# Contents

## Review

**Machine Learning Applications in Mental Health and Substance Use Research Among the LGBTQ2S+ Population: Scoping Review** *(e28962)*

Anasua Kundu, Michael Chaiton, Rebecca Billington, Daniel Grace, Rui Fu, Carmen Logie, Bruce Baskerville, Christina Yager, Nicholas Mitsakakis, Robert Schwartz .......................................................... 3

## Original Papers

**The Role of Physicians in Digitalizing Health Care Provision: Web-Based Survey Study** *(e31527)*

Anja Burmann, Max Tischler, Mira Faßbach, Sophie Schneitler, Sven Meister .......................................................... 16

**Risk Factors Associated With Nonfatal Opioid Overdose Leading to Intensive Care Unit Admission: A Cross-sectional Study** *(e32851)*

Avijit Mitra, Hiba Ahsan, Wenjun Li, Weisong Liu, Robert Kerns, Jack Tsai, William Becker, David Smelson, Hong Yu .......................................................... 33

**A Neural Network Approach for Understanding Patient Experiences of Chronic Obstructive Pulmonary Disease (COPD): Retrospective, Cross-sectional Study of Social Media Content** *(e26272)*

Tobe Freeman, Raul Rodriguez-Esteban, Juergen Gottowik, Xing Yang, Veit Erpenbeck, Mathias Leddin .......................................................... 44

**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** *(e28763)*

Kei Teramoto, Toshihiro Takeda, Naoki Mihara, Yoshie Shimai, Shirou Manabe, Shigeki Kuwata, Hiroshi Kondoh, Yasushi Matsumura .......................................................... 51

**Toward Personalized Web-Based Cognitive Rehabilitation for Patients With Ischemic Stroke: Elo Rating Approach** *(e28090)*

Alejandro García-Rudolph, Eloy Opisso, Jose Tormos, Vince Madai, Dietmar Frey, Helard Becerra, John Kelleher, Montserrat Bernabeu Guitart, Jaume López .......................................................... 61

**Health Information Needs of Young Chinese People Based on an Online Health Community: Topic and Statistical Analysis** *(e30356)*

Jie Wang, Xin Wang, Lei Wang, Yan Peng .......................................................... 77

**Visualizing Knowledge Evolution Trends and Research Hotspots of Personal Health Data Research: Bibliometric Analysis** *(e31142)*

Jianxia Gong, Vikrant Sihag, Qingxia Kong, Lindu Zhao .......................................................... 89
Event Prediction Model Considering Time and Input Error Using Electronic Medical Records in the Intensive Care Unit: Retrospective Study (e26426)
MinDong Sung, Sangchul Hahn, Chang Han, Jung Lee, Jayoung Lee, Jinkyu Yoo, Jay Heo, Young Kim, Kyung Chung. .......................... 103

Local Differential Privacy in the Medical Domain to Protect Sensitive Information: Algorithm Development and Real-World Validation (e26914)
MinDong Sung, Dongchul Cha, Yu Park. .................................................. 115

Accurate Prediction of Stroke for Hypertensive Patients Based on Medical Big Data and Machine Learning Algorithms: Retrospective Study (e30277)
Yujie Yang, Jing Zheng, Zhenzhen Du, Ye Li, Yunpeng Cai. .......................... 123

A Blockchain-Based Dynamic Consent Architecture to Support Clinical Genomic Data Sharing (ConsentChain): Proof-of-Concept Study (e27816)
Faisal Albalwy, Andrew Brass, Angela Davies. ........................................ 138

Jeongmin Kim, Hakyung Lim, Jae-Hyeon Ahn, Kyoung Lee, Kwang Lee, Kyo Koo. .................................................. 162

Stroke Outcome Measurements From Electronic Medical Records: Cross-sectional Study on the Effectiveness of Neural and Nonneural Classifiers (e29120)

Corrigenda and Addenda

Correction: Use of Deep Learning to Predict Acute Kidney Injury After Intravenous Contrast Media Administration: Prediction Model Development Study (e34411)
Donghwan Yun, Semin Cho, Yong Kim, Dong Kim, Kook-Hwan Oh, Kwon Joo, Yon Kim, Seung Han. ................................. 136
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).
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].

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Number of studies, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Countries</strong></td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td>7 (63)</td>
</tr>
<tr>
<td>China</td>
<td>2 (18)</td>
</tr>
<tr>
<td>Sweden</td>
<td>1 (9)</td>
</tr>
<tr>
<td>Australia</td>
<td>1 (9)</td>
</tr>
<tr>
<td><strong>Years published</strong></td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>5 (45)</td>
</tr>
<tr>
<td>2020</td>
<td>6 (54)</td>
</tr>
<tr>
<td><strong>Field of study</strong></td>
<td></td>
</tr>
<tr>
<td>Mental health (n=9)</td>
<td></td>
</tr>
<tr>
<td>Suicide or self-injury</td>
<td>2 (18)</td>
</tr>
<tr>
<td>Depression</td>
<td>2 (18)</td>
</tr>
<tr>
<td>Mood or affect processes</td>
<td>3 (27)</td>
</tr>
<tr>
<td>Minority stress</td>
<td>1 (9)</td>
</tr>
<tr>
<td>Gender incongruence</td>
<td>1 (9)</td>
</tr>
<tr>
<td>Substance use (n=2)</td>
<td></td>
</tr>
<tr>
<td>Tobacco</td>
<td>1 (9)</td>
</tr>
<tr>
<td>Poppers or alkyl nitrites</td>
<td>1 (9)</td>
</tr>
<tr>
<td><strong>Target population</strong></td>
<td></td>
</tr>
<tr>
<td>Sexual minorities: male (gay, MSM, bisexual)</td>
<td>5 (45)</td>
</tr>
<tr>
<td>Sexual minorities: female (lesbian, bisexual)</td>
<td>2 (18)</td>
</tr>
<tr>
<td>Transgender or gender minorities</td>
<td>3 (27)</td>
</tr>
<tr>
<td>LGBT/LGBTQ+</td>
<td>3 (27)</td>
</tr>
<tr>
<td><strong>Domains of ML</strong></td>
<td></td>
</tr>
<tr>
<td>Web content analysis</td>
<td>6 (55)</td>
</tr>
<tr>
<td>Prediction modeling</td>
<td>4 (36)</td>
</tr>
<tr>
<td>Imaging study</td>
<td>1 (9)</td>
</tr>
<tr>
<td><strong>Type of ML</strong></td>
<td></td>
</tr>
<tr>
<td>Supervised</td>
<td>9 (82)</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>3 (27)</td>
</tr>
<tr>
<td>Deep</td>
<td>1 (9)</td>
</tr>
<tr>
<td><strong>ML algorithms</strong></td>
<td></td>
</tr>
<tr>
<td>LDA</td>
<td>3 (27)</td>
</tr>
<tr>
<td>RF</td>
<td>2 (18)</td>
</tr>
<tr>
<td>SVM</td>
<td>2 (18)</td>
</tr>
<tr>
<td>CNN</td>
<td>1 (9)</td>
</tr>
<tr>
<td>MLP</td>
<td>1 (9)</td>
</tr>
<tr>
<td>NB</td>
<td>1 (9)</td>
</tr>
<tr>
<td>Penalized regression (LASSO, elastic net regularized regression, ridge regression)</td>
<td>2 (18)</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>1 (9)</td>
</tr>
<tr>
<td>Characteristics</td>
<td>Number of studies, n (%)</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td><strong>Boosting</strong></td>
<td>3 (27)</td>
</tr>
<tr>
<td>(XGBoost(^m), AdaBoost(^n), GBM(^o))</td>
<td></td>
</tr>
<tr>
<td>Classification tree</td>
<td>2 (18)</td>
</tr>
<tr>
<td><strong>Feature selection</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>7 (64)</td>
</tr>
<tr>
<td>No</td>
<td>4 (36)</td>
</tr>
<tr>
<td><strong>Discussed model performance</strong></td>
<td></td>
</tr>
<tr>
<td>Used performance metrics</td>
<td>9 (82)</td>
</tr>
<tr>
<td>Didn’t use performance metrics</td>
<td>1 (9)</td>
</tr>
<tr>
<td>Didn’t discuss performance</td>
<td>1 (9)</td>
</tr>
<tr>
<td><strong>Method of validation</strong></td>
<td></td>
</tr>
<tr>
<td>Hold-out</td>
<td>2 (18)</td>
</tr>
<tr>
<td>Cross-validation</td>
<td>7 (64)</td>
</tr>
<tr>
<td>External validation</td>
<td>2 (18)</td>
</tr>
<tr>
<td>Unspecified</td>
<td>4 (36)</td>
</tr>
</tbody>
</table>

\(^a\)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 nitrates 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 Haimsam 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


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.

**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:


URL: https://medinform.jmir.org/2021/11/e28962
doi: 10.2196/28962
PMID: 35652059

https://medinform.jmir.org/2021/11/e28962 JMIR Med Inform 2021 | vol. 9 | iss. 11 | e28962 | p.14  
(page number not for citation purposes)
The Role of Physicians in Digitalizing Health Care Provision: Web-Based Survey Study

Anja Burmann1,2, MSc; Max Tischler3,4; Mira Faßbach4,5; Sophie Schneitler4,6,7, Dr med; Sven Meister1,2, Prof Dr

1Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany
2Witten/Herdecke University, Witten, Germany
3Hautärzte am Markt, Dortmund, Germany
4Bündnis Junge Ärzte, Berlin, Germany
5Helios Klinikum Duisburg, Duisburg, Germany
6Saarland University Hospital, Homburg, Germany
7German 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, noncentrized, 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, procedural, 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 incomplete 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).

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

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 $\alpha$=.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.

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, 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.

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 responses 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.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Completely unknown</th>
<th>Basically known</th>
<th>Completely understood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital health apps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digital health care act</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eHealth act</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telemedicine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telematics infrastructure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electronic patient record</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electronic health record</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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).

<table>
<thead>
<tr>
<th>Area</th>
<th>Worsening, n (%)</th>
<th>No change, n (%)</th>
<th>Improving, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early detection of diseases</td>
<td>51 (4)</td>
<td>549 (43.69)</td>
<td>674 (52.91)</td>
</tr>
<tr>
<td>Medical quality</td>
<td>156 (12.24)</td>
<td>422 (33.12)</td>
<td>696 (54.63)</td>
</tr>
<tr>
<td>Access to knowledge</td>
<td>10 (0.78)</td>
<td>128 (10.05)</td>
<td>1136 (89.17)</td>
</tr>
<tr>
<td>Treatment of rare diseases</td>
<td>8 (0.63)</td>
<td>250 (19.62)</td>
<td>1016 (79.75)</td>
</tr>
<tr>
<td>Administration</td>
<td>265 (20.80)</td>
<td>207 (16.25)</td>
<td>802 (62.95)</td>
</tr>
<tr>
<td>Patient adherence</td>
<td>91 (7.14)</td>
<td>691 (54.24)</td>
<td>492 (38.62)</td>
</tr>
<tr>
<td>Physician-patient relationship</td>
<td>397 (31.16)</td>
<td>672 (52.75)</td>
<td>205 (16.09)</td>
</tr>
<tr>
<td>Interdisciplinary collaboration</td>
<td>78 (6.12)</td>
<td>334 (26.22)</td>
<td>862 (67.66)</td>
</tr>
<tr>
<td>Attractiveness of the profession</td>
<td>237 (18.60)</td>
<td>644 (50.55)</td>
<td>393 (30.85)</td>
</tr>
<tr>
<td>Medical research</td>
<td>20 (1.57)</td>
<td>231 (18.13)</td>
<td>1023 (80.30)</td>
</tr>
</tbody>
</table>

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% (376/1274) 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 $\eta$ 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 $\eta$ 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.

