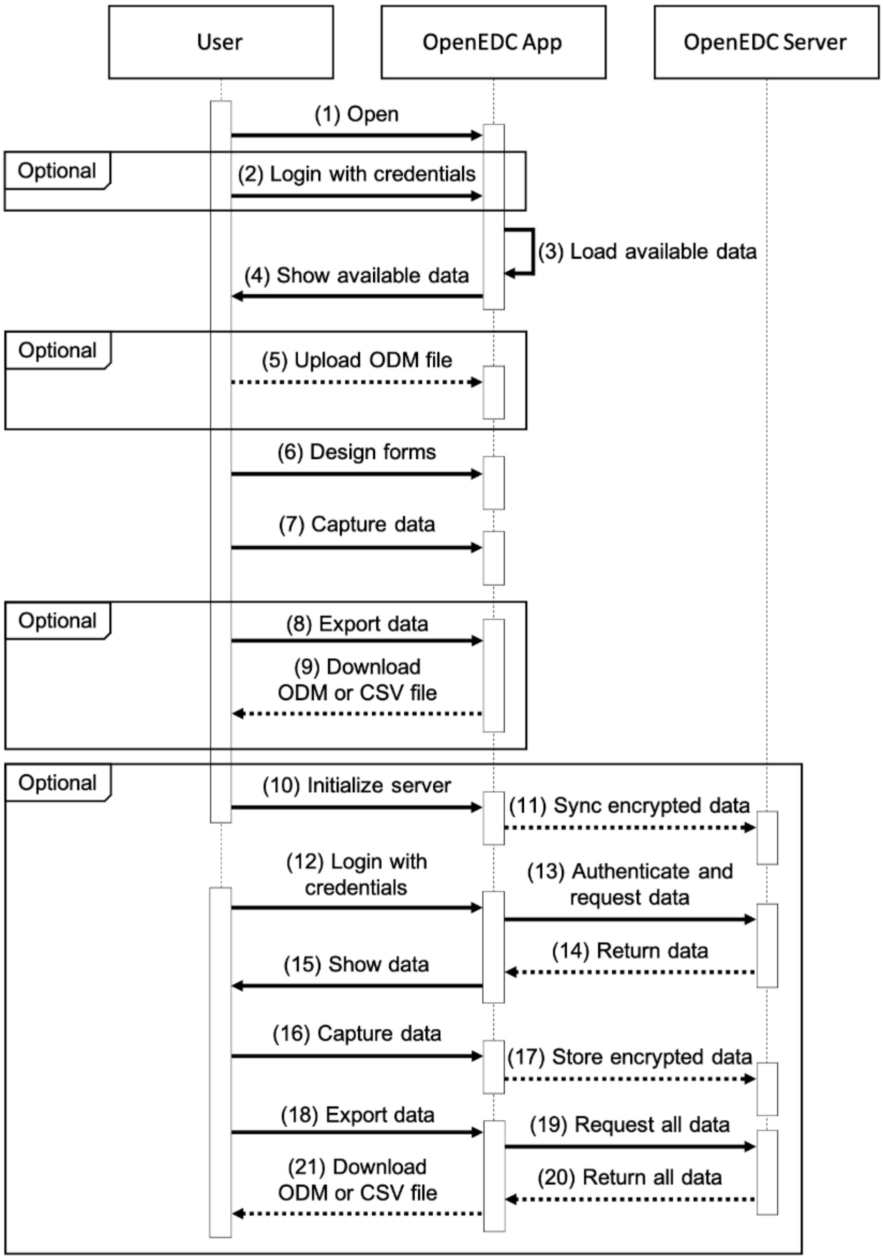
What is your height? 1.78 m

Demographic questions

Optional Client-Server Architecture

To accomplish the *availability* requirement of distributed real-time system access for multiuser and multicentric research studies, we developed an optional OpenEDC server [46]. A connection to the server can be established even after the data are captured locally using the stand-alone OpenEDC web app. All data are synchronized with the host server. From this time on, new user accounts can be created, and collaboratively captured data are centrally stored on the OpenEDC server. Figure 3 shows a sequence diagram of this usage scenario.

Figure 3. Sequence diagram of a typical use scenario with OpenEDC. In this example, the stand-alone OpenEDC web application is used to design electronic case report forms and capture data. A Clinical Data Interchange Standards Consortium Operational Data Model file can be uploaded to reuse metadata or import clinical data. Optionally, the user can initialize an empty OpenEDC server with locally stored data. This enables the user to set up a multiuser system and conduct multicentric research studies. EDC: electronic data capture; ODM: operational data model.



The server provides authentication and authorization services to address the remaining *security* requirements. Clients can authenticate against a representational state transfer application programming interface [47]. When authenticated, the server verifies that a user has the required authorization to access the requested application programming interface end point. End points are available for reading, storing, and removing metadata, clinical subject data, or administrative data. Each end point enforces different user rights. All data is further transferred and stored encrypted. The OpenEDC web app enforces an encrypted https over a transport layer security connection. In addition, all data are encrypted before transfer from the client to the server and can only be decrypted by an authorized client when received back from the server. This ensures that even people with logical or physical server access cannot read data without permission. Local password encryption can also be used to encrypt data when the OpenEDC web app is not connected to a server.

Evaluation

Usability is a subjectively perceived characteristic of a particular context and user [29]. The SUS was used to estimate the usability of OpenEDC. In total, 16 persons were asked to participate in the test, of which 50% (8/16) were women and 50% (8/16) were men. All participants were digital health or biomedical domain experts with an average work experience of 5.3 years (SD 3.1 years) in medical research projects.

OpenEDC achieved a mean usability score of 83.1 (SD 9.6) out of 100. Men rated it slightly lower than women with an average score of 82.5 (SD 11.7) compared with 83.8 (SD 7.6). Two additional open questions were answered by 75% (12/16) of the participants. They provided very heterogeneous suggestions for improvement, with most being related to the user interface and few to functionality. Interface-related suggestions were shortcut buttons for frequently used functions, more noticeable highlighting of inputs with implausible data, and a larger visual difference between the metadata and clinical data view. Introducing simple statistics for data completeness and patient enrollment, labeling conditionally unavailable items in the CSV export, and improving support for older browsers were suggestions related to functionality. Most participants stated that they liked the clear user interface and the performance of the system.

Discussion

Principal Findings

This paper describes the implementation process of OpenEDC, an open-source and standard-compliant EDC system for medical research. We conducted a requirements analysis to identify the academic and regulatory demands for digital data collection. After implementation, we performed a usability evaluation to obtain feedback from the users. OpenEDC achieved a mean usability score of 83.1, which can be considered user-friendly [31]. OpenEDC is available worldwide without installation or configuration requirements. It focuses on cross-platform support for desktop and mobile devices to allow the collection of increasingly important patient-reported outcomes.

Strengths and Limitations

OpenEDC is based on the CDISC ODM standard, yielding several advantages. Metadata and clinical research data can be imported and exported without constraints in a nonproprietary format and without vendor lock-in effects. Investigators may also download eCRFs from public metadata registries, such as the Portal of Medical Data Models [48], to swiftly create databases for data capture. Therefore, we hope to encourage the reuse of metadata, foster compatibility of medical research, and ultimately support open science [11]. In contrast, the CDISC ODM is relatively limited when it comes to the visual representation of items. For example, multiple-choice questions need to be implemented as lists of Boolean items. Moreover, it is impossible to uniformly distinguish between single-choice items rendered as radio buttons or as a drop-down list or to label multiple items with the same predefined choices as a Likert scale. Other systems that support the CDISC ODM often work around this limitation by extending or modifying the standard. However, this can render a system incompatible with other systems. Research Electronic Data Capture (REDCap) [21], for example, imports and exports REDCap-specific CDISC ODM files but often fails when attempting to import standard-compliant files. Currently, OpenEDC favors standard conformance over nonspecified input types.

Other standards exist for exchanging metadata and clinical research data. For example, Fast Healthcare Interoperability Resources (FHIR) from Health Level 7 (HL7) is increasingly adopted to exchange electronic health records and other information in the medical domain [49]. Resources constitute the fundamental building block and are also available for eCRF metadata and their associated clinical data, called Questionnaire and Questionnaire Response, respectively [50]. Although these are rather unspecific by default, the structured data capture (SDC) implementation guide targets improved interoperability [51]. In contrast to the CDISC ODM, however, the HL7 FHIR SDC does not provide a holistic archive format, including users, sites, audit trail records, and their relationships to captured information. Moreover, it does not support the definition of longitudinal events or multiple languages by default. These limitations made HL7 FHIR alone unsuitable for the requirements of the present EDC system. However, as it is widely used to exchange more discrete parts of data, we prepared to implement support for the import and export of HL7 FHIR SDC Questionnaire and Questionnaire Response resources at the time of writing. A similar approach can be adopted for ISO/IEC 11179 [52] and ISO/TS (International Organization for Standardization Technical Specification) 21526 [53]. Both are metadata standards, with the ISO/TS 21526 explicitly targeting the health care sector [53]. However, as they do not provide a specific data format to support the syntactic compatibility of information [54], integration is currently not prioritized.

OpenEDC is publicly available for the creation of local studies. The app is available via the web for desktop and mobile devices, whereas data storage occurs locally and encrypted. This architecture allows researchers to benefit from metadata-driven digital data collection without an information technology department, web server configuration issues, or device constraints. In addition, it leaves data sovereignty to the investigator, rather than a third-party infrastructure or server provider. While this approach offers advantages in terms of flexibility, it also has some drawbacks. It is generally helpful to have a dedicated computer scientist who can make educated decisions about data security, data backup, and metadata design concerns. Moreover, it may be beneficial for a study's sustainability to have a contact person for technical problems and issues. However, it is worth noting that an information technology specialist can still be employed when using OpenEDC. In particular, when an OpenEDC server must be configured, for example, for projects with multiple users and sites, knowledge in setting up a web server is important. In our opinion, OpenEDC's architecture is particularly useful for investigator-initiated studies and enables researchers to set up and test databases before information technology support and infrastructure investments have to be made.

Comparison With Prior Work

Other EDC systems also exist. One of the most frequently used EDC systems is REDCap [21]. REDCap provides various functions that are not present in OpenEDC, such as an extensive admin server dashboard, support for surveys that can be sent via a link to participants, and a module for randomization. While a systematic comparison is beyond the scope of this work, there are some aspects in which OpenEDC has advantages. For example, although REDCap is free to use, it is strictly licensed and not open-source, requires a web server, and is not standard-compliant, as it uses a customized CDISC ODM syntax. It is worth noting that open-source EDC systems also exist. Examples include the OpenClinica Community Edition [55], Open Data Kit (ODK) ecosystem [56], the Rare Disease

Registry Framework [57], and the Open-Source Registry System for Rare Diseases [58]. OpenClinica and ODK are established systems with functionalities that are absent in OpenEDC. For example, OpenClinica provides double data entry and a query management system. ODK provides more input types, such as sliders, as well as widgets for image capturing and drawing. However, OpenClinica Community Edition requires a web server, form design via Microsoft Excel, and is not suitable for smartphones or tablet computers. On the other hand, data capture using ODK is designed only for Android mobile devices. While OpenClinica and ODK are multipurpose EDC systems, the Rare Disease Registry Framework and Open-Source Registry System for Rare Diseases specifically target registries for rare diseases. Similar to OpenEDC, the 2 systems address technically underresourced settings and foster metadata reuse. However, both lack offline mobile device support and a standardized export of metadata and clinical research data.

Future Work

Future work is necessary. The main objective was to ensure the applicability of OpenEDC to a wide range of research projects. However, literature-based requirements analysis was influenced by the demands of a large-scale medical register. Rarely mentioned requirements were not included if they were not required by the internal project. Examples of rarely mentioned but deferred demands are integrated query management as well as document storage and report functionalities. In addition, although OpenEDC complies with relevant laws and regulations, including 21 Code of Federal Regulations Part 11 and GDPR, a computer system validation required for interventional trials has not yet been conducted. Validating an EDC system is also trial-specific and requires activities by the investigator or sponsor. Currently, we see OpenEDC's distinct advantages for observational and translational research studies by clinicians rather than commercial clinical trials. We hope it is a valuable first step toward an openly available, standard-compliant, and mobile EDC system. We plan to develop OpenEDC further and use it in prospective studies. To expand the support for varying study protocols, unavailable functions stated earlier should be added. We hope for contributions from the research community, as we have published OpenEDC under the MIT open-source license.

Conclusions

We showed that it is possible to develop an EDC system for use without upfront investment and preservation of data sovereignty. The primary focus was on standard compliance to foster metadata reuse, interoperable research data, and open science. Future work is necessary to extend the system's functionality and prove its robustness in large-scale studies. OpenEDC is publicly available and released under the MIT open-source license.

Authors' Contributions

LG designed and implemented OpenEDC and wrote the manuscript. SH conceived and reviewed the manuscript. MD designed the overall concept, supervised the work, and reviewed the manuscript.

Conflicts of Interest

None declared.

References

1. Saczynski JS, McManus DD, Goldberg RJ. Commonly used data-collection approaches in clinical research. Am J Med 2013 Nov;126(11):946-950 [FREE Full text] [doi: 10.1016/j.amjmed.2013.04.016] [Medline: 24050485]
2. Beam AL, Kohane IS. Big data and machine learning in health care. J Am Med Assoc 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
3. El Emam K, Jonker E, Sampson M, Krleza-Jerić K, Neisa A. The use of electronic data capture tools in clinical trials: web-survey of 259 Canadian trials. J Med Internet Res 2009 Mar 09;11(1):e8 [FREE Full text] [doi: 10.2196/jmir.1120] [Medline: 19275984]
4. Welker JA. Implementation of electronic data capture systems: barriers and solutions. Contemp Clin Trials 2007 May;28(3):329-336. [doi: 10.1016/j.cct.2007.01.001] [Medline: 17287151]
5. Fleming S, Barsdorf AI, Howry C, O'Gorman H, Coons SJ. Optimizing electronic capture of clinical outcome assessment data in clinical trials: the case of patient-reported endpoints. Ther Innov Regul Sci 2015 Nov;49(6):797-804. [doi: 10.1177/2168479015609102] [Medline: 30222384]
6. Rorie DA, Flynn RW, Grieve K, Doney A, Mackenzie I, MacDonald TM, et al. Electronic case report forms and electronic data capture within clinical trials and pharmacoepidemiology. Br J Clin Pharmacol 2017 Sep;83(9):1880-1895 [FREE Full text] [doi: 10.1111/bcp.13285] [Medline: 28276585]
7. Pavlović I, Kern T, Miklavcic D. Comparison of paper-based and electronic data collection process in clinical trials: costs simulation study. Contemp Clin Trials 2009 Jul;30(4):300-316. [doi: 10.1016/j.cct.2009.03.008] [Medline: 19345286]
8. Walther B, Hossin S, Townend J, Abernethy N, Parker D, Jeffries D. Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data. PLoS One 2011;6(9):e25348 [FREE Full text] [doi: 10.1371/journal.pone.0025348] [Medline: 21966505]

9. Paudel D, Ahmed M, Pradhan A, Lal DR. Successful use of tablet personal computers and wireless technologies for the 2011 Nepal Demographic and Health Survey. Glob Health Sci Pract 2013 Aug;1(2):277-284 [FREE Full text] [doi: 10.9745/GHSP-D-12-00056] [Medline: 25276539]
10. Dugas M. Design of case report forms based on a public metadata registry: re-use of data elements to improve compatibility of data. Trials 2016 Nov 29;17(1):566 [FREE Full text] [doi: 10.1186/s13063-016-1691-8] [Medline: 27899162]
11. Ross JS, Lehman R, Gross CP. The importance of clinical trial data sharing: toward more open science. Circ Cardiovasc Qual Outcomes 2012 Mar 01;5(2):238-240 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.112.965798] [Medline: 22438465]
12. Ohmann C, Kuchinke W, Canham S, Lauritsen J, Salas N, Schade-Brittinger C, et al. Standard requirements for GCP-compliant data management in multinational clinical trials. Trials 2011 Mar 22;12:85 [FREE Full text] [doi: 10.1186/1745-6215-12-85] [Medline: 21426576]
13. Franklin JD, Guidry A, Brinkley JF. A partnership approach for Electronic Data Capture in small-scale clinical trials. J Biomed Inform 2011 Dec;44 Suppl 1:103-108 [FREE Full text] [doi: 10.1016/j.jbi.2011.05.008] [Medline: 21651992]
14. Shah J, Rajgor D, Pradhan S, McCready M, Zaveri A, Pietrobon R. Electronic data capture for registries and clinical trials in orthopaedic surgery: open source versus commercial systems. Clin Orthop Relat Res 2010 Oct;468(10):2664-2671 [FREE Full text] [doi: 10.1007/s11999-010-1469-3] [Medline: 20635174]
15. Meyer J, Fredrich D, Piegsa J, Habes M, van den Berg N, Hoffmann W. A mobile and asynchronous electronic data capture system for epidemiologic studies. Comput Methods Programs Biomed 2013 Jun;110(3):369-379. [doi: 10.1016/j.cmpb.2012.10.015] [Medline: 23195493]
16. ODM-XML. Clinical Data Interchange Standards Consortium (CDISC). URL: https://www.cdisc.org/standards/data-exchange/odm [accessed 2021-07-07]
17. OpenEDC - Web Application. URL: https://openedc.org [accessed 2021-07-07]
18. OpenEDC. GitHub Repository. URL: https://github.com/imi-muenster/OpenEDC [accessed 2021-07-07]
19. ISO/IEC. Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models. ISO/IEC 25010:2011 2017:A. [doi: 10.5220/0005097303630368]
20. Krishnankutty B, Bellary S, Kumar NB, Moodahadu LS. Data management in clinical research: an overview. Indian J Pharmacol 2012 Mar;44(2):168-172 [FREE Full text] [doi: 10.4103/0253-7613.93842] [Medline: 22529469]
21. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
22. Ngari MM, Waithira N, Chilengi R, Njuguna P, Lang T, Fegan G. Experience of using an open source clinical trials data management software system in Kenya. BMC Res Notes 2014 Nov 26;7:845 [FREE Full text] [doi: 10.1186/1756-0500-7-845] [Medline: 25424974]
23. Dillon DG, Pirie F, Rice S, Pomilla C, Sandhu MS, Motala AA, African Partnership for Chronic Disease Research (APCDR). Open-source electronic data capture system offered increased accuracy and cost-effectiveness compared with paper methods in Africa. J Clin Epidemiol 2014 Dec;67(12):1358-1363 [FREE Full text] [doi: 10.1016/j.jclinepi.2014.06.012] [Medline: 25135245]
24. Style S, Beard BJ, Harris-Fry H, Sengupta A, Jha S, Shrestha BP, et al. Experiences in running a complex electronic data capture system using mobile phones in a large-scale population trial in southern Nepal. Glob Health Action 2017;10(1):1330858 [FREE Full text] [doi: 10.1080/16549716.2017.1330858] [Medline: 28613121]
25. Philipp B, Christian F, Martin D. x4T-EDC: A prototype for study documentation based on the single source concept. In: Proceedings of the 24th Medical Informatics in Europe Conference (MIE2012). 2012 Presented at: 24th Medical Informatics in Europe Conference (MIE2012); Aug 26-29, 2012; Pisa, Italy URL: https://www.wi.uni-muenster.de/research/publications/82782
26. King C, Hall J, Banda M, Beard J, Bird J, Kazembe P, et al. Electronic data capture in a rural African setting: evaluating experiences with different systems in Malawi. Glob Health Action 2014;7:25878 [FREE Full text] [doi: 10.3402/gha.v7.25878] [Medline: 25363364]
27. Morak J, Schwetz V, Hayn D, Fruhwald F, Schreier G. Electronic data capture platform for clinical research based on mobile phones and near field communication technology. Annu Int Conf IEEE Eng Med Biol Soc 2008;2008:5334-5337. [doi: 10.1109/IEMBS.2008.4650419] [Medline: 19163922]
28. Alshamrani A, Bahattab A. A comparison between three systems development life cycle models: waterfall model, spiral model, and incremental/iterative model. Int J Comput Sci Issues 2015;12(1):106-111 [FREE Full text]
29. Brooke J. SUS - A quick and dirty usability scale. In: Usability Evaluation in Industry. Boca Raton, Florida, United States: CRC Press; 1996:189-194.
30. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum-Comput Interact 2008 Jul 30;24(6):574-594. [doi: 10.1080/10447310802205776]
31. Bangor A, Kortum P, MILLER J, Miller J, Bangor AW, Kortum PT, et al. Determining what individual SUS scores mean: adding an adjective rating scale. J Usab Stud 2009;4(3):114-123 [FREE Full text]

32. Greulich L. An introduction to OpenEDC – simple and standardized electronic data capture for medical research. OpenEDC. URL: https://www.youtube.com/watch?v=5cV8cvCXMgg [accessed 2021-07-07]
33. Vengadeswaran A, Neuhaus P, Hegselmann S, Storf H, Kadioglu D. Semantically annotated metadata: interconnecting samply.MDR and MDM-portal. Stud Health Technol Inform 2019 Sep 03;267:86-92. [doi: 10.3233/SHTI190810] [Medline: 31483259]
34. Lin C, Wu N, Liou D. A multi-technique approach to bridge electronic case report form design and data standard adoption. J Biomed Inform 2015 Mar;53:49-57 [FREE Full text] [doi: 10.1016/j.jbi.2014.08.013] [Medline: 25200473]
35. Define-XML. Clinical Data Interchange Standards Consortium (CDISC). URL: https://www.cdisc.org/standards/data-exchange/define-xml [accessed 2021-07-07]
36. Providing regulatory submissions in electronic format - standardized study data : guidance for industry. US Food & Drug Administration. 2021. URL: https://www.fda.gov/media/82716/download [accessed 2021-07-07]
37. Hegselmann S, Storck M, Geßner S, Neuhaus P, Varghese J, Dugas M. A web service to suggest semantic codes based on the MDM-portal. Stud Health Technol Inform 2018;253:35-39. [Medline: 30147036]
38. Regulation (EU) 2016/679 of the European Parliament and of the Council. Official Journal of the European Union. 2016. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN [accessed 2021-07-07]
39. Goddard M. The EU General Data Protection Regulation (GDPR): European Regulation that has a global impact. Int J Mark Res 2017 Nov 01;59(6):703-705. [doi: 10.2501/ijmr-2017-050]
40. Albrecht J. How the GDPR will change the world. Eur Data Protect Law Rev 2016;2(3):287-289. [doi: 10.21552/edpl/2016/3/4]
41. Title 21, Part 11 - Electronic records; electronic signatures. Code of Federal Regulations. URL: https://www.ecfr.gov/cgi-bin/text-idx?node=pt21.1.11 [accessed 2021-07-07]
42. Progressive web apps (PWAs). MDN Web Docs. URL: https://developer.mozilla.org/en-US/docs/Web/Progressive_web_apps [accessed 2021-07-07]
43. Service worker API. MDN Web Docs. URL: https://developer.mozilla.org/en-US/docs/Web/API/Service_Worker_API [accessed 2021-01-15]
44. Web components. MDN Web Docs. URL: https://developer.mozilla.org/en-US/docs/Web/Web_Components [accessed 2021-07-07]
45. JavaScript modules. MDN Web Docs. URL: https://developer.mozilla.org/de/docs/Web/JavaScript [accessed 2021-07-07]
46. OpenEDC Server. GitHub Repository. URL: https://github.com/imi-muenster/OpenEDC-Server [accessed 2021-07-07]
47. Fielding RT. Architectural styles and the design of network-based software architectures. Doctoral Dissertation, University of California, Irvine. 2000. URL: https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm [accessed 2021-10-18]
48. Portal of Medical Data Models (MDM-Portal). URL: https://medical-data-models.org/?lang=en [accessed 2021-07-07]
49. 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] [Medline: 31483254]
50. Resource index. HL7 FHIR. URL: https://www.hl7.org/fhir/resourcelist.html [accessed 2021-07-07]
51. SDC. HL7 FHIR. URL: http://hl7.org/fhir/us/sdc/ [accessed 2021-07-07]
52. Ngouongo SM, Löbe M, Stausberg J. The ISO/IEC 11179 norm for metadata registries: does it cover healthcare standards in empirical research? J Biomed Inform 2013 Apr;46(2):318-327 [FREE Full text] [doi: 10.1016/j.jbi.2012.11.008] [Medline: 23246614]
53. Ulrich H, Kock-Schoppenhauer A, Drenkhahn C, Löbe M, Ingenerf J. Analysis of ISO/TS 21526 towards the extension of a standardized query API. Stud Health Technol Inform 2020 Nov 23;275:202-206. [doi: 10.3233/SHTI200723] [Medline: 33227769]
54. Ulrich H, Kern J, Tas D, Kock-Schoppenhauer AK, Ückert F, Ingenerf J, et al. QLMDR: a GraphQL query language for ISO 11179-based metadata repositories. BMC Med Inform Decis Mak 2019 Mar 18;19(1):45 [FREE Full text] [doi: 10.1186/s12911-019-0794-z] [Medline: 30885183]
55. OpenClinica - Community Edition. URL: https://www.openclinica.com/community-edition-open-source-edc/ [accessed 2021-07-07]
56. Open Data Kit. URL: https://opendatakit.org [accessed 2021-07-07]
57. Bellgard MI, Render L, Radochonski M, Hunter A. Second generation registry framework. Source Code Biol Med 2014;9:14 [FREE Full text] [doi: 10.1186/1751-0473-9-14] [Medline: 24982690]
58. Storf H, Schaaf J, Kadioglu D, Göbel J, Wagner TO, Ückert F. [Registries for rare diseases : OSSE - an open-source framework for technical implementation]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2017 May;60(5):523-531. [doi: 10.1007/s00103-017-2536-7] [Medline: 28289778]

Abbreviations

CDISC: Clinical Data Interchange Standards Consortium
eCRF: electronic case report form
EDC: electronic data capture

FHIR: Fast Health care Interoperability Resources
GDPR: General Data Protection Regulation
HL7: Health Level 7
ISO/IEC: International Organization for Standardization/International Electrotechnical Commission
ODK: Open Data Kit
ODM: operational data model
REDCap: Research Electronic Data Capture
SDC: structured data capture
SUS: system usability scale

Edited by R Kukafka; submitted 29.03.21; peer-reviewed by K Mann, X Jing, D Kadioglu; comments to author 22.05.21; revised version received 13.07.21; accepted 28.09.21; published 19.11.21.

Please cite as:
Greulich L, Hegselmann S, Dugas M
An Open-Source, Standard-Compliant, and Mobile Electronic Data Capture System for Medical Research (OpenEDC): Design and Evaluation Study
JMIR Med Inform 2021;9(11):e29176
URL: https://medinform.jmir.org/2021/11/e29176
doi:10.2196/29176
PMID:34806987

Original Paper

Optimal Triage for COVID-19 Patients Under Limited Health Care Resources With a Parsimonious Machine Learning Prediction Model and Threshold Optimization Using Discrete-Event Simulation: Development Study

Jeongmin Kim[1*], BBA; Hakyung Lim[1*], MSc; Jae-Hyeon Ahn[1], PhD; Kyoung Hwa Lee[2], MD, MMS; Kwang Suk Lee[3*], MD, MMS; Kyo Chul Koo[3*], MD, PhD

[1]College of Business, Korea Advanced Institute of Science and Technology, Seoul, Republic of Korea

[2]Division of Infectious Disease, Department of Internal Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea

[3]Department of Urology, Yonsei University College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Kyo Chul Koo, MD, PhD
Department of Urology
Yonsei University College of Medicine
211 Eonju-ro, Gangnam-gu
Seoul, 135-720
Republic of Korea
Phone: 82 01099480342
Fax: 82 0234628887
Email: gckoo@yuhs.ac

Abstract

Background: The COVID-19 pandemic has placed an unprecedented burden on health care systems.

Objective: We aimed to effectively triage COVID-19 patients within situations of limited data availability and explore optimal thresholds to minimize mortality rates while maintaining health care system capacity.

Methods: A nationwide sample of 5601 patients confirmed with COVID-19 until April 2020 was retrospectively reviewed. Extreme gradient boosting (XGBoost) and logistic regression analysis were used to develop prediction models for the maximum clinical severity during hospitalization, classified according to the World Health Organization Ordinal Scale for Clinical Improvement (OSCI). The recursive feature elimination technique was used to evaluate the maintenance of model performance when clinical and laboratory variables were eliminated. Using populations based on hypothetical patient influx scenarios, discrete-event simulation was performed to find an optimal threshold within limited resource environments that minimizes mortality rates.

Results: The cross-validated area under the receiver operating characteristic curve (AUROC) of the baseline XGBoost model that utilized all 37 variables was 0.965 for OSCI ≥6. Compared to the baseline model's performance, the AUROC of the feature-eliminated model that utilized 17 variables was maintained at 0.963 with statistical insignificance. Optimal thresholds were found to minimize mortality rates in a hypothetical patient influx scenario. The benefit of utilizing an optimal triage threshold was clear, reducing mortality up to 18.1%, compared with the conventional Youden index.

Conclusions: Our adaptive triage model and its threshold optimization capability revealed that COVID-19 management can be achieved via the cooperation of both the medical and health care management sectors for maximum treatment efficacy. The model is available online for clinical implementation.

(JMIR Med Inform 2021;9(11):e32726) doi:10.2196/32726

KEYWORDS

COVID-19; decision support techniques; machine learning; prediction; triage

Introduction

The high incidences of infection, critical illness, and mortality due to COVID-19 have placed unprecedented burdens on international health care systems. In response, the World Health Organization (WHO) guidelines have recommended that all countries prepare for infection surges in their health care facilities and implement appropriate triage protocols [1]. Unfortunately, these guidelines fail to provide a one-size-fits-all approach that works for individual regions while accounting for unique outbreak surges.

Numerous prognostic models have been developed to ensure effective triage for COVID-19 patients [2-7]. While these models exhibit modest predictive accuracy, their generalizability has been questioned due to their confinement to single clinical outcome measures and reductions in their discrimination performance when using insufficient data. Most importantly, the classification thresholds of these prediction models, which are crucial for ensuring effective resource utilization by health care systems, have been neglected, thereby limiting their practicality. To overcome these models' shortcomings, combing multi-institutional data with advanced prediction models, such as those using machine learning and simulation modeling, is needed.

COVID-19 is associated with significant disruptions to most health care infrastructures. Therefore, an adjustable risk stratification model that considers the resource availability of various regions, as well as one that identifies patients who will likely require hospitalization and intensive care, will help to reduce these systems' burdens. In this study, we propose an adaptive triage model that takes into account deficits in established health care resources due to the COVID-19 pandemic. Our study has several main contributions. The first contribution is a powerful and interpretable prediction model using extreme gradient boosting (XGBoost) and Shapley additive explanations (SHAP) that provides accurate prognoses to facilitate preemptive treatments, thereby ensuring improvements in patient survival outcomes. The second contribution is the ability to apply the model with readily available assessment parameters using the recursive feature elimination (RFE) technique, thereby maintaining its reliability in data-limited environments [8,9]. The third contribution is the consideration of resource availability at either the facility or national level relative to varying patient influx volumes by employing the discrete-event simulation (DES) technique.

Our study objectives were 3-fold. First, we sought to develop a baseline prediction model with an explanatory feature for triaging COVID-19 patients. Second, based on this model, we aimed to utilize the RFE technique to develop feature-eliminated models that would help ensure efficient resource utilization under limited data availability. Finally, we set out to develop an adaptive triage model using the DES technique to assist in efficient resource utilization under limited health care resources.

Methods

Ethics Statement

This study was approved by an institutional ethics committee (2020-0883-001) and the Korea Disease Control and Prevention Agency (KDCA) epidemiological survey and analysis committee (20201120_4a). All study procedures complied with the 1946 Declaration of Helsinki and its 2008 update.

Patient Cohort

We retrospectively retrieved the demographic, clinical, laboratory, and disease outcome records of 5628 patients who were confirmed with SARS‐CoV‐2 by real-time reverse transcription-polymerase chain reaction using nasopharyngeal/oropharyngeal swab or sputum specimens until April 2020. The data were collected and comprehensively managed by the KDCA. Among 10,774 patients consecutively diagnosed with COVID-19 within this time frame, data on 52.2% (5628/10,774) of the patient population were publicized for research purposes after excluding patients with any missing data. The database did not account for the location of diagnosis within Korea. The database included patients who had been treated and released from quarantine or hospitalization, as well as those who died from COVID-19 sequelae. The criteria for patient release included obtaining 2 consecutive negative results at least 24 hours apart and an asymptomatic status. Among the 5628 patients, 27 patients with missing clinical severity data were excluded, resulting in a final development cohort of 5601 patients.

Covariates and Outcome Definitions

Baseline data collected at each patient's diagnosis were used for model development. Demographic data included patient age, sex, systolic and diastolic blood pressure, heart rate, body temperature, and BMI. Medical comorbidities included hypertension, diabetes mellitus, heart failure, cardiovascular disease, asthma, chronic kidney disease, chronic obstructive pulmonary disease, chronic liver disease, autoimmune disease, dementia, malignancy, and pregnancy. Clinical findings included a history of fever (temperature ≥37.5˚C), cough, sputum production, myalgia, fatigue, sore throat, rhinorrhea, dyspnea, vomiting, nausea, diarrhea, headache, and altered consciousness. Laboratory data included hemoglobin, hematocrit, white blood cell count, %leukocyte, and platelet count. Each patient's maximum clinical severity during quarantine or hospitalization was classified according to the WHO Ordinal Scale for Clinical Improvement (OSCI) [10].

Statistical Analysis

Model Development

Multivariate logistic regression (LR) and XGBoost were used to select the best performing prediction model using all available clinical and laboratory data [11]. The models were developed and cross-validated using data from 5037 (89.9%) patients and were then revalidated using a hold-out cohort of 564 (10.1%) patients. Performance metrics were calculated using 10-fold cross-validation to avoid any overfitting. Model development was performed using the *caret* package in R Statistical Package (version 4.0.5; R Project for Statistical Computing). The best performing model derived from XGBoost was defined as Model 1 and was used as a baseline model for RFE.

Variable Elimination

The RFE technique was used to evaluate the extent of the maintenance of model performance when various predictors were eliminated. RFE was performed for the following 2 models that incorporated all clinical data with and without laboratory data: Model 1 (clinical data with laboratory data) and Model 2 (clinical data without laboratory data). SHAP was used to rank each variable based on its significance to the models for its desirable properties, including local accuracy, missingness, and consistency [12]. At each RFE iteration, the lowest-ranked feature was eliminated, the model was refitted, and its performance was assessed using 10-fold cross-validation. The feature-eliminated models (Model 3: limited clinical data with laboratory data and Model 4: limited clinical data without laboratory data) were then selected at a point wherein the number of features was minimized while differences in area under the receiver operating characteristic curve (AUROC) values remained statistically insignificant. The 4 classification models were revalidated with the hold-out cohort to avoid any overfitting. Analysis was performed using *caret* and the *SHAPforXGBOOST* package in R.

Model Interpretation and Comparison

To interpret Model 1, we used SHAP as it provides visible post-hoc interpretability to black-box machine learning models [12]. Patient-specific plots were created by aggregating the SHAP score of each variable for a specific prediction.

The hyperparameters of the XGBoost algorithm were optimized to maximize its AUROC values using a simple grid search with 10-fold cross-validation. Accuracy, AUROC, sensitivity, positive predictive value (PPV), and negative predictive value (NPV) were calculated at 90% specificity using the *pROC* package in R. CIs of the performance measures were then calculated using a stratified bootstrap method with 2000 replicates.

Threshold Optimization

DES and Patient Influx Generation

The DES technique replicates complex behaviors and interactions among individuals, populations, and their environments. Therefore, it has been widely used to form more effective clinical decisions to minimize mortality rates under medical resource constraints [13]. Thus, we applied DES to identify the optimal threshold within limited medical resource environments that minimizes mortality rates, as calculated by *n* (*total deaths*) / *n* (*total patients*), using the *simmer* R package.

First, we ran a simulation using different COVID-19 historical epidemic patient influx scenarios (H1, H2, H3, and H4) that were observed between February 2020 and February 2021 (Multimedia Appendix 1) [14]. Second, hypothetical patient influx scenarios were created using the susceptible-infectious-recovered (SIR) model for disease spread [15]. The total population calculated was fixed at 60,000, considering that the largest historical influx observed in South Korea was H4 (58,654 cumulative patients). We defined initial conditions at time t=0, S(0), I(0), and R(0), and I(0) and R(0) were fixed at 6 and 0, respectively. The recovery rate gamma was set at 0.05 because the average COVID-19 recovery time was 20.1 days [16]. The transmission rate beta ranged between 0.75 and 5 when generating influxes with different R0 (basic reproduction rate) levels. The number of newly confirmed patients per day was obtained from the SIR modeling data (Multimedia Appendix 2).

Probability Generation

Out-of-fold prediction results of the 10-fold cross-validation were aggregated to generate an empirical probability distribution of the disease severity probability. We used the results of Model 3 because of its high performance and its potential use in instances of limited diagnostic tools. Inverse transformation sampling was performed on the empirical probability distribution function, which was approximated using Gaussian kernel density estimation and linear interpolation [17]. The process was performed separately for severe and nonsevere patients, with sampled probabilities being randomly matched with generated patient influx rates while maintaining the prevalence of severe patients. The prediction probability distribution of the out-of-fold samples and the generated prediction probability distribution are presented in Multimedia Appendix 3.

Simulation Scenarios

Patients with a severe disease probability above the threshold are directed to the intensive care unit (ICU), with admission to this unit then being dependent on its current capacity. Rejected patients are directed to the general ward along with those who have a severe disease probability below the threshold. The probability of severe disease patients dying while in the ICU was 0.507, while it was 0.990 for those outside of the ICU [18]. We assumed that nonsevere patients would survive regardless of ICU admission. Patient deaths were categorized as follows: resource-independent deaths, wherein severe patients died despite ICU care (type I); resource-dependent deaths, wherein severe patients died due to ICU unavailability (type II); and threshold-dependent deaths, wherein severe patients died after being incorrectly classified as “nonsevere” and subsequently directed to the general ward (type III).

The maximum capacity of the ICU was established as 504 beds based on the number of isolation beds under negative pressure [14]. To estimate the distribution of length of stay, we used a previously suggested gamma distribution with a shape parameter of 1.5488 and a rate parameter of 0.1331 for those who died, and with a shape parameter of 0.8904 and a rate parameter of 0.0477 for those who survived to approximate the median and IQR [18,19]. Simulations were repeated 20 times for each influx scenario to ensure robustness.

Results

Patient Characteristics

Descriptive characteristics of the training and hold-out cohorts are provided in Tables 1 and 2. A total of 5330 (95.2%) patients exhibited nonsevere disease symptoms with an OSCI value <6, while 271 (4.8%) exhibited severe disease symptoms with an OSCI value ≥6.

Table 1. Demographic characteristics.

Variable	Total cohort (N=5601)		Training cohort (N=5037)		Hold-out cohort (N=564)		*P* value[a]
	Value, n (%) or mean (SD)	Missing data, %	Value, n (%) or mean (SD)	Missing data, %	Value, n (%) or mean (SD)	Missing data, %	
Age (years)		0.0%		0.0%		0.0%	.41
0-9	66 (1.2%)		61 (1.2%)		5 (0.9%)		
10-19	205 (3.7%)		185 (3.7%)		20 (3.6%)		
20-29	1110 (19.8%)		988 (19.6%)		122 (21.6%)		
30-39	564 (10.1%)		513 (10.2%)		51 (9.0%)		
40-49	739 (13.2%)		652 (12.9%)		87 (15.4%)		
50-59	1141 (20.4%)		1039 (20.6%)		102 (18.1%)		
60-69	907 (16.2%)		809 (16.1%)		98 (17.4%)		
70-79	545 (9.7%)		495 (9.8%)		50 (8.9%)		
≥80	324 (5.8%)		295 (5.9%)		29 (5.1%)		
Sex (male)	2310 (41.2%)	0.0%	2073 (41.2%)	0.0%	237 (42.0%)	0.0%	.73
BMI (kg/m^2)		21.4%		21.5%		20.9%	.65
<18.5	259 (4.6%)		236 (4.7%)		23 (4.1%)		
18.5-22.9	1854 (33.1%)		1666 (33.1%)		188 (33.3%)		
23.0-24.9	1035 (18.5%)		929 (18.4%)		106 (18.8%)		
25.0-29.9	1045 (18.7%)		938 (18.6%)		107 (19.0%)		
≥30	207 (3.7%)		185 (3.7%)		22 (3.9%)		
Medical history							
Diabetes mellitus	688 (12.3%)	0.1%	620 (12.3%)	0.1%	68 (12.1%)	0.0%	.92
Hypertension	1198 (21.4%)	0.1%	1087 (21.6%)	0.1%	111 (19.7%)	0.0%	.32
Heart failure	59 (1.1%)	0.1%	52 (1.0%)	0.1%	7 (1.2%)	0.0%	.81
Cardiovascular disease	179 (3.2%)	0.3%	156 (3.1%)	0.3%	23 (4.1%)	0.4%	.26
Asthma	128 (2.3%)	0.1%	118 (2.3%)	0.1%	10 (1.8%)	0.0%	.48
Chronic obstructive pulmonary disease	40 (0.7%)	0.1%	38 (0.8%)	0.1%	2 (0.4%)	0.0%	.43
Chronic kidney disease	55 (1.0%)	0.1%	48 (1.0%)	0.1%	7 (1.2%)	0.0%	.67
Malignancy	145 (2.6%)	0.1%	134 (2.7%)	0.1%	11 (2.0%)	0.0%	.39
Chronic liver disease	83 (1.6%)	5.8%	75 (1.6%)	5.7%	8 (1.5%)	6.7%	>.99
Autoimmune disease	38 (0.7%)	5.9%	32 (0.7%)	5.8%	6 (1.1%)	6.9%	.37
Dementia	224 (4.2%)	5.9%	203 (4.3%)	5.8%	21 (3.7%)	6.7%	.81

[a] Differences between groups were analyzed using the Welch *t* test for continuous variables, the Mann-Whitney *U* test for ordinal variables, the chi-square test for categorical variables with frequencies above 5, and the Fisher exact test for categorical variables with frequencies below 5. Two-sided *P* values are reported.

Table 2. Clinical characteristics.

Variable	Total cohort (N=5601)		Training cohort (N=5037)		Hold-out cohort (N=564)		*P* value[a]
	Value, n (%) or mean (SD)	Missing data, %	Value, n (%) or mean (SD)	Missing data, %	Value, n (%) or mean (SD)	Missing data, %	
Systolic blood pressure (mmHg)		2.5%		2.5%		2.7%	.60
<120	1306 (23.3%)		1177 (23.4%)		129 (22.9%)		
120-129	1138 (20.3%)		1012 (20.1%)		126 (22.3%)		
130-139	1084 (19.4%)		977 (19.4%)		107 (19.0%)		
140-159	1418 (25.3%)		1281 (25.4%)		137 (24.3%)		
≥160	513 (9.2%)		463 (9.2%)		50 (8.9%)		
Diastolic blood pressure (mmHg)		2.5%		2.5%		2.7%	.04
<80	2102 (37.5%)		1878 (37.3%)		224 (39.7%)		
80-89	1797 (32.1%)		1601 (31.8%)		196 (34.8%)		
90-99	1056 (18.9%)		971 (19.3%)		85 (15.1%)		
≥100	504 (9.0%)		460 (9.1%)		44 (7.8%)		
Heart rate (bpm)	85.8 (SD 15.1)	2.3%	85.8 (SD 15.0)	2.3%	86.3 (SD 15.4)	2.5%	.47
Body temperature (°C)	36.9 (SD 0.6)	0.7%	36.9 (SD 0.6)	0.8%	37.0 (SD 0.6)	0.7%	.86
Symptoms							
Fever	1302 (23.3%)	0.1%	1168 (23.2%)	0.1%	134 (23.8%)	0.0%	.80
Cough	2331 (41.6%)	0.1%	2103 (41.8%)	0.1%	228 (40.4%)	0.0%	.58
Sputum	1611 (28.8%)	0.1%	1460 (29.0%)	0.1%	151 (26.8%)	0.0%	.29
Sore throat	872 (15.6%)	0.1%	779 (15.5%)	0.1%	93 (16.5%)	0.0%	.57
Rhinorrhea	617 (11.0%)	0.1%	560 (11.1%)	0.1%	57 (10.1%)	0.0%	.51
Myalgia	920 (16.4%)	0.1%	820 (16.3%)	0.1%	100 (17.7%)	0.0%	.41
Fatigue	233 (4.2%)	0.1%	207 (4.1%)	0.1%	26 (4.6%)	0.0%	.65
Shortness of breath	665 (11.9%)	0.1%	608 (12.1%)	0.1%	57 (10.1%)	0.0%	.19
Headache	963 (17.2%)	0.1%	873 (17.3%)	0.1%	90 (16.0%)	0.0%	.45
Altered consciousness	35 (0.6%)	0.1%	31 (0.6%)	0.1%	4 (0.7%)	0.0%	.78
Vomiting	244 (4.4%)	0.1%	210 (4.2%)	0.1%	34 (6.0%)	0.0%	.05
Diarrhea	516 (9.2%)	0.1%	457 (9.1%)	0.1%	59 (10.5%)	0.0%	.32
Laboratory values							
Hemoglobin (g/dL)	13.3 (SD 1.8)	27.2%	13.3 (SD 1.8)	26.7%	13.2 (SD 1.8)	31.6%	.41
Hematocrit (%)	39.2 (SD 5.0)	27.2%	39.3 (SD 4.9)	26.7%	39.1 (SD 5.2)	31.7%	.56
Lymphocyte proportion (%)	29.2 (SD 11.7)	27.6%	29.3 (SD 11.7)	27.1%	28.2 (SD 11.0)	32.1%	.09
Platelet count (/μL)	236,697 (SD 82,897)	27.1%	236,776 (SD 82,534)	26.7%	235,943 (SD 86,395)	31.4%	.86
White blood cell count (/μL)	6126 (SD 2824)	27.1%	6121 (SD 2841)	26.7%	6167 (SD 2666)	31.4%	.75
WHO OSCI[b] ≥6	271 (4.8%)	0.0%	242 (4.8%)	0.0%	29 (5.1%)	0.0%	.80
Pregnancy	19 (0.3%)	0.4%	17 (0.3%)	0.3%	2 (0.4%)	0.5%	>.99
Pregnancy weeks	0.05 (SD 1.1)	0.4%	0.06 (SD 1.1)	0.4%	0.03 (SD 0.5)	0.5%	.40

[a] Differences between groups were analyzed using the Welch *t* test for continuous variables, the Mann-Whitney *U* test for ordinal variables, the chi-square test for categorical variables with frequencies above 5, and the Fisher exact test for categorical variables with frequencies below 5. Two-sided *P* values are reported.

[b] WHO OSCI: World Health Organization Ordinal Scale for Clinical Improvement.

Model Performance

The cross-validated AUROC values of the XGBoost and LR models were 0.965 (95% CI 0.958-0.972) and 0.938 (95% CI 0.911-0.959), respectively (*P*=.04). We chose the XGBoost model as our baseline Model 1 since it outperformed the LR model across all performance measures. Regarding the AUROC, we also examined XGBoost's outperformance across 4 different severity endpoints (Multimedia Appendix 4). An online clinical decision-support system based on Model 3 is provided for clinical implementation [20].

Model Interpretability

According to SHAP, age and lymphocyte count were the most important risk factors for predicting disease severity of OSCI ≥6 (Figure 1). Patient age, lymphocyte proportion, platelet count, BMI, hematocrit, and heart rate all exhibited nonlinear influences in predicting disease severity (Figure 2). In addition to the overall impact of each feature on the model's output, SHAP provides patient-specific influences of each variable on the predicted disease severity (Multimedia Appendix 5).

Figure 1. Relationships between each feature and Shapley additive explanations (SHAP) values. Summary plot in which each dot point represents the SHAP value of a patient in the data set used to construct the developed model. The dots are plotted for every feature used to fit the baseline model, excluding 2 features (pregnancy and number of weeks pregnant) that were not selected for the developed model. The SHAP values are displayed in rank order, based on their feature importance, along the y-axis as calculated by averaging the absolute SHAP values of each dot. A point's location on the x-axis shows its impact on the predictive output of the model. Purple indicates a relatively high feature value, while yellow represents a relatively low feature value. Grey dots represent missing values. COPD: chronic obstructive pulmonary disease; WBC: white blood cell.

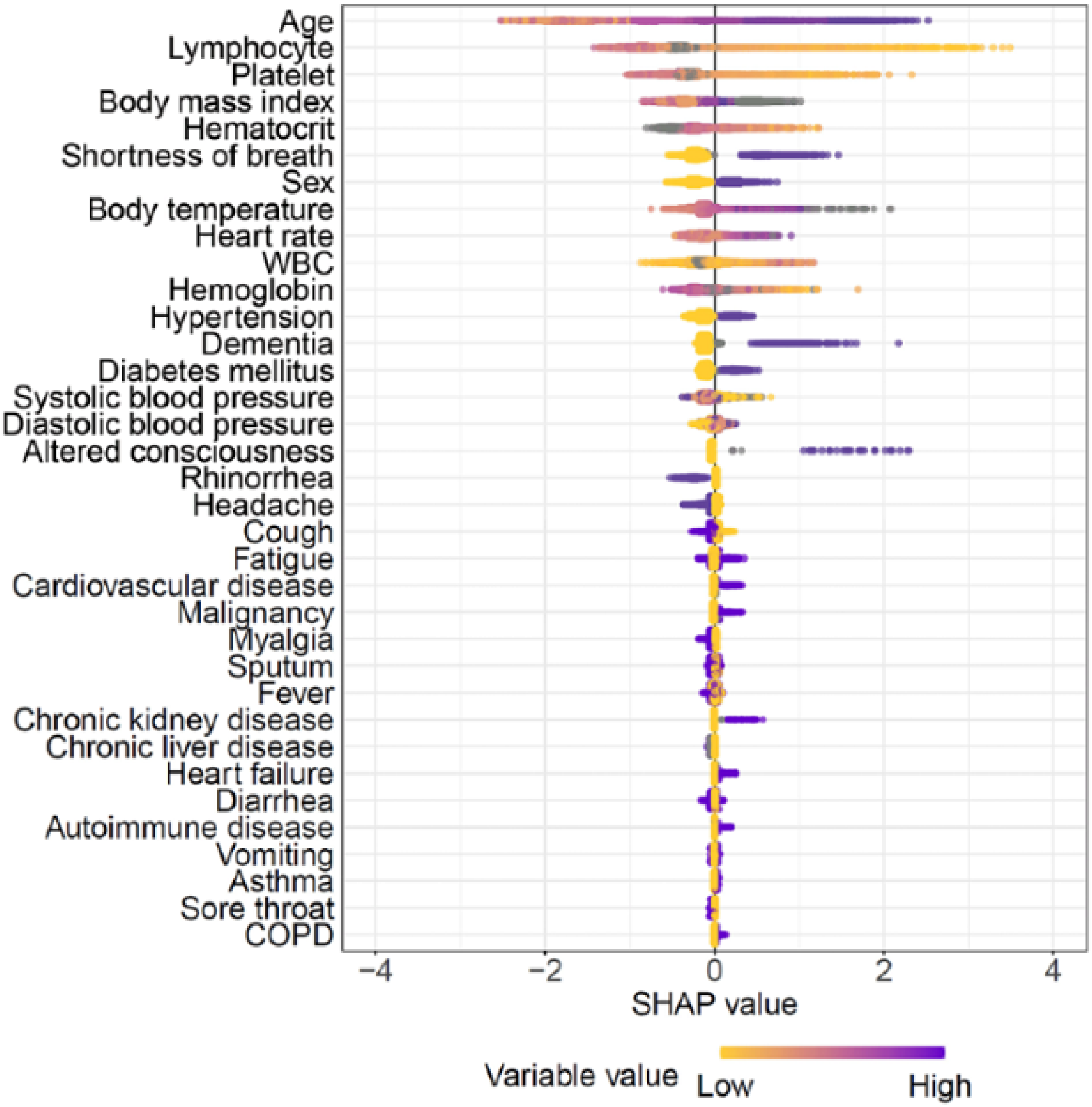
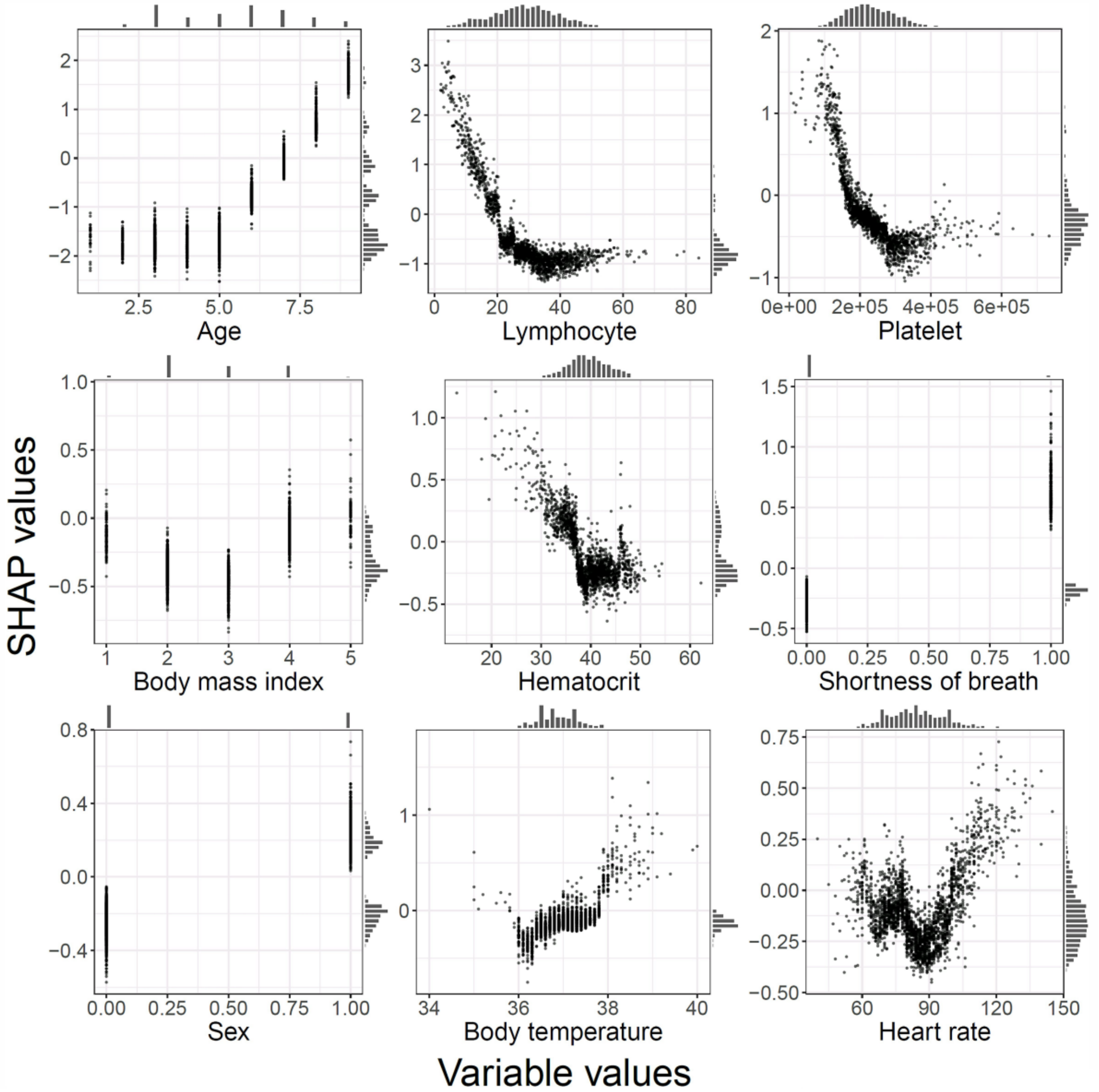


Figure 2. Relationships between each feature and Shapley additive explanations (SHAP) values. Dependence plots for each of the top 9 important features, including patient age, lymphocyte proportion, platelet count, BMI, hematocrit, shortness of breath, sex, body temperature, and heart rate. Each scatter plot shows the impact of each feature on the predictions made by the study model. The x-axis represents the variables' values, and the y-axis represents their SHAP values. The inflection points indicate the nonlinear impact of a feature on the model's prediction.



Predictive Performance Under Limited Data Availability

An AUROC of 0.965 (95% CI 0.958-0.972) was obtained with Model 1, which included all 37 variables. Notably, a reduction in its performance was found to be insignificant when 20 variables were eliminated, resulting in Model 3 (Multimedia Appendix 6 and Multimedia Appendix 7). Model 1 achieved both a sensitivity and specificity greater than 90%. Model 3 achieved a sensitivity of 88% and a PPV of 31% at the specificity level of 90%. Model 3 still outperformed the LR model regarding all performance measures.

An AUROC of 0.946 (95% CI 0.936-0.956) was obtained with Model 2, which included 32 variables. The reduction in performance was found to be insignificant when 21 variables were eliminated, resulting in Model 4 (Multimedia Appendix 7 and Multimedia Appendix 8). Models 2 and 4 achieved sensitivities of 84% and 81%, respectively, at a fixed specificity level of 90% (Table 3). Significant differences in AUROCs were observed when laboratory variables were excluded in these models, which implied that the laboratory variables had a solid discriminative power (all *P*≤.01).

Table 3. Comparison of model performance.

Model	Number of variables	AUROC[a], value (95% CI)	Specificity, value (95% CI)	Sensitivity, value (95% CI)	Accuracy, value (95% CI)	PPV[b], value (95% CI)	NPV[c], value (95% CI)
1	37	0.965 (0.958-0.972)	0.900 (0.892-0.909)	0.905 (0.868-0.942)	0.900 (0.892-0.908)	0.314 (0.295-0.335)	0.995 (0.993-0.997)
2	32	0.946 (0.936-0.956)	0.900 (0.891-0.908)	0.839 (0.793-0.884)	0.897 (0.888-0.905)	0.297 (0.276-0.319)	0.991 (0.988-0.994)
3	17	0.963 (0.955-0.971)	0.900 (0.892-0.908)	0.884 (0.839-0.921)	0.899 (0.891-0.907)	0.309 (0.289-0.329)	0.994 (0.991-0.996)
4	11	0.942 (0.931-0.953)	0.901 (0.892-0.909)	0.810 (0.756-0.860)	0.896 (0.888-0.904)	0.291 (0.270-0.313)	0.989 (0.987-0.992)

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

[b] PPV: positive predictive value.

[c] NPV: negative predictive value.

The AUROCs of Models 1 and 2 for the held-out cohorts were 0.958 (95% CI 0.924-0.991) and 0.943 (95% CI 0.901-0.985), respectively, which were both indifferent from the cross-validation results (*P*=.66 and *P*=.89, respectively). The AUROCs of Models 3 and 4 for the held-out cohorts were 0.949 (95% CI 0.906-0.990) and 0.941 (95% CI 0.903-0.978), respectively, and were also indifferent from the cross-validation results (*P*=.54 and *P*=.95, respectively). The indifferences between the cross-validation and hold-out results revealed that all models had a degree of generalizability to unseen data (Multimedia Appendix 9). Detailed results and the selected variables used at each step of the RFE are presented in Multimedia Appendix 7 and Multimedia Appendix 10.

Optimal Triage Under Limited Resource Availability

The overall DES workflow is illustrated in Figure 3. Mortality rates were minimized at thresholds of 0.1, 0.01, 0.04, and 0.24 for H1, H2, H3, and H4, respectively (Multimedia Appendix 11). The mortality rates showed a convex shape in accordance with these thresholds (Multimedia Appendix 12).