<table>
<thead>
<tr>
<th>Category</th>
<th>Noticeable benefits</th>
<th>Untapped potential</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Association with age (η)</td>
<td>Group 1, n (%) (N=672)</td>
</tr>
<tr>
<td>Data quality/readability</td>
<td>0.08</td>
<td>486 (72.32)</td>
</tr>
<tr>
<td>Data availability</td>
<td>0.05</td>
<td>449 (66.82)</td>
</tr>
<tr>
<td>Data generation</td>
<td>0.08</td>
<td>372 (55.36)</td>
</tr>
<tr>
<td>Transparency</td>
<td>0.00</td>
<td>160 (23.81)</td>
</tr>
<tr>
<td>Patient participation</td>
<td>0.03</td>
<td>115 (17.11)</td>
</tr>
<tr>
<td>Work structuring</td>
<td>0.11</td>
<td>201 (29.91)</td>
</tr>
<tr>
<td>Reconciliation of family and working life</td>
<td>0.09</td>
<td>168 (25)</td>
</tr>
<tr>
<td>Data quality/readability</td>
<td>0.26 *</td>
<td>601 (89.43)</td>
</tr>
<tr>
<td>Data availability</td>
<td>0.27 *</td>
<td>638 (94.94)</td>
</tr>
<tr>
<td>Data generation</td>
<td>0.16</td>
<td>614 (91.37)</td>
</tr>
<tr>
<td>Transparency</td>
<td>0.23 *</td>
<td>533 (79.32)</td>
</tr>
<tr>
<td>Patient participation</td>
<td>0.25 *</td>
<td>518 (77.08)</td>
</tr>
<tr>
<td>Work structuring</td>
<td>0.25 *</td>
<td>534 (79.46)</td>
</tr>
<tr>
<td>Reconciliation of family and working life</td>
<td>0.27 *</td>
<td>478 (71.13)</td>
</tr>
</tbody>
</table>

*aItalics 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.

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

^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 ($\eta = 0.26$) 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 ($\eta = 0.29$ and $\eta = 0.23$, respectively; Table 5.)

<table>
<thead>
<tr>
<th>Professional usage fields of mobile apps</th>
<th>Association with age ($\eta$)</th>
<th>Group 1, n (%) (N=672)</th>
<th>Group 2, n (%) (N=602)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>0.03</td>
<td>148 (41.49)</td>
<td>152 (56.93)</td>
</tr>
<tr>
<td>Training</td>
<td>0.13</td>
<td>195 (42.76)</td>
<td>122 (45.69)</td>
</tr>
<tr>
<td>Information on pharmaceuticals</td>
<td>$0.29^a$</td>
<td>354 (77.63)</td>
<td>162 (60.67)</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>$0.23^a$</td>
<td>264 (57.89)</td>
<td>122 (45.69)</td>
</tr>
</tbody>
</table>

$^a$ 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 $\eta$ coefficient ($\eta=0.35$) showed a weak association with the age.

The recommendation rate of mobile health apps to patients did not relate noticeably with respondent age ($\eta=0.02$). 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 $\eta=0.21$ 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 ($\eta=0.21$).

As an important source of information, the internet ($\eta=0.14$), public bodies ($\eta=0.09$), medical societies ($\eta=0.15$), and developers ($\eta=0.04$) 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%; $\eta=0.26$).

**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” ($\eta=0.30$), “attractiveness of the profession” ($\eta=0.28$), “administration” ($\eta=0.27$), “patient adherence” ($\eta=0.27$), “physician-patient relationship” ($\eta=0.25$), “early detection of diseases” ($\eta=0.21$), and “interdisciplinary collaboration” ($\eta=0.21$).
The subsequent rating of the personal attitude towards upcoming changes through digitalization also showed a weak association with the age ($\eta=0.36$), 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, 38.69%; group 2: 116/602, 19.27%) showed an association, albeit a weak one, with age ($\eta=0.27$ and $\eta=0.21$, 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 ($\eta=0.02$), 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 BJA 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 ]


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]


Abbreviations

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

©Anja Burmann, Max Tischler, Mira Faßbach, Sophie Schneitler, Sven Meister. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 11.11.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Risk Factors Associated With Nonfatal Opioid Overdose Leading to Intensive Care Unit Admission: A Cross-sectional Study

Avijit Mitra¹, MSc; Hiba Ahsan¹, BTech; Wenjun Li²,³, PhD; Weisong Liu⁴, PhD; Robert D Kerns⁵,⁶,⁷, PhD; Jack Tsa⁹,¹⁰, PhD; William Becker⁸,¹¹, MD; David A Smelson³,¹², PsyD; Hong Yu¹,³,⁴,¹³, PhD

¹College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA, United States
²Department of Public Health, University of Massachusetts Lowell, Lowell, MA, United States
³Center for Healthcare Organization and Implementation Research, Veterans Affairs Bedford Healthcare System, Bedford, MA, United States
⁴Department of Computer Science, University of Massachusetts Lowell, Lowell, MA, United States
⁵Department of Psychiatry, Yale University School of Medicine, New Haven, CT, United States
⁶Department of Neurology, Yale University School of Medicine, New Haven, CT, United States
⁷Department of Psychology, Yale University School of Medicine, New Haven, CT, United States
⁸Pain Research, Informatics, Multimorbidities and Education Center, Veterans Affairs Connecticut Healthcare System, West Haven, CT, United States
⁹School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, United States
¹⁰National Center on Homelessness Among Veterans, United States Department of Veterans Affairs, Tampa, FL, United States
¹¹Department of Internal Medicine, Yale University School of Medicine, New Haven, CT, United States
¹²Department of Psychiatry, University of Massachusetts Chan Medical School, Worcester, MA, United States
¹³Department 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 2.09, 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.
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.

Figure 1. Data selection process.

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.

<table>
<thead>
<tr>
<th>SBDH Variable</th>
<th>Description and example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housing insecurity</td>
<td>Lack housing or stable shelter. Example: <em>homeless</em>, living with friends.</td>
</tr>
<tr>
<td>Yes</td>
<td>Has access to housing. Example: <em>lives in [<strong>location</strong>]</em> by himself.</td>
</tr>
<tr>
<td>No</td>
<td>Patient has no source of income or lost job. Example: Patient used to work for the state lottery system, currently <em>unemployed</em>.</td>
</tr>
<tr>
<td>Unemployment</td>
<td>Patient has employment or some source of income. Example: He works for [<strong>Company</strong>*].</td>
</tr>
<tr>
<td>Social isolation</td>
<td>Lack of social support or community engagement. Example: Lives <em>alone</em> in [<strong>Location</strong>].</td>
</tr>
<tr>
<td>Yes</td>
<td>Presence of social support. Example: He is <em>married</em> and <em>lives</em> with his wife.</td>
</tr>
<tr>
<td>No</td>
<td>Patient currently consumes alcohol. Example: two glasses of <em>wine</em> per night and 3 bottles over the weekend.</td>
</tr>
<tr>
<td>Alcohol use</td>
<td>Patient has a history of alcohol consumption. Example: He has a history of <em>alcohol abuse</em>.</td>
</tr>
<tr>
<td>Current</td>
<td>Patient never consumed alcohol. Example: She denies any <em>alcohol use</em>.</td>
</tr>
<tr>
<td>Former</td>
<td>Patient never consumed tobacco. Example: She is a <em>nonsmoker</em>.</td>
</tr>
<tr>
<td>Illicit drug use</td>
<td>Patient uses non-prescribed controlled substance. Example: occasional <em>marijuana use</em>.</td>
</tr>
<tr>
<td>Current</td>
<td>Patient has a history of using non-prescribed controlled substance. Example: Has a h/o of <em>cocaine and marijuana abuse</em>.</td>
</tr>
<tr>
<td>Former</td>
<td>Patient never used non-prescribed controlled substance, e.g., cocaine, marijuana. Example: Does not drink alcohol or use recreational <em>drugs</em>.</td>
</tr>
</tbody>
</table>

*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 ($\chi^2=10.39; P=.24$). 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.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Overall (n=48,869)</th>
<th>With OD&lt;sup&gt;a&lt;/sup&gt; (n=171)</th>
<th>Without OD (n=48,698)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong>&lt;sup&gt;b&lt;/sup&gt; (years), n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;40</td>
<td>4715 (9.65)</td>
<td>62 (36.26)</td>
<td>4653 (9.55)</td>
</tr>
<tr>
<td>40-64</td>
<td>18,878 (38.63)</td>
<td>92 (53.80)</td>
<td>18,786 (38.58)</td>
</tr>
<tr>
<td>&gt;64</td>
<td>25,276 (51.72)</td>
<td>17 (9.94)</td>
<td>25,259 (51.87)</td>
</tr>
<tr>
<td><strong>Gender</strong>&lt;sup&gt;b&lt;/sup&gt;, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>27,436 (56.14)</td>
<td>100 (58.48)</td>
<td>27,336 (56.13)</td>
</tr>
<tr>
<td>Female</td>
<td>21,433 (43.86)</td>
<td>71 (41.52)</td>
<td>21,362 (43.87)</td>
</tr>
<tr>
<td><strong>Race/ethnicity</strong>&lt;sup&gt;b&lt;/sup&gt;, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>35,058 (71.74)</td>
<td>127 (74.27)</td>
<td>34,931 (71.73)</td>
</tr>
<tr>
<td>Black</td>
<td>4694 (9.61)</td>
<td>13 (7.60)</td>
<td>4681 (9.61)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1664 (3.40)</td>
<td>8 (4.68)</td>
<td>1656 (3.40)</td>
</tr>
<tr>
<td>Other</td>
<td>7453 (15.25)</td>
<td>23 (13.45)</td>
<td>7430 (15.26)</td>
</tr>
<tr>
<td><strong>Marital status</strong>&lt;sup&gt;b&lt;/sup&gt;, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>23,378 (47.84)</td>
<td>42 (24.56)</td>
<td>23,336 (47.92)</td>
</tr>
<tr>
<td>Divorced</td>
<td>3664 (7.50)</td>
<td>22 (12.87)</td>
<td>3642 (7.48)</td>
</tr>
<tr>
<td>Widowed</td>
<td>7018 (14.36)</td>
<td>6 (3.51)</td>
<td>7012 (14.40)</td>
</tr>
<tr>
<td>Single</td>
<td>12,329 (25.23)</td>
<td>78 (45.61)</td>
<td>12,251 (25.16)</td>
</tr>
<tr>
<td>Unknown</td>
<td>2480 (5.07)</td>
<td>23 (13.45)</td>
<td>2457 (5.04)</td>
</tr>
<tr>
<td><strong>Clinical variables</strong>&lt;sup&gt;b&lt;/sup&gt;, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug use disorder</td>
<td>1493 (3.06)</td>
<td>80 (46.78)</td>
<td>1413 (2.90)</td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td>1009 (2.06)</td>
<td>28 (16.37)</td>
<td>981 (2.01)</td>
</tr>
<tr>
<td>Tobacco use disorder</td>
<td>3274 (6.70)</td>
<td>20 (11.70)</td>
<td>3254 (6.68)</td>
</tr>
<tr>
<td>Major depressive disorder</td>
<td>298 (0.61)</td>
<td>7 (4.09)</td>
<td>291 (0.60)</td>
</tr>
<tr>
<td>Alcohol use disorder</td>
<td>4121 (8.43)</td>
<td>37 (21.64)</td>
<td>4084 (8.39)</td>
</tr>
<tr>
<td>Cirrhosis</td>
<td>2431 (4.97)</td>
<td>19 (11.11)</td>
<td>2412 (4.95)</td>
</tr>
<tr>
<td>COPD&lt;sup&gt;c&lt;/sup&gt;</td>
<td>5674 (11.61)</td>
<td>18 (10.53)</td>
<td>5656 (11.61)</td>
</tr>
<tr>
<td>Renal insufficiency</td>
<td>8158 (16.69)</td>
<td>12 (7.02)</td>
<td>8146 (16.73)</td>
</tr>
<tr>
<td>**Social determinant&lt;sup&gt;d&lt;/sup&gt;: insurance provider, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private</td>
<td>15,371 (31.45)</td>
<td>43 (25.15)</td>
<td>15,328 (31.48)</td>
</tr>
<tr>
<td>Medicaid</td>
<td>4307 (8.81)</td>
<td>60 (35.09)</td>
<td>4247 (8.72)</td>
</tr>
<tr>
<td>Medicare</td>
<td>27,365 (56.00)</td>
<td>48 (28.07)</td>
<td>27,317 (56.09)</td>
</tr>
<tr>
<td>Government (others)</td>
<td>1324 (2.71)</td>
<td>14 (8.19)</td>
<td>1310 (2.69)</td>
</tr>
<tr>
<td>Self-pay</td>
<td>502 (1.03)</td>
<td>6 (3.50)</td>
<td>496 (1.02)</td>
</tr>
<tr>
<td>**Social determinant&lt;sup&gt;d&lt;/sup&gt;: housing insecurity, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>402 (0.82)</td>
<td>10 (5.85)</td>
<td>392 (0.80)</td>
</tr>
<tr>
<td>No</td>
<td>27,119 (55.49)</td>
<td>92 (53.80)</td>
<td>27,027 (55.50)</td>
</tr>
<tr>
<td>Unknown</td>
<td>21,348 (43.69)</td>
<td>69 (40.35)</td>
<td>21,279 (43.70)</td>
</tr>
<tr>
<td>**Social determinant&lt;sup&gt;d&lt;/sup&gt;: unemployment, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>3559 (7.28)</td>
<td>37 (21.64)</td>
<td>3522 (7.22)</td>
</tr>
<tr>
<td>No</td>
<td>12,671 (25.93)</td>
<td>31 (18.13)</td>
<td>12,640 (25.96)</td>
</tr>
</tbody>
</table>
Variables | Overall (n=48,869) | With OD\(^d\) (n=171) | Without OD\(^d\) (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 (24.66) | 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).