We can infer that as the death rate increases, the threshold should be raised when a large increase is accompanied. While the association between mortality rates and triage thresholds across various patient influx scenarios is inferable through an analysis of historical influx data, it is impractical to draw general conclusions from this information. For example, looking at Multimedia Appendix 11, an upward trend in the optimal threshold and optimized mortality rate occurred when comparing H2, H3, and H4, wherein there was a clear increase in the patient influx volume. However, it is difficult to infer this information when comparing H1 with H3 or H4 because of differences in their multidimensional characteristics, including duration, maximum daily patients, and cumulative patients. To further support our results, we performed additional simulations using patient flow data that were generated using the SIR model with varying R0s.

The DES using hypothetical patient influxes revealed that the optimal threshold ranged from 0.02 to 0.66, while the respective minimized mortality rates ranged from 0.017 (1.7%) to 0.042 (4.2%) (Multimedia Appendix 13). The optimal threshold values and minimized mortality rates for each R0 showed that a larger R0 value tends to result in increases in both of these variables. The optimal threshold is increased along with the R0 values to increase precision for severe patients while fully utilizing the ICU. The optimized mortality rates were increased due to an increased proportion of deaths outside the ICU resulting from a larger volume of patient influx. The benefits of utilizing an optimal triage threshold were clear when compared with the conventional Youden Index (J-index) as a benchmark value, which was 0.013. Decreased mortality rates ([J-index mortality rate – optimized mortality rate] / J-index mortality rate) were notably large in a magnitude ranging from 6.1% to 18.1% (Figure 4). Detailed data are provided in Table 4.

Figure 3. Simulation workflow. Diagram showing how medical resources can be allocated among COVID-19 patients according to the machine learning–based triage system. Patients with a prediction probability exceeding a certain threshold are first triaged to an intensive care unit (ICU) that is currently under its total capacity. Conversely, patients are directed to a general ward if the ICU's capacity is full or if their severity prediction probability is lower than the threshold. Type I deaths represent those occurring in the ICU. Type II and III deaths represent those of patients who have been directed to the general ward due to ICU unavailability and because they were found to have a disease severity probability lower than the threshold, respectively. We used simulations to obtain the optimal threshold wherein the mortality rate (n [total deaths] / n [total patients] = n [type I death + type II death + type III death] / n [total patients]) is minimized.

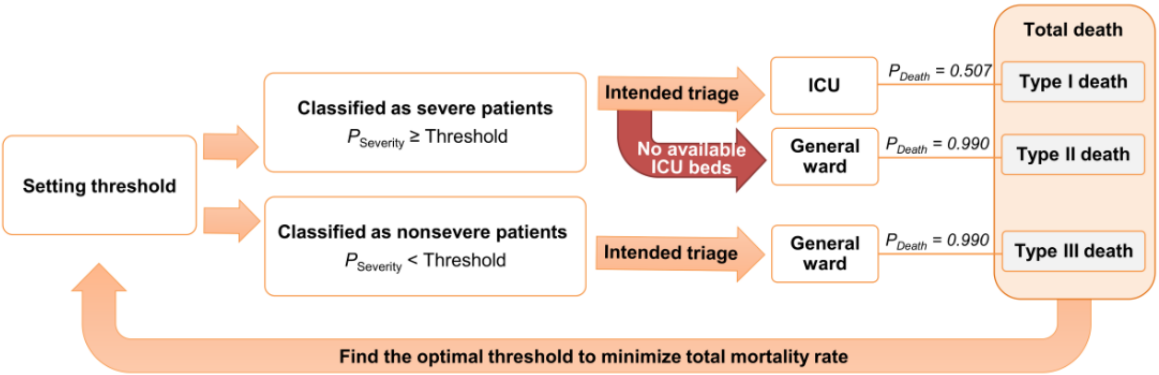


Figure 4. Optimized results of the patient triage simulations for hypothetical influx. Decreased mortality rate = (J-index mortality rate − optimized mortality rate) / J-index mortality rate.

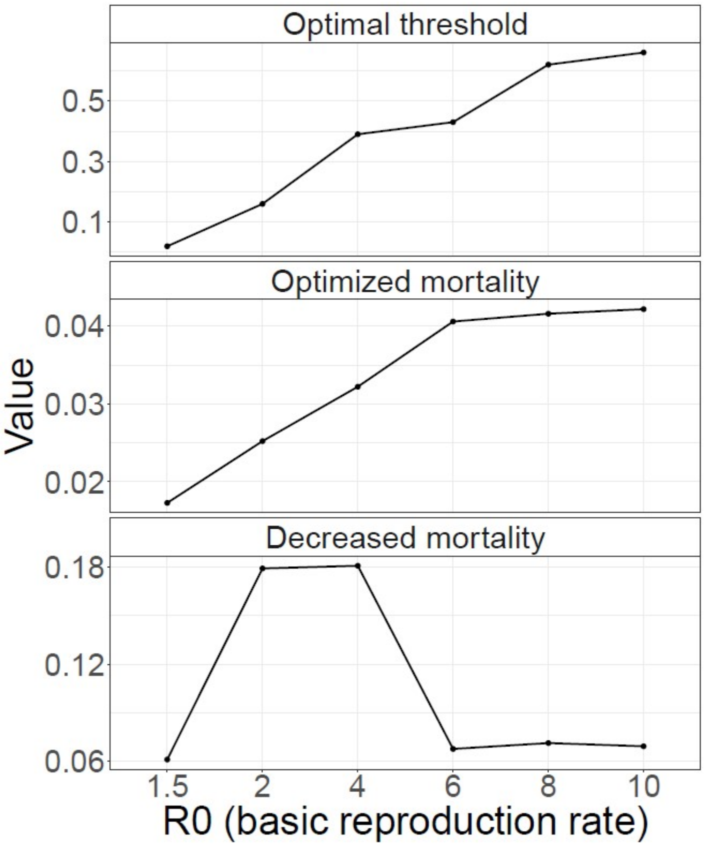


Table 4. Optimized threshold and its benefits on mortality outcomes according to patient influx settings.

Influx	Optimal threshold	Optimized mortality rate	Decreased mortality rate[a]
H[b]1	0.10	0.022	0.298
H2	0.01	0.015	0.047
H3	0.04	0.019	0.146
H4	0.24	0.031	0.209
R0[c]=1.5	0.02	0.017	0.061
R0=2	0.16	0.025	0.179
R0=4	0.39	0.032	0.181
R0=6	0.43	0.041	0.068
R0=8	0.62	0.042	0.071
R0=10	0.66	0.042	0.069

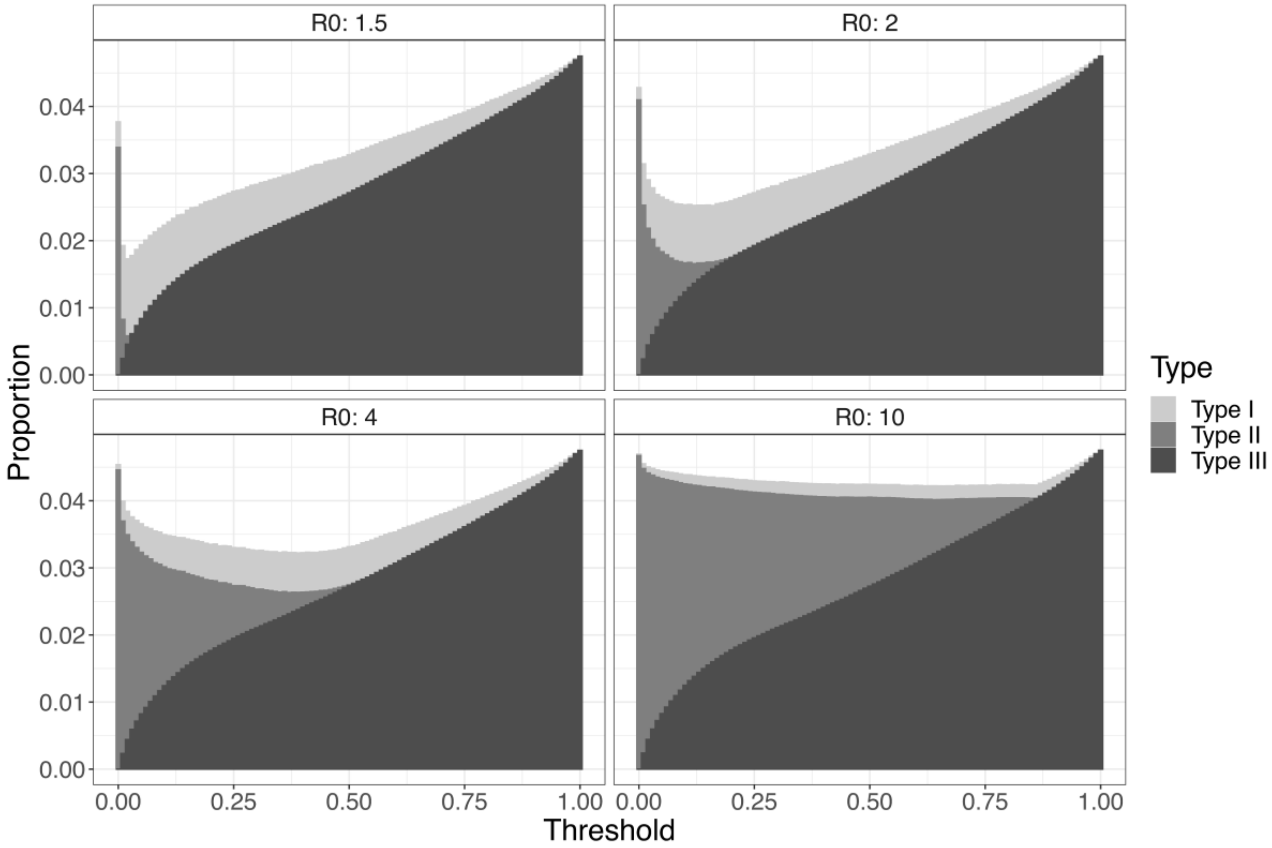
[a] Decreased mortality rate: (J-index mortality rate – optimized mortality rate) / J-index mortality rate.

[b] H: historical epidemic patient influx scenario.

[c] R0: basic reproduction rate.

We observed a convex relationship for mortality rates in accordance with the thresholds in Figure 5. The mortality rate was minimized at a point where type I death, which had the lowest P_{death} (50.7%), was maximized in proportion to total death. For example, when R0 was 1.5, the proportion of type I deaths was maximized at the optimal threshold, accounting for 66.4% of all deaths. However, a threshold that is too low leads to inadequate capacity exhaustion with misclassified nonsevere patients. Consequently, the resulting limited capacity for actual severe patients then decreases the proportion of type I deaths and increases those of type II deaths. Conversely, a threshold that is too high would result in unnecessary rejection for severe patients, which then decreases the proportion of type I deaths and increases those of type III deaths.

Figure 5. Mortality rates in hypothetical patient influxes are decomposed by death subtype at each threshold. The x-axis represents the threshold, and the y-axis represents the stacked proportion of each death subtype to the total number of patients, calculated as n (death subtype) / n (total patients) at each threshold. R0: basic reproduction rate.



In situations of excessively high R0 values and increased ICU demand, increasing the triage threshold to reject more patients will still deplete the ICU capacity. Therefore, adjusting the threshold will mostly result in trade-offs between the numbers of threshold- and capacity-dependent rejections, limiting the influence of threshold adjustment on minimizing patient mortality. In situations of sufficiently low R0 values, the effect of threshold optimization is reduced along with its necessity. Nonetheless, the large reduction in mortality rates among the remaining influxes highlights the substantial benefits of optimizing the patient triage threshold under resource constraints.

Code Availability

The code used to develop and evaluate this study's models is available online [21].

Discussion

Principal Findings

A distinctive feature of our Model 1 is its high discriminative power with an AUROC that exceeded 0.97 in both cross-validation and hold-out settings. Previous prediction models for determining the clinical deterioration of COVID-19 patients have reported predictive accuracies ranging from 0.77 to 0.91 [2-5]. Additionally, these models require specific diagnostic data, including laboratory data, peripheral oxygen saturation, or radiographic findings, to maintain their predictive accuracies. Moreover, to what extent the performance abilities of these models are maintained during the partial absence of data has not been studied. Given this unmet clinical need, we developed Model 1. In addition, we confirmed that our feature-eliminated models maintained an adequate discriminative power even in the partial absence of data. The advantages of our feature-eliminated models include not only their increased generalizability to unseen data, but also their applicability within scenarios wherein there is limited medical data. We have uploaded Model 3 online to be implemented in clinical practice. Given the acute exacerbation of pneumonia in COVID-19 patients, our model can also be used to re-evaluate hospitalized patients in the short term, so that individuals whose clinical manifestations are likely to worsen can be identified as early as possible [22].

A noteworthy feature of our model is its ability to discriminate between patient-specific factors contributing to disease exacerbation and their individual contributions using SHAP values. Current COVID-19 treatment guidelines provide recommendations based on the average-risk patient under limited available insights into their disease stage [10]. These recommendations provide a one-size-fits-all approach to all patients, which is problematic for those with more complex or atypical disease presentations. Our model obviates the need for arbitrary patient risk groupings and is therefore useful in maximizing survival odds based on individual risk stratification. Furthermore, our model can be integrated into electronic medical record systems, which utilize coding algorithms, as a notification system that helps in the early identification of disease exacerbation risk factors.

The validity of our model is supported by the high consistency between the results of its interpretation using SHAP and previously reported prognosticators of COVID-19 severity [23-28]. We noted that old age, followed by lymphopenia and thrombocytopenia, exhibited the highest Shapley values for disease exacerbation. We presume that age interacts with relevant features in older adults, including poor functional performance and increased frailty, which are associated with adverse outcomes and increased mortality among patients with respiratory syndromes [29]. Our findings also support literature indicating that lymphopenia plays an important role in COVID-19 exacerbation [25-28]. Lymphopenia is characterized by the lowering of lymphocytes due to injured alveolar epithelial cells and is commonly observed in COVID-19 patients [30]. Consistent with previous studies, thrombocytopenia was also found to be associated with adverse COVID-19 outcomes [26,31]. It has been suggested that a reduction or morphological alternation in the pulmonary capillary bed exerts pathological platelet defragmentation because the lungs are a platelet release site with mature megakaryocytes [32]. Our prediction model supports the notion that the early identification of COVID-19 infection, before a hematological crisis occurs, is necessary for ensuring a better prognosis.

There is no existing study that has examined COVID-19 severity prediction models in an attempt to provide an explicit solution for the delivery of optimal triage using threshold modification that accounts for limited resource availability. We employed DES in our Model 3 to examine discrimination thresholds that are usable in an adaptive manner across various patient influx scenarios and the related health care resource availability. Our simulations revealed that applying the optimal thresholds of both historical and generated patient influxes will minimize the mortality rate of each patient influx scenario. Our hypothesis is supported by the significant differences found in mortality rates between the J-index and our optimized thresholds when applied to the expected patient influx volumes. This observation supports the potential usability of our model to substantially reduce COVID-19 mortality rates through the appropriate and effective adjustment of triage thresholds.

Limitations

One limitation of our study is its incorporation of a single national cohort of Asian ethnicity with a relatively small sample size, which impacts the generalizability of our findings. External validation using a more multiethnic population is thus needed to determine if a similar discrimination performance occurs among other ethnic groups. However, to ensure our model's robustness, we implemented 10-fold cross-validation with additional confirmation using a hold-out cohort. Second, the triage threshold was evaluated using a simulation. Simulations do not yield concrete answers and are unable to assess all kinds of potential situations [33]. Third, the applicability of utilizing SHAP values to discriminate patient-specific contributing factors for disease exacerbation has not been prospectively validated. Whether the early identification of disease exacerbation risk factors and their individual contributions can result in a better prognosis would need to be validated after the implementation of our online system into clinical practice. Lastly, clinical data, including self-reported measurements, may not be objectively interpreted, and models utilizing these parameters should be interpreted cautiously.

Conclusions

We developed and validated a robust prediction model, with an explanatory feature, that offers an effective means of enhancing the efficiency of COVID-19 triage. We further proposed an adaptive triage model that utilizes both patient influx volume and the capacity of a health care system to minimize mortality rates within the scope of resource limits. Our model has the potential for effective application because it is available online for patients and providers in both inpatient and outpatient settings. Overall, our results imply that COVID-19 treatment plans need to integrate both medical and health care management expertise to guarantee maximum efficacy.

Acknowledgments

We would like to thank the Korea Disease Control and Prevention Agency (KDCA) and the Health Information Managers in the participating hospitals for their assistance in collecting the medical records utilized in this study. This study was supported through the Infection Prevention Strategy Development Program of Korea (HW20C2103). The funding source had no role in the study design, analyses, or interpretation of the data; in the writing of the manuscript; or in the decision to submit the article for publication.

Authors' Contributions

JMK, HKL, and KCK contributed to the conception and writing of the original draft. JMK, HKL, KHL, and KSL contributed to the acquisition, analysis, and interpretation of the study data. JHA and KCK contributed to the review, editing, and supervision of this study. All authors have read and approved the submitted version of this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Epidemic incidence curves of historical patient influxes of COVID-19 in South Korea.
[DOCX File , 84 KB - medinform_v9i11e32726_app1.docx]

Multimedia Appendix 2
Susceptible-infectious-recovered (SIR) simulated patient influx.
[DOCX File , 77 KB - medinform_v9i11e32726_app2.docx]

Multimedia Appendix 3
Prediction probability distribution graph of patients in the out-of-fold samples.
[DOCX File , 88 KB - medinform_v9i11e32726_app3.docx]

Multimedia Appendix 4
Performance of the models according to the World Health Organization Ordinal Scale for Clinical Improvement.
[DOCX File , 31 KB - medinform_v9i11e32726_app4.docx]

Multimedia Appendix 5
Patient-specific Shapley additive explanations (SHAP) plots.
[DOCX File , 204 KB - medinform_v9i11e32726_app5.docx]

Multimedia Appendix 6
Changes to the model’s performance after applying recursive feature elimination (RFE) (Model 1).
[DOCX File , 81 KB - medinform_v9i11e32726_app6.docx]

Multimedia Appendix 7
Areas under the receiver operating characteristic curve (AUROCs) at each step of the recursive feature elimination (RFE) and <italic>P</italic> values of differences in AUROC values for Models 1 and 2.
[DOCX File , 32 KB - medinform_v9i11e32726_app7.docx]

Multimedia Appendix 8
Changes to the model’s performance after applying recursive feature elimination (RFE) (Model 2).
[DOCX File , 73 KB - medinform_v9i11e32726_app8.docx]

Multimedia Appendix 9
Performance of the models in the hold-out cohort.
[DOCX File , 29 KB - medinform_v9i11e32726_app9.docx]

Multimedia Appendix 10
Order of feature importance for each model.
[DOCX File , 33 KB - medinform_v9i11e32726_app10.docx]

Multimedia Appendix 11
Optimized results of the patient triage simulations for the historical influx.
[DOCX File , 80 KB - medinform_v9i11e32726_app11.docx]

Multimedia Appendix 12
Mortality rates of the historical patient influx scenarios according to each threshold and at each threshold across different scenarios.
[DOCX File , 90 KB - medinform_v9i11e32726_app12.docx]

Multimedia Appendix 13
Mortality rates of the hypothetical patient influx scenarios according to each threshold and at each threshold across different scenarios.
[DOCX File , 84 KB - medinform_v9i11e32726_app13.docx]

References

1. Critical preparedness, readiness and response actions for COVID-19: interim guidance, 7 March 2020. World Health Organization. 2020. URL: https://apps.who.int/iris/handle/10665/331422 [accessed 2021-06-01]
2. Gupta RK, Harrison EM, Ho A, Docherty AB, Knight SR, van Smeden M, ISARIC4C Investigators. Development and validation of the ISARIC 4C Deterioration model for adults hospitalised with COVID-19: a prospective cohort study. Lancet Respir Med 2021 Apr;9(4):349-359 [FREE Full text] [doi: 10.1016/S2213-2600(20)30559-2] [Medline: 33444539]
3. Mejía-Vilet JM, Córdova-Sánchez BM, Fernández-Camargo DA, Méndez-Pérez RA, Morales-Buenrostro LE, Hernández-Gilsoul T. A Risk Score to Predict Admission to the Intensive Care Unit in Patients with COVID-19: the ABC-GOALS score. Salud Publica Mex 2020 Dec 22;63(1, ene-feb):1-11. [doi: 10.21149/11684] [Medline: 33021362]
4. Ji D, Zhang D, Xu J, Chen Z, Yang T, Zhao P, et al. Prediction for Progression Risk in Patients With COVID-19 Pneumonia: The CALL Score. Clin Infect Dis 2020 Sep 12;71(6):1393-1399 [FREE Full text] [doi: 10.1093/cid/ciaa414] [Medline: 32271369]
5. Knight SR, Ho A, Pius R, Buchan I, Carson G, Drake TM, ISARIC4C investigators. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. BMJ 2020 Sep 09;370:m3339 [FREE Full text] [doi: 10.1136/bmj.m3339] [Medline: 32907855]
6. Gao Y, Cai G, Fang W, Li H, Wang S, Chen L, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. Nat Commun 2020 Oct 06;11(1):5033 [FREE Full text] [doi: 10.1038/s41467-020-18684-2] [Medline: 33024092]
7. Bolourani S, Brenner M, Wang P, McGinn T, Hirsch JS, Barnaby D, Northwell COVID-19 Research Consortium. A Machine Learning Prediction Model of Respiratory Failure Within 48 Hours of Patient Admission for COVID-19: Model Development and Validation. J Med Internet Res 2021 Feb 10;23(2):e24246 [FREE Full text] [doi: 10.2196/24246] [Medline: 33476281]
8. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 2002;46:389-422 [FREE Full text] [doi: 10.1023/A:1012487302797]
9. Karnon J, Stahl J, Brennan A, Caro JJ, Mar J, Möller J, ISPOR-SMDM Modeling Good Research Practices Task Force. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--4. Value Health 2012 Sep;15(6):821-827 [FREE Full text] [doi: 10.1016/j.jval.2012.04.013] [Medline: 22999131]
10. World Health Organization. 2020. URL: https://www.who.int/blueprint/priority-diseases/key-action/COVID-19_Treatment_Trial_Design_Master_Protocol_synopsis_Final_18022020.pdf [accessed 2021-06-01]
11. 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, California, U.S.A p. 785-794. [doi: 10.1145/2939672.2939785]
12. Lundberg S, Lee S. A unified approach to interpreting model predictions. 2017 Presented at: 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, California, USA.
13. Pidd M. Computer Simulation in Management Science. Hoboken, New Jersey, USA: John Wiley and Sons; 2006.

14. Infectious Disease Portal. Korea Disease Control and Prevention Agency. URL: http://www.kdca.go.kr/npt/biz/npp/portal/nppIssueIcdMain.do [accessed 2021-06-01]
15. Hethcote H. The Mathematics of Infectious Diseases. SIAM Rev 2000 Jan;42(4):599-653 [FREE Full text] [doi: 10.1137/s0036144500371907]
16. Lee Y, Hong CM, Kim DH, Lee TH, Lee J. Clinical Course of Asymptomatic and Mildly Symptomatic Patients with Coronavirus Disease Admitted to Community Treatment Centers, South Korea. Emerg Infect Dis 2020 Oct;26(10):2346-2352 [FREE Full text] [doi: 10.3201/eid2610.201620] [Medline: 32568662]
17. Caflisch RE. Monte Carlo and quasi-Monte Carlo methods. Acta Numerica 2008 Nov 07;7:1-49. [doi: 10.1017/s0962492900002804]
18. Wood RM, Pratt AC, Kenward C, McWilliams CJ, Booton RD, Thomas MJ, et al. The Value of Triage during Periods of Intense COVID-19 Demand: Simulation Modeling Study. Med Decis Making 2021 May;41(4):393-407. [doi: 10.1177/0272989X21994035] [Medline: 33560181]
19. COVID-19 Report. ICNARC. URL: https://www.icnarc.org/Our-Audit/Audits/Cmp/Reports [accessed 2021-06-01]
20. COVID-19 Patient Severity Prediction Using XGBoost. COVID-19 Severity. URL: http://covid19severity.duckdns.org/ [accessed 2021-10-22]
21. Study of optimal triage for severe COVID-19 patients to minimize mortality rate. GitHub. URL: https://github.com/minkim88/Optimal-Triage-COVID-19 [accessed 2021-10-22]
22. Haimovich AD, Ravindra NG, Stoytchev S, Young HP, Wilson FP, van Dijk D, et al. Development and Validation of the Quick COVID-19 Severity Index: A Prognostic Tool for Early Clinical Decompensation. Ann Emerg Med 2020 Oct;76(4):442-453 [FREE Full text] [doi: 10.1016/j.annemergmed.2020.07.022] [Medline: 33012378]
23. Ye J, Zhang X, Zhu F, Tang Y. Application of a prediction model with laboratory indexes in the risk stratification of patients with COVID-19. Exp Ther Med 2021 Mar 05;21(3):182 [FREE Full text] [doi: 10.3892/etm.2021.9613] [Medline: 33488791]
24. Hu L, Chen S, Fu Y, Gao Z, Long H, Ren HW, et al. Risk Factors Associated With Clinical Outcomes in 323 Coronavirus Disease 2019 (COVID-19) Hospitalized Patients in Wuhan, China. Clin Infect Dis 2020 Nov 19;71(16):2089-2098 [FREE Full text] [doi: 10.1093/cid/ciaa539] [Medline: 32361738]
25. Lassau N, Ammari S, Chouzenoux E, Gortais H, Herent P, Devilder M, et al. Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. Nat Commun 2021 Jan 27;12(1):634 [FREE Full text] [doi: 10.1038/s41467-020-20657-4] [Medline: 33504775]
26. Lippi G, Plebani M, Henry B. Thrombocytopenia is associated with severe coronavirus disease 2019 (COVID-19) infections: A meta-analysis. Clin Chim Acta 2020 Jul;506:145-148 (forthcoming) [FREE Full text] [doi: 10.1016/j.cca.2020.03.022] [Medline: 32178975]
27. Henry BM, Lippi G. Chronic kidney disease is associated with severe coronavirus disease 2019 (COVID-19) infection. Int Urol Nephrol 2020 Jun 28;52(6):1193-1194 [FREE Full text] [doi: 10.1007/s11255-020-02451-9] [Medline: 32222883]
28. Ruan Q, Yang K, Wang W, Jiang L, Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. Intensive Care Med 2020 May 3;46(5):846-848 [FREE Full text] [doi: 10.1007/s00134-020-05991-x] [Medline: 32125452]
29. Wu G, Yang P, Xie Y, Woodruff HC, Rao X, Guiot J, et al. Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study. Eur Respir J 2020 Aug;56(2):2001104 [FREE Full text] [doi: 10.1183/13993003.01104-2020] [Medline: 32616597]
30. Chan JF, Yuan S, Kok K, To KK, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The Lancet 2020 Feb 15;395(10223):514-523 [FREE Full text] [doi: 10.1016/S0140-6736(20)30154-9] [Medline: 31986261]
31. Goodall JW, Reed TAN, Ardissino M, Bassett P, Whittington AM, Cohen DL, et al. Risk factors for severe disease in patients admitted with COVID-19 to a hospital in London, England: a retrospective cohort study. Epidemiol Infect 2020 Oct 13;148:e251 [FREE Full text] [doi: 10.1017/S0950268820002472] [Medline: 33046155]
32. Yang M, Ng MH, Li CK. Thrombocytopenia in patients with severe acute respiratory syndrome (review). Hematology 2005 Apr 04;10(2):101-105. [doi: 10.1080/10245330400026170] [Medline: 16019455]
33. Tulsian P, Pandey V. Quantitative Techniques: Theory and Problems. Delhi, India: Pearson India; 2006.

Abbreviations

AUROC: area under the receiver operating characteristic curve
DES: discrete-event simulation
ICU: intensive care unit
KDCA: Korea Disease Control and Prevention Agency
LR: logistic regression
NPV: negative predictive value
OSCI: Ordinal Scale for Clinical Improvement
PPV: positive predictive value

R0: basic reproduction rate
RFE: recursive feature elimination
SHAP: Shapley additive explanations
SIR: susceptible-infectious-recovered
WHO: World Health Organization
XGBoost: extreme gradient boosting

Edited by G Eysenbach; submitted 10.08.21; peer-reviewed by M Salimi, X Cheng, R Sampath Kumaran; comments to author 31.08.21; revised version received 17.09.21; accepted 03.10.21; published 02.11.21.

Please cite as:
Kim J, Lim H, Ahn JH, Lee KH, Lee KS, Koo KC
Optimal Triage for COVID-19 Patients Under Limited Health Care Resources With a Parsimonious Machine Learning Prediction Model and Threshold Optimization Using Discrete-Event Simulation: Development Study
JMIR Med Inform 2021;9(11):e32726
URL: https://medinform.jmir.org/2021/11/e32726
doi:10.2196/32726
PMID:34609319

Original Paper

Assessing the Performance of a New Artificial Intelligence–Driven Diagnostic Support Tool Using Medical Board Exam Simulations: Clinical Vignette Study

Niv Ben-Shabat[1,2,3], MD, MPH; Ariel Sloma[1,3], MD; Tomer Weizman[3,4], MD; David Kiderman[5], MD; Howard Amital[1,2,6], MD, MHA

[1]Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel

[2]Department of Medicine 'B', Sheba Medical Center, Ramat Gan, Israel

[3]Kahun Medical Ltd, Tel-Aviv, Israel

[4]The Rappaport Faculty of Medicine, Technion Israel Institute of Technology, Haifa, Israel

[5]Hadassah Faculty of Medicine, The Hebrew University, Jerusalem, Israel

[6]The Zabludowicz Center for Autoimmune Diseases, Sheba Medical Center, Ramat Gan, Israel

Corresponding Author:
Niv Ben-Shabat, MD, MPH
Department of Medicine 'B'
Sheba Medical Center
Sheba Road 2
Ramat Gan, 52621
Israel
Phone: 972 3 530 2652
Fax: 972 3 535 4796
Email: nivben7@gmail.com

Abstract

Background: Diagnostic decision support systems (DDSS) are computer programs aimed to improve health care by supporting clinicians in the process of diagnostic decision-making. Previous studies on DDSS demonstrated their ability to enhance clinicians' diagnostic skills, prevent diagnostic errors, and reduce hospitalization costs. Despite the potential benefits, their utilization in clinical practice is limited, emphasizing the need for new and improved products.

Objective: The aim of this study was to conduct a preliminary analysis of the diagnostic performance of "Kahun," a new artificial intelligence-driven diagnostic tool.

Methods: Diagnostic performance was evaluated based on the program's ability to "solve" clinical cases from the United States Medical Licensing Examination Step 2 Clinical Skills board exam simulations that were drawn from the case banks of 3 leading preparation companies. Each case included 3 expected differential diagnoses. The cases were entered into the Kahun platform by 3 blinded junior physicians. For each case, the presence and the rank of the correct diagnoses within the generated differential diagnoses list were recorded. Each diagnostic performance was measured in two ways: first, as diagnostic sensitivity, and second, as case-specific success rates that represent diagnostic comprehensiveness.

Results: The study included 91 clinical cases with 78 different chief complaints and a mean number of 38 (SD 8) findings for each case. The total number of expected diagnoses was 272, of which 174 were different (some appeared more than once). Of the 272 expected diagnoses, 231 (87.5%; 95% CI 76-99) diagnoses were suggested within the top 20 listed diagnoses, 209 (76.8%; 95% CI 66-87) were suggested within the top 10, and 168 (61.8%; 95% CI 52-71) within the top 5. The median rank of correct diagnoses was 3 (IQR 2-6). Of the 91 expected diagnoses, 62 (68%; 95% CI 59-78) of the cases were suggested within the top 20 listed diagnoses, 44 (48%; 95% CI 38-59) within the top 10, and 24 (26%; 95% CI 17-35) within the top 5. Of the 91 expected diagnoses, in 87 (96%; 95% CI 91-100), at least 2 out of 3 of the cases' expected diagnoses were suggested within the top 20 listed diagnoses; 78 (86%; 95% CI 79-93) were suggested within the top 10; and 61 (67%; 95% CI 57-77) within the top 5.

Conclusions: The diagnostic support tool evaluated in this study demonstrated good diagnostic accuracy and comprehensiveness; it also had the ability to manage a wide range of clinical findings.

(JMIR Med Inform 2021;9(11):e32507) doi:10.2196/32507

KEYWORDS

diagnostic decision support systems; diagnostic support; medical decision-making; medical informatics; artificial intelligence; Kahun; decision support

Introduction

Background

Diagnostic decision support systems (DDSS) are computer programs that aim to improve healthcare and minimize diagnostic errors by supporting healthcare professionals in the process of diagnostic decision-making [1-3]. These processes, both in general and specifically in medicine, are influenced by cognitive biases [4,5], difficulty estimating pre- or posttest probabilities [6,7], and the experience level of the caregiver [8]. The currently available DDSS vary greatly in terms of knowledge base source and curation, algorithmic complexity, available features, and user interface [9-12]. However, all DDSS generally work by providing diagnostic suggestions based on a patient's specific data. Previous studies have demonstrated the ability of DDSS to enhance clinicians' diagnostic skills [2,3,13,14], prevent diagnostic errors [14], and reduce hospitalization costs [15]. However, no effect regarding patient-related outcomes has been reported yet [16,17]. Despite the potential benefits of DDSS and the fact that the first products were introduced decades ago [1,10,11,18], they are not yet widely accepted in the medical community and are not used routinely in clinical practice [17,19]. The factors proposed to be responsible for this state include negative perceptions and biases of practitioners, poor accuracy of the available tools, inherent tendency to prefer sensitivity over specificity, lack of standardized nomenclature, and poor usability and integration into the practitioner's workflow [16,19-22]. These facts emphasize the need for new products harnessing recent advances in the data science field.

About the Diagnostic Support System Evaluated

In this study, we evaluated the diagnostic performance of Kahun (Kahun Medical Ltd), a new diagnostic support tool for healthcare practitioners, freely available to use online or as a mobile app. Kahun enables users to input a wide range of findings concerning their patients and, in turn, generates: (1) a differential-diagnoses (DDX) list, ranked according to likelihood; (2) stronger and weaker findings alongside a graph of clinical associations for each suggested diagnosis, all with direct references; and (3) further options for diagnostic workup with evidence-based justifications aimed to refine the DDX, to exclude life-threatening cases, and to reach a definitive diagnosis. A video demonstrating the use of the platform for a standard patient is presented in Multimedia Appendix 1. A series of step-by-step screenshots portraying the different panels and functions of the mobile app is presented in Multimedia Appendix 2.

Kahun's knowledge base is a structured, scalable, quantitative knowledge graph designed to model both ontological and empirical medical knowledge as they appear in evidence-based literature. To combine aspects of semantic knowledge graphs with empirical and probabilistic relationships, Kahun adopts the techniques of causal graphs and probabilistic graphing models. The platform's sources of knowledge include core clinical journals and established medical textbooks of internal medicine, as well as ontological poly-hierarchies such as the Systematized Nomenclature of Medicine (SNOMED) and the Logical Observation Identifiers Names and Codes (LOINC) [23]. Each data point is referenced back to the original source, thus enabling the assignment of different weights for each data point according to the strength of evidence of its source. Data from these sources are curated using a model that transforms textual representations into structured interconnections between medical concepts found in the text; these connections point to the specific cohorts and cite the statistical metrics provided by the source. The knowledge base is continuously being updated and growing all the time. It currently contains over 10,000 concepts, alongside 20,000,000 facts and metrics cataloged from over 50,000 referenced sources.

Given a set of findings, the Kahun core algorithm processes information from the structured knowledge base to support the clinical reasoning process. The goal of the algorithm is to highlight all relevant knowledge in the context of a specific patient. Hence, the system is always dealing with a "cohort of one," meaning a cohort representing patients that match all known attributes of the presented patient. The algorithm can synthesize and transform metrics, where valid (eg, using published sensitivity and likelihood ratio to compute the specificity of a test). Most often, metrics must be estimated despite missing data in the literature. In such cases, the algorithm will estimate probabilities, which are an extension of existing facts and in harmony with other published metrics. The transparency at the heart of the knowledge graph allows all such estimates to be explained, using clinical reasoning, and referenced back to their sources. The Kahun system goes through a constant process of quality assurance, carried out by a combination of medical experts and automated tools. Internal tools provide an on-demand view of knowledge per medical concept (eg, disease, clinical finding, and more), and test reports are produced for the clinical reasoning given patient presentations. Both are tested continuously against data sets of medical cases.

Objectives

The goal of this study was to test the diagnostic accuracy of Kahun in terms of its ability to suggest the expected diagnosis in a series of cases from the United States Medical Licensing Examination (USMLE) Step 2 Clinical Skills board exam simulations. This is meant to be a preliminary evaluation of the platform, aimed at providing an initial indication regarding its diagnostic capability and general practicality. Further investigations are planned to evaluate its influence on practitioners' skills and behavior in both simulated and real-life settings, with the end goal of demonstrating its effect on healthcare quality measures and patient-related outcomes.

Methods

Case Selection

Cases were extracted from the case banks of 3 leading USMLE board exams preparation companies: UWorld, Amboss, and FirstAid. All cases available for subscribed users were drawn and checked for eligibility. Each case included a summary of the patient’s clinical findings (demographics, medical and family history, medications, habits, symptoms, and signs) and 3 “correct” DDX that are expected to be suggested. The cases were reviewed by 3 physicians, who are registered specialists in emergency medicine, rheumatology, and internal medicine, with at least 5 years of practicing experience. Each case was assigned to a medical discipline based on its chief complaint. Cases from the disciplines of pediatrics, obstetrics, trauma, and psychiatry were excluded if at least 2 reviewers allocated these cases to such groups.

Procedures and Design

A group of 3 junior physicians, interns in internal medicine from a tertiary hospital in Israel, were recruited to enter the clinical findings of the selected cases into the Kahun platform. To avoid biases and simulate use by an inexperienced user, the selected physicians had no prior experience using Kahun. They were blinded to the correct diagnoses, and the only guidance they received was a short online tutorial video (Multimedia Appendix 1). For each case, the presence and the rank of the correct diagnoses within the generated DDX list were recorded.

Statistical Analysis

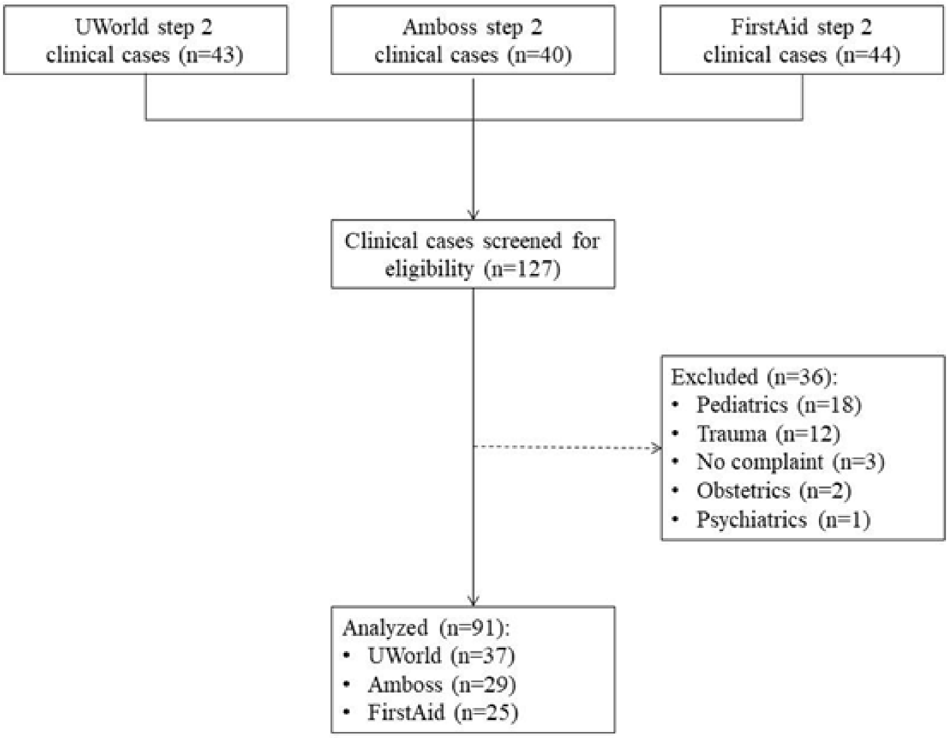
A case was considered successful if Kahun listed the correct diagnosis within the top 5, 10, and 20 places of the generated DDX list, which includes a maximum of the 20 most likely diagnoses. Diagnostic performance was measured in two ways. First, as sensitivity, calculated as the total number of the expected DDX appropriately suggested (within the top 5, 10, and 20 of the listed diagnoses), divided by the total number of the expected diagnoses in all cases. This analysis was further stratified according to organ system. Second, the comprehensiveness of the DDX list was measured and calculated as the number of cases with 1/3, 2/3, and 3/3 of the expected DDX appropriately suggested (within the top 5, 10, and 20 listed diagnoses), divided by the total number of cases. Statistical analysis was performed using the commercial software SPSS (for Windows, version 26.0, IBM Corp). The 95% CIs were calculated assuming binomial distribution.

Results

Characteristics of Cases

A total of 127 cases were screened from Amboss (n=40), FirstAid (n=44), and UWorld (n=44); 36 cases were excluded because they were classified as pediatric (n=18), trauma (n=12), obstetric (n=2), and psychiatric (n=1) cases or were routine-checkup cases without a chief complaint (n=3). The remaining 91 cases, Amboss (n=29), FirstAid (n=25), and UWorld (n=37), were analyzed in the study (Figure 1).

Figure 1. Case selection flow chart.



Each case was provided with 5 (n=1), 4 (n=1), 3 (n=85), or 2 (n=4) correct diagnoses, resulting in a total of 272 tested diagnoses of which 174 were unique (some diagnoses appeared in more than 1 case). The most common expected diagnosis was hypothyroidism (n=6), followed by adverse drug reaction, pelvic inflammatory disease, hyperthyroidism, pneumonia, and depressive disorder (n=5). The distribution of diagnoses according to organ systems and basic success rates is presented in Table 1. The best diagnostic sensitivity rates were demonstrated for diagnoses related to the digestive system (54/55, 98.2%) and to the genitourinary system (35/36, 97.2%), while the worst were demonstrated for autoimmune or inflammatory diagnoses (8/13, 61.5%). Diagnostic accuracy did not fall below 50% in any category. Overall, 845 different findings (both positive and negative) were entered into Kahun in the test, with a mean number of 39.8 (SD 8) findings for each case.

Table 1. Distribution of case diagnoses according to organ system and specific accuracy rates.

Organ system	Accuracy[a] n/N (%)	95% CI
Cardiovascular	18/21 (85)	71-100
Respiratory	14/18 (77)	59-97
Gastrointestinal	54/55 (98)	95-100
Genitourinary	35/36 (97)	92-100
Infectious	26/30 (86)	75-99
Nervous	22/27 (81)	67-96
Musculoskeletal	4/5 (80)	45-100
Ear-nose-throat	12/16 (75)	54-96
Autoimmune or inflammatory	8/13 (61)	35-88
Endocrine/metabolic/drugs	19/29 (65)	48-83
Psychiatric	17/19 (89)	76-100
Other	2/3 (67)	13-100

[a]Within the top 20 listed diagnoses.

Diagnostic Sensitivity Rates

Diagnostic sensitivity rates are presented in Table 2. Out of the total 272 expected diagnoses, 231 (87.5%) diagnoses were accurately suggested within the top 20 listed diagnoses (95% CI 76-99), of which 209 (76.8%) were listed within the top 10 (95% CI 66-87), and 168 (61.8%) listed within the top 5 (95% CI 52-71). There was no statistical significance in the difference of sensitivities between the different case sources. The median rank of correct diagnoses was 3 (IQR 2-6).

Table 2. Diagnostic sensitivity.

Company name	Correctly suggested diagnoses					
	Within top 5 listed diagnoses		Within top 10 listed diagnoses		Within top 20 listed diagnoses	
	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI
Total (N=272)	168 (61.8)	52-71	209 (76.8)	66-87	238 (87.5)	76-99
Amboss (n=87)	57 (65.5)	49-83	72 (82.8)	64-100	79 (90.8)	71-100
FirstAid (n=76)	43 (56.6)	40-73	56 (73.7)	54-93	61 (80.3)	60-100
UWorld (n=109)	68 (62.4)	48-77	81 (74.3)	58-90	98 (89.9)	72-100

Diagnostic Comprehensiveness

Case-specific success rates are presented in Table 3. In 62 (68%) out of 91 cases (95% CI 59-78), all of the cases' expected diagnoses were suggested within the top 20 listed diagnoses; in 44 (48%; 95% CI 38-59), they were listed within the top 10 diagnoses; and in 24 (26%; 95% CI 17-35), within the top 5 diagnoses. In 87 (96%) out of 91 cases (95% CI 91-100), at least 2 out of 3 of the cases' expected diagnoses were suggested within the top 20 listed diagnoses; in 78 (86%; 95% CI 79-93) within the top 10 listed diagnoses; and in 61 (67%; 95% CI 57-77) within the top 5 listed diagnoses.

Table 3. Case-specific success rates.

Top diagnoses	Rate of correctly suggested diagnoses per case (n=91)					
	3/3[a]		≥2/3[b]		≥1/3[c]	
	Cases, n (%)	95% CI	Cases, n (%)	95% CI	Cases, n (%)	95% CI
Within top 5 listed diagnoses	24 (26)	17-35	61 (67)	57-77	84 (92)	87-98
Within top 10 listed diagnoses	44 (48)	38-59	78 (86)	79-93	88 (97)	93-100
Within top 20 listed diagnoses	62 (68)	59-78	87 (96)	91-100	90 (99)	97-100

[a] Including cases with 2/2, 4/4, and 5/5 correct diagnoses.

[b] Including a case with 4/5 correct diagnoses.

[c] Including cases with 1/2 and 2/4 correct diagnoses.

Discussion

Principal Results

In this study, we evaluated the diagnostic performance of Kahun, a new open-access DDSS, based on its ability to suggest the expected diagnoses in simulated board exam cases. Overall, Kahun demonstrated good diagnostic sensitivity and comprehensiveness in managing these cases. Moreover, the system demonstrated its ability to manage a wide range of patient-related findings and to reach a wide range of accurate diagnoses from different fields of medicine.

Comparison to Previous Studies

The general literature addressing computer-assisted diagnosis is vast. However, when we narrow the scope to commercially available systems that adhere to the definition of DDSS (as established by Bond et al [24]) and those that are targeted for general practice rather than a specific field or condition, only a handful of original studies regarding diagnostic accuracy remain [1,12,24,25]. Similar to our study, all of these studies used a structured clinical case model to evaluate diagnostic systems. Of the studies we reviewed, 3 used cases from different case banks [12,24,25], while 1 used structured cases based on real patients [1]. Unlike our study, all of these [1,12,24,25] defined accuracy as the retrieval rate of a single "gold standard" diagnosis in the top 20 or 30 differential diagnoses generated by the tested tool. None of the studies [1,12,24,25] reported the mean rank of correct diagnoses or the number of findings the system was able to include, except for the study by Graber et al [25] on ISABEL (Isabel Healthcare), which used 3 to 6 key findings for each case. Regarding diagnostic sensitivity, a recent comprehensive metanalysis [26], covering 36 original studies, reported a pooled sensitivity of 70% (95% CI 63-77) overall, and 68% (95% CI 61-74) in studies with stronger methodological quality ratings. The highest accuracy rate was observed for ISABEL, which demonstrated a pooled sensitivity of 89% (95% CI 83-94) with a high heterogeneity between studies [26]. Importantly, the studies in which ISABEL demonstrated the highest accuracy rates defined success as the tool's ability to output the correct diagnosis in a DDX list containing the 30 most likely diagnoses, as opposed to the 20 diagnoses in our study [25]. A recent study [12], comparing Doknosis, DXplain (Massachusetts General Hospital), and ISABEL, analyzed diagnostic accuracy on a data set including cases from the UWorld case bank, which was also used in our study. In this analysis, the best sensitivity rate observed was 47%. Given these findings, it is safe to assume that the diagnostic sensitivity observed in our study falls in the upper range of what was previously demonstrated by the existing systems. Clearly, no direct comparison between the products could be made in our study.

Strengths

In this study, we used structured clinical cases that simulate the USMLE Step 2 board exams to evaluate a new diagnostic support tool. These cases have the advantage of being principal cases, which are frequently encountered in primary care and emergency department settings. Moreover, they are designated for the level of junior physicians and medical students, who are populations that were demonstrated to benefit the most from using DDSS [3]. An additional advantage was the fact that each case had 3 "correct" diagnoses rather than a single final diagnosis. This more accurately reflects the true nature of these systems: to serve as valuable resources in the hands of the physician by providing reliable and reasoned case-specific diagnostic and workup suggestions, rather than serving as a "Greek oracle" predicting the correct diagnosis [3,13]. This approach also enabled us to assess the comprehensiveness [1,3,13] of the DDX quality. The cases were entered into the platform by first-time users, which increased the platform's external validity by allowing an extrapolation of the results to those of an "average" user; it also enabled the study to reflect on the instinctive nature of the diagnostic system. This procedure was performed while the subjects were blinded to the correct diagnoses, thus reducing the chance of response bias.

Limitations

Our study has several limitations. First, it was designed to assess the accuracy of Kahun in an ideal environment, which does not reflect the stressful and time-limiting working environment of a junior clinician in the primary care clinic, emergency department, or internal medicine department settings. Moreover, the patient summaries used in this study were already somewhat processed and do not account for the clinician's judgment regarding the relevancy of certain findings or the ability to produce and interpret findings from a physical examination. Another shortcoming for this type of comparison is that it measures the accuracy of the diagnostic tool itself, rather than its ability to augment the user's informed decision-making, which is perhaps a more valuable measure of performance [1,3,13]. For these reasons, caution needs to be taken when extrapolating the results to performance in an actual clinical setting. The clinical cases selected in this study were based on the USMLE board exams, which, although diverse, are less representative of the rare or unique cases usually depicted in case-report studies. Furthermore, they do not include laboratory and imaging findings and, therefore, do not measure the ability of Kahun to handle these findings. Finally, regarding the platform itself, Kahun is currently not set up to manage patients in pediatrics, trauma, obstetrics, and psychiatry settings. Therefore, we were forced to exclude these cases from the analysis. Nevertheless, it is important to note that Kahun was able to generate DDX from these fields with similar accuracy rates.

Conclusions

Kahun is a new diagnostic tool that demonstrates an acceptable level of diagnostic accuracy and comprehensiveness. Further studies are warranted to evaluate its contribution to the physician's decision-making process, to the quality of healthcare, and to the clinical outcomes of the patients, including direct comparison to other DDSS.

Acknowledgments

We would like to thank the cofounder and chief marketing officer of Kahun Medical Ltd, Michal Tzuchman-Katz, MD, for providing software-related information and technical support for this study. All of the participants in the study were volunteers. No grants or any other funds were received for the purpose of the study. NBS, AS, and TW are part-time employees at Kahun Medical Ltd.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Tutorial video demonstrating a standard patient's run in Kahun's platform.
[MP4 File (MP4 Video), 1994 KB - medinform_v9i11e32507_app1.mp4]

Multimedia Appendix 2
A series of screenshots portraying panels and functions of the mobile app.
[PPTX File , 1578 KB - medinform_v9i11e32507_app2.pptx]

References

1. Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, et al. Performance of Four Computer-Based Diagnostic Systems. N Engl J Med 1994 Jun 23;330(25):1792-1796. [doi: 10.1056/nejm199406233302506]
2. Berner ES, Maisiak RS, Cobbs CG, Taunton OD. Effects of a decision support system on physicians' diagnostic performance. J Am Med Inform Assoc 1999 Sep 01;6(5):420-427 [FREE Full text] [doi: 10.1136/jamia.1999.0060420] [Medline: 10495101]
3. Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, et al. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. JAMA 1999 Nov 17;282(19):1851-1856. [doi: 10.1001/jama.282.19.1851] [Medline: 10573277]
4. Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. Science 1974 Sep 27;185(4157):1124-1131. [doi: 10.1126/science.185.4157.1124] [Medline: 17835457]
5. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak 2016 Nov 03;16(1):138 [FREE Full text] [doi: 10.1186/s12911-016-0377-1] [Medline: 27809908]
6. Morgan DJ, Pineles L, Owczarzak J, Magder L, Scherer L, Brown JP, et al. Accuracy of Practitioner Estimates of Probability of Diagnosis Before and After Testing. JAMA Intern Med 2021 Jun 01;181(6):747-755 [FREE Full text] [doi: 10.1001/jamainternmed.2021.0269] [Medline: 33818595]
7. Whiting PF, Davenport C, Jameson C, Burke M, Sterne JAC, Hyde C, et al. How well do health professionals interpret diagnostic information? A systematic review. BMJ Open 2015 Jul 28;5(7):e008155 [FREE Full text] [doi: 10.1136/bmjopen-2015-008155] [Medline: 26220870]
8. Friedman CP, Gatti GG, Franz TM, Murphy GC, Wolf FM, Heckerling PS, et al. Do physicians know when their diagnoses are correct? Implications for decision support and error reduction. J Gen Intern Med 2005 Apr;20(4):334-339 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.30145.x] [Medline: 15857490]
9. Ramnarayan P, Tomlinson A, Rao A, Coren M, Winrow A, Britto J. ISABEL: a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation. Arch Dis Child 2003 May;88(5):408-413 [FREE Full text] [doi: 10.1136/adc.88.5.408] [Medline: 12716712]

10. Barnett GO. DXplain: An Evolving Diagnostic Decision-Support System. JAMA 1987 Jul 03;258(1):67-74. [doi: 10.1001/jama.1987.03400010071030]
11. Miller R, Masarie FE, Myers JD. Quick medical reference (QMR) for diagnostic assistance. MD Comput 1986;3(5):34-48. [Medline: 3537611]
12. Müller L, Gangadharaiah R, Klein SC, Perry J, Bernstein G, Nurkse D, et al. An open access medical knowledge base for community driven diagnostic decision support system development. BMC Med Inform Decis Mak 2019 Apr 27;19(1):93 [FREE Full text] [doi: 10.1186/s12911-019-0804-1] [Medline: 31029130]
13. Ramnarayan P, Kapoor RR, Coren M, Nanduri V, Tomlinson AL, Taylor PM, et al. Measuring the Impact of Diagnostic Decision Support on the Quality of Clinical Decision Making: Development of a Reliable and Valid Composite Score. J Am Med Inform Assoc 2003 Nov 01;10(6):563-572. [doi: 10.1197/jamia.m1338]
14. Ramnarayan P, Roberts GC, Coren M, Nanduri V, Tomlinson A, Taylor PM, et al. Assessment of the potential impact of a reminder system on the reduction of diagnostic errors: a quasi-experimental study. BMC Med Inform Decis Mak 2006 Apr 28;6(1):22 [FREE Full text] [doi: 10.1186/1472-6947-6-22] [Medline: 16646956]
15. Elkin PL, Liebow M, Bauer BA, Chaliki S, Wahner-Roedler D, Bundrick J, et al. The introduction of a diagnostic decision support system (DXplain™) into the workflow of a teaching hospital service can decrease the cost of service for diagnostically challenging Diagnostic Related Groups (DRGs). Int J Med Inform 2010 Nov;79(11):772-777 [FREE Full text] [doi: 10.1016/j.ijmedinf.2010.09.004] [Medline: 20951080]
16. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005 Mar 09;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
17. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
18. Bergeron B. Iliad: a diagnostic consultant and patient simulator. MD Comput 1991;8(1):46-53. [Medline: 1822085]
19. Berner ES. Diagnostic decision support systems: why aren't they used more and what can we do about it? 2006 Nov Presented at: AMIA Annual Symposium Proceedings; November 11-15, 2006; Washington DC p. 1167-1168 URL: https://knowledge.amia.org/amia-55142-a2006a-1.620145/t-003-1.622242/f-001-1.622243/a-493-1.622256/an-493-1.622257?qr=1
20. Kawamoto K, Lobach DF. Clinical Decision Support Provided within Physician Order Entry Systems: A Systematic Review of Features Effective for Changing Clinician Behavior. 2003 Nov Presented at: AMIA Annual Symposium; November 8-12, 2003; Washington DC p. 361-365 URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480005/
21. O'Sullivan D, Fraccaro P, Carson E, Weller P. Decision time for clinical decision support systems. Clin Med (Lond) 2014 Aug 06;14(4):338-341 [FREE Full text] [doi: 10.7861/clinmedicine.14-4-338] [Medline: 25099829]
22. Shibl R, Lawley M, Debuse J. Factors influencing decision support system acceptance. Decision Support Systems 2013 Jan;54(2):953-961. [doi: 10.1016/j.dss.2012.09.018]
23. Schuyler P, Hole W, Tuttle M, Sherertz D. The UMLS Metathesaurus: representing different views of biomedical concepts. Bull Med Libr Assoc 1993 Apr;81(2):217-222 [FREE Full text] [Medline: 8472007]
24. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: an evaluation of currently available computer programs. J Gen Intern Med 2012 Feb 26;27(2):213-219 [FREE Full text] [doi: 10.1007/s11606-011-1804-8] [Medline: 21789717]
25. Graber ML, Mathew A. Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med 2008 Jan 19;23 Suppl 1(S1):37-40 [FREE Full text] [doi: 10.1007/s11606-007-0271-8] [Medline: 18095042]
26. Riches N, Panagioti M, Alam R, Cheraghi-Sohi S, Campbell S, Esmail A, et al. The Effectiveness of Electronic Differential Diagnoses (DDX) Generators: A Systematic Review and Meta-Analysis. PLoS One 2016;11(3):e0148991 [FREE Full text] [doi: 10.1371/journal.pone.0148991] [Medline: 26954234]

Abbreviations

DDSS: diagnostic decision support system
DDX: differential diagnoses
LOINC: Logical Observation Identifiers Names and Codes
SNOMED: Systematized Nomenclature of Medicine
USMLE: United States Medical Licensing Examination

Edited by J Hefner, C Lovis; submitted 30.07.21; peer-reviewed by R De Carvalho, A Benis; comments to author 28.09.21; revised version received 20.10.21; accepted 20.10.21; published 30.11.21.

Please cite as:
Ben-Shabat N, Sloma A, Weizman T, Kiderman D, Amital H
Assessing the Performance of a New Artificial Intelligence–Driven Diagnostic Support Tool Using Medical Board Exam Simulations: Clinical Vignette Study
JMIR Med Inform 2021;9(11):e32507
URL: https://medinform.jmir.org/2021/11/e32507
doi:10.2196/32507
PMID:34672262

Original Paper

Stroke Outcome Measurements From Electronic Medical Records: Cross-sectional Study on the Effectiveness of Neural and Nonneural Classifiers

Bruna Stella Zanotto[1,2*], MSc, PharmD; Ana Paula Beck da Silva Etges[1,3*], Eng, MSc, PhD; Avner dal Bosco[3*], MSc; Eduardo Gabriel Cortes[4*], MSc; Renata Ruschel[1*], PT; Ana Claudia De Souza[5*], MD, PhD; Claudio M V Andrade[6*], MSc; Felipe Viegas[6*], MSc; Sergio Canuto[6*], MSc, PhD; Washington Luiz[6*], MSc; Sheila Ouriques Martins[5*], MSc, MD, PhD; Renata Vieira[7*], MSc, PhD; Carisi Polanczyk[1,2*], MSc, MD, PhD; Marcos André Gonçalves[6*], MSc, PhD

[1]National Institute of Health Technology Assessment - INCT/IATS (CNPQ 465518/2014-1), Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

[2]Graduate Program in Epidemiology, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

[3]School of Technology, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil

[4]Graduate Program of Computer Science, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

[5]Brazilian Stroke Network, Hospital Moinhos de Vento, Porto Alegre, Brazil

[6]Computer Science Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[7]Centro Interdisciplinar de História, Culturas e Sociedades (CIDEHUS), Universidade de Évora, Évora, Portugal

[*]all authors contributed equally

Corresponding Author:
Marcos André Gonçalves, MSc, PhD
Computer Science Department
Universidade Federal de Minas Gerais
Avenue Antônio Carlos, 6627
Belo Horizonte, 31270-901
Brazil
Phone: 55 3134095860
Email: mgoncalv@dcc.ufmg.br

Abstract

Background: With the rapid adoption of electronic medical records (EMRs), there is an ever-increasing opportunity to collect data and extract knowledge from EMRs to support patient-centered stroke management.