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

**Variables**

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

\(^a\)OR: odds ratio.
\(^b\)aOR: adjusted odds ratio.
\(^c\)Ref: Reference.
\(^d\)COPD: chronic obstructive pulmonary disorder.
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_v911e32851_app1.docx ]

Multimedia Appendix 2
Natural language processing model training and evaluation.
[DOCX File , 23 KB - medinform_v911e32851_app2.docx ]

References


30. medspacy. GitHub. URL: https://github.com/medspacy/medspacy [accessed 2021-10-16]


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

©Avijit Mitra, Hiba Ahsan, Wenjun Li, Weisong Liu, Robert D Kerns, Jack Tsai, William Becker, David A Smelson, Hong Yu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 08.11.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
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 Freeman1,2, BSc, DPhil; Raul Rodriguez-Esteban1, PhD; Juergen Gottowik1, PhD; Xing Yang3, PhD; Veit Johannes Erpenbeck1, MD; Mathias Leddin1, PhD

1Roche Pharma Research and Early Development, Pharma Research and Early Development Informatics, Roche Innovation Center Basel, F. Hoffmann–La Roche Ltd, Basel, Switzerland
2wordup development AG, CH-8006, Zurich, Switzerland
3Roche Pharma Research and Early Development, Pharma Research and Early Development Informatics, Roche Innovation Center Little Falls, NJ, United States
4Roche 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.
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

KEYWORDS

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

![Topics mentioned in association with COPD](https://medinform.jmir.org/2021/11/e26272)

Breathless  
Cough  
Mucus
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


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%20Disease%22%5BMeSH%5D+Terms%5D [accessed 2021-10-08]


15. Global Initiative for Chronic Obstructive Lung Disease (GOLD). URL: https://goldcopd.org/ [accessed 2021-10-08]


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 Teramoto1,2*, MS; Toshihiro Takeda1*, MD, PhD; Naoki Mihara3, MD, PhD; Yoshie Shimai1, PhD; Shirou Manabe1, MS; Shigeki Kuwata4, PhD; Hiroshi Kondoh2, MD, PhD; Yasushi Matsumura1, MD, PhD

1Department of Medical Informatics, Graduate School of Medicine, Osaka University, Suita, Japan
2Division of Medical Informatics, Tottori University Hospital, Yonago, Japan
3Department of Medical Informatics, National Cancer Center Hospital, Tokyo, Japan
4Department 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% (27/30 patients) 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.

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.

**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.

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th>Criterion details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation of ALT(^a) after initiation of medication</td>
<td>Peak ALT(^b) &gt; ULN(^c) of ALT and Peak ALT ≥ ALT (before start of medication)(^d) × 2</td>
</tr>
<tr>
<td>Decrease of ALT after cessation of medication</td>
<td>ALT (after cessation of medication)(^e) &lt; Max ALT(^f) × 0.5 or ALT (after cessation of medication) &lt; ALT (ULN)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exclusion criteria</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Decrease of ALT during the medication period but after the day of peak ALT</td>
<td>ALT (during medication period)(^g) &lt; peak ALT × 0.25 or ALT (during medication period) &lt; ULN of ALT</td>
</tr>
<tr>
<td>Liver injury induced by nondrug causes</td>
<td>Previous liver injury(^h), viral hepatitis(^i), or surgical operation(^j)</td>
</tr>
</tbody>
</table>

\(^a\) 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 bile 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.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>OUH(^b)</th>
<th>NCVC(^b)</th>
<th>TUH(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturer</td>
<td>NEC Corp.</td>
<td>NEC Corp.</td>
<td>IBM Corp.</td>
</tr>
<tr>
<td>CPOE(^d) database model</td>
<td>Oracle</td>
<td>Oracle</td>
<td>Database 2</td>
</tr>
<tr>
<td>Data range (mm/dd/yy)</td>
<td>04/01/00-12/01/12</td>
<td>04/01/00-02/01/14</td>
<td>0/01-03-09/01/13</td>
</tr>
<tr>
<td>Patients, n</td>
<td>1,028,852</td>
<td>251,143</td>
<td>307,944</td>
</tr>
<tr>
<td>CPOE records, n</td>
<td>20,447,443</td>
<td>8,128,059</td>
<td>9,360,281</td>
</tr>
</tbody>
</table>

\(^a\) 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).

<table>
<thead>
<tr>
<th>Values</th>
<th>Aspirin</th>
<th>Clopidogrel</th>
<th>Ticlopidine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients, n (%a)</td>
<td>40,938 (2.58)</td>
<td>10,263 (0.65)</td>
<td>6224 (0.39)</td>
</tr>
<tr>
<td>CPOEb records, n</td>
<td>668,765</td>
<td>88,520</td>
<td>104,003</td>
</tr>
<tr>
<td>CPOE records per patient, mean</td>
<td>16.3</td>
<td>8.6</td>
<td>16.7</td>
</tr>
<tr>
<td>Medication history records, n</td>
<td>58,873</td>
<td>12,224</td>
<td>8,854</td>
</tr>
<tr>
<td>Medication history records per patient, mean</td>
<td>1.4</td>
<td>1.2</td>
<td>1.4</td>
</tr>
</tbody>
</table>

aPercentage of the study population treated with the target medicine/electronic medical record–registered population (1,587,939 patients).
bCPOE: 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.

<table>
<thead>
<tr>
<th></th>
<th>Aspirin</th>
<th>Clopidogrel</th>
<th>Ticlopidine</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target patient distribution, n</strong></td>
<td>26,059</td>
<td>5076</td>
<td>3974</td>
</tr>
<tr>
<td><strong>DILI</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(MAX(^b) ALT(^c) &gt; ULN(^d))</td>
<td>868 (3.33%)</td>
<td>188 (3.70%)</td>
<td>226 (5.69%)</td>
</tr>
<tr>
<td>MAX ALT ≥ 80 IU/L</td>
<td>341 (0.95%)</td>
<td>69 (0.93%)</td>
<td>83 (1.43%)</td>
</tr>
<tr>
<td>MAX ALT ≥ 200 IU/L</td>
<td>93 (0.36%)</td>
<td>22 (0.43%)</td>
<td>26 (0.65%)</td>
</tr>
</tbody>
</table>

\(^a\)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.

<table>
<thead>
<tr>
<th>Medical record values</th>
<th>Aspirin</th>
<th>Clopidogrel</th>
<th>Ticlopidine</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADEs(^a) with ALT(^b) elevation, n</td>
<td>27</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>Caused by target medicine</td>
<td>8</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Caused by concomitant medicine</td>
<td>11</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Offending medicine not specified</td>
<td>8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Other causes of liver injury, n</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total, n</strong></td>
<td>30</td>
<td>9</td>
<td>4</td>
</tr>
</tbody>
</table>

\(^a\)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. 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.

<table>
<thead>
<tr>
<th>Values</th>
<th>Aspirin</th>
<th>Clopidogrel</th>
<th>Ticlopidine</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OUH$^a$</td>
<td>TUH$^b$</td>
<td>OUH</td>
<td>TUH</td>
</tr>
<tr>
<td>Target patients</td>
<td>7611</td>
<td>4002</td>
<td>1266</td>
<td>951</td>
</tr>
<tr>
<td>RUCAM$^c$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First$^d$ screening</td>
<td>5</td>
<td>28</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>ADE$^e,f$</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>ADE medication history$^g$</td>
<td>26</td>
<td>10</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

$^a$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%; clopidogrel: 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 disease 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


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

©Kei Teramoto, Toshihiro Takeda, Naoki Mihara, Yoshie Shimai, Shirou Manabe, Shigeki Kuwata, Hiroshi Kondoh, Yasushi Matsumura. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 01.11.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
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 R² 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), 31.3% (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...
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 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 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, gnosis, 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 \( \theta_i \) 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:

\[
\text{which is used to update the ratings as follows, based on the Bradley-Terry model [38]:}
\]

where \( K \) is a constant parameter that controls how quickly \( \theta \) changes, with large \( K \) values resulting in \( \theta \) changing quickly and small \( K \) values resulting in \( \theta \) 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>
<thead>
<tr>
<th>Variables</th>
<th>Admission</th>
<th>Discharge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>51 (9)</td>
<td>N/A(^b)</td>
</tr>
<tr>
<td>Age &lt;65 years, n (%)</td>
<td>270 (93.8)</td>
<td>N/A</td>
</tr>
<tr>
<td>Males, n (%)</td>
<td>195 (67.7)</td>
<td>N/A</td>
</tr>
<tr>
<td>Marital status (married), n (%)</td>
<td>180 (62.5)</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Educational level, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read and write</td>
<td>9 (3.1)</td>
<td>N/A</td>
</tr>
<tr>
<td>Primary</td>
<td>114 (50)</td>
<td>N/A</td>
</tr>
<tr>
<td>Secondary</td>
<td>88 (30.5)</td>
<td>N/A</td>
</tr>
<tr>
<td>Higher</td>
<td>66 (22.9)</td>
<td>N/A</td>
</tr>
<tr>
<td>NIHSS(^c), median (IQR)</td>
<td>11 (7-15)</td>
<td>N/A</td>
</tr>
<tr>
<td>TMT(^d) A, mean (SD)</td>
<td>82 (64)</td>
<td>N/A</td>
</tr>
<tr>
<td>TMT B, mean (SD)</td>
<td>157 (90)</td>
<td>N/A</td>
</tr>
<tr>
<td>PMR(^e), mean (SD)</td>
<td>27 (12)</td>
<td>N/A</td>
</tr>
<tr>
<td>VC(^f)–CUBS, mean (SD)</td>
<td>23 (12)</td>
<td>N/A</td>
</tr>
<tr>
<td>VP(^g)–IMAGES, mean (SD)</td>
<td>17 (3)</td>
<td>N/A</td>
</tr>
<tr>
<td>VP–WAIS III(^h), mean (SD)</td>
<td>37 (15)</td>
<td>N/A</td>
</tr>
<tr>
<td>Barcelona test–repetition, mean (SD)</td>
<td>9 (1)</td>
<td>N/A</td>
</tr>
<tr>
<td>Barcelona test–denomination, mean (SD)</td>
<td>13 (1)</td>
<td>N/A</td>
</tr>
<tr>
<td>Barcelona test–comprehension, mean (SD)</td>
<td>15 (1)</td>
<td>N/A</td>
</tr>
<tr>
<td>Barcelona test–DIGITS, mean (SD)</td>
<td>3 (1)</td>
<td>4 (1)</td>
</tr>
<tr>
<td>Barcelona test–DIGITS, median (IQR)</td>
<td>4 (3-4)</td>
<td>4 (3-5)</td>
</tr>
<tr>
<td>RAVLT(^i) 075, mean (SD)</td>
<td>37 (10)</td>
<td>43 (11)</td>
</tr>
<tr>
<td>RAVLT075, median (IQR)</td>
<td>37 (30-45)</td>
<td>44 (35-52)</td>
</tr>
<tr>
<td>RAVLT015, mean (SD)</td>
<td>6 (3)</td>
<td>8 (3)</td>
</tr>
<tr>
<td>RAVLT015, median (IQR)</td>
<td>7 (5-9)</td>
<td>9 (6-12)</td>
</tr>
<tr>
<td>RAVLT015R, mean (SD)</td>
<td>10 (4)</td>
<td>11 (3)</td>
</tr>
<tr>
<td>RAVLT015R, median (IQR)</td>
<td>12 (8-14)</td>
<td>13 (10-14)</td>
</tr>
<tr>
<td>Length of stay (days), mean (SD)</td>
<td>88 (36)</td>
<td>N/A</td>
</tr>
<tr>
<td>Length of stay (days), median (IQR)</td>
<td>84 (55-113)</td>
<td>N/A</td>
</tr>
<tr>
<td>Time since onset to rehab admission (days), mean (SD)</td>
<td>55 (35)</td>
<td>N/A</td>
</tr>
<tr>
<td>Time since onset to rehab admission (days), median (IQR)</td>
<td>43 (29-75)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

\(^a\)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).

<table>
<thead>
<tr>
<th>Player</th>
<th>Glicko rating (deviation)</th>
<th>Games</th>
<th>Win</th>
<th>Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>id19</td>
<td>2565 (146.26)</td>
<td>20</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>id20</td>
<td>2565 (146.26)</td>
<td>20</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>id16</td>
<td>2450 (124.22)</td>
<td>20</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>id14</td>
<td>2368 (118.61)</td>
<td>20</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>id17</td>
<td>2364 (124.24)</td>
<td>20</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>id15</td>
<td>2293 (113.49)</td>
<td>20</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>id18</td>
<td>2273 (119.90)</td>
<td>20</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>id12</td>
<td>2240 (104.33)</td>
<td>20</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>id10</td>
<td>2195 (97.57)</td>
<td>20</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>id13</td>
<td>2190 (99.44)</td>
<td>20</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>id9</td>
<td>2177 (103.78)</td>
<td>20</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>id11</td>
<td>2143 (95.22)</td>
<td>20</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>id7</td>
<td>2122 (101.43)</td>
<td>20</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>id8</td>
<td>2112 (91.53)</td>
<td>20</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>id3</td>
<td>2035 (86.81)</td>
<td>20</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Maze</td>
<td>1999 (36.77)</td>
<td>400</td>
<td>118</td>
<td>282</td>
</tr>
<tr>
<td>id2</td>
<td>1981 (87.77)</td>
<td>20</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>id5</td>
<td>1965 (88.01)</td>
<td>20</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>id4</td>
<td>1960 (87.11)</td>
<td>20</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>id6</td>
<td>1957 (87.46)</td>
<td>20</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>id1</td>
<td>1930 (89.33)</td>
<td>20</td>
<td>7</td>
<td>13</td>
</tr>
</tbody>
</table>
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).

<table>
<thead>
<tr>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of task executions</td>
<td>44,814</td>
</tr>
<tr>
<td>Executions per patient, mean (SD)</td>
<td>155 (113.2)</td>
</tr>
<tr>
<td>Total number of sessions</td>
<td>5008</td>
</tr>
<tr>
<td>Sessions executed per patient, mean (SD)</td>
<td>17 (11.5)</td>
</tr>
<tr>
<td>Tasks executed per session per patient, mean (SD)</td>
<td>9 (4.4)</td>
</tr>
<tr>
<td>Total number of memory tasks executed</td>
<td>18,183</td>
</tr>
<tr>
<td>Total number of executive functioning tasks executed</td>
<td>14,061</td>
</tr>
<tr>
<td>Total number of attention tasks executed</td>
<td>8062</td>
</tr>
<tr>
<td>Total number of gnosias tasks executed</td>
<td>1795</td>
</tr>
<tr>
<td>Total number of calculus tasks executed</td>
<td>1695</td>
</tr>
<tr>
<td>Total number of orientation tasks executed</td>
<td>741</td>
</tr>
<tr>
<td>Total number of language tasks executed</td>
<td>261</td>
</tr>
<tr>
<td>Total number of social cognition tasks executed</td>
<td>16</td>
</tr>
<tr>
<td>Memory task results, mean (SD)</td>
<td>53.1 (36.4)</td>
</tr>
<tr>
<td>Executive functioning tasks results, mean (SD)</td>
<td>49.6 (38.7)</td>
</tr>
<tr>
<td>Attention task results, mean (SD)</td>
<td>59.4 (36.7)</td>
</tr>
<tr>
<td>Gnosias task results, mean (SD)</td>
<td>74.4 (30.8)</td>
</tr>
<tr>
<td>Calculus task results, mean (SD)</td>
<td>72.9 (35.8)</td>
</tr>
<tr>
<td>Orientation task results, mean (SD)</td>
<td>75.6 (38.0)</td>
</tr>
<tr>
<td>Language task results, mean (SD)</td>
<td>55.5 (38.4)</td>
</tr>
<tr>
<td>Social cognition task results, mean (SD)</td>
<td>56.7 (37.1)</td>
</tr>
</tbody>
</table>

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).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Low Elo (n=12,431)</th>
<th>Mid Elo (n=12,449)</th>
<th>High Elo (n=13,297)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex (female), n (%)</td>
<td>4396 (35.36)</td>
<td>4607 (37.01)</td>
<td>3793 (28.52)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Age (years) when starting rehabilitation, mean (SD)</td>
<td>52 (8)</td>
<td>51 (8)</td>
<td>48 (10)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>DIGITS at admission, mean (SD)</td>
<td>3.4 (0.9)</td>
<td>3.8 (0.9)</td>
<td>4.0 (1.0)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>RAVLT&lt;sup&gt;a&lt;/sup&gt; at admission, mean (SD)</td>
<td>37 (9)</td>
<td>36 (10)</td>
<td>38 (10)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>RAVLT015 at admission, mean (SD)</td>
<td>6 (3)</td>
<td>6 (3)</td>
<td>7 (3)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>RAVLT015R at admission, mean (SD)</td>
<td>10 (4)</td>
<td>10 (4)</td>
<td>11 (4)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Length of stay (days), mean (SD)</td>
<td>104 (37)</td>
<td>105 (35)</td>
<td>106 (40)</td>
<td>.05</td>
</tr>
<tr>
<td>Executed tasks, mean (SD)</td>
<td>245 (129)</td>
<td>222 (122)</td>
<td>241 (124)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Obtained results in tasks, mean (SD)</td>
<td>37 (36)</td>
<td>56 (36)</td>
<td>68 (33)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>DIGITS at discharge, mean (SD)</td>
<td>3.5 (1.0)</td>
<td>4.5 (0.9)</td>
<td>4.4 (0.8)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>RAVLT075 at discharge, mean (SD)</td>
<td>42 (11)</td>
<td>45 (11)</td>
<td>45 (12)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>RAVLT015 at discharge, mean (SD)</td>
<td>8 (3)</td>
<td>9 (3)</td>
<td>9 (4)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>RAVLT015R at discharge, mean (SD)</td>
<td>11 (3)</td>
<td>12 (3)</td>
<td>12 (3)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>DIGITS IMP&lt;sup&gt;b&lt;/sup&gt; (yes), n (%)</td>
<td>2859 (22.99)</td>
<td>7059 (56.7)</td>
<td>5353 (40.26)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>RAVLT075 IMP (yes), n (%)</td>
<td>5308 (42.69)</td>
<td>8356 (67.12)</td>
<td>7484 (56.28)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>RAVLT015 IMP (yes), n (%)</td>
<td>6136 (49.36)</td>
<td>7325 (58.84)</td>
<td>7482 (56.26)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>RAVLT015R IMP (yes), n (%)</td>
<td>6802 (54.72)</td>
<td>6683 (53.68)</td>
<td>7132 (53.64)</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

**Task difficulty level, n (%)**

- Level 1: 4812 (38.71), 3997 (32.11), 3536 (26.59), <.01
- Level 2: 3999 (32.17), 3857 (30.98), 4346 (32.68), <.01
- Level 3: 3620 (29.12), 4595 (36.91), 5415 (40.72), <.01

<sup>a</sup>RAVLT: Rey Auditory Verbal Learning Test.

<sup>b</sup>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 $\beta$ (95% CI), standard $\beta$, Durbin-Watson (D-W) test, variance inflation factor, and $R^2$ and adjusted $R^2$ for RAVLT075 and DIGITS at discharge (N=288).