Objective: This study aims to compare the effectiveness of state-of-the-art automatic text classification methods in classifying data to support the prediction of clinical patient outcomes and the extraction of patient characteristics from EMRs.

Methods: Our study addressed the computational problems of information extraction and automatic text classification. We identified essential tasks to be considered in an ischemic stroke value-based program. The 30 selected tasks were classified (manually labeled by specialists) according to the following value agenda: tier 1 (achieved health care status), tier 2 (recovery process), care related (clinical management and risk scores), and baseline characteristics. The analyzed data set was retrospectively extracted from the EMRs of patients with stroke from a private Brazilian hospital between 2018 and 2019. A total of 44,206 sentences from free-text medical records in Portuguese were used to train and develop 10 supervised computational machine learning methods, including state-of-the-art neural and nonneural methods, along with ontological rules. As an experimental protocol, we used a 5-fold cross-validation procedure repeated 6 times, along with *subject-wise sampling*. A heatmap was used to display comparative result analyses according to the best algorithmic effectiveness (F1 score), supported by statistical significance tests. A feature importance analysis was conducted to provide insights into the results.

Results: The top-performing models were support vector machines trained with lexical and semantic textual features, showing the importance of dealing with noise in EMR textual representations. The support vector machine models produced statistically superior results in 71% (17/24) of tasks, with an F1 score >80% regarding care-related tasks (patient treatment location, fall risk, thrombolytic therapy, and pressure ulcer risk), the process of recovery (ability to feed orally or ambulate and communicate), health care status achieved (mortality), and baseline characteristics (diabetes, obesity, dyslipidemia, and smoking status). Neural methods were largely outperformed by more traditional nonneural methods, given the characteristics of the data set. Ontological rules were also effective in tasks such as baseline characteristics (alcoholism, atrial fibrillation, and coronary artery disease) and the Rankin scale. The complementarity in effectiveness among models suggests that a combination of models could enhance the results and cover more tasks in the future.

Conclusions: Advances in information technology capacity are essential for scalability and agility in measuring health status outcomes. This study allowed us to measure effectiveness and identify opportunities for automating the classification of outcomes of specific tasks related to clinical conditions of stroke victims, and thus ultimately assess the possibility of proactively using these machine learning techniques in real-world situations.

(JMIR Med Inform 2021;9(11):e29120) doi:10.2196/29120

KEYWORDS

natural language processing; stroke; outcomes; electronic medical records; EHR; electronic health records; text processing; data mining; text classification; patient outcomes

Introduction

Background

Stroke is the second leading cause of mortality and disability-adjusted life years globally [1,2]. The outcomes of stroke can vary greatly, and timely assessment is essential for optimal management. As such, there has been an increasing interest in the use of automated machine learning (ML) techniques to track stroke outcomes, with the hope that such methods could make use of large, routinely collected data sets and deliver accurate, personalized prognoses [3]. However, studies applying ML methods to stroke, although published regularly, have focused mostly on stroke imaging applications [4-6] and structured data retrieval [3]. Few studies have addressed the unstructured textual portion of electronic medical records (EMRs) as the primary source of information.

Indeed, the use of EMR data in the last decade has led to promising findings in population health research, such as patient-use stratification [7], treatment-effectiveness evaluation [8], early detection of diseases [9], and predictive modeling [10]. However, dealing with EMR data is often labor intensive [11] and challenging because of the lack of standardization in data entry, changes in coding procedures over time, and the impact of missing information [9,12-14]. The information technology (IT) gap between automated data collection from EMRs and improving the quality of care has been described in the literature as a decelerator of value initiatives [15-18].

With recent advances in IT, several groups have attempted to apply natural language processing (NLP) to the text analysis of EMRs to achieve early diagnosis of multiple conditions, such as peripheral arterial disease [19], asthma [20], multiple sclerosis [21], and heart failure [22]. In these studies, NLP was used to find specific words or phrases in a predefined dictionary that described the symptoms or signs of each disease [14,21,23].

Objectives

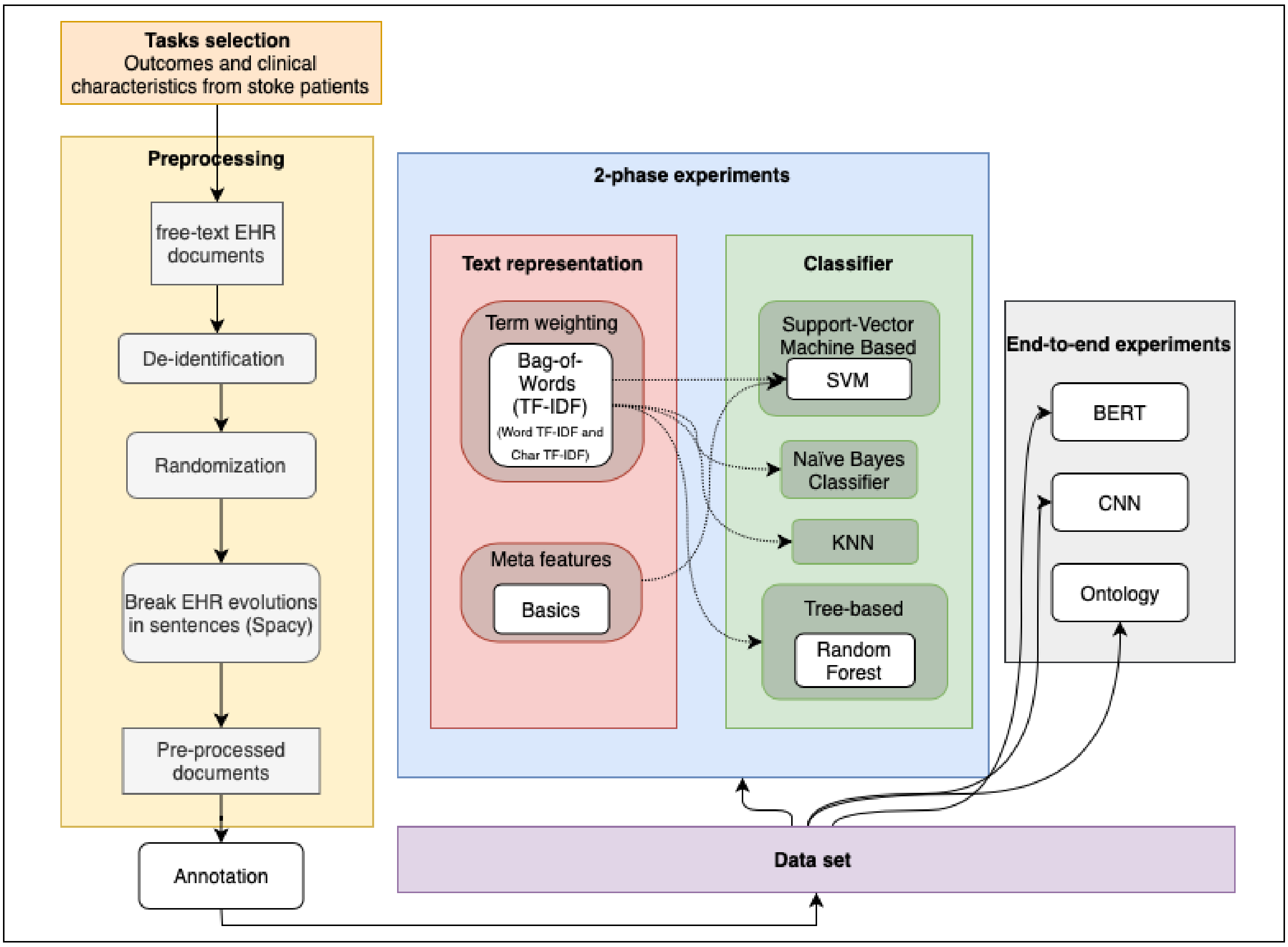
Generating value for the patient as the central guide requires advances in strategies to automate the capturing of data that will allow managers to assess the quality of service delivery to patients [24,25]. Accordingly, our research aims to compare the effectiveness of state-of-the-art automatic text classification methods in classifying data to support the prediction of clinical patient outcomes and the extraction of patient characteristics from EMR sentences. With stroke as our case study application, our specific goal is to investigate the capability of these methods to automatically identify, with reasonable effectiveness, the outcomes and clinical characteristics of patients from EMRs that may be considered in a stroke outcome measurement program.

Methods

Overview

This study faced a computational problem related to information extraction and free-text classification. As presented in Figure 1, the dotted lines represent the union of the text representative technique that was used with each classifier in the two-phase experiments. Our study was generally organized into four stages: (1) task selection; (2) study design, preprocessing, and data annotation; (3) definition of automatic text classification methods; and (4) experimental evaluation (experimental protocol, setup, and analysis of results).

Figure 1. Study architecture. BERT: bidirectional encoder representation from transformers; CNN: convolutional neural network; EHR: electronic health record; KNN: K-nearest neighbor; SVM: support vector machine; TF-IDF: term frequency-inverted document frequency.



Task Selection

A literature review and multidisciplinary expert interviews (n=8) were used to define specific outcome dimensions and measures that may be considered in an outcome measurement program for ischemic stroke. The outcome identification step was based on adhering to value agenda element dimensions to cover the tiers of the outcome hierarchy [26], such as functionality dimensions, the recovery process, and outcomes that matter to patients. These dimensions included risk events, achieved health care status, and stroke outcome scales, such as the National Institutes of Health Stroke Scale (NIHSS) and the modified Rankin scale (mRS) [27,28].

Study Design and Data Annotation

We retrospectively built a database of medical records from a digital hospital system. The database covered 2 years of patients hospitalized for ischemic stroke. The hospital is a private institution of excellence in southern Brazil. The EMR system used was the MV Soul (Recife). Since 2017, the hospital has introduced the ICHOM standard sets' data collection routine for different clinical pathways and created an office for institutional values. To examine the stroke pathway, data were collected on October 15, 2015. In 2019, the hospital incorporated the Angel Awards Program [29], which was certified as a platinum category at the end of the first year. This study was approved by the hospital ethics committee (CAAE: 29694720000005330).

Medical records of patients were submitted to preprocessing using the spaCy Python library (Python Software Foundation; Python Language Reference, version 2.7) [30] to stratify texts into sentences. A total of 44,206 EMR sentences were obtained from 188 patients. The approach followed a hypothesis for managing unbalanced data, such as electronic health records, which assumes that relevant information to be retrieved from EMRs encompasses a small space of words delimited as sentences, and the residual is noise [31-33]. During the text stratification process, spaCy [30] uses rule-based algorithms that set the sentence limits according to the patterns of characters, thereby delimiting its beginning and end. The names of patients and medical staff were identified, thus removing all confidential information from the data set. The preprocessed textual sentence was represented in a vector of words that disregarded grammar and word order but maintained their multiplicity.

For sentence annotation (intratask class labeling), we developed annotation guidelines that provided an explicit definition of each task, its classes (response options), and examples to be identified in the documents. This guideline is written in Portuguese and is available upon request.

Two annotators independently reviewed the preprocessed text documents (44,206 sentences) and had the percent agreement between them measured by κ, which was higher than 0.61 (substantial agreement) [34]. Task-level disagreements were resolved by consensus determination by 2 annotators, with assistance from a committee composed of experts (APE, ACS, MP, KBR, and CAP).

Each task could have two or more output answers, depending on the meaning of the sentence. Examples of an EMR and the annotation process can be seen in Multimedia Appendices 1 and 2. Task details in terms of class and sentence distribution are shown in Multimedia Appendix 3 and demonstrate the highly imbalanced nature of the tasks with most of the sentences belonging to the NI (noninformative) class. This makes it a very hard endeavor from an ML perspective. Subsequently, we evaluated the impact of this imbalance in the experimental results.

Automatic Text Classification Methods

As presented in the study design, the ML methods were divided into two categories: two-phase methods and end-to-end (E2E) methods [35]. The first category of methods consisted of approaches whose document (ie, sentence) representation was intrinsically independent of the classification algorithm used to predict the class. In other words, the classifier used to predict the class of documents was not used in the construction phase of the document representation. In terms of text representations, we considered three alternatives, namely traditional term-weighting alternatives (term frequency-inverted document frequency [TFIDF]); weighting based on word and character (n-gram) frequency; and recent representations based on meta-features, which capture statistical information from a document's neighborhood and have obtained state-of-the-art effectiveness in recent benchmarks [35-39].

As two-phase classification algorithms, we exploited support vector machines (SVMs), which are still considered the most robust nonneural network text classification algorithm [35,39,40], random forests (RF), K-nearest neighbor (KNN), and naïve Bayes classifier (NBC), to address the most popular algorithms in terms of classification and retrieval of text information [41-44].

In contrast, E2E methods use a discriminative classifier function to transform the document representation space into a new and more informed (usually more reduced and compact) space and use this classifier to predict the document class. In general, these approaches use an iterative process of representation, classification, evaluation, and parameter adaptation (eg, transform, predict, evaluate loss function, and backpropagate, respectively). For E2E classifiers, we exploited two neural architectures, namely convolutional neural networks (CNNs), which exploit textual patterns such as word co-occurrences, and bidirectional encoder representation from transformers (BERT), which exploits attention mechanisms and constitute the current state-of-the-art in many NLP tasks.

Finally, we exploited a rule-based classifier specialized for the tasks at hand (stroke tasks, represented in the ontology web language [OWL]). The rule-based knowledge model was developed using logical conditions built alongside domain specialists [45]. This technique has shown effectiveness equivalent to that of some ML classification models in certain domains without the need for a large amount of data and training time, which are commonly required by supervised methods [46-49]. In contrast, it is heavily dependent on the specialists and the coverage of the rules on the text expressions. More details about each of the exploited algorithms are provided in Multimedia Appendix 4 [3,35,37,39,41-45,50-63].

The two-phase methods used in this research are referred to as the representation technique combined with the classification algorithm, as follows: word-TFIDF and character-TFIDF combined with SVM (SVM+W+C), Bag-of-Words (BoW) combined with SVM (SVM+BoW), meta-features combined with SVM (meta-features), word-TFIDF combined with SVM (SVM+Word-TFIDF), character-TFIDF combined with SVM (SVM+Chard-TFIDF), Word-TFIDF combined with random forest (RF+Word-TFIDF), word-TFIDF combined with KNN (KNN+Word-TFIDF), and word-TFIDF combined with naïve Bayes (Naïve Bayes+Word-TFIDF). In contrast to TFIDF, BoW explores only the frequency of terms (term frequency) and not the frequency of terms in the collection (IDF component). The E2E methods are simply called CNN and BERT, and the ontological method is called OWL.

Experimental Evaluation

Overview

The experimental process consisted of testing different classification methods with sets of annotated data to assess and compare their performances (effectiveness). The experimental procedure, described in Multimedia Appendix 5, consisted of four phases: (1) representing the free-text sentences as numerical vectors, (2) the training and tuning process (in a validation set) by means of a folded cross-validation procedure, (3) the execution of the classification algorithms in the test set and effectiveness assessment, and (4) the synthesis of the results in a heatmap table.

A classification model was developed for each task. Each task resulted in an individual automatic classification model for the training and testing process of the model. As an experimental protocol, we used a five-fold cross-validation procedure repeated six times (resulting in 30 test samples). We also exploited *subject-wise cross-validation* in the sense that the information from the same patient was always assigned to the same fold to test the ability of the model to predict new data that was not used in the learning process. These procedures address potential problems, such as overfitting and selection bias [64], and produce results that are more reliable.

To evaluate the ability to classify the relevant Brazilian-Portuguese medical free-text records correctly, we used the Macro-F1 score (equation 1). This metric is based on a *confusion matrix* and is defined as follows:

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. Precision (positive predictive value) = TP / TP + FP = the number of returned hits that were true positive. Recall (sensibility) = TP / TP + FN = is the fraction of the total number of true positives retrieved.

The F1 measure is calculated for each class. Macro-F1 summarizes the classification effectiveness by averaging F1 values for all classes. Macro-F1 is one of the most popular aggregated evaluation metrics for the classifier evaluation of unbalanced or skewed data sets [42,65,66]. Macro-F1 is especially suitable for imbalanced data sets, as the effectiveness of each individual class contributes equally to producing a final score. For instance, in a task with four classes, in which one of them is NI, if all classes are predicted as NI, the Macro-F1 score will be no higher than 0.25 (F1 of 1 for NI and 0 for the three other classes). Accuracy or any other evaluation measure focused on the instance, instead of the class effectiveness, would produce a very high score (close to 1 in this particular case).

To compare the average results of our cross-validation experiments, we assessed statistical significance by using a paired two-tailed *t* test with 95% CIs. To account for multiple tests, we adopted the Friedman-Nemenyi test [67] with Bonferroni correction for multiple comparisons of mean rank sums. The Friedman test was used to compare multiple methods.

We consider that making the data and the code used in our experimental protocol available to others is potentially useful for reproducibility and for use in other studies. Both the code and data will be available upon request. The mood-specific parameter tuning details are presented in Multimedia Appendix 6.

Experimental Analysis

The experiments aimed to provide relationships between the classification methods and the tasks, allowing for connecting the best methods with each outcome measure or patient characteristics. Considering that the model's results can influence health decision-making in some way, the F1 score thresholds may vary depending on the type of class and the imbalance of the data. We reported the results by means of a heatmap, adopting a red color for F1<20%, a gradual color scale from orange to yellow for 21%<F1<79%, and green for F1>80% [68-71]. Tasks (represented by the lines) were ordered by the average of the performed models, whereas the ordering of the columns shows the rank position of each method according to the statistical analysis.

For the sake of the fairness of the comparison, the OWL technique should not be and is not directly compared and ranked herein along with the other ML models described above that require a combination of text representations with trained classification algorithms. OWL rules were designed to work with the entire corpus (including the test) and were not designed for generalization. Instead, they are built to work well in the specific domain or task for which they were created. In any case, for reasons of practical application and as a research exercise, as a secondary analysis, we compared (later) the OWL technique with the ML model ranked as the best based on the Friedman test. This analysis allowed us to identify the weaknesses and strengths of both approaches (generalized ML models vs domain or task-specific ontological rules) in the contrasting tasks.

Moreover, we performed a feature selection analysis [72,73]. This technique is used to rank the most informative features of each task according to the information theory criteria. In particular, we used SelectKBest (Python Software Foundation; Python Language Reference, version 2.7) with the chi-square, which is independent of the classification algorithms used [74]. This final analysis helps in understanding how ML can help with outcome measurements for the stroke care pathway, potentially boosting advances in quality indicator automation.

Finally, to complete the analysis and evaluate the impact of the highly skewed distribution, especially toward the NI class, we ran an experiment in which we performed a random undersampling process for all considered tasks (we used the RandomUnderSampler Phyton library [75]). In detail, we randomly selected the same number of training random examples of the NI as the number of instances of the second largest (non-NI) class of a given task. We then reran all ML classifiers (the ontology method is not affected by this process as it has no training) in all 24 tasks, considering as the training set the reduced (undersampled) NI training samples along with the same (unchanged) previous samples for the other classes. We did that for all six rounds of five-fold cross-validation of our experimental procedure, changing the seed for selection in each round, resulting in six different NI reduced training sets. The test folds in all cases remain unchanged, meaning that we keep the same skewed distribution as in the original data set, as we do not know the class of the test instances.

Results

Tasks Selection

Discussions with experts in the stroke care pathway allowed us to define 30 tasks that were considered feasible to extract from EMRs. For the first tier, the standard sets were usually defined to evaluate the clinical stroke outcomes that were used, including the mRS [27] and the NIHSS scales [76], in addition to traditional outcomes such as mortality and pain level. For tier 2, the ICHOM standard set developed for ischemic stroke was used [77], which considers measures of mobility, ability to communicate, ability to feed orally, the ability to understand, and measures and scales of strength level. Indicators of the hospitalization care process used in the institution were also included, such as rating scales and risk events tracked by fall risk, pressure ulcer risk, fall events during hospitalization, infection indicators, intracranial hemorrhage, therapy care (thrombolytic, thrombectomy, or both), and the location of the patient during the inpatient path [78]. Finally, baseline characteristics important for tracking the population and further risk-adjusted analysis were included [79], such as high blood pressure, smoking status, coronary artery disease, atrial fibrillation, diabetes, prior stroke, active cancer, alcoholism, obesity, and dyslipidemia. Each category, containing the tasks and their respective classes, is presented in Table 1.

Table 1. Eligible tasks for analysis and classification rules.

Tasks	Number of classes	Supporting information for classes
Health care status achieved (tier 1)		
Rankin	8	0-6 NI[a]
National Institutes of Health Stroke Scale	42	1-41 NI
Death	3	Absence of vital signs Vital signs present NI
Process of recovery (tier 2)		
Mobility level	16	1-15 NI
Self-care	3	Able Unable NI
Pain	4	No pain Low to intermediate pain Intense pain NI
Strength	7	0-5 NI
Paresis	3	Yes No NI
Ability to feed orally	3	Yes No NI
Ability to communicate	4	Yes No Poorly or symptomatic NI
Ability of understanding	4	Yes No Poorly or symptomatic NI
Ability to ambulate	4	Yes No Poorly or symptomatic NI
Treatment or care related		
Thrombolytic therapy	3	No delta Yes NI
Thrombectomy	3	No delta Yes NI

Tasks	Number of classes	Supporting information for classes
Location	4	• Emergency room • ICU[b] • Inpatient unit • NI
Infection indication	3	• Yes • No • NI
Intracranial hemorrhage	3	• Yes • No • NI
Fall risk	4	• Low risk • Moderate risk • High risk • NI
Pressure ulcer risk	4	• Low risk • Moderate risk • High risk • NI
Fall event during inpatient	3	• Yes • No • NI
Baseline characteristics		

Tasks	Number of classes	Supporting information for classes
High blood pressure	3	• Yes • No • NI
Smoking status	4	• Yes • No • Former • NI
Coronary artery disease	3	• Yes • No • NI
Atrial fibrillation	3	• Yes • No • NI
Diabetes	3	• Yes • No • NI
Prior stroke	3	• Yes • No • NI
Cancer	3	• Yes • No • NI
Alcoholism	4	• Yes • No • Former • NI
Obesity	3	• Yes • No • NI
Dyslipidemia	3	• Yes • No • NI

[a] NI: noninformative.

[b] ICU: intensive care unit.

After the identification of all tasks and the annotation process, the analysis proceeded only with tasks that had substantial (0.61>κ>0.80) and almost perfect (κ≥0.81) agreement between annotators [34]. A total of six tasks were excluded from the final analysis because of moderate or fair agreement or disagreement: (1) active cancer information, (2) strength level, (3) intracranial hemorrhage, (4) ability to understand, (5) self-care, and (6) fall events during inpatient visits. All documents were labeled by the annotators, and the median κ regarding the 24 remaining tasks was 0.74 (IQR 0.65-0.89; substantial agreement).

Patient Characteristics

The descriptive characteristics of patients, including previous comorbidities, NIHSS score, and clinical care, are presented in Table 2.

Table 2. Descriptive characteristics of the patients.

Characteristics	Patients with ischemic stroke evaluated (n=188)	
	Values, median (range)	Values, n (%)
Age (years)	79 (68-87)	N/A[a]
LOS[b] (days)	6 (4-12)	N/A
Sex		
Female	N/A	100 (53)
Male	N/A	88 (47)
Comorbidities		
Previous stroke	N/A	38 (20)
Previous coronary artery disease	N/A	12 (6)
Atrial fibrillation	N/A	33 (18)
Diabetes	N/A	53 (28)
Hypertension	N/A	125 (66)
Smoking status	N/A	15 (8)
Alcoholism	N/A	4 (2)
Treatment and care related		
Antithrombotic therapy	N/A	131 (70)
Thrombolysis with rtPA[c]	N/A	38 (20)
Thrombectomy	N/A	12 (6)
Thrombolysis and thrombectomy	N/A	7(4)
NIHSS[d]		
<8	N/A	147 (78)
>8 and <15	N/A	24 (13)
>15	N/A	17 (9)

[a] N/A: not applicable.

[b] LOS: length of stay.

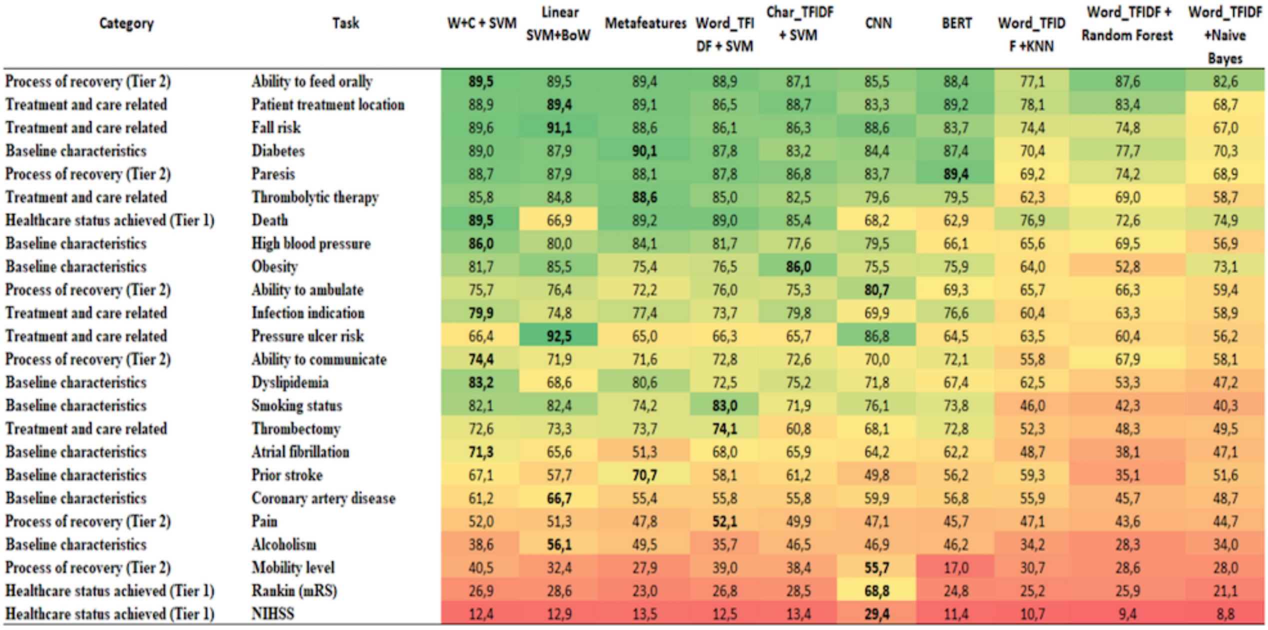
[c] rtPA: alteplase.

[d] NIHSS: National Institutes of Health Stroke Scale.

Experimental Results

The Macro-F1 values for each of the 24 tasks using the 10 compared models are shown in Figure 2. Considering each task separately, there is no single method that always dominates, and there is no agreement on a unique category of tasks that perform better. The ML models SVM+W+C and SVM+BoW were the best and most consistent techniques used in this data set. Both techniques use term-weighting representations that are used alongside SVM classifiers. The latter simply exploits within-document word term frequencies (term frequency), whereas the former, in addition to exploiting data set–oriented term statistics (IDF), also builds character-based n-gram representations of the words in the vocabulary. The character-based n-grams, despite increasing the vocabulary size and sparsity, help to deal with misspellings and word variations that are common in EMRs, which might explain the SVM+W+C good results.

Figure 2. Results of Macro-F1 for each task and comparative models (expressed in percentage). BERT: bidirectional encoder representation from transformers; CNN: convolutional neural network; mRS: Modified Rankin Score; NIHSS: National Institutes of Health Stroke Scale; SVM+BoW: support vector machine plus Bag-of-Words; TFIDF: term frequency-inverted document frequency; W+C+SVM: word-term frequency-inverted document frequency and character-term frequency-inverted document frequency combined with support vector machine.



Category	Task	W+C + SVM	Linear SVM+BoW	Metafeatures	Word_TFIDF + SVM	Char_TFIDF + SVM	CNN	BERT	Word_TFIDF +KNN	Word_TFIDF + Random Forest	Word_TFIDF +Naive Bayes
Process of recovery (Tier 2)	Ability to feed orally	**89,5**	89,5	89,4	88,9	87,1	85,5	88,4	77,1	87,6	82,6
Treatment and care related	Patient treatment location	88,9	**89,4**	89,1	86,5	88,7	83,3	89,2	78,1	83,4	68,7
Treatment and care related	Fall risk	89,6	**91,1**	88,6	86,1	86,3	88,6	83,7	74,4	74,8	67,0
Baseline characteristics	Diabetes	89,0	87,9	**90,1**	87,8	83,2	84,4	87,4	70,4	77,7	70,3
Process of recovery (Tier 2)	Paresis	88,7	87,9	88,1	87,8	86,8	83,7	**89,4**	69,2	74,2	68,9
Treatment and care related	Thrombolytic therapy	85,8	84,8	**88,6**	85,0	82,5	79,6	79,5	62,3	69,0	58,7
Healthcare status achieved (Tier 1)	Death	**89,5**	66,9	89,2	89,0	85,4	68,2	62,9	76,9	72,6	74,9
Baseline characteristics	High blood pressure	**86,0**	80,0	84,1	81,7	77,6	79,5	66,1	65,6	69,5	56,9
Baseline characteristics	Obesity	81,7	85,5	75,4	76,5	**86,0**	75,5	75,9	64,0	52,8	73,1
Process of recovery (Tier 2)	Ability to ambulate	75,7	76,4	72,2	76,0	75,3	**80,7**	69,3	65,7	66,3	59,4
Treatment and care related	Infection indication	**79,9**	74,8	77,4	73,7	79,8	69,9	76,6	60,4	63,3	58,9
Treatment and care related	Pressure ulcer risk	66,4	**92,5**	65,0	66,3	65,7	86,8	64,5	63,5	60,4	56,2
Process of recovery (Tier 2)	Ability to communicate	**74,4**	71,9	71,6	72,8	72,6	70,0	72,1	55,8	67,9	58,1
Baseline characteristics	Dyslipidemia	**83,2**	68,6	80,6	72,5	75,2	71,8	67,4	62,5	53,3	47,2
Baseline characteristics	Smoking status	82,1	82,4	74,2	**83,0**	71,9	76,1	73,8	46,0	42,3	40,3
Treatment and care related	Thrombectomy	72,6	73,3	73,7	**74,1**	60,8	68,1	72,8	52,3	48,3	49,5
Baseline characteristics	Atrial fibrillation	**71,3**	65,6	51,3	68,0	65,9	64,2	62,2	48,7	38,1	47,1
Baseline characteristics	Prior stroke	67,1	57,7	**70,7**	58,1	61,2	49,8	56,2	59,3	35,1	51,6
Baseline characteristics	Coronary artery disease	61,2	**66,7**	55,4	55,8	55,8	59,9	56,8	55,9	45,7	48,7
Process of recovery (Tier 2)	Pain	52,0	51,3	47,8	**52,1**	49,9	47,1	45,7	47,1	43,6	44,7
Baseline characteristics	Alcoholism	38,6	**56,1**	49,5	35,7	46,5	46,9	46,2	34,2	28,3	34,0
Process of recovery (Tier 2)	Mobility level	40,5	32,4	27,9	39,0	38,4	**55,7**	17,0	30,7	28,6	28,0
Healthcare status achieved (Tier 1)	Rankin (mRS)	26,9	28,6	23,0	26,8	28,5	**68,8**	24,8	25,2	25,9	21,1
Healthcare status achieved (Tier 1)	NIHSS	12,4	12,9	13,5	12,5	13,4	**29,4**	11,4	10,7	9,4	8,8

The SVM+W+C model excels in tasks belonging to different categories, such as the ability to feed orally (Tier 2: the process of recovery), with an F1 score of 89.5% (95% CI 89.2%-89.8%); death (tier 1: health care status achieved), with an F1 score of 89.5% (95% CI 87.5%-92.5%); and high blood pressure and dyslipidemia (the baseline characteristics of patients), with F1 scores of 86% (95% CI 83.8%-88.2%) and 83.2% (95% CI 77%-89%), respectively. SVM+BoW, in turn, excels in tasks belonging to the treatment- or care-related categories, such as patient location during treatment (F1 score 89.4%; 95% CI 88%-91%), fall risk (F1 score 91.1%; 95% CI 90.1%-92.1%), and pressure ulcer risk (F1 score 92.5; 95% CI 91.5%-93.5%). The meta-features model, which also exploits SVM as a classifier but uses a completely different text representation, was on average, the third-best placed ML model to cover more tasks with good effectiveness, except in tasks such as diabetes (F1 score 90.1%; 95% CI 88.8%-91.4%) and thrombolytic therapy (F1 score 88.6%; 95% CI 87.5%-90.1%), in which it was the sole winner model (best performer with no ties). The models that used SVM but exploited either only word- or character-based representations came in the fourth and fifth places, losing to methods that exploited both representations in a conjugated way.

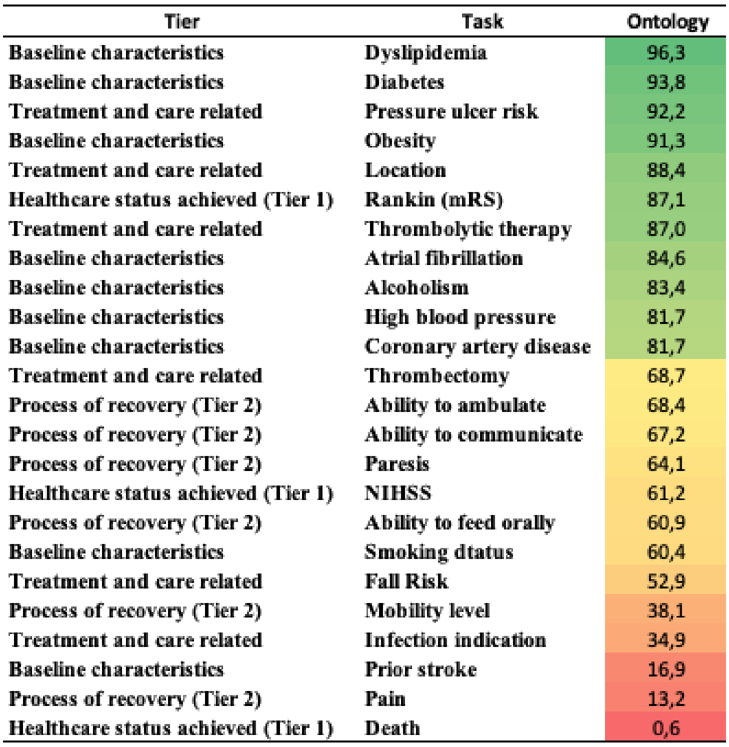
The neural methods CNN and BERT were grouped in the middle, with only moderate effectiveness in most tasks. This outcome is mostly due to the lack of sufficient training data for the optimal deployment of these methods. Indeed, previous work has demonstrated that neural solutions are not adequate for tasks with low to moderate training data, and they can only outperform other more traditional ML methods in text classification tasks when presented with massive amounts of training [35,39], which is generally uncommon in the health domain.

Regarding the effectiveness of the tasks, patient characteristics and care-related process tasks produced better effectiveness. Five of them are examples of good adherence with multiple models, including patient treatment location, fall risk, thrombolytic therapy, diabetes, and paresis, all with multiple models with high effectiveness. Tasks related to measures of mobility, ability to communicate, ability to ambulate, and pain did not achieve high Macro-F1 values in most models.

The tasks with many classes, such as NIHSS (42 classes), mobility level (n=16), and Rankin (n=8), performed worse, regardless of the model. This outcome is mostly due to issues related to the very skewed distribution (high imbalance) found in our unstructured real-life data set. Indeed, the high percentage of NI in the document penalizes effectiveness, mainly for the minor classes, which are captured more faithfully by the Macro-F1 score. However, properly dealing with such an imbalance is not a simple task, as discussed next. Finally, as the sentence length was very similar across tasks and classes, this factor did not affect the results, that is, we could not infer any significant relationship between the mean number of words per sentence and the Macro-F1 scores of the models.

Figure 3 provides information regarding the effectiveness of the OWL classifier. In general, the OWL effectiveness is similar to that of the best ML models, with 11 tasks having a Macro-F1 score higher than 80%. The most interesting issue is that most of the best-performing tasks by OWL *do not coincide* with the best ones produced by the ML models in Figure 2. For instance, the OWL classifier performed very well on the patient's baseline characteristics tasks, such as NIHSS and mRS scale, precisely the ones in which the ML models performed poorly. Overall, the OWL strategy was more robust in the tasks in which the ML models suffered from a scarcity of examples and high imbalance. On the contrary, OWL suffered on tasks that were much more passible in interpretation and had more text representations from those for which they were built [49,80]. For instance, in the *death* task, despite good within-annotator agreement, we believe that due to a variety of clinical terms in the clinical text used to describe multiple clinical concepts, the rules initially created failed to reflect the understanding of a noninformative sentence versus a sentence that reports the vital signs of patients, which penalized the OWL model.

Figure 3. Effectiveness results for the ontology-based model. mRS: Modified Rankin Score; NIHSS: National Institutes of Health Stroke Scale.



Tier	Task	Ontology
Baseline characteristics	Dyslipidemia	96,3
Baseline characteristics	Diabetes	93,8
Treatment and care related	Pressure ulcer risk	92,2
Baseline characteristics	Obesity	91,3
Treatment and care related	Location	88,4
Healthcare status achieved (Tier 1)	Rankin (mRS)	87,1
Treatment and care related	Thrombolytic therapy	87,0
Baseline characteristics	Atrial fibrillation	84,6
Baseline characteristics	Alcoholism	83,4
Baseline characteristics	High blood pressure	81,7
Baseline characteristics	Coronary artery disease	81,7
Treatment and care related	Thrombectomy	68,7
Process of recovery (Tier 2)	Ability to ambulate	68,4
Process of recovery (Tier 2)	Ability to communicate	67,2
Process of recovery (Tier 2)	Paresis	64,1
Healthcare status achieved (Tier 1)	NIHSS	61,2
Process of recovery (Tier 2)	Ability to feed orally	60,9
Baseline characteristics	Smoking dtatus	60,4
Treatment and care related	Fall Risk	52,9
Process of recovery (Tier 2)	Mobility level	38,1
Treatment and care related	Infection indication	34,9
Baseline characteristics	Prior stroke	16,9
Process of recovery (Tier 2)	Pain	13,2
Healthcare status achieved (Tier 1)	Death	0,6

A direct comparison between OWL and the best ML method is presented in Figures 4 and 5, in which Figure 4 represents the tasks in which OWL performed better than the best ML model for the same tasks and Figure 5 represents the tasks with higher F1 scores in the ML model against OWL. SVM+W+C has a considerable advantage over the other ML strategies, as the strategy of choice to be compared in the vast majority of cases. The best tasks performed by the best model in each case, either SVM+W+C or OWL, do not coincide. Indeed, there is a potential complementarity between ML and alternatives.

Figure 4. Best performed tasks in Ontology versus top-ranked model. mRS: Modified Rankin Score; NIHSS: National Institutes of Health Stroke Scale; SVM: support vector machine; W+C+SVM: word- term frequency-inverted document frequency and character- term frequency-inverted document frequency combined with support vector machine.

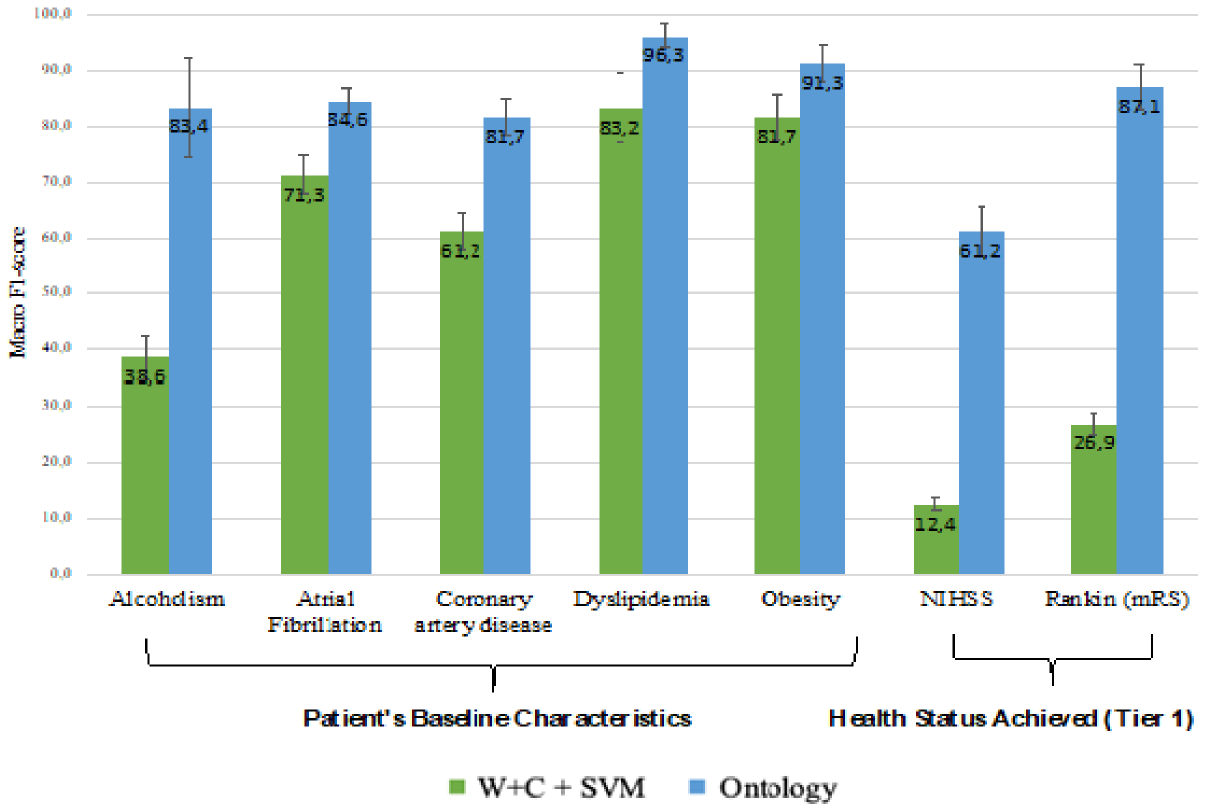
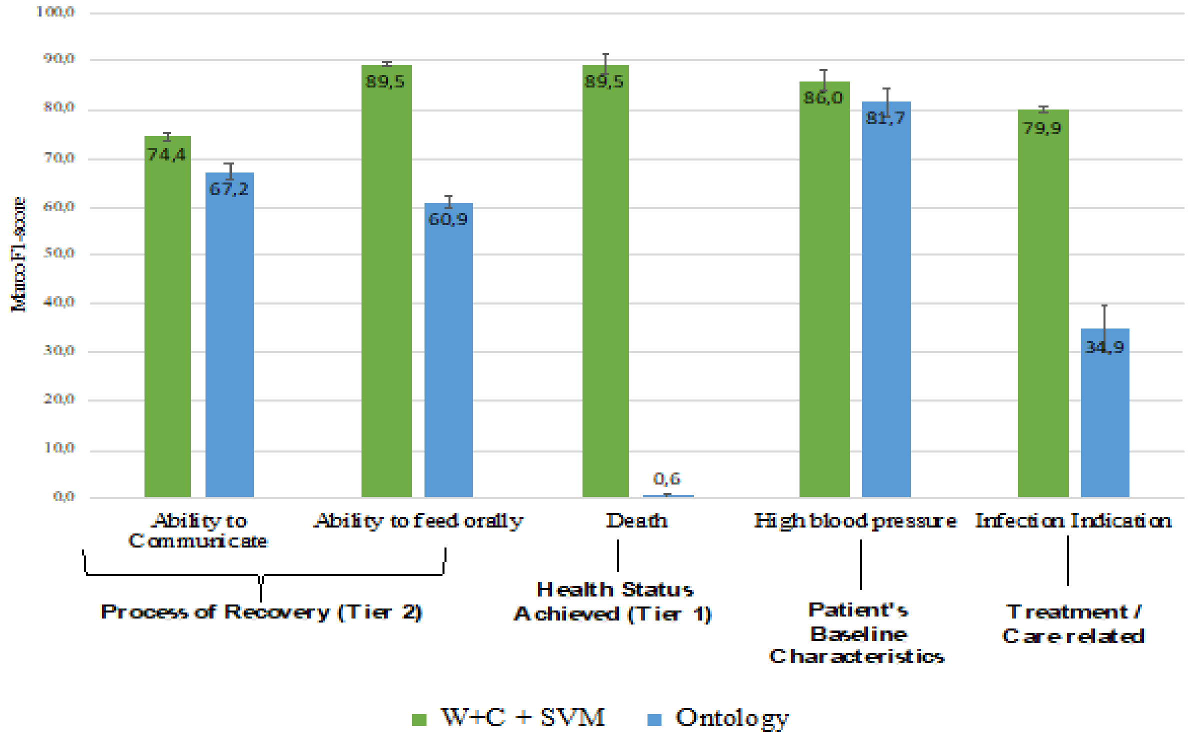


Figure 5. Best performed tasks in the top-ranked model versus Ontology. SVM: support vector machine; W+C: word-term frequency-inverted document frequency and character-term frequency-inverted document frequency.



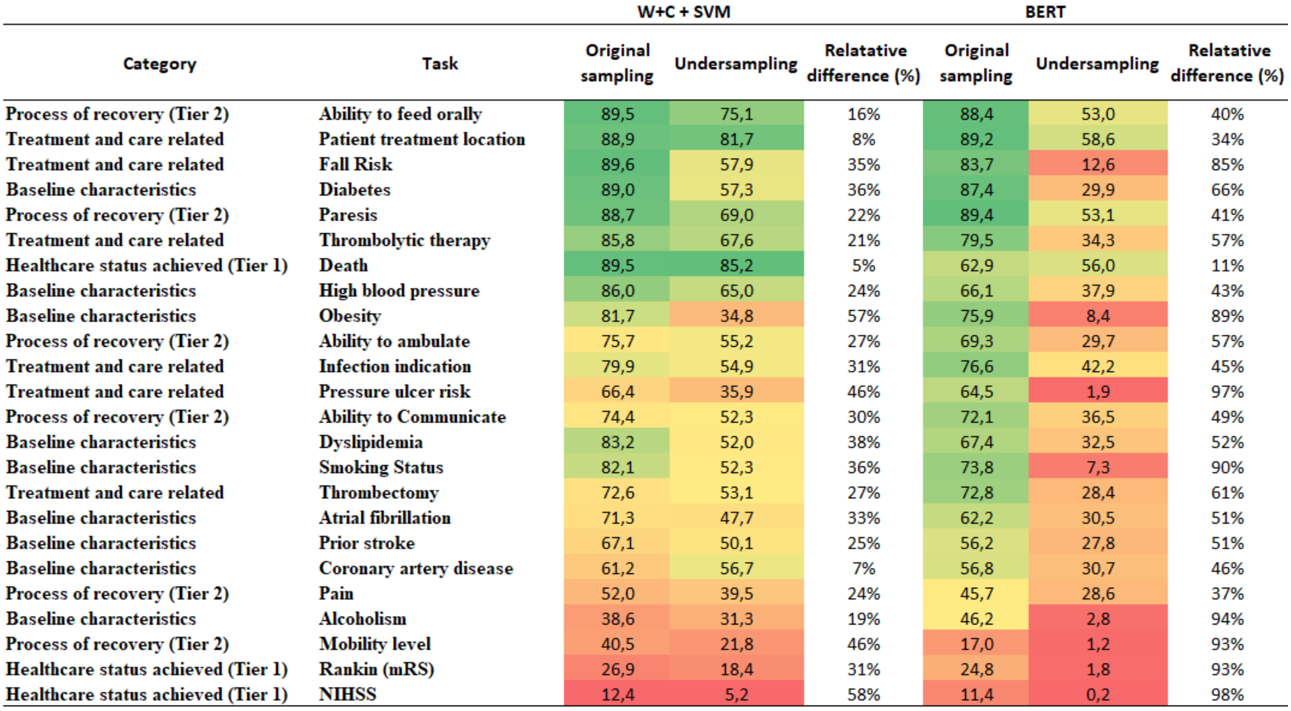
Effect of Class Imbalance on the Results—Undersampling

As we have discussed, all our tasks are extremely skewed, in the sense that the NI (noninformed; majority) class dominates over the other (minority) classes, where the useful information really lies. This imbalance occurs in a proportion that can achieve 1:1000 examples in the minority class to the majority class for some tasks.

This imbalance may cause bias in the training data set influencing some of the experimented ML algorithms toward giving priority to NI class, ultimately undermining the classification of the minority classes on which predictions are most important. One approach to addressing the problem of class imbalance is to randomly resample the training data set. A simple, yet effective approach to deal with the problem is to randomly delete examples from the majority class, a technique known as random undersampling [81].

The results of this experiment are shown in Figure 6, which compares the performance of the classifiers in scenarios with and without undersampling. For the sake of space, we only show the results for the best nonneural (W+C+SVM) and neural (BERT) classifiers, but the results are similar for all tested classifiers (Multimedia Appendix 7).

Figure 6. Results of Macro-F1 score in the undersampling sample, expressed by percentage. mRS: Modified Rankin Score; NIHSS: National Institutes of Health Stroke Scale; SVM: support vector machine; W+C: word- term frequency-inverted document frequency and character- term frequency-inverted document frequency.



		W+C + SVM			BERT		
Category	Task	Original sampling	Undersampling	Relatative difference (%)	Original sampling	Undersampling	Relatative difference (%)
Process of recovery (Tier 2)	Ability to feed orally	89,5	75,1	16%	88,4	53,0	40%
Treatment and care related	Patient treatment location	88,9	81,7	8%	89,2	58,6	34%
Treatment and care related	Fall Risk	89,6	57,9	35%	83,7	12,6	85%
Baseline characteristics	Diabetes	89,0	57,3	36%	87,4	29,9	66%
Process of recovery (Tier 2)	Paresis	88,7	69,0	22%	89,4	53,1	41%
Treatment and care related	Thrombolytic therapy	85,8	67,6	21%	79,5	34,3	57%
Healthcare status achieved (Tier 1)	Death	89,5	85,2	5%	62,9	56,0	11%
Baseline characteristics	High blood pressure	86,0	65,0	24%	66,1	37,9	43%
Baseline characteristics	Obesity	81,7	34,8	57%	75,9	8,4	89%
Process of recovery (Tier 2)	Ability to ambulate	75,7	55,2	27%	69,3	29,7	57%
Treatment and care related	Infection indication	79,9	54,9	31%	76,6	42,2	45%
Treatment and care related	Pressure ulcer risk	66,4	35,9	46%	64,5	1,9	97%
Process of recovery (Tier 2)	Ability to Communicate	74,4	52,3	30%	72,1	36,5	49%
Baseline characteristics	Dyslipidemia	83,2	52,0	38%	67,4	32,5	52%
Baseline characteristics	Smoking Status	82,1	52,3	36%	73,8	7,3	90%
Treatment and care related	Thrombectomy	72,6	53,1	27%	72,8	28,4	61%
Baseline characteristics	Atrial fibrillation	71,3	47,7	33%	62,2	30,5	51%
Baseline characteristics	Prior stroke	67,1	50,1	25%	56,2	27,8	51%
Baseline characteristics	Coronary artery disease	61,2	56,7	7%	56,8	30,7	46%
Process of recovery (Tier 2)	Pain	52,0	39,5	24%	45,7	28,6	37%
Baseline characteristics	Alcoholism	38,6	31,3	19%	46,2	2,8	94%
Process of recovery (Tier 2)	Mobility level	40,5	21,8	46%	17,0	1,2	93%
Healthcare status achieved (Tier 1)	Rankin (mRS)	26,9	18,4	31%	24,8	1,8	93%
Healthcare status achieved (Tier 1)	NIHSS	12,4	5,2	58%	11,4	0,2	98%

As it can been seen, the undersampling process caused major losses in both classifiers. Such losses occurred across all tasks, varying from 5% of Macro-F1 score reduction (death) to 58% (NIHSS) for W+C+SVM, and 11% (death) to 98% (NIHSS) of Macro-F1 effectiveness loss in BERT. The largest losses for the neural method were expected, as this type of classifier is more sensitive to the amount of training. However, to a certain degree, all the classifiers suffered major losses after the undersampling process. These results may be attributed to the largest difference in class distribution between training and testing and the inevitable loss of information that comes after the removal of training instances after undersampling.

These phenomena can be better seen when we look at the individual values of F1, precision, and recall of the classes of the tasks. Table 3 shows an example of the tasks of infection indication, thrombolytic therapy, and ability to communicate with the W+C+SVM classifier. As we can see, all classes have a reduced F1 in the undersampling scenario. This is mainly due to a large reduction in the precision of the classes. This happens because W+C+SVM misclassifies NI instances as belonging to some of the relevant classes. As the classifier is obliged to categorize a sentence in one of the existing classes, the lack of information about the fact that a sentence does not have useful information for assigning the sentence in one of the classes of interest confounds the classifier. In other words, the negative information about the NI (eg, frequent words in NI sentences that help to characterize this class but that are also shared by some non-NI instances, and whose frequency was altered by the undersampling) is in fact useful information for avoiding false positives, which may cause many problems in a real scenario, including false alarms, waste of resources, and distrust of the automatic methods.

Table 3. Comparison of undersampling and original sampling in terms of precision, recall, and Macro-F1 score (W+C+SVM model).

Class	Undersampling			Original sampling		
	Precision	Recall	F1 (%)[a]	Precision	Recall	F1 (%)[a]
Infection indicative						
−1	1	0.96	98	0.99	1	99
0	0.39	0.89	54	0.88	0.75	81
1	0.28	0.82	42	0.68	0.53	59
Thrombolytic therapy						
−1	1	0.98	99	1	1	100
0	0.32	0.62	42	0.69	0.52	59
1	0.31	0.91	47	0.89	0.91	90
Ability to communicate						
−1	1	0.96	98	0.99	1	100
0	0.34	0.63	44	0.9	0.26	40
1	0.35	0.81	49	0.76	0.64	69
2	0.32	0.93	48	0.82	0.8	81

[a] Macro-F1 score (W+C+SVM model).

Feature Importance

For the tasks presented in Textbox 1 (alcoholism, atrial fibrillation, coronary artery disease, dyslipidemia, obesity, NIHSS, Rankin [mRs], infection indicators, high blood pressure, death, ability to feed orally, and ability to communicate), we present the top 10 clinical features (ie, words) used in the task prediction in Textbox 1, which means the 10 features with higher contribution to task prediction. This analysis helps to better understand the divergence between approaches. It is worth noting that in the tasks in which the ML models performed better (second column), the top-ranked features were all related to the semantics of the task. For instance, considering the *death* task as an example, the ML model was able to identify important features for the task, which produced a higher information gain than the OWL model. Indeed, for *death*, only three features of the 10 most relevant explicitly use the word *death*, but most features are somewhat related to this outcome. This finding suggests data quality issues (vocabulary coverage) that may drastically influence the effectiveness of the OWL strategy, which exploits only rules that explicitly contain the word *death* (or related ones) but no other terms. However, for the features in the first column, in which the OWL models were better, there were still features with considerable contributions that were not directly related to the information sought. For example, to mention the NIHSS task, rule-based knowledge models built alongside clinical domain vocabulary specialists may be the best alternative.

Textbox 1. Top 10 clinical indicators for task prediction models and feature importance. In parenthesis, the translation to English language is indicated, where there may be misspellings in the original writing that are also indicated.

Alcoholism

- etilismo (alcoholism)
- etilista (alcoholic)
- fumo (smoke)
- históira (story with misspelling in the original)
- álcool (alcohol)
- cart
- osteoartrose (osteoarthritis)
- ttu (short for transurethral resection of the prostate)
- tabagismo (smoking)
- cesária (cesarean)

Atrial fibrillation

- fa (short for atrial fibrillation)
- comorbidades (comorbidities)
- acfa (short for atrial fibrillation)
- paroxística (paroxysmal)
- has (short for high blood pressure)
- anticoagulado (anticoagulated)
- depressão (depression)
- indeterminado (indeterminate)
- digoxina (digoxin)
- institucionalizada (institutionalized)

Coronary artery disease

- cardiopatia (heart disease)
- isquêmica (ischemic)
- actp (short for percutaneous transluminal coronary angioplasty)
- dp
- crm (short for myocardial revascularization surgery)
- iam (short for acute myocardial infarction)
- 2014
- infarto (short for acute myocardial infarction)
- mm
- sf

Dyslipidemia

- dislipidemia (dyslipidemia)
- comorbidades (comorbidities)
- 1hora
- cesária (cesarean)
- morbidades (morbidities)
- puerpera (puerperal)
- has (short for high blood pressure)

- fêmur (fêmur)
- tep
- previas (previous)

Obesity

- BMI (short for body mass index)
- obesidade (obesity)
- m²
- 1994
- lipschitz
- eutrofia
- altura (height)
- peso (weight)
- estatura (stature)
- obesa (obese)

National Institutes of Health Stroke Scale

- nihss
- súbito (sudden)
- asistolia (asystolia)
- sens
- territ
- suboclusiva (subocclusive)
- perg
- mecania (mecanic with mispelling in the original)
- severo (severe)
- visto (seen)

Ability to communicate

- afasia (afasia)
- comunicativa (talkative)
- disartria (dysarthria)
- comunicativo (talkative)
- colóquio (colloquium)
- verbalizando (verbalizing)
- alerta (alert)
- verbaliza (verbalizes)
- expressão (expression)
- hemiparesia (hemiparesis)

Ability to feed orally

- vo (short for orally)
- sne (short for nasoenteral probe)
- dieta (diet)
- pastosa (pasty)
- gastrostomia (gastrostomy)

- enteral (enteral)
- aceitação (acceptance)
- semi (semi)
- exclusiva (exclusive)
- polimérica (polymeric diet)

Death

- óbito (death)
- constato (i’ve verified)
- leito (bed)
- ar (air)
- estável (stable)
- ambiente (environment or room)
- no
- doação (donation)
- obito (death with misspelling in the original)
- óbito (death with misspelling in the original)

High blood pressure

- has (short for high blood pressure)
- dm (short for diabetes)
- dislipidemia (dyslipidemia)
- dm2 (short for diabetes type 2)
- comorbidades (comorbidities)
- fa (short for atrial fibrillation)
- artrite (arthritis)
- definitivo (definitive)
- reumatoide (rheumatoid)
- demencial (dementia)

Infection indication

- afebril (afebrile)
- flogísticos (phlogistic)
- sinais (signs)
- cefuroxima (cefuroxime)
- inserção (insertion)
- tax
- klebsiella (klebsiella)
- d0 (short for day 0)
- atb (short for antibiotics)
- azitromicina (azithromycin)

Modified Rankin Score

- rankin
- mrankin
- demência (dementia)

- caminha (walks)
- corversa (talks)
- alimenta (feed)
- alzheimer
- aparentes (apparent)
- comer (eat)
- mrk (mrs with misspelling in the original)

Discussion

Principal Findings

The study intended to recognize the path and opportunities that may be advanced in terms of the technological capacity to support the outcome measurement process for the stroke care pathway. Real-world sentences from ischemic stroke EMRs were used to develop automatic models using ML and NLP techniques. It was possible to identify that SVM+W+C and SVM+BoW were the most effective models to be used to classify characteristics of a patient and process of care based on the extraction of Brazilian-Portuguese free-text data from the EMRs of patients. Ontological rules were also effective in this task, and perhaps even more importantly, most of the best-performing tasks with the OWL and ML models did not coincide. This outcome opens up the opportunity to exploit such complementarities to improve the coverage of tasks when implementing a real solution for outcome management or even to improve the individual effectiveness of each alternative by means of ensemble techniques such as stacking [82].