<table>
<thead>
<tr>
<th>Variables</th>
<th>$\beta$ (95% CI)</th>
<th>Standard $\beta$</th>
<th>1/VIF$^a$</th>
<th>$P$ value</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 predictors of RAVLT$^b$ 075 at discharge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elo rating</td>
<td>.01 (.01 to .02)</td>
<td>.09</td>
<td>0.95</td>
<td>.02</td>
<td>0.55</td>
<td>0.54</td>
</tr>
<tr>
<td>RAVLT075 at admission</td>
<td>.76 (.67 to .85)</td>
<td>.66</td>
<td>0.92</td>
<td>&lt;.001</td>
<td>0.55</td>
<td>0.54</td>
</tr>
<tr>
<td>LOS$^c$</td>
<td>.04 (.01 to .06)</td>
<td>.13</td>
<td>0.98</td>
<td>.002</td>
<td>0.55</td>
<td>0.54</td>
</tr>
<tr>
<td>Sex</td>
<td>−2.48 (--4.48 to −0.48)</td>
<td>−0.10</td>
<td>0.93</td>
<td>.01</td>
<td>0.55</td>
<td>0.54</td>
</tr>
<tr>
<td>Age (years)</td>
<td>−0.09 (−0.19 to .01)</td>
<td>−0.07</td>
<td>0.92</td>
<td>.06</td>
<td>0.55</td>
<td>0.54</td>
</tr>
<tr>
<td>D-W=$^d$=1.89</td>
<td>N/A$^e$</td>
<td>N/A</td>
<td>N/A</td>
<td>.37</td>
<td>0.55</td>
<td>0.54</td>
</tr>
<tr>
<td>Model 2 predictors of DIGITS at discharge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elo rating</td>
<td>.00 (.00 to .00)</td>
<td>.10</td>
<td>0.88</td>
<td>.02</td>
<td>0.44</td>
<td>0.43</td>
</tr>
<tr>
<td>DIGITS at admission</td>
<td>.63 (.54 to .72)</td>
<td>.63</td>
<td>0.91</td>
<td>&lt;.001</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>LOS</td>
<td>.00 (.00 to .00)</td>
<td>.05</td>
<td>0.98</td>
<td>.22</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>Sex</td>
<td>.04 (-0.15 to .23)</td>
<td>.01</td>
<td>0.95</td>
<td>.67</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>Age (years)</td>
<td>.00 (.00 to .01)</td>
<td>.02</td>
<td>0.94</td>
<td>.53</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>D-W=$^d$=2.01</td>
<td>N/A$^e$</td>
<td>N/A</td>
<td>N/A</td>
<td>.95</td>
<td>0.44</td>
<td>0.44</td>
</tr>
</tbody>
</table>

$^a$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 $\beta$, odds ratio (95% CI), variance inflation factor for RAVLT075, and DIGITS improvement at discharge (N=288).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Odds ratio (95% CI)</th>
<th>$\beta$ coefficients</th>
<th>1/VIF&lt;sup&gt;a&lt;/sup&gt;</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model 3&lt;sup&gt;b&lt;/sup&gt; predictors of RAVLT&lt;sup&gt;c&lt;/sup&gt; 075 improvement at discharge</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rating</td>
<td>1.00 (1.00-1.00)</td>
<td>.61</td>
<td>0.93</td>
<td>.02</td>
</tr>
<tr>
<td>RAVLT075 at admission</td>
<td>0.95 (0.92-0.97)</td>
<td>−0.95</td>
<td>0.88</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>LOS&lt;sup&gt;d&lt;/sup&gt;</td>
<td>1.00 (1.00-1.01)</td>
<td>.67</td>
<td>0.98</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Sex</td>
<td>0.64 (0.37-1.11)</td>
<td>−0.40</td>
<td>0.92</td>
<td>.12</td>
</tr>
<tr>
<td>Age</td>
<td>0.97 (0.94-0.99)</td>
<td>−0.52</td>
<td>0.92</td>
<td>.04</td>
</tr>
<tr>
<td><strong>Model 4&lt;sup&gt;e&lt;/sup&gt; predictors of DIGITS improvement at discharge</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rating</td>
<td>1.00 (0.99-1.00)</td>
<td>.57</td>
<td>0.86</td>
<td>.06</td>
</tr>
<tr>
<td>DIGITS at admission</td>
<td>0.38 (0.26-0.52)</td>
<td>−2.04</td>
<td>0.86</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>LOS</td>
<td>1.00 (1.00-1.01)</td>
<td>.71</td>
<td>0.97</td>
<td>.01</td>
</tr>
<tr>
<td>Sex</td>
<td>1.02 (0.57-1.85)</td>
<td>.02</td>
<td>0.96</td>
<td>.92</td>
</tr>
<tr>
<td>Age</td>
<td>1.00 (0.97-1.02)</td>
<td>.01</td>
<td>0.95</td>
<td>.95</td>
</tr>
</tbody>
</table>

<sup>a</sup>VIF: variance inflation factor.

<sup>b</sup>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.

<sup>c</sup>RAVLT: Rey Auditory Verbal Memory Test.

<sup>d</sup>LOS: length of stay.

<sup>e</sup>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 Gutmman, 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=0.9 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 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.

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 descendant, 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.

References


48. 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

©Alejandro Garcia-Rudolph, Eloy Opisso, Jose M Tormos, Vince Istvan Madai, Dietmar Frey, Helard Becerra, John D Kelleher, Montserrat Bernabeu Guitart, Jaume López. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 10.11.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Health Information Needs of Young Chinese People Based on an Online Health Community: Topic and Statistical Analysis

Jie Wang¹,², PhD; Xin Wang³, PhD; Lei Wang¹, MS; Yan Peng¹, PhD

¹School of Management, Capital Normal University, Beijing, China
²State key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China
³Department 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. Health 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, \( Q = \{ q_1, q_2, \ldots, q_{|Q|} \} \). 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 \( C_1, C_2, \ldots, C_k \), 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 \( R_{n \times n} \), and then \( R_{n \times n} \) was projected to the eigenvector space \( R_{n \times k}^{p} \) corresponding to the first \( k \)-dominant eigenvalues whose cumulative contribution rate was \( \lambda \). 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 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,” “ 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.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Frequency, n (%) (N=1400)</th>
<th>Concrete content</th>
<th>Keywords (top 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>625 (44.64)</td>
<td>Consult for treatment of menstrual irregularities and medications</td>
<td>Delay, how to treat, how to do, dysmenorrhea, causes</td>
</tr>
<tr>
<td>Pathology</td>
<td>84 (6.00)</td>
<td>Consult for types, etiology, and pathology of menstrual irregularities</td>
<td>What is going on, menstruation, bleeding, causes, brown</td>
</tr>
<tr>
<td>Symptom</td>
<td>242 (17.29)</td>
<td>Consult for signs, symptoms, and tests of menstrual irregularities</td>
<td>Delay, leucorrhea, examination, symptoms, feelings</td>
</tr>
<tr>
<td>Pregnancy</td>
<td>189 (13.50)</td>
<td>Counseling whether menstrual irregularities are associated with pregnancy</td>
<td>Pregnancy, have sexual intercourse, birth control pills, boyfriend, safety period</td>
</tr>
<tr>
<td>Diet</td>
<td>260 (18.57)</td>
<td>Consult for menstrual irregularities, dietary contraindications and precautions, and suitability of certain foods</td>
<td>What to eat, food, conditioning, diet, brown sugar</td>
</tr>
</tbody>
</table>

Table 2. Topic extraction results for weight loss.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Frequency, n (%) (N=3381)</th>
<th>Concrete content</th>
<th>Keywords (top 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet</td>
<td>2371 (70.11)</td>
<td>Counseling on coarse grain cereals that help lose weight, as well as consumption effects</td>
<td>Potatoes, sweet potatoes, corn, red beans, oats</td>
</tr>
<tr>
<td>Coarse grain</td>
<td>411 (12.16)</td>
<td>Counseling on coarse grain cereals that help lose weight, as well as consumption effects</td>
<td>Apples, fruits, bananas, cucumbers, bitter gourd</td>
</tr>
<tr>
<td>Fruits and vegetables</td>
<td>487 (14.40)</td>
<td>Counseling on fruits and vegetables that help you lose weight and how they work</td>
<td>Yogurt, honey water, milk, diet tea, coffee</td>
</tr>
<tr>
<td>Beverages</td>
<td>353 (10.44)</td>
<td>Counseling on various types of beverages that help with weight loss and how well they work</td>
<td></td>
</tr>
<tr>
<td>Weight loss recipes</td>
<td>1120 (33.13)</td>
<td>Counseling on healthy recipes that help to lose weight</td>
<td>What to eat, how to eat, food, effect, dieting</td>
</tr>
<tr>
<td>Surgery</td>
<td>614 (18.16)</td>
<td>Counseling on various surgical weight loss methods, effects, and costs.</td>
<td>Diet, treatment, liposuction, thin face pin, surgery, lipolysis, effect</td>
</tr>
<tr>
<td>Pathology</td>
<td>396 (11.71%)</td>
<td>Counseling on the causes of obesity and weight loss methods</td>
<td>Obesity, getting fat, causing, sweets, why</td>
</tr>
</tbody>
</table>

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 ($X^2 = 2591.7; P < .001$).

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 ($X^2 = 113.7; P < .001$), vitiligo ($X^2 = 100.6; P < .001$), weight loss ($X^2 = 49.0; P < .001$), and depression ($X^2 = 88.7; P < .001$) 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 ($X^2 = 44.4; P < .001$), influenza ($X^2 = 81.1; P < .001$), vitiligo ($X^2 = 64.2; P < .001$), weight loss ($X^2 = 157.5; P < .001$), and depression ($X^2 = 30.2; P < .001$) 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.

References


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]


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
Visualizing Knowledge Evolution Trends and Research Hotspots of Personal Health Data Research: Bibliometric Analysis

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.”
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.

<table>
<thead>
<tr>
<th>Study</th>
<th>Research question</th>
<th>Sample size</th>
<th>Time range</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archer et al [17]</td>
<td>PHR design, functionality, implementation, application, outcomes, and benefits</td>
<td>130</td>
<td>Unlimited-2010</td>
<td>Systematic review</td>
</tr>
<tr>
<td>Van Panhuis et al [19]</td>
<td>Barriers to data sharing</td>
<td>65</td>
<td>Unlimited-2013</td>
<td>Systematic review</td>
</tr>
<tr>
<td>Kruse et al [18]</td>
<td>Adoption factors of EHRs</td>
<td>31</td>
<td>2012-2015</td>
<td>Systematic review</td>
</tr>
<tr>
<td>Maher et al [20]</td>
<td>Ethical issues in passive data collection</td>
<td>48</td>
<td>Unlimited-2018</td>
<td>Systematic review</td>
</tr>
<tr>
<td>Mehta and Pandit [22]</td>
<td>Big data analytics in PHD</td>
<td>58</td>
<td>2013-2018</td>
<td>Systematic review</td>
</tr>
<tr>
<td>Wang et al [28]</td>
<td>Evolution of publication hotspots in EHRs</td>
<td>17,678</td>
<td>1957-2016</td>
<td>Bibliometric method</td>
</tr>
<tr>
<td>Guo et al [23]</td>
<td>Status, hotspots of EHR</td>
<td>5095</td>
<td>2005-2010</td>
<td>Bibliometric method</td>
</tr>
<tr>
<td>Qian et al [29]</td>
<td>Landscape, hot topics, trends of EHRs</td>
<td>13,438</td>
<td>1900-2019</td>
<td>Bibliometric method</td>
</tr>
<tr>
<td>Lin et al [26]</td>
<td>Status of EMR research in China</td>
<td>1752</td>
<td>1999-2012</td>
<td>Bibliometric method</td>
</tr>
<tr>
<td>Karampela et al [2]</td>
<td>Publication source, publication year, research topic</td>
<td>246</td>
<td>Unlimited-2018</td>
<td>Systematic mapping study</td>
</tr>
<tr>
<td>This study</td>
<td>Knowledge evolution trajectory of PHD, including EHR, PHR, and EMR</td>
<td>8281</td>
<td>2009-2018</td>
<td>Bibliometric method</td>
</tr>
</tbody>
</table>

aPHR: personal health record.
bEHR: electronic health record.
cPHD: personal health data.
dEMR: electronic medical record.
Table 2. Mapping questions.