One of the good practices that the literature has demonstrated to increase the success of ML algorithms applied to health care is the inclusion of a clinical background in the annotation process [83]. The availability of training data is critical in obtaining good results, thus indicating that variations in clinical terms found in the clinical text could be specific to the type and source of clinical notes that may not have been captured in an available resource. The results from our feature importance analysis are consistent with other study results [21,68,76,83-85] concerning many clinical terms applied to multiple clinical concepts, although there are specific patterns based on semantic types that can help. In general, it is difficult to determine the correct concept when a clinical term normalizes to multiple concepts, and this issue can penalize the effectiveness of the model [86,87].

Our effectiveness results agree with the literature [83,88], in which a Macro-F1 score >80% is considered a successful extraction of medical records. Even though there is still a need to cover more tasks related to ICHOM patient-reported outcome measures [3,74,76,85], we hypothesized that these tasks comprise a feeling state, and the lack of normalization of data contained in EMRs may explain the fact that these task categories did not perform very well [70,89]. Medical records related to baseline characteristics and care processes typically contain much more structured data (eg, numerical values for tasks) than medical patient-reported outcomes, which focus more on unstructured data [83,90]. This issue has been explored in previous studies on EMR-based clinical quality measures [22,82], in which it is suggested that these kinds of data (for baseline characteristics and care-related processes) have the potential to be scaled in other clinical conditions, such as cardiovascular and endocrine conditions [83].

Previous studies have found various advantages of EMR compared with traditional paper records [91]. However, as reported by Ausserhofer et al [12], care workers do not find them useful for guaranteeing safe care and treatment because of the difficulty of tracking clinical and quality measures. The same authors have discussed the importance of having IT capability to track care workers' documentation while increasing safety and quality of care. They emphasized that this approach is important for addressing EMR data collection issues that have been historically extracted via manual review by clinical experts, leading to scalability and cost issues [83,85,90]. In our study, it was possible to demonstrate that for the stroke care pathway, the use of ML models to measure clinical outcomes remains a challenge, but the technology has the potential to support the extraction of relevant patient characteristics and care-process information.

Despite the challenges regarding the accuracy of the outcome measures, promising approaches regarding baseline characteristics and care-related process data have been achieved. This may be the first step toward unlocking the full potential of EMR data [83]. The usefulness of having baseline characteristics tracked is to assist disease prevalence studies and identify opportunities to guide political decisions about the public health sector [13,92,93], automatize eligibility of patients for clinical research [84], and feed risk assessment tools [94]. On the contrary, care-related process metrics boost the opportunity to improve decision-making with new technologies, maintain the effectiveness of treatments, and encourage alternative remuneration models [17,92,95].

The next step would be to invest in the automation of tasks at the patient level that support the control of the progression of patients in real-time during stroke episodes. In a similar manner, it would be useful to identify opportunities to improve the EMR data quality, such as the implementation of quality software with dynamic autocompletes with normalized terms register. The use of NLP for quality measures also adds to the capture of large amounts of clinical data from EMRs [82]. The products of NLP and mixed methods pipelines could potentially impact a number of clinical areas and could facilitate appropriate care by feeding hospital outcome indicators and data to support epidemiological studies or value-based programs [82].

Limitations

This study had several limitations. For clinical NLP method development to advance further globally and to become an integral part of clinical outcome research or have a natural place in clinical practice, there are still challenges ahead. Our work is based on the EMR of a single center, with a limited number of annotated patients. Thus, further work is needed to test this approach in EMRs from different centers with different patients, who may use different languages for clinical documentation. We have no access to data from exams or hospital indicators, which is the reason why our infection identification, for example, was based on any report of antibiotic use, typical symptoms of infection, or tests described. We were unable to find data samples that included all the risk factors that were discovered in the literature. It would be worth conducting a future study with a larger and different data set with more features to examine whether the findings of this research are still valid. Finally, the design focused on sentences can be significantly influenced by the NI data volume—if a patient smokes, this will probably be reflected in just one sentence, maybe two, and for all of the others, you will have NI. One possible approach would be to use hierarchy models to first classify whether a sentence is relevant and then evolve to classification algorithms to predict classes. Then, the entire record can inform the prediction of the outcome of patients, instead of saying whether a specific sentence indicates a task.

Regarding the undersampling experiment, more intelligent strategies such as choosing the *most positive of the negative samples* or Tomek links [81] should be tested for better effectiveness. We leave this for future work and suggest practical purposes to maintain the original distribution, whereas more effective strategies are not further studied.

Conclusions

This study is innovative in that it considered many and diverse types of automatic classifiers (neural, nonneural, and ontological) using a large real-world data set containing thousands of textual sentences from real-world EMRs and a large number of tasks (n=24) with multiple classes using Brazilian-Portuguese unstructured free-text EMR databases. The effectiveness of these models demonstrated a better result when used to classify care processes and patient characteristics than patient-reported outcomes, which suggests that advances in intelligence in informational technology for clinical outcomes are still a gap in the scalability of outcome measurements in health care. Future research should explore the development of mixed methods to increase task effectiveness. Advances in IT capacity have proved to be essential for the scalability and agility of the ability to measure health outcomes and how it reflects on its external validation to support health real-time quality measurement indicators.

Conflicts of Interest

The authors disclose receipt of the following financial support for the research, authorship, and publication of this paper: this research was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES)–Finance Code 001 and National Council for Scientific and Technological Development (CNPq 465518/2014-1 and others), Research Support Foundation of the State of Minas Gerais (FAPEMIG), Google, and NVIDIA Corporation.

Multimedia Appendix 1
Example of an evolution on the electronic medical record.
[DOCX File , 14 KB - medinform_v9i11e29120_app1.docx]

Multimedia Appendix 2
Example of the annotation process.
[DOCX File , 19 KB - medinform_v9i11e29120_app2.docx]

Multimedia Appendix 3
Data set characteristics.
[DOCX File , 20 KB - medinform_v9i11e29120_app3.docx]

Multimedia Appendix 4
Details of the automatic text classification methods.
[DOCX File , 28 KB - medinform_v9i11e29120_app4.docx]

Multimedia Appendix 5
Experimental procedure.
[PNG File , 97 KB - medinform_v9i11e29120_app5.png]

Multimedia Appendix 6
Experimental protocol details—specific parameter tuning.

[DOCX File , 15 KB - medinform_v9i11e29120_app6.docx]

Multimedia Appendix 7
Results of F1 score from the random undersampling experiment. BERT: bidirectional encoder representation from transformers; BoW: Bag-of-Words; KNN: K-nearest neighbor; mRS: Modified Rankin Score; NIHSS: National Institutes of Health Stroke Scale; SVM: support vector machine; TFIDF: term frequency-inverted document frequency; W+C: word- term frequency-inverted document frequency and character- term frequency-inverted document frequency.
[PNG File , 308 KB - medinform_v9i11e29120_app7.png]

References

1. GBD 2016 Stroke Collaborators. Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol 2019 May;18(5):439-458 [FREE Full text] [doi: 10.1016/S1474-4422(19)30034-1] [Medline: 30871944]
2. Findings From the Global Burden of Disease Study 2017. Institute for Health Metrics and Evaluation (IHME). 2018. URL: http://www.healthdata.org/sites/default/files/files/policy_report/2019/GBD_2017_Booklet.pdf [accessed 2021-10-11]
3. Wang W, Kiik M, Peek N, Curcin V, Marshall I, Rudd A, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLoS One 2020 Jun 12;15(6):e0234722 [FREE Full text] [doi: 10.1371/journal.pone.0234722] [Medline: 32530947]
4. Kamal H, Lopez V, Sheth SA. Machine learning in acute ischemic stroke neuroimaging. Front Neurol 2018 Nov 8;9:945 [FREE Full text] [doi: 10.3389/fneur.2018.00945] [Medline: 30467491]
5. Feng R, Badgeley M, Mocco J, Oermann EK. Deep learning guided stroke management: a review of clinical applications. J Neurointerv Surg 2018 Apr;10(4):358-362 [FREE Full text] [doi: 10.1136/neurintsurg-2017-013355] [Medline: 28954825]
6. Lee E, Kim Y, Kim N, Kang D. Deep into the brain: artificial intelligence in stroke imaging. J Stroke 2017 Sep;19(3):277-285 [FREE Full text] [doi: 10.5853/jos.2017.02054] [Medline: 29037014]
7. Wodchis WP, Austin PC, Henry DA. A 3-year study of high-cost users of health care. Can Med Asso J 2016 Feb 16;188(3):182-188 [FREE Full text] [doi: 10.1503/cmaj.150064] [Medline: 26755672]
8. Markatou M, Don PK, Hu J, Wang F, Sun J, Sorrentino R, et al. Case-based reasoning in comparative effectiveness research. IBM J Res Dev 2012 Sep;56(5):4:1-4:12. [doi: 10.1147/JRD.2012.2198311]
9. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 2010 Jun;48(6 Suppl):106-113. [doi: 10.1097/MLR.0b013e3181de9e17] [Medline: 20473190]
10. Chechulin Y, Nazerian A, Rais S, Malikov K. Predicting patients with high risk of becoming high-cost healthcare users in Ontario (Canada). Health Care Policy 2014 Feb 26;9(3):68-79. [doi: 10.12927/hcpol.2014.23710]
11. 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 [FREE Full text] [doi: 10.1038/s41598-017-05778-z] [Medline: 28729710]
12. Ausserhofer D, Favez L, Simon M, Zúñiga F. Electronic health record use in Swiss nursing homes and its association with implicit rationing of nursing care documentation: multicenter cross-sectional survey study. JMIR Med Inform 2021 Mar 02;9(3):e22974 [FREE Full text] [doi: 10.2196/22974] [Medline: 33650983]
13. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61-81 [FREE Full text] [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]
14. Fernandes M, Sun H, Jain A, Alabsi HS, Brenner LN, Ye E, et al. Classification of the disposition of patients hospitalized with COVID-19: reading discharge summaries using natural language processing. JMIR Med Inform 2021 Mar 10;9(2):e25457 [FREE Full text] [doi: 10.2196/25457] [Medline: 33449908]
15. Porter M, Lee T. The strategy that will fix health care. Harvard Business Review. 2013. URL: https://hbr.org/2013/10/the-strategy-that-will-fix-health-care [accessed 2021-09-07]
16. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak 2018 Jun 22;18(1):44 [FREE Full text] [doi: 10.1186/s12911-018-0620-z] [Medline: 29929496]
17. Glaser J. It’s time for a new kind of electronic health record. Harvard Bussiness Review. 2020. URL: https://hbr.org/2020/06/its-time-for-a-new-kind-of-electronic-health-record [accessed 2021-09-07]
18. Carberry K, Landman Z, Xie M, Feeley T, Henderson J, Fraser C. Incorporating longitudinal pediatric patient-centered outcome measurement into the clinical workflow using a commercial electronic health record: a step toward increasing value for the patient. J Am Med Inform Assoc 2016 Jan;23(1):88-93 [FREE Full text] [doi: 10.1093/jamia/ocv125] [Medline: 26377989]
19. Afzal N, Sohn S, Abram S, Liu H, Kullo IJ, Arruda-Olson AM. Identifying peripheral arterial disease cases using natural language processing of clinical notes. IEEE EMBS Int Conf Biomed Health Inform 2016 Feb;2016:126-131 [FREE Full text] [doi: 10.1109/BHI.2016.7455851] [Medline: 28111640]

20. Wi C, Sohn S, Rolfes MC, Seabright A, Ryu E, Voge G, et al. Application of a natural language processing algorithm to asthma ascertainment. An automated chart review. Am J Respir Crit Care Med 2017 Aug 15;196(4):430-437 [FREE Full text] [doi: 10.1164/rccm.201610-2006OC] [Medline: 28375665]
21. Chase HS, Mitrani LR, Lu GG, Fulgieri DJ. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC Med Inform Decis Mak 2017 Feb 28;17(1):24 [FREE Full text] [doi: 10.1186/s12911-017-0418-4] [Medline: 28241760]
22. Garvin JH, Kim Y, Gobbel GT, Matheny ME, Redd A, Bray BE, et al. Automating quality measures for heart failure using natural language processing: a descriptive study in the department of veterans affairs. JMIR Med Inform 2018 Jan 15;6(1):e5 [FREE Full text] [doi: 10.2196/medinform.9150] [Medline: 29335238]
23. Dai H, Lee Y, Nekkantti C, Jonnagaddala J. Family history information extraction with neural attention and an enhanced relation-side scheme: algorithm development and validation. JMIR Med Inform 2020 Dec 01;8(12):e21750 [FREE Full text] [doi: 10.2196/21750] [Medline: 33258777]
24. Lee TH. Putting the value framework to work. N Engl J Med 2010 Dec 23;363(26):2481-2483. [doi: 10.1056/NEJMp1013111] [Medline: 21142527]
25. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010 Aug 5;363(6):501-504. [doi: 10.1056/NEJMp1006114] [Medline: 20647183]
26. Porter ME, Larsson S, Lee TH. Standardizing patient outcomes measurement. N Engl J Med 2016 Feb 11;374(6):504-506. [doi: 10.1056/NEJMp1511701] [Medline: 26863351]
27. Wilson JL, Hareendran A, Grant M, Baird T, Schulz UG, Muir KW, et al. Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified Rankin Scale. Stroke 2002 Sep;33(9):2243-2246. [doi: 10.1161/01.str.0000027437.22450.bd] [Medline: 12215594]
28. Lyden PD, Lu M, Levine SR, Brott TG, Broderick J, NINDS rtPA Stroke Study Group. A modified National Institutes of Health Stroke Scale for use in stroke clinical trials: preliminary reliability and validity. Stroke 2001 Jun;32(6):1310-1317. [doi: 10.1161/01.str.32.6.1310] [Medline: 11387492]
29. Caso V, Zakaria M, Tomek A, Mikulik R, Martins S, Nguyen T, et al. Improving stroke care across the world: the ANGELS Initiative. CNS - Oruen Ltd. 2018. URL: https://www.oruen.com/wp-content/uploads/2018/12/Review-article-4.pdf [accessed 2021-09-07]
30. Honnibal M, Montani I. Industrial-strength natural language processing. spaCy. URL: https://spacy.io [accessed 2021-09-07]
31. Klie J, Bugert M, Boullosa B, de Castilho RE, Gurevych I. The INCEpTION platform: machine-assisted and knowledge-oriented interactive annotation. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. 2018 Aug 01 Presented at: 27th International Conference on Computational Linguistics: System Demonstrations; August, 2018; Santa Fe, New Mexico p. 5-9 URL: https://aclanthology.org/C18-2002/ [doi: 10.18653/v1/d18-2022]
32. Manning C, Raghawan P, Schutze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press; 2008:1-506.
33. Manning C, Schutze H. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press; 1999:1-720.
34. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005 May;37(5):360-363 [FREE Full text] [Medline: 15883903]
35. Cunha W, Mangaravite V, Gomes C, Canuto S, Resende E, Nascimento C, et al. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: a comprehensive comparative study. Inf Process Manag 2021 May;58(3):102481. [doi: 10.1016/j.ipm.2020.102481]
36. Canuto S, Salles T, Rosa TC, Couto T, Gonçalves MA. Similarity-based synthetic document representations for meta-feature generation in text classification. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019 Jan 01 Presented at: SIGIR '19: The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval; Jul 21-25, 2019; Paris France p. 355-364. [doi: 10.1145/3331184.3331239]
37. Canuto S, Salles T, Gonçalves M, Rocha L, Ramos G, Gonçalves G. On efficient meta-level features for effective text classification. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. 2014 Jan 01 Presented at: CIKM '14: 2014 ACM Conference on Information and Knowledge Management; Nov 3-7, 2014; Shanghai China p. 1709-1718. [doi: 10.1145/2661829.2662060]
38. Canuto S, Sousa DX, Goncalves MA, Rosa TC. A thorough evaluation of distance-based meta-features for automated text classification. IEEE Trans Knowl Data Eng 2018 Mar 27;30(12):2242-2256. [doi: 10.1109/tkde.2018.2820051]
39. Cunha W, Canuto S, Viegas F, Salles T, Gomes C, Mangaravite V, et al. Extended pre-processing pipeline for text classification: on the role of meta-feature representations, sparsification and selective sampling. Inf Process Manag 2020 Jul;57(4):102263. [doi: 10.1016/j.ipm.2020.102263]
40. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
41. Breiman L. Random forests. Mach Learn 2001 Oct 1;45(1):5-32. [doi: 10.1023/A:1010933404324]

42. Kowsari K, Meimandi KJ, Heidarysafa M, Mendu S, Barnes L, Brown D. Text classification algorithms: a survey. Information 2019 Apr 23;10(4):150. [doi: 10.3390/info10040150]
43. Larson RR. Introduction to information retrieval. J Am Soc Inf Sci Technol 2009 Oct 19;61(4):852-853. [doi: 10.1002/asi.21234]
44. Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001 Dec;17(12):1131-1142. [doi: 10.1093/bioinformatics/17.12.1131] [Medline: 11751221]
45. Almeida MB, Bax MP. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. Ci Inf 2003 Dec;32(3):7-20. [doi: 10.1590/s0100-19652003000300002]
46. Allahyari M, Kochut K, Janik M. Ontology-based text classification into dynamically defined topics. In: Proceedings of the IEEE International Conference on Semantic Computing. 2014 Jan 01 Presented at: IEEE International Conference on Semantic Computing; Jun 16-18, 2014; Newport Beach, CA, USA p. 273-278. [doi: 10.1109/icsc.2014.51]
47. Chi N, Lin K, Hsieh S. Using ontology-based text classification to assist job hazard analysis. Adv Eng Inf 2014 Oct;28(4):381-394. [doi: 10.1016/j.aei.2014.05.001]
48. Garla VN, Brandt C. Ontology-guided feature engineering for clinical text classification. J Biomed Inform 2012 Oct;45(5):992-998 [FREE Full text] [doi: 10.1016/j.jbi.2012.04.010] [Medline: 22580178]
49. Wang B, McKay R, Abbass H, Barlow M. A comparative study for domain ontology guided feature extraction. Australian Computer Society. 2003. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.3384&rep=rep1&type=pdf [accessed 2021-09-07]
50. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag 1988 Jan;24(5):513-523. [doi: 10.1016/0306-4573(88)90021-0]
51. Andrade CM, Gonçalves MA. Combining representations for effective citation classification. In: Proceedings of The International Workshop on Mining Scientific Publications. 2020 Presented at: The International Workshop on Mining Scientific Publications; Aug 2020; Wuhan, China.
52. Cortes EG, Woloszyn V, Barone DA. When, where, who, what or why? A hybrid model to question answering systems. In: Computational Processing of the Portuguese Language. Cham: Springer; 2018.
53. Viegas F, Rocha L, Resende E, Salles T, Martins W, Freitas MF, et al. Exploiting efficient and effective lazy Semi-Bayesian strategies for text classification. Neurocomput 2018 Sep 13;307:153-171. [doi: 10.1016/j.neucom.2018.04.033]
54. Fei Y. Simultaneous Support Vector selection and parameter optimization using Support Vector Machines for sentiment classification. In: Proceedings of the 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS). 2016 Presented at: 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS); Aug 26-28, 2016; Beijing, China. [doi: 10.1109/ICSESS.2016.7883015]
55. Shen Y. Selection incentives in a performance-based contracting system. Health Serv Res 2003 Apr;38(2):535-552 [FREE Full text] [doi: 10.1111/1475-6773.00132] [Medline: 12785560]
56. Georgakopoulos SV, Tasoulis SK, Vrahatis AG, Plagianakos VP. Convolutional neural networks for toxic comment classification. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence. 2018 Presented at: SETN '18: 10th Hellenic Conference on Artificial Intelligence; Jul 9-12, 2018; Patras Greece. [doi: 10.1145/3200947.3208069]
57. Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017 Presented at: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 30 - Aug 4, 2017; Vancouver, Canada. [doi: 10.18653/v1/P17-1052]
58. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the conference of the North American chapter of the association for computational linguistics: Human language technologies. 2019 Presented at: Proceedings of the conference of the North American chapter of the association for computational linguistics: Human language technologies; Jun,2019; Minneapolis, Minnesota.
59. Gomez-Perez A, Corcho O, Fernández-López M. Ontological Engineering With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web. London: Springer; 2004.
60. Han EH, Karypis G. Centroid-based document classification: analysis and experimental results. In: Principles of Data Mining and Knowledge Discovery. Berlin, Heidelberg: Springer; 2000.
61. Manevitz LM, Yousef M. One-class svms for document classification. J Mach Learn Res 2002 Jan 3;2:139-154. [doi: 10.5555/944790.944808]
62. Layeghian Javan S, Sepehri MM, Aghajani H. Toward analyzing and synthesizing previous research in early prediction of cardiac arrest using machine learning based on a multi-layered integrative framework. J Biomed Inform 2018 Dec;88:70-89 [FREE Full text] [doi: 10.1016/j.jbi.2018.10.008] [Medline: 30389440]
63. Salles T, Gonçalves M, Rodrigues V, Rocha L. Improving random forests by neighborhood projection for effective text classification. Inf Syst 2018 Sep;77:1-21. [doi: 10.1016/j.is.2018.05.006]
64. Cawley G, Talbot N. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 2010;11:2079-2107 [FREE Full text]

65. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical Natural Language Processing for health outcomes research: overview and actionable suggestions for future advances. J Biomed Inform 2018 Dec;88:11-19 [FREE Full text] [doi: 10.1016/j.jbi.2018.10.005] [Medline: 30368002]
66. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015 Mar 4;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
67. Zar JH. Biostatistical Analysis, 5th Edition. London, UK: Pearson; 2010.
68. Reys AD, Silva D, Severo D, Pedro S, de Sousa e Sá MM, Salgado GA. Predicting multiple ICD-10 codes from Brazilian-Portuguese clinical notes. In: Cerri R, Prati RC, editors. Intelligent Systems. Cham: Springer; 2020.
69. Lee GH, Shin S. Federated learning on clinical benchmark data: performance assessment. J Med Internet Res 2020 Oct 26;22(10):e20891 [FREE Full text] [doi: 10.2196/20891] [Medline: 33104011]
70. Kate RJ. Clinical term normalization using learned edit patterns and subconcept matching: system development and evaluation. JMIR Med Inform 2021 Jan 14;9(1):e23104 [FREE Full text] [doi: 10.2196/23104] [Medline: 33443483]
71. Lee DH, Yetisgen M, Vanderwende L, Horvitz E. Predicting severe clinical events by learning about life-saving actions and outcomes using distant supervision. J Biomed Inform 2020 Jul;107:103425. [doi: 10.1016/j.jbi.2020.103425] [Medline: 32348850]
72. Sheikhpour R, Sarram MA, Gharaghani S, Chahooki MA. A survey on semi-supervised feature selection methods. Pattern Recognit 2017 Apr;64:141-158. [doi: 10.1016/j.patcog.2016.11.003]
73. Diao X, Huo Y, Yan Z, Wang H, Yuan J, Wang Y, et al. An application of machine learning to etiological diagnosis of secondary hypertension: retrospective study using electronic medical records. JMIR Med Inform 2021 Jan 25;9(1):e19739 [FREE Full text] [doi: 10.2196/19739] [Medline: 33492233]
74. Zhang Y, Zhou Y, Zhang D, Song W. A stroke risk detection: improving hybrid feature selection method. J Med Internet Res 2019 Apr 02;21(4):e12437 [FREE Full text] [doi: 10.2196/12437] [Medline: 30938684]
75. Guillaume LF, Christos K, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 2017 Jan;18(1):559-563. [doi: 10.5555/3122009.3122026]
76. Kogan E, Twyman K, Heap J, Milentijevic D, Lin JH, Alberts M. Assessing stroke severity using electronic health record data: a machine learning approach. BMC Med Inform Decis Mak 2020 Jan 08;20(1):8 [FREE Full text] [doi: 10.1186/s12911-019-1010-x] [Medline: 31914991]
77. Healthcare Improvement - Patient-Reported Outcomes. ICHOM. URL: https://www.ichom.org/ [accessed 2021-09-07]
78. Freeman D, Barret K, Nordan L, Spaulding A, Kaplan R, Karney M. Lessons from Mayo clinic's redesign of stroke care. Harvard Business Review. 2018. URL: https://hbr.org/2018/10/lessons-from-mayo-clinics-redesign-of-stroke-care [accessed 2021-09-07]
79. Feigin VL, Krishnamurthi R. Stroke is largely preventable across the globe: where to next? Lancet 2016 Aug 20;388(10046):733-734. [doi: 10.1016/S0140-6736(16)30679-1] [Medline: 27431357]
80. Zhou P, El-Gohary N. Ontology-based multilabel text classification of construction regulatory documents. J Comput Civ Eng 2015 Sep;30(4):04015058. [doi: 10.1061/(asce)cp.1943-5487.0000530]
81. Chawla NV. Data mining for imbalanced datasets: an overview. In: Maimon O, Rokach L, editors. Data Mining and Knowledge Discovery Handbook. US: Springer; 2010.
82. Weiskopf NG, Khan FJ, Woodcock D, Dorr DA, Cigarroa JE, Cohen AM. A mixed methods task analysis of the implementation and validation of EHR-based clinical quality measures. AMIA Annu Symp Proc 2017 Feb 10;2016:1229-1237 [FREE Full text] [Medline: 28269920]
83. 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] [Medline: 31066697]
84. Ling A, Kurian A, Caswell-Jin J, Sledge G, Shah N, Tamang S. Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. JAMIA Open 2019 Sep 18;2(4):528-537 [FREE Full text] [doi: 10.1093/jamiaopen/ooz040] [Medline: 32025650]
85. Wang SV, Rogers JR, Jin Y, Bates DW, Fischer MA. Use of electronic healthcare records to identify complex patients with atrial fibrillation for targeted intervention. J Am Med Inform Assoc 2017 Mar 01;24(2):339-344. [doi: 10.1093/jamia/ocw082] [Medline: 27375290]
86. Ali A, Shamsuddin S, Ralescu A. Classification with class imbalance problem: a review. Int J Advance Soft Compu Appl 2013 Nov;5(3):176-204 [FREE Full text]
87. Li D, Liu C, Hu SC. A learning method for the class imbalance problem with medical data sets. Comput Biol Med 2010 May;40(5):509-518. [doi: 10.1016/j.compbiomed.2010.03.005] [Medline: 20347072]
88. Geng W, Qin X, Yang T, Cong Z, Wang Z, Kong Q, et al. Model-based reasoning of clinical diagnosis in integrative medicine: real-world methodological study of electronic medical records and natural language processing methods. JMIR Med Inform 2020 Dec 21;8(12):e23082 [FREE Full text] [doi: 10.2196/23082] [Medline: 33346740]

89. Ridgway JP, Uvin A, Schmitt J, Oliwa T, Almirol E, Devlin S, et al. Natural language processing of clinical notes to identify mental illness and substance use among people living with HIV: retrospective cohort study. JMIR Med Inform 2021 Mar 10;9(3):e23456 [FREE Full text] [doi: 10.2196/23456] [Medline: 33688848]
90. Liao KP, Ananthakrishnan AN, Kumar V, Xia Z, Cagan A, Gainer VS, et al. Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. PLoS One 2015 Aug 24;10(8):e0136651 [FREE Full text] [doi: 10.1371/journal.pone.0136651] [Medline: 26301417]
91. Kruse CS, Mileski M, Alaytsev V, Carol E, Williams A. Adoption factors associated with electronic health record among long-term care facilities: a systematic review. BMJ Open 2015 Jan 28;5(1):e006615 [FREE Full text] [doi: 10.1136/bmjopen-2014-006615] [Medline: 25631311]
92. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
93. Bugnon B, Geissbuhler A, Bischoff T, Bonnabry P, von Plessen C. Improving primary care medication processes by using shared electronic medication plans in Switzerland: lessons learned from a participatory action research study. JMIR Form Res 2021 Jan 07;5(1):e22319 [FREE Full text] [doi: 10.2196/22319] [Medline: 33410753]
94. Nakatani H, Nakao M, Uchiyama H, Toyoshiba H, Ochiai C. Predicting inpatient falls using natural language processing of nursing records obtained from Japanese electronic medical records: case-control study. JMIR Med Inform 2020 Apr 22;8(4):e16970 [FREE Full text] [doi: 10.2196/16970] [Medline: 32319959]
95. Dafny L, Lee T. Health care needs real competition. Harvard Business Review (Competitive Strategy). 2016. URL: https://hbr.org/2016/12/health-care-needs-real-competition [accessed 2021-09-07]

Abbreviations

BERT: bidirectional encoder representation from transformers
BoW: Bag-of-Words
CNN: convolutional neural network
EMR: electronic medical record
IT: information technology
KNN: K-nearest neighbor
ML: machine learning
NIHSS: National Institutes of Health Stroke Scale
NLP: natural language processing
OWL: ontology web language
SVM: support vector machine
TFIDF: term frequency-inverted document frequency

Edited by T Hao; submitted 29.03.21; peer-reviewed by M Fernandes, X Diao; comments to author 30.05.21; revised version received 27.06.21; accepted 05.08.21; published 01.11.21.

Please cite as:
Zanotto BS, Beck da Silva Etges AP, dal Bosco A, Cortes EG, Ruschel R, De Souza AC, Andrade CMV, Viegas F, Canuto S, Luiz W, Ouriques Martins S, Vieira R, Polanczyk C, André Gonçalves M
Stroke Outcome Measurements From Electronic Medical Records: Cross-sectional Study on the Effectiveness of Neural and Nonneural Classifiers
JMIR Med Inform 2021;9(11):e29120
URL: https://medinform.jmir.org/2021/11/e29120
doi:10.2196/29120
PMID:34723829

Original Paper

A Semiautomated Chart Review for Assessing the Development of Radiation Pneumonitis Using Natural Language Processing: Diagnostic Accuracy and Feasibility Study

Jordan McKenzie[1], BSc; Rasika Rajapakshe[2,3], PhD; Hua Shen[4], MMath, PhD; Shan Rajapakshe[5], BSc, MSc; Angela Lin[3,6], MD

[1]Northern Medical Program, Faculty of Medicine, University of British Columbia, Prince George, BC, Canada

[2]Medical Physics, BC Cancer, Kelowna, BC, Canada

[3]Department of Surgery, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

[4]Department of Mathematics and Statistics, University of Calgary, Calgary, AB, Canada

[5]Island Medical Program, Faculty of Medicine, University of British Columbia, Victoria, BC, Canada

[6]Radiation Oncology, BC Cancer, Kelowna, BC, Canada

Corresponding Author:
Angela Lin, MD
Radiation Oncology
BC Cancer
399 Royal Avenue
Kelowna, BC, V1Y 5L3
Canada
Phone: 1 250 712 3979
Email: angela.lin@bccancer.bc.ca

Abstract

Background: Health research frequently requires manual chart reviews to identify patients in a study-specific cohort and examine their clinical outcomes. Manual chart review is a labor-intensive process that requires significant time investment for clinical researchers.