<table>
<thead>
<tr>
<th>Question and ID</th>
<th>Mapping question</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>MQ1a: References</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MQ1.1</td>
<td>How does the references co-citation network shape?</td>
<td>To understand the main topics and the development of research topics in PHD.</td>
</tr>
<tr>
<td>MQ1.2</td>
<td>How has the knowledge cluster evolved?</td>
<td>To identify which PHD topic has the most longevity and the newest hotspot.</td>
</tr>
<tr>
<td>MQ1.3</td>
<td>What are the citation bursts of reference networks?</td>
<td>To explore the emerging PHD research topic characterized by articles.</td>
</tr>
<tr>
<td>MQ2: Keywords</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MQ2.1</td>
<td>What are the keyword bursts in recent years?</td>
<td>To explore the emerging research interests in PHD characterized by keywords.</td>
</tr>
<tr>
<td>MQ3: Disciplines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MQ3.1</td>
<td>What does the discipline categories co-occurrence net-</td>
<td>To identify the trends of discipline categories that are involved in PHD.</td>
</tr>
<tr>
<td></td>
<td>work shape?</td>
<td></td>
</tr>
<tr>
<td>MQ3.2</td>
<td>What are the discipline categories bursts?</td>
<td>To explore the discipline categories that increased abruptly in PHD.</td>
</tr>
</tbody>
</table>

aMQ: mapping question.
bPHD: 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:

\[ TS(Topic)=\text{“personal health data” OR “personal health record” OR “electronic health record” OR “electronic medical record”} \]
\[ \text{AND DT(Document Types)=“Articles” OR “Review”} \]
\[ \text{AND PY(Year Published)= (2009-2018).} \]

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).
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.

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].

**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 $\geq 0.7$ indicate high similarity among articles in the same cluster, while modularity Q values $\geq 0.6662$ 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 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.

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.

<table>
<thead>
<tr>
<th>Mean year</th>
<th>Cluster ID</th>
<th>Size</th>
<th>Silhouette</th>
<th>Label (LLR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>15</td>
<td>10</td>
<td>0.984</td>
<td>User groups perspective</td>
</tr>
<tr>
<td>2005</td>
<td>3</td>
<td>69</td>
<td>0.805</td>
<td>Clinical decision support</td>
</tr>
<tr>
<td>2005</td>
<td>5</td>
<td>62</td>
<td>0.815</td>
<td>Information technology diffusion</td>
</tr>
<tr>
<td>2005</td>
<td>10</td>
<td>31</td>
<td>0.777</td>
<td>Clinical documentation</td>
</tr>
<tr>
<td>2006</td>
<td>1</td>
<td>72</td>
<td>0.872</td>
<td>Integrative review</td>
</tr>
<tr>
<td>2006</td>
<td>13</td>
<td>18</td>
<td>0.947</td>
<td>Medication reconciliation issue</td>
</tr>
<tr>
<td>2007</td>
<td>4</td>
<td>64</td>
<td>0.838</td>
<td>Quality requirement</td>
</tr>
<tr>
<td>2007</td>
<td>6</td>
<td>61</td>
<td>0.809</td>
<td>Contingency factor</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>70</td>
<td>0.804</td>
<td>EHR systems</td>
</tr>
<tr>
<td>2010</td>
<td>7</td>
<td>61</td>
<td>0.847</td>
<td>Clinical decision support system</td>
</tr>
<tr>
<td>2010</td>
<td>12</td>
<td>30</td>
<td>0.833</td>
<td>Electronic health information exchange</td>
</tr>
<tr>
<td>2011</td>
<td>0</td>
<td>95</td>
<td>0.849</td>
<td>Genomic era</td>
</tr>
<tr>
<td>2011</td>
<td>11</td>
<td>31</td>
<td>0.936</td>
<td>Frequency</td>
</tr>
<tr>
<td>2013</td>
<td>8</td>
<td>51</td>
<td>0.893</td>
<td>Developing knowledge resource</td>
</tr>
<tr>
<td>2013</td>
<td>9</td>
<td>43</td>
<td>0.950</td>
<td>Allocating scarce resource</td>
</tr>
</tbody>
</table>

aThe connected components in cluster 14 are less than the default value (K=25), so CiteSpace did not report 14 [39].

bThe average year of the articles in a cluster.

cThe number of articles in each cluster.

dLLR: log-likelihood ratio.
eEHR: 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.
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).

Multimedia Appendix 2
A list of articles that contributed to clusters #8 and #9.

Multimedia Appendix 3
Keywords co-occurrence network.

Multimedia Appendix 4
The 56 keywords with the strongest burst (2009-2018).

Multimedia Appendix 5
The 15 disciplines with the strongest burst (2009–2018).

References


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

https://medinform.jmir.org/2021/11/e31142
JMIR Med Inform 2021 | vol. 9 | iss. 11 | e31142 | p.101
(page number not for citation purposes)
Event Prediction Model Considering Time and Input Error Using Electronic Medical Records in the Intensive Care Unit: Retrospective Study

MinDong Sung1*, MD; Sangchul Hahn2*, MSc; Chang Hoon Han3*, MD; Jung Mo Lee3, MD; Jayoung Lee2, MD, PhD; Jinkyu Yoo4, MSc; Jay Heo4, MSc; Young Sam Kim5, MD, PhD; Kyung Soo Chung5, MD, PhD

1Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea
2AITRICS, Inc, Seoul, Republic of Korea
3Division of Pulmonology, Department of Internal Medicine, National Health Insurance Service Ilsan Hospital, Goyang-si, Republic of Korea
4Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
5Division 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 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, medicsurgical, 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 (NHMC 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·kg\(^{-1}\)·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: T1 (near future), T2 (mid-term future), and T3 (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 Ti (i ∈ {1, 2, 3}), we preprocessed the data as positive and negative instances. Specifically, we randomly chose some of the time points within Ti from the event onset for positive instances and some of the time points within Ti from randomly chosen time points for negative instances.

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 F1 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.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Development cohort (n=21,732)</th>
<th>Validation cohort (n=2487)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients, n (%)</td>
<td>20,053 (92.27)</td>
<td>2362 (94.97)</td>
</tr>
<tr>
<td>Age in years, mean (SD)</td>
<td>60.97 (15.2)</td>
<td>69.05 (14.13)</td>
</tr>
<tr>
<td>Sex, male, n (%)</td>
<td>12,416 (57.13)</td>
<td>1405 (56.49)</td>
</tr>
<tr>
<td>Death, n (%)</td>
<td>783 (3.6)</td>
<td>209 (8.40)</td>
</tr>
<tr>
<td>Sepsis, n (%)</td>
<td>679 (3.12)</td>
<td>243 (9.77)</td>
</tr>
<tr>
<td>Acute kidney injury, n (%)</td>
<td>1,978 (9.1)</td>
<td>287 (11.54)</td>
</tr>
<tr>
<td>Length of ICU stay (days), mean (SD)</td>
<td>3.23 (19.15)</td>
<td>2.99 (3.65)</td>
</tr>
<tr>
<td>APACHE II score, mean (SD)</td>
<td>11.57 (5.04)</td>
<td>16.21 (7.25)</td>
</tr>
<tr>
<td>SOFA score, mean (SD)</td>
<td>3.66 (3.01)</td>
<td>4.11 (1.04)</td>
</tr>
</tbody>
</table>

ICU admission\textsuperscript{d}

- MICU\textsuperscript{e}: 3138 (14.44) 606 (24.37)
- SICU\textsuperscript{f}: 4604 (21.19) 1141 (45.88)
- CCU\textsuperscript{g}: 5172 (23.79) 740 (29.75)
- HICU\textsuperscript{h}: 3335 (15.35) —\textsuperscript{i}
- NCU\textsuperscript{j}: 5483 (25.23) —

\textsuperscript{a}ICU: intensive care unit.
\textsuperscript{b}APACHE: Acute Physiology and Chronic Health Evaluation.
\textsuperscript{c}SOFA: Sequential Organ Failure Assessment.
\textsuperscript{d}Patient could have multiple admissions to the ICU; the sum of the types of ICU admissions exceeds 100%.
\textsuperscript{e}MICU: medical intensive care unit.
\textsuperscript{f}SICU: surgical intensive care unit
\textsuperscript{g}CCU: critical care unit.
\textsuperscript{h}HICU: high intensity care unit.
\textsuperscript{i}Not available.
\textsuperscript{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. Additionally, 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.
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 which 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

Figure 5. Screenshot of the application on which the model was deployed.

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 which 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


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

©MinDong Sung, Sangchul Hahn, Chang Hoon Han, Jung Mo Lee, Jayoung Lee, Jin Kyu Yoo, Jay Heo, Young Sam Kim, Kyung Soo Chung. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 04.11.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Local Differential Privacy in the Medical Domain to Protect Sensitive Information: Algorithm Development and Real-World Validation

MinDong Sung¹*, MD; Dongchul Cha¹,²*, MD; Yu Rang Park¹, PhD

¹Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea
²Department 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 $10^{-3}$ and $10^{-5}$, 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 (IPA) and discretization postprocessing to provide high-fidelity data while preserving the privacy of the original data.
The Value of \( \varepsilon \) for Local Differential Privacy

Dwork et al [22] defined \( \varepsilon \)-differential privacy as a randomized function. For adjacent data \( Y_1 \) and \( Y_2 \), function \( \kappa \) is \((\varepsilon, \delta)\)-differentially private if

\[
P([\kappa(y_1) \in S] \leq \varepsilon \cdot P([\kappa(y_2) \in S] + \delta
\]

where \( S \subseteq \text{Range}(\kappa) \). 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 \( f_{\varepsilon,b}(y) \) as:

\[
f_{\varepsilon,b}(y) = \frac{1}{2} \cdot \frac{1}{b} \cdot \exp\left(-\frac{|y-q|}{b}\right)
\]

where

\[
\varepsilon = 0, b (\text{lower bound}) \text{ as } -1, u (\text{upper bound}) \text{ as } 1, \text{ and } \Delta Q \text{ as } 2 \text{ in our experiments and adjusted } \varepsilon \text{ to measure the effect of the privacy changes.}
\]

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 \( \Delta \) intervals. Therefore, for given perturbed data \( y \), we obtain the following:

\[
\text{After calculating } k, \text{ the Bernoulli probability } p \text{ was sampled such that}
\]

\[
\text{which is the distance between two adjacent possibilities. Finally, we discretized the perturbed data } y \text{ concerning the Bernoulli probability } p \text{ such that}
\]

\[
\text{where } \square \text{ denotes the Bernoulli distribution function.}
\]

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 \( \varepsilon \) 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, \( \text{pH} \), hematocrit, creatinine, albumin, oxygen pressure, \( \text{CO}_2 \) 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 \( \varepsilon = 0.1, \Delta = 0 \) (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 \( \epsilon = 0.1, \delta = 0. \)

**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 \( \epsilon \) was small (Figure 3B). As \( \epsilon \) increased, all perturbed values approached their original values for both continuous and categorical variables (Figures 3A and 3B).

**Figure 3.** \( \epsilon \) 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 \( \epsilon \), 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 \( \epsilon \) caused severe data perturbation, resulting in an accuracy that was near the chance level. Increasing the value of \( \epsilon \) 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 $\epsilon$. 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, $\epsilon$-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]. $\epsilon$-differential privacy also suffered from severe utility loss under a low $\epsilon$, 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 $\epsilon$ 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 $\epsilon$ was low (Figure 3). As $\epsilon$ 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 $\epsilon$-differential privacy, we may consider providing the $\epsilon$ 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 $\epsilon$ 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 $\epsilon$ 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 $\epsilon$ value, which determines the degree of perturbation, should be set to apply to the algorithm. In this study, a value of $\epsilon$ 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


22. Dwork C, Rothblum GN. Concentrated differential privacy. ArXiv Preprint posted online on March 6, 2016 [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
Accurate Prediction of Stroke for Hypertensive Patients Based on Medical Big Data and Machine Learning Algorithms: Retrospective Study

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.
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.
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.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Positive cases (N=8,827), n (%)</th>
<th>Negative cases (N=42,088), n (%)</th>
<th>Negative cases after sampling N=11,126, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender, male</td>
<td>5251 (59.49)</td>
<td>25990 (61.75)</td>
<td>6174 (55.49)</td>
</tr>
<tr>
<td>Age, years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30-40</td>
<td>414 (4.69)</td>
<td>5843 (13.88)</td>
<td>522 (4.69)</td>
</tr>
<tr>
<td>40-50</td>
<td>1746 (19.78)</td>
<td>17342 (41.20)</td>
<td>2204 (19.81)</td>
</tr>
<tr>
<td>50-60</td>
<td>2104 (23.84)</td>
<td>10415 (24.75)</td>
<td>2656 (23.87)</td>
</tr>
<tr>
<td>60-70</td>
<td>2462 (27.89)</td>
<td>5448 (12.94)</td>
<td>3108 (27.93)</td>
</tr>
<tr>
<td>70-85</td>
<td>2088 (23.65)</td>
<td>2636 (6.26)</td>
<td>2636 (23.69)</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Positive cases (N=8,827)</th>
<th>Negative cases (N=11,126)</th>
<th>P value^a</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender, n (%), male</td>
<td>5,251 (59.49)</td>
<td>6174 (55.49)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Age, mean (SD), years</td>
<td>60.21 (11.88)</td>
<td>59.73 (11.94)</td>
<td>.005</td>
</tr>
<tr>
<td>Years_after_hypertension, mean (SD), years</td>
<td>6.25 (5.64)</td>
<td>6.78 (5.27)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Lifestyle (current or previous), n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td>768 (8.70)</td>
<td>1233 (11.08)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Drink</td>
<td>1000 (11.33)</td>
<td>1643 (14.77)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Family history, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAM_hypertension</td>
<td>239 (2.71)</td>
<td>489 (4.40)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>FAM_diabetes</td>
<td>57 (0.65)</td>
<td>116 (1.04)</td>
<td>.002</td>
</tr>
<tr>
<td><strong>Physical examination, mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBP^b, mmHg</td>
<td>133.76 (13.42)</td>
<td>131.33 (10.02)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>DBP^c, mmHg</td>
<td>81.93 (9.56)</td>
<td>80.17 (7.45)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>PPD^d, mmHg</td>
<td>52.16 (10.59)</td>
<td>51.15 (8.81)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Trend characteristics, mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_followup_1year</td>
<td>4.13 (3.66)</td>
<td>5.89 (3.84)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>SBP_max, mmHg</td>
<td>140.29 (14.56)</td>
<td>142.77 (13.46)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>SBP_min, mmHg</td>
<td>127.17 (14.09)</td>
<td>122.81 (10.21)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>SBP_mean mmHg</td>
<td>133.20 (11.58)</td>
<td>131.75 (7.90)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>DBP_max, mmHg</td>
<td>86.47 (9.69)</td>
<td>89.10 (8.51)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>DBP_min, mmHg</td>
<td>76.67 (10.16)</td>
<td>73.42 (7.36)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>DBP_mean, mmHg</td>
<td>81.35 (8.17)</td>
<td>80.71 (5.80)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>PPD_max, mmHg</td>
<td>58.41 (12.02)</td>
<td>61.48 (10.83)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>PPD_min, mmHg</td>
<td>46.01 (11.05)</td>
<td>41.69 (8.28)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>PPD_mean, mmHg</td>
<td>51.89 (8.75)</td>
<td>51.04 (6.26)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>HR^e_max, times/min</td>
<td>78.57 (7.08)</td>
<td>79.57 (7.11)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>HR_min, times/min</td>
<td>74.29 (6.68)</td>
<td>72.97 (5.97)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>SBP_delta_mean, mmHg</td>
<td>4.37 (3.53)</td>
<td>4.01 (3.17)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>DBP_delta_mean, mmHg</td>
<td>3.46 (2.44)</td>
<td>3.24 (2.10)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>PPD_delta_mean, mmHg</td>
<td>4.30 (3.08)</td>
<td>4.04 (2.68)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>HR_delta_mean, times/min</td>
<td>1.23 (1.83)</td>
<td>1.08 (1.51)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Medical history, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior cardiovascular diseases</td>
<td>176 (1.99)</td>
<td>11 (0.1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>53 (0.6)</td>
<td>16 (0.14)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Atherosclerosis</td>
<td>488 (5.53)</td>
<td>358 (3.22)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>sleep disorder</td>
<td>99 (1.12)</td>
<td>475 (4.27)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Dizziness and headache</td>
<td>1094 (12.39)</td>
<td>1804 (16.21)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Malaise and fatigue</td>
<td>6 (0.07)</td>
<td>55 (0.49)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Giddiness</td>
<td>9 (0.10)</td>
<td>55 (0.49)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Migraine</td>
<td>7 (0.08)</td>
<td>38 (0.34)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Antihypertensive treatment</td>
<td>8551 (96.87)</td>
<td>10905 (98.01)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>AUC(^a)</th>
<th>Accuracy</th>
<th>Recall</th>
<th>F1-score</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.8544</td>
<td>0.7726</td>
<td>0.7141</td>
<td>0.7354</td>
<td>0.8191</td>
</tr>
<tr>
<td>SVM(^b)</td>
<td>0.8898</td>
<td>0.8112</td>
<td>0.7844</td>
<td>0.7861</td>
<td>0.8325</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.8956</td>
<td>0.8343</td>
<td>0.8157</td>
<td>0.8133</td>
<td>0.8490</td>
</tr>
<tr>
<td>XGBoost(^c)</td>
<td>0.9220</td>
<td>0.8478</td>
<td>0.8512</td>
<td>0.8319</td>
<td>0.8451</td>
</tr>
</tbody>
</table>

\(^a\)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.
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


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

©Yujie Yang, Jing Zheng, Zhenzhen Du, Ye Li, Yunpeng Cai. Originally published in JMIR Medical Informatics (https://medinform.jmir.org). 10.11.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Correction: Use of Deep Learning to Predict Acute Kidney Injury After Intravenous Contrast Media Administration: Prediction Model Development Study

Donghwan Yun1,2, MD; Semin Cho2, MD; Yong Chul Kim2, MD, PhD; Dong Ki Kim2, MD, PhD; Kook-Hwan Oh2, MD, PhD; Kwon Wook Joo2, MD, PhD; Yon Su Kim1,2, MD, PhD; Seung Seok Han1,2, MD, PhD

1Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, Republic of Korea
2Department 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
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
A Blockchain-Based Dynamic Consent Architecture to Support Clinical Genomic Data Sharing (ConsentChain): Proof-of-Concept Study

Faisal Albalwy1,2,3, BSc, MS; Andrew Brass1,3, BSc, PhD; Angela Davies3, BSc, PhD

1Department of Computer Science, University of Manchester, Manchester, United Kingdom
2Department of Computer Science, College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia
3Division 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
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.
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.

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 participants 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].

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.

**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.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype</td>
<td>GNE</td>
</tr>
<tr>
<td>Phenotype</td>
<td>PHE</td>
</tr>
<tr>
<td>Metadata</td>
<td>MEA</td>
</tr>
</tbody>
</table>

### Table 2. Code representing the role in consent element.

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>TRT</td>
</tr>
<tr>
<td>Research</td>
<td>REH</td>
</tr>
<tr>
<td>Clinical</td>
<td>CLL</td>
</tr>
</tbody>
</table>

### Table 3. Code representing the purpose in consent element.

<table>
<thead>
<tr>
<th>Role</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinician</td>
<td>CLN</td>
</tr>
<tr>
<td>Researcher</td>
<td>REE</td>
</tr>
<tr>
<td>Bioinformatician</td>
<td>BIN</td>
</tr>
</tbody>
</table>
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.

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 Hyperledger 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. Choudhury 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 Dwarma 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 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.

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. A portion of the patient’s smart contract code.
Figure 7. An illustrative example of patient smart contract code.

```solidity
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 ∧ patientWalletAddress ≠ 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 ∧ dataCreatorWalletAddress ≠ null then
Create newDataCreatorSmartContract
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 ∧ dataCreatorWalletAddress ≠ null ∧
dataRequesterPUK ≠ 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

Textbox 5. Pseudocode of registering new patient.

Algorithm 1: createNewPatientContracter
Input: caller, patientWalletAddress
Output: smartContractAddress
If caller=admin ∧ patientWalletAddress ≠ null then
Create newPatientSmartContract
Set newPatientSmartContract owner to patientWalletAddress
Output newPatientSmartContract 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

```plaintext
Algorithm 4: setConsent
Input: caller, dataType, role, purpose
Output: status
CONSENT←mapping
If caller=contractOwner∧dataType ≠ null ∧ role ≠ null ∧ purpose ≠ null, then
h←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.

```plaintext
Algorithm 5: cancelConsent
Input: caller, dataType, role, purpose
Output: status
CONSENT←mapping
If caller=contractOwner∧dataType ≠ null ∧ role ≠ null ∧ purpose ≠ null, then
h←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.

```plaintext
Algorithm 6: checkConsent
Input: dataType, role, purpose
Output: status
CONSENT←mapping
If dataType ≠ null ∧ role ≠ null ∧ purpose ≠ null, then
h←hash(dataType, role, purpose)
r←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 (e.g., 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 ∧ dataHash ≠ 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.**

<table>
<thead>
<tr>
<th>Algorithm 9: requestAccessToken</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> caller, dataProfileId, ticketId</td>
</tr>
<tr>
<td><strong>Output:</strong> tokenId</td>
</tr>
<tr>
<td>DATAPROFILE ← mapping</td>
</tr>
<tr>
<td>If caller=contractOwner ( \land ) dataProfileId ≠ null ( \land ) ticketId ≠ null, then</td>
</tr>
<tr>
<td>( d \leftarrow \text{DATAPROFILE}.\text{return(dataProfileId)} )</td>
</tr>
<tr>
<td>dataCreator ← ( d.\text{dataCreatorSmartContract} )</td>
</tr>
<tr>
<td>patient ← ( d.\text{patientSmartContract} )</td>
</tr>
<tr>
<td>if patient.ticket[ticketId].status=true then</td>
</tr>
<tr>
<td>(\text{token} \leftarrow \text{dataCreator.createAccessToken(caller, dataProfileId)} )</td>
</tr>
<tr>
<td>Token.status=true</td>
</tr>
<tr>
<td>Output token.id</td>
</tr>
<tr>
<td>Else</td>
</tr>
<tr>
<td>revert smart contract state and show an error message</td>
</tr>
<tr>
<td>Else</td>
</tr>
<tr>
<td>revert smart contract state and show an error message</td>
</tr>
</tbody>
</table>

**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.
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 blockchain 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**


45. Albalwy et al. JMIR MEDICAL INFORMATICS 2021 | vol. 9 | iss. 11 | e27816 | p.158 https://medinform.jmir.org/2021/11/e27816
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]


73. The ProvableTM blockchain oracle for modern DApps. Provable. URL: https:// provable.xyz/ [accessed 2021-01-14]


https://medinform.jmir.org/2021/11/e27816

JMIR Med Inform 2021 | vol. 9 | iss. 11 | e27816 | p.159

(page number not for citation purposes)


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

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

©Faisal Albalwy, Andrew Brass, Angela Davies. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 03.11.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

Jeongmin Kim¹*, BBA; Hakyung Lim¹, MSc; Jae-Hyeon Ahn¹, PhD; Kyoung Hwa Lee², MD, MMS; Kwang Suk Lee³*, MD, MMS; Kyo Chul Koo³*, MD, PhD

¹College of Business, Korea Advanced Institute of Science and Technology, Seoul, Republic of Korea
²Division of Infectious Disease, Department of Internal Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea
³Department 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.
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 \( n (\text{total deaths}) / n (\text{total patients}) \), 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 $\geq 6$.