Objective: This study aims to evaluate the feasibility and accuracy of an assisted chart review program, using an in-house rule-based text-extraction program written in Python, to identify patients who developed radiation pneumonitis (RP) after receiving curative radiotherapy.

Methods: A retrospective manual chart review was completed for patients who received curative radiotherapy for stage 2-3 lung cancer from January 1, 2013 to December 31, 2015, at British Columbia Cancer, Kelowna Centre. In the manual chart review, RP diagnosis and grading were recorded using the Common Terminology Criteria for Adverse Events version 5.0. From the charts of 50 sample patients, a total of 1413 clinical documents were obtained for review from the electronic medical record system. The text-extraction program was built using the Natural Language Toolkit Python platform (and regular expressions, also known as RegEx). Python version 3.7.2 was used to run the text-extraction program. The output of the text-extraction program was a list of the full sentences containing the key terms, document IDs, and dates from which these sentences were extracted. The results from the manual review were used as the gold standard in this study, with which the results of the text-extraction program were compared.

Results: Fifty percent (25/50) of the sample patients developed grade ≥1 RP; the natural language processing program was able to ascertain 92% (23/25) of these patients (sensitivity 0.92, 95% CI 0.74-0.99; specificity 0.36, 95% CI 0.18-0.57). Furthermore, the text-extraction program was able to correctly identify all 9 patients with grade ≥2 RP, which are patients with clinically significant symptoms (sensitivity 1.0, 95% CI 0.66-1.0; specificity 0.27, 95% CI 0.14-0.43). The program was useful for distinguishing patients with RP from those without RP. The text-extraction program in this study avoided unnecessary manual review of 22% (11/50) of the sample patients, as these patients were identified as grade 0 RP and would not require further manual review in subsequent studies.

Conclusions: This feasibility study showed that the text-extraction program was able to assist with the identification of patients who developed RP after curative radiotherapy. The program streamlines the manual chart review further by identifying the key sentences of interest. This work has the potential to improve future clinical research, as the text-extraction program shows promise in performing chart review in a more time-efficient manner, compared with the traditional labor-intensive manual chart review.

(JMIR Med Inform 2021;9(11):e29241) doi:10.2196/29241

KEYWORDS

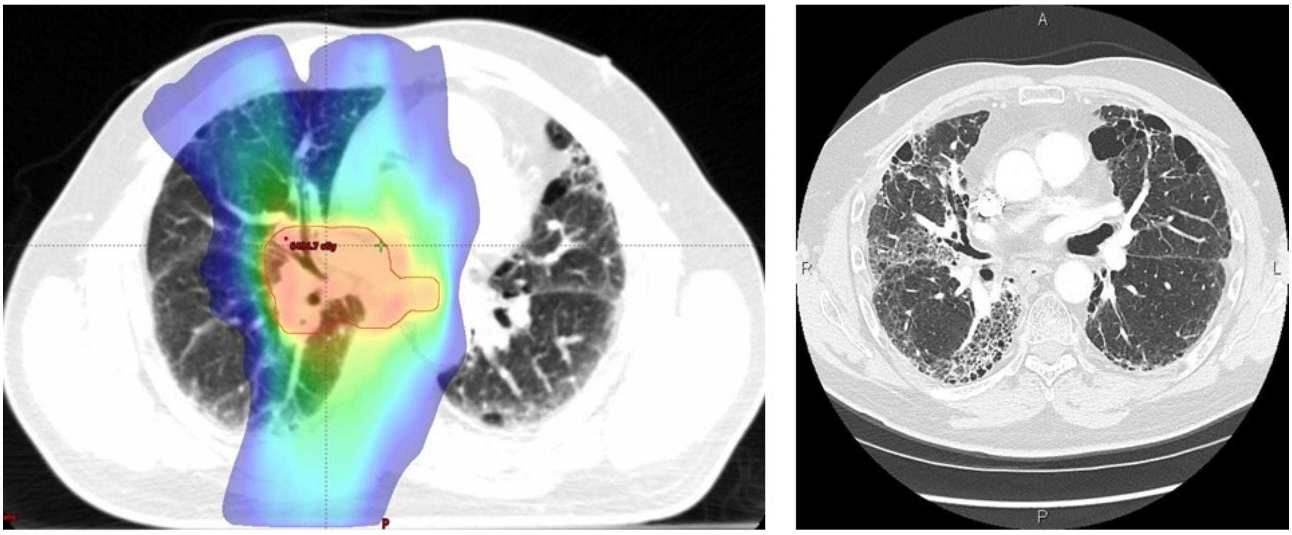
chart review; natural language processing; text extraction; radiation pneumonitis; lung cancer; radiation therapy; python; electronic medical record; accuracy

Introduction

Background

Retrospective chart reviews require the analysis of pre-existing clinical data to answer a research question. To identify the patient cohort of interest, researchers often need to use certain inclusion criteria to scan a large *database*. After the patient cohort is identified, data abstraction begins, and a number of patient variables can be collected [1-3]. For example, cancer research frequently uses chart reviews to examine the outcomes and specific side effects of therapies. Radiation pneumonitis (RP) is a potential side effect of radiation therapy (RT) in patients with lung cancer, which can lead to permanent lung damage visible on radiography (Figure 1) [4,5]. Patients with RP may develop supplemental oxygen dependence and have a lower quality of life; as such, it is an important outcome to consider after RT and important to understand factors that may increase or decrease the likelihood of its development [4]. Of the patients with lung cancer treated with RT, it is expected that approximately 10% to 20% will develop moderate to severe RP [6-9]. Although RP fatality is uncommon, it still occurs in 1.9% of those affected [10]. For selecting a cohort of patients who developed symptomatic RP, the charts of patients with stage 2-3 lung cancer who received curative RT during the study period must be reviewed. In a typical manual chart review, this would involve researchers going through patient charts and looking for evidence and severity of RP diagnosis based on the Common Terminology Criteria for Adverse Events (CTCAE) version 5.0 [9]. This process takes significant human resources and time to identify the patient cohort of interest [11,12]. The time requirement is amplified in cohorts that have a small representation in the larger data set, where a much larger data set is necessary to be reviewed to find a significant number of rare events [12]. This decreases the chart review productivity, where a high percentage of the chart review process will be unfruitful in identifying patients for the cohort and can be seen as a loss of valuable research time. Our goal is to use a computer program developed in-house to assist in the identification of the cohort of interest and move toward an automated chart review process.

Figure 1. Color wash of the dose distribution on a radiation therapy planning computed tomography for a patient with lung cancer (left). The blue edge represents the 20 gray dose line, which is the recognized dose associated with increased risk of radiation pneumonitis. The same patient’s 3-month follow-up computed tomography image showed opacity indicating a partial filling of the air spaces in the lungs. These radiologic changes are representative of radiation pneumonitis in the radiation field (right).



The most recent and sophisticated computer programs designed to assist in chart review studies have implemented natural language processing (NLP) [13-16]. NLP is a computer model that can manipulate a document's narrative text and speech, also known as natural language, and export it in a structured format for analysis [16]. This type of modeling is necessary because of the nature of electronic medical records (EMRs). Typically, patient charts in the EMR are written in a narrative text format, which is more difficult for a computer program to extract information from compared with a structured charting system that is arranged in tables [17]. It has been estimated that up to 80% of health care data are in an unstructured narrative format within most EMR systems [18]. Using an NLP computer algorithm as a tool could enable a chart review to be completed in less time with less human resources.

Objective

The objective of this study is to evaluate the feasibility and accuracy of an in-house developed rule-based text-extraction program written in Python to identify patients with lung cancer who developed RP after receiving curative RT. This rule-based text-extraction program written in Python is the first stage of developing a more robust NLP program. RP is an important factor to consider with respect to RT and serves as a marker for treatment-specific variables and allows us to evaluate the use of the text-extraction program. Specifically, the focus of identification in this study is on clinically significant cases of grade ≥2 RP. RP is graded by severity; if the patient's quality of life is affected by shortness of breath and cough, it is grade ≥2, whereas grade 1 RP is only seen on imaging and is not associated with any symptoms (Textbox 1) [9].

Textbox 1. Radiation pneumonitis (RP) grading based on Common Terminology Criteria for Adverse Events version 5.0.

Grade 0

- No RP present

Grade 1

- Asymptomatic; clinical or diagnostic observations only; intervention not indicated

Grade 2

- Symptomatic; medical intervention indicated; limiting instrumental activities of daily living

Grade 3

- Severe symptoms; limiting self-care activities of daily living; oxygen indicated

Grade 4

- Life-threatening respiratory compromise; urgent intervention indicated (eg, tracheotomy or intubation)

Grade 5

- Death

Methods

Recruitment

The study population included a sample subset of 50 patients, from those who received curative RT for stage 2-3 non–small cell lung cancer from January 1, 2013, to December 31, 2015, at British Columbia (BC) Cancer Kelowna. The sample subset was designed to represent the proportions of RP grades in the literature [6-8,10,19]. However, there is a lack of consensus on the proportions of RP grades among patients treated with RT, most likely because of the numerous variables identified in contributing to RP development, including age, RT dose, concurrent chemotherapy, and underlying comorbidities such as chronic obstructive pulmonary disease [6-8,10,19]. The sample subset represents the proportion of RP grades most likely to be encountered in a larger randomized data set. Once the proportions of RP grade were decided on for the cohort based on the literature, simple random sampling without replacement was done on the manually reviewed cohort.

Data Exclusion

Patients who underwent surgery after radiation treatment were excluded. Patients who received palliative radiation and patients with small cell lung cancer were excluded.

Workflow

A manual chart review was completed by reviewing patient charts from the institutional EMR at BC Cancer Kelowna. The manual chart review results served as the definitive diagnosis, with which the assisted chart review program was compared. In the manual chart review, RP diagnosis and grading were recorded using CTCAE version 5.0 (Textbox 1) [9].

The in-house text-extraction program was built using the Natural Language Toolkit Python platform (and regular expressions, also known as RegEx). Patient charts were extracted from the BC Cancer EMR system and were subsequently formatted into the American Standard Code for Information Interchange text files to be compatible with the text-extraction program. From the charts of 50 sample patients, a total of 1413 clinical documents (clinical notes and radiology reports) were obtained for review. The reports from the BC Cancer EMR system were obtained by either direct conversion to text format documents or were printed in PDF and then converted to text format using the open-source Python Tesseract optical character recognition program. This step of obtaining and converting the documents to text format from the BC Cancer EMR system was necessary, as the text-extraction program input requires text format documents. Python version 3.7.2 was used to run the assisted chart review text-extraction program. The terms *pneumonitis*, *radiation pneumonitis*, *radiation induced lung injury*, and *fibrosis* were used as key terms for the assisted chart review. These key terms were chosen by the radiation oncologist contributing clinical expertise in this study, and they represent terminology that a physician would use to identify RP in dictated reports. The output of the text-extraction program was a list of full sentences containing the key terms, along with the document IDs and dates from which these sentences were extracted. The text-extraction program was designed to search through all the charts and extract the whole sentence that contained the key terms. If a sentence was extracted from a patient's chart, the patient was identified as having RP. The text-extraction program organized the extracted information, identified the patients, and indicated the exact documents containing the key terms. The results from the text-extraction program were then compared with those from the manual chart review.

If the text-extraction program is shown to be feasible and accurate, a more expedited manual chart review can be performed using the results of the text-extraction program in future studies. Patients with no key terms identified in their charts will be designated as grade 0 RP, and no further chart review of these patients will need to be completed. For the patients identified by the text-extraction program to have RP, the sentences containing the key terms can be reviewed manually, first to confirm that these patients are correctly identified as having RP, and then to grade the RP severity in an expedited manner. Thus, there is an opportunity to improve the text-extraction program specificity during this sentence review process by correcting the false-positive cases.

Statistical Analysis

The comparison between the manual chart review and text-extraction program output was viewed and analyzed in 2 different ways: the first approach considered the diseased state to be grade ≥1 RP, and the second approach considered the diseased state to be grade ≥2 RP, with grade 1 RP classified as a healthy state as the patients with grade 1 RP had no clinical symptoms. The text-extraction program was designed to look for any grade of RP when searching through the patient charts, so this lends itself to being able to perform well during the first analysis. However, grade 1 RP is only visible radiographically and thus is not clinically relevant to a patient's further care. Thus, we wanted to look at how well the assisted chart review system was able to identify patients with symptomatic RP. Statistical analyses were performed using SAS software version 9.4.

Results

Text-Extraction Program Output

The results of the text-extraction program used to identify patients with RP of any grade are shown in Tables 1 and 2. The text-extraction program was able to ascertain 92% (23/25) of patients who developed grade ≥1 RP (sensitivity 0.92, 95% CI 0.74-0.99; specificity 0.36, 95% CI 0.18-0.57). The results of the text-extraction program used to identify patients with symptomatic RP, that is, grade ≥2, is shown in Table 3. The text-extraction program was able to correctly identify all 9 patients with grade ≥2 RP (sensitivity 1.0, 95% CI 0.66-1.0; specificity 0.27, 95% CI 0.14-0.43). Both analyses revealed that the text-extraction program was capable of significantly differentiating between the diseased and healthy groups.

Table 1. The assisted chart review text-extraction program results and the accuracy for each RP grade.

RP severity (grade)	Total, N	Correctly identified, n (%)
0	25	9 (36)
1	16	14 (88)
2	7	7 (100)
3	2	2 (100)

Table 2. The assisted chart review text-extraction program results for differentiating between patients with radiation pneumonitis (RP) of grade 0 (healthy) versus those with RP of grade ≥1 (diseased).

Text-extraction program findings	Manual chart review finding		
	Healthy (grade 0 RP), n (%)	Diseased (grade ≥1 RP), n (%)	Total, N
Healthy (grade 0 RP)	9 (18)	2 (4)	11
Diseased (grade ≥1 RP)	16 (32)	23 (46)	39
Total	25 (50)	25 (50)	50

Table 3. The assisted chart review text-extraction program results looking at the ability to distinguish between patients with radiation pneumonitis (RP) of grade ≤1 (healthy) and those with of grade ≥2 (diseased).

Text-extraction program findings	Manual chart review finding		
	Healthy (grade ≤1 RP), n (%)	Diseased (grade ≥2 RP), n (%)	Total, N
Healthy (grade ≤1 RP)	11 (22)	0 (0)	11
Diseased (grade ≥2 RP)	30 (60)	9 (18)	39
Total	41 (82)	9 (18)	50

The text-extraction program missed 2 patients with grade 1 RP. Upon further review, the 2 patients with grade 1 RP that the text-extraction program *missed* were found to truly have grade 0 RP but were incorrectly labeled as patients with RP because of human error in the manual chart review. If we correct for this human error, the sensitivity improves to 1.0 for the text-extraction program's ability to identify grade ≥1 RP.

Clinical Utility

In our cohort, each patient's chart consisted of an average of 28 clinical documents that make up their chart, with a range of 15 to 150 documents. The average time spent during the manual chart review of one patient's chart was 30 minutes. Therefore, the manual chart review of the 50-patient cohort took 25 hours. In comparison, the assisted chart review text-extraction program processed the 1413 clinical documents and exported the results in <5 minutes.

The use of the text-extraction program in this study would be to avoid unnecessary manual review of 22% (11/50) of the sample, including their electronic documents (198/1413, 14%), as these patients were identified as not having RP and thus would not require any manual review. It will also streamline the rest of the manual review as key sentences with the key terms are identified, thus further reducing the number of clinical documents necessary for the manual review to confirm that the patient should be included in the cohort.

Discussion

Principal Findings

The text-extraction program was able to identify patients with RP with high sensitivity but, unfortunately, low specificity. This can assist in the identification of a patient cohort of interest in a more efficient manner.

The text-extraction program correctly identified 2 patients with grade 0 RP that the manual chart review incorrectly identified. Similar findings have been reported in the literature, where one study found that their automated chart review outperformed their manual chart review as the human reviewer missed the correct classification on manual evaluation of the chart [11]. Therefore, although the gold standard for assessing the accuracy of the text-extraction program in this study is manual chart review, the process is very tedious and not guaranteed to be perfect because of human error [11,20]. This highlights a potential advantage of the text-extraction program at being more accurate than the human-led manual chart review.

The utility of the text-extraction program in this study would be to perform a rapid scan of a larger data set of documents and avoid unnecessary manual review of many of the non-RP patient charts. The program is able to use key terms, such as RP or fibrosis, to return a list of patients with those terms in the patient charts. This will significantly cut down on the number of charts that the manual review will include. This is mainly because of the fact that even if a patient does develop RP, most of their charts do not include any indication of their diagnosis. The computer program organizes the extracted information into which patient and which exact chart, thus further reducing the amount of chart review that is necessary to manually review to confirm that the patient should be included in our cohort.

The end goal of using text-extraction programs to perform chart reviews is to save the researcher time and effort of combing through patient charts to form a cohort in which to begin studying a clinical outcome. Our text-extraction program was able to output its results in <5 minutes compared with the 25 hours it took the manual chart review control to create the RP cohort.

Limitations

A limitation to implementing this assisted chart review program is its current high false-positive rate, leading to unnecessary chart review of patients with no RP. The development of automated chart reviews must consider the balance between NLP program accuracy (no diseased cases missed) versus the amount of time saved by confidently eliminating true RP grade 0 patients in the review. Designing the key terms was an important process to balance the accuracy of the text-extraction program versus the time saved using the text-extraction program. Selecting broad key terms is important to capture all patients who may fall into our cohort; however, more specific key terms would better rule out patients not within the study cohort. Our goal was to maximize the sensitivity of the text-extraction program by including broad terms so as to not miss any patients with the diseased state initially, as the sentence output of the text-extraction program allows for a truncated chart review to improve the specificity. This means that the possible time saved in this feasibility study was less as more false-positive RP patients were identified. Future work is underway to improve the specificity of the text-extraction program with a larger sample.

Another limitation of our work is the small sample size of 50 patients. This sample group was used as a proof of concept for our in-house developed text-extraction program. This study's results will guide further refinement of the text-extraction program and validation with a larger sample of patients.

The rule-based text-extraction program used in this study still requires human involvement in a number of steps. The clinical documents in the BC Cancer EMR system had to be obtained manually rather than automatically, which continues to pose a barrier in making chart review research as time efficient as possible.

In addition, it is important to point out that expert opinions were necessary to identify the key terms to be used in the text-extraction program. This is not only another human involvement requirement but also indicates that the results are dependent on the quality of the expert. In addition, this makes the program less generalizable to other cohorts without a new expert to create the proper key terms for each specific cohort.

Comparison With Previous Work

Other studies have used NLP programs to assist with chart reviews in many scopes of medicine, including respirology, cardiology, and neurosurgery, and now our cancer research to identify patients who developed RP [21-24]. NLP has different applications in medical research, such as identifying patient cohorts such as our study and similar studies that identified cohorts of progressive heart failure and patients with asthma [21,22]. Other studies have used NLP programs to extract specific clinical features from clinical charts, such as tuberculosis patient factors and radiology characteristics of glioblastoma [23,24]. Our use of an NLP program to extract information based on key terms to reduce the amount of chart review necessary is similar to the study by Cao et al [25], where they used search terms to identify medical errors through patient charts. This allowed their group to reduce the number of charts that needed to be reviewed, from 286,000 discharge summaries to 2744 discharge summaries that were found to contain the search terms [25]. This meant that the Cao et al [25] manual review only had to be done on <1% of the initial data set. Reducing the number of charts to review saves many hours of manual chart review and would greatly increase the speed at which the review could be completed. Thus, an assisted chart review program opens the possibility of expanding the study, including a much larger data set that would be impractical to review manually. Our study adds to the existing literature on this topic by supporting the validity of NLP programs; it demonstrates the ability to further analyze an identified patient cohort based on variables of interest, such as illness severity.

Conclusions

In conclusion, the NLP-based text-extraction program used in this study is a feasible and valuable method for identifying patients who developed RP after curative radiotherapy. First, the text-extraction program helped save chart review time by completely eliminating patient charts identified with grade 0 RP. Second, the text-extraction program extracted key sentences from patient charts and allowed for an efficient review of relevant phrases, should this be needed to grade patients' RP severity without having to peruse the rest of their charts. For example, in a quick scan, a researcher would be able to read only the sentences with the identified keyword in a patient's chart instead of sifting through many full documents.

The analysis revealed that the text-extraction program was capable of significantly differentiating between diseased and healthy groups. Compared with the manual chart review of the 50-patient cohort that took 25 hours, the text-extraction program was able to process all the charts in <5 minutes and exported the list of patients that had RP mentioned somewhere in their chart.

This work has the potential to improve future clinical research as the text-extraction program shows promise in performing chart review in a more time- and effort-efficient manner compared with the traditional manual chart review. The text-extraction program is available by contacting the authors (RR).

Acknowledgments

This study received funding support from the British Columbia Cancer Foundation and the University of British Columbia Faculty of Medicine. The authors thank Jinying Wu for statistical analysis support.

Conflicts of Interest

None declared.

References

1. Worster A, Haines T. Advanced statistics: understanding medical record review (MRR) studies. Acad Emerg Med 2004 Feb;11(2):187-192 [FREE Full text] [Medline: 14759964]
2. Panacek EA. Performing chart review studies. Air Med J 2007;26(5):206-210. [doi: 10.1016/j.amj.2007.06.007] [Medline: 17765825]
3. Gearing RE, Mian IA, Barber J, Ickowicz A. A methodology for conducting retrospective chart review research in child and adolescent psychiatry. J Can Acad Child Adolesc Psychiatry 2006 Aug;15(3):126-134 [FREE Full text] [Medline: 18392182]
4. Weytjens R, Erven K, De Ruysscher RD. Radiation pneumonitis: occurrence, prediction, prevention and treatment. Belg J Med Oncol 2013 Sep;7(4):105-110 [FREE Full text]
5. Kainthola A, Haritwal T, Tiwari M, Gupta N, Parvez S, Tiwari M, et al. Immunological aspect of radiation-induced pneumonitis, current treatment strategies, and future prospects. Front Immunol 2017 May 2;8:506 [FREE Full text] [doi: 10.3389/fimmu.2017.00506] [Medline: 28512460]

6. Liang B, Yan H, Tian Y, Chen X, Yan L, Zhang T, et al. Dosiomics: extracting 3d spatial features from dose distribution to predict incidence of radiation pneumonitis. Front Oncol 2019 Apr 12;9:269 [FREE Full text] [doi: 10.3389/fonc.2019.00269] [Medline: 31032229]
7. Moreno M, Aristu J, Ramos LI, Arbea L, López-Picazo JM, Cambeiro M, et al. Predictive factors for radiation-induced pulmonary toxicity after three-dimensional conformal chemoradiation in locally advanced non-small-cell lung cancer. Clin Transl Oncol 2007 Sep;9(9):596-602. [doi: 10.1007/s12094-007-0109-1] [Medline: 17921108]
8. Anthony GJ, Cunliffe A, Castillo R, Pham N, Guerrero T, Armato SG, et al. Incorporation of pre-therapy F-FDG uptake data with CT texture features into a radiomics model for radiation pneumonitis diagnosis. Med Phys 2017 Jul;44(7):3686-3694 [FREE Full text] [doi: 10.1002/mp.12282] [Medline: 28422299]
9. Common Terminology Criteria for Adverse Events (CTCAE) Version 5.0. NIH National Cancer Institute. URL: https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm#ctc_50
10. Palma DA, Senan S, Tsujino K, Barriger RB, Rengan R, Moreno M, et al. Predicting radiation pneumonitis after chemoradiation therapy for lung cancer: an international individual patient data meta-analysis. Int J Radiat Oncol Biol Phys 2013 Feb 01;85(2):444-450 [FREE Full text] [doi: 10.1016/j.ijrobp.2012.04.043] [Medline: 22682812]
11. Duz M, Marshall JF, Parkin T. Validation of an improved computer-assisted technique for mining free-text electronic medical records. JMIR Med Inform 2017 Jun 29;5(2):e17 [FREE Full text] [doi: 10.2196/medinform.7123] [Medline: 28663163]
12. Dipaola F, Gatti M, Pacetti V, Bottaccioli AG, Shiffer D, Minonzio M, et al. Artificial intelligence algorithms and natural language processing for the recognition of syncope patients on emergency department medical records. J Clin Med 2019 Oct 14;8(10):1677 [FREE Full text] [doi: 10.3390/jcm8101677] [Medline: 31614982]
13. Hardjojo A, Gunachandran A, Pang L, Abdullah MR, Wah W, Chong JW, et al. Validation of a natural language processing algorithm for detecting infectious disease symptoms in primary care electronic medical records in singapore. JMIR Med Inform 2018 Jun 11;6(2):e36 [FREE Full text] [doi: 10.2196/medinform.8204] [Medline: 29907560]
14. Zhou L, Suominen H, Gedeon T. Adapting state-of-the-art deep language models to clinical information extraction systems: potentials, challenges, and solutions. JMIR Med Inform 2019 Apr 25;7(2):e11499 [FREE Full text] [doi: 10.2196/11499] [Medline: 31021325]
15. Zheng S, Jabbour SK, O'Reilly SE, Lu JJ, Dong L, Ding L, et al. Automated information extraction on treatment and prognosis for non-small cell lung cancer radiotherapy patients: clinical study. JMIR Med Inform 2018 Feb 01;6(1):e8 [FREE Full text] [doi: 10.2196/medinform.8662] [Medline: 29391345]
16. Usui M, Aramaki E, Iwao T, Wakamiya S, Sakamoto T, Mochizuki M. Extraction and standardization of patient complaints from electronic medication histories for pharmacovigilance: natural language processing analysis in japanese. JMIR Med Inform 2018 Sep 27;6(3):e11021 [FREE Full text] [doi: 10.2196/11021] [Medline: 30262450]
17. Tignanelli CJ, Silverman GM, Lindemann EA, Trembley AL, Gipson JC, Beilman G, et al. Natural language processing of prehospital emergency medical services trauma records allows for automated characterization of treatment appropriateness. J Trauma Acute Care Surg 2020 May;88(5):607-614. [doi: 10.1097/TA.0000000000002598] [Medline: 31977990]
18. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol 2020 Feb;145(2):463-469 [FREE Full text] [doi: 10.1016/j.jaci.2019.12.897] [Medline: 31883846]
19. Inoue A, Kunitoh H, Sekine I, Sumi M, Tokuuye K, Saijo N. Radiation pneumonitis in lung cancer patients: a retrospective study of risk factors and the long-term prognosis. Int J Radiat Oncol Biol Phys 2001 Mar 01;49(3):649-655. [doi: 10.1016/s0360-3016(00)00783-5] [Medline: 11172945]
20. Chan L, Beers K, Yau AA, Chauhan K, Duffy A, Chaudhary K, et al. Natural language processing of electronic health records is superior to billing codes to identify symptom burden in hemodialysis patients. Kidney Int 2020 Feb;97(2):383-392 [FREE Full text] [doi: 10.1016/j.kint.2019.10.023] [Medline: 31883805]
21. Kaur H, Sohn S, Wi C, Ryu E, Park MA, Bachman K, et al. Automated chart review utilizing natural language processing algorithm for asthma predictive index. BMC Pulm Med 2018 Feb 13;18(1):34 [FREE Full text] [doi: 10.1186/s12890-018-0593-9] [Medline: 29439692]
22. Lindvall C, Forsyth A, Barzilay R, Tulsky J. Natural language processing: an opportunity to make chart data come alive in palliative care research (FR481A). J Pain Symptom Manag 2017 Feb 1;53(2):385 [FREE Full text] [doi: 10.1016/j.jpainsymman.2016.12.164]
23. Petch J, Batt J, Murray J, Mamdani M. Extracting clinical features from dictated ambulatory consult notes using a commercially available natural language processing tool: pilot, retrospective, cross-sectional validation study. JMIR Med Inform 2019 Nov 01;7(4):e12575 [FREE Full text] [doi: 10.2196/12575] [Medline: 31682579]
24. Senders JT, Cho LD, Calvachi P, McNulty JJ, Ashby JL, Schulte IS, et al. Automating clinical chart review: an open-source natural language processing pipeline developed on free-text radiology reports from patients with glioblastoma. JCO Clin Cancer Inform 2020 Jan;4:25-34 [FREE Full text] [doi: 10.1200/CCI.19.00060] [Medline: 31977252]
25. Cao H, Stetson P, Hripcsak G. Assessing explicit error reporting in the narrative electronic medical record using keyword searching. J Biomed Inform 2003;36(1-2):99-105 [FREE Full text] [doi: 10.1016/s1532-0464(03)00058-3] [Medline: 14552851]

Abbreviations

BC: British Columbia
CTCAE: Common Terminology Criteria for Adverse Events
EMR: electronic medical record
NLP: natural language processing
RP: radiation pneumonitis
RT: radiation therapy

Edited by T Hao, Z Huang, B Tang; submitted 30.03.21; peer-reviewed by M Gatti, H Park; comments to author 19.05.21; revised version received 18.06.21; accepted 05.08.21; published 12.11.21.

Please cite as:
McKenzie J, Rajapakshe R, Shen H, Rajapakshe S, Lin A
A Semiautomated Chart Review for Assessing the Development of Radiation Pneumonitis Using Natural Language Processing: Diagnostic Accuracy and Feasibility Study
JMIR Med Inform 2021;9(11):e29241
URL: https://medinform.jmir.org/2021/11/e29241
doi:10.2196/29241
PMID:34766919

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org