### Table 1. Demographic characteristics.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Total cohort (N=5601)</th>
<th>Training cohort (N=5037)</th>
<th>Hold-out cohort (N=564)</th>
<th>$P$ value$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value, n (%) or mean (SD)</td>
<td>Missing data, %</td>
<td>Value, n (%) or mean (SD)</td>
<td>Missing data, %</td>
<td></td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-9</td>
<td>66 (1.2%)</td>
<td>61 (1.2%)</td>
<td>5 (0.9%)</td>
<td>.41</td>
</tr>
<tr>
<td>10-19</td>
<td>205 (3.7%)</td>
<td>185 (3.7%)</td>
<td>20 (3.6%)</td>
<td></td>
</tr>
<tr>
<td>20-29</td>
<td>1110 (19.8%)</td>
<td>988 (19.6%)</td>
<td>122 (21.6%)</td>
<td></td>
</tr>
<tr>
<td>30-39</td>
<td>564 (10.1%)</td>
<td>513 (10.2%)</td>
<td>51 (9.0%)</td>
<td></td>
</tr>
<tr>
<td>40-49</td>
<td>739 (13.2%)</td>
<td>652 (12.9%)</td>
<td>87 (15.4%)</td>
<td></td>
</tr>
<tr>
<td>50-59</td>
<td>1141 (20.4%)</td>
<td>1039 (20.6%)</td>
<td>102 (18.1%)</td>
<td></td>
</tr>
<tr>
<td>60-69</td>
<td>907 (16.2%)</td>
<td>809 (16.1%)</td>
<td>98 (17.4%)</td>
<td></td>
</tr>
<tr>
<td>70-79</td>
<td>545 (9.7%)</td>
<td>495 (9.8%)</td>
<td>50 (8.9%)</td>
<td></td>
</tr>
<tr>
<td>≥80</td>
<td>324 (5.8%)</td>
<td>295 (5.9%)</td>
<td>29 (5.1%)</td>
<td></td>
</tr>
<tr>
<td><strong>Sex (male)</strong></td>
<td></td>
<td></td>
<td></td>
<td>.73</td>
</tr>
<tr>
<td>0.9%</td>
<td>2310 (41.2%)</td>
<td>2073 (41.2%)</td>
<td>237 (42.0%)</td>
<td></td>
</tr>
<tr>
<td>21.4%</td>
<td>21.5%</td>
<td>20.9%</td>
<td>20.9%</td>
<td></td>
</tr>
<tr>
<td><strong>BMI (kg/m$^2$)</strong></td>
<td></td>
<td></td>
<td></td>
<td>.65</td>
</tr>
<tr>
<td>&lt;18.5</td>
<td>259 (4.6%)</td>
<td>236 (4.7%)</td>
<td>23 (4.1%)</td>
<td></td>
</tr>
<tr>
<td>18.5-22.9</td>
<td>1854 (33.1%)</td>
<td>1666 (33.1%)</td>
<td>188 (33.3%)</td>
<td></td>
</tr>
<tr>
<td>23.0-24.9</td>
<td>1035 (18.5%)</td>
<td>929 (18.4%)</td>
<td>106 (18.8%)</td>
<td></td>
</tr>
<tr>
<td>25.0-29.9</td>
<td>1045 (18.7%)</td>
<td>938 (18.6%)</td>
<td>107 (19.0%)</td>
<td></td>
</tr>
<tr>
<td>≥30</td>
<td>207 (3.7%)</td>
<td>185 (3.7%)</td>
<td>22 (3.9%)</td>
<td></td>
</tr>
<tr>
<td><strong>Medical history</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>688 (12.3%)</td>
<td>620 (12.3%)</td>
<td>68 (12.1%)</td>
<td>.92</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1198 (21.4%)</td>
<td>1087 (21.6%)</td>
<td>111 (19.7%)</td>
<td>.32</td>
</tr>
<tr>
<td>Heart failure</td>
<td>59 (1.1%)</td>
<td>52 (1.0%)</td>
<td>7 (1.2%)</td>
<td>.81</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>179 (3.2%)</td>
<td>156 (3.1%)</td>
<td>23 (4.1%)</td>
<td>.26</td>
</tr>
<tr>
<td>Asthma</td>
<td>128 (2.3%)</td>
<td>118 (2.3%)</td>
<td>10 (1.8%)</td>
<td>.48</td>
</tr>
<tr>
<td>Chronic obstructive pulmo-</td>
<td>40 (0.7%)</td>
<td>38 (0.8%)</td>
<td>2 (0.4%)</td>
<td>.43</td>
</tr>
<tr>
<td>Chronic kidney disease</td>
<td>55 (1.0%)</td>
<td>48 (1.0%)</td>
<td>7 (1.2%)</td>
<td>.67</td>
</tr>
<tr>
<td>Malignancy</td>
<td>145 (2.6%)</td>
<td>134 (2.7%)</td>
<td>11 (2.0%)</td>
<td>.39</td>
</tr>
<tr>
<td>Chronic liver disease</td>
<td>83 (1.6%)</td>
<td>75 (1.6%)</td>
<td>8 (1.5%)</td>
<td>6.7%</td>
</tr>
<tr>
<td>Autoimmune disease</td>
<td>38 (0.7%)</td>
<td>32 (0.7%)</td>
<td>6 (1.1%)</td>
<td>.37</td>
</tr>
<tr>
<td>Dementia</td>
<td>224 (4.2%)</td>
<td>203 (4.3%)</td>
<td>21 (3.7%)</td>
<td>.81</td>
</tr>
</tbody>
</table>

$^a$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>
<thead>
<tr>
<th>Variable</th>
<th>Total cohort (N=5601)</th>
<th>Training cohort (N=5037)</th>
<th>Hold-out cohort (N=564)</th>
<th>P value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Value, n (%) or mean (SD)</td>
<td>Missing data, %</td>
<td>Value, n (%) or mean (SD)</td>
<td>Missing data, %</td>
</tr>
<tr>
<td><strong>Systolic blood pressure (mmHg)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;120</td>
<td>1306 (23.3%)</td>
<td>2.5%</td>
<td>1177 (23.4%)</td>
<td>2.5%</td>
</tr>
<tr>
<td>120-129</td>
<td>1138 (20.3%)</td>
<td>2.5%</td>
<td>1012 (20.1%)</td>
<td>2.5%</td>
</tr>
<tr>
<td>130-139</td>
<td>1084 (19.4%)</td>
<td>2.5%</td>
<td>977 (19.4%)</td>
<td>2.5%</td>
</tr>
<tr>
<td>140-159</td>
<td>1418 (25.3%)</td>
<td>2.5%</td>
<td>1281 (25.4%)</td>
<td>2.5%</td>
</tr>
<tr>
<td>≥160</td>
<td>513 (9.2%)</td>
<td>2.5%</td>
<td>463 (9.2%)</td>
<td>2.5%</td>
</tr>
<tr>
<td><strong>Diastolic blood pressure (mmHg)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;80</td>
<td>2102 (37.5%)</td>
<td>2.5%</td>
<td>1878 (37.3%)</td>
<td>2.5%</td>
</tr>
<tr>
<td>80-89</td>
<td>1797 (32.1%)</td>
<td>2.5%</td>
<td>1601 (31.8%)</td>
<td>2.5%</td>
</tr>
<tr>
<td>90-99</td>
<td>1056 (18.9%)</td>
<td>2.5%</td>
<td>971 (19.3%)</td>
<td>2.5%</td>
</tr>
<tr>
<td>≥100</td>
<td>504 (9.0%)</td>
<td>2.5%</td>
<td>460 (9.1%)</td>
<td>2.5%</td>
</tr>
<tr>
<td><strong>Heart rate (bpm)</strong></td>
<td>85.8 (SD 15.1)</td>
<td>2.3%</td>
<td>85.8 (SD 15.0)</td>
<td>2.3%</td>
</tr>
<tr>
<td><strong>Body temperature (°C)</strong></td>
<td>36.9 (SD 0.6)</td>
<td>0.7%</td>
<td>36.9 (SD 0.6)</td>
<td>0.8%</td>
</tr>
<tr>
<td><strong>Symptoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fever</td>
<td>1302 (23.3%)</td>
<td>0.1%</td>
<td>1168 (23.2%)</td>
<td>0.1%</td>
</tr>
<tr>
<td>Cough</td>
<td>2331 (41.6%)</td>
<td>0.1%</td>
<td>2103 (41.8%)</td>
<td>0.1%</td>
</tr>
<tr>
<td>Sputum</td>
<td>1611 (28.8%)</td>
<td>0.1%</td>
<td>1460 (29.0%)</td>
<td>0.1%</td>
</tr>
<tr>
<td>Sore throat</td>
<td>872 (15.6%)</td>
<td>0.1%</td>
<td>779 (15.5%)</td>
<td>0.1%</td>
</tr>
<tr>
<td>Rhinorrhea</td>
<td>617 (11.0%)</td>
<td>0.1%</td>
<td>560 (11.1%)</td>
<td>0.1%</td>
</tr>
<tr>
<td>Myalgia</td>
<td>920 (16.4%)</td>
<td>0.1%</td>
<td>820 (16.3%)</td>
<td>0.1%</td>
</tr>
<tr>
<td>Fatigue</td>
<td>233 (4.2%)</td>
<td>0.1%</td>
<td>207 (4.1%)</td>
<td>0.1%</td>
</tr>
<tr>
<td>Shortness of breath</td>
<td>665 (11.9%)</td>
<td>0.1%</td>
<td>608 (12.1%)</td>
<td>0.1%</td>
</tr>
<tr>
<td>Headache</td>
<td>963 (17.2%)</td>
<td>0.1%</td>
<td>873 (17.3%)</td>
<td>0.1%</td>
</tr>
<tr>
<td>Altered consciousness</td>
<td>35 (0.6%)</td>
<td>0.1%</td>
<td>31 (0.6%)</td>
<td>0.1%</td>
</tr>
<tr>
<td>Vomiting</td>
<td>244 (4.4%)</td>
<td>0.1%</td>
<td>210 (4.2%)</td>
<td>0.1%</td>
</tr>
<tr>
<td>Diarrhea</td>
<td>516 (9.2%)</td>
<td>0.1%</td>
<td>457 (9.1%)</td>
<td>0.1%</td>
</tr>
<tr>
<td><strong>Laboratory values</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemoglobin (g/dL)</td>
<td>13.3 (SD 1.8)</td>
<td>27.2%</td>
<td>13.3 (SD 1.8)</td>
<td>26.7%</td>
</tr>
<tr>
<td>Hematocrit (%)</td>
<td>39.2 (SD 5.0)</td>
<td>27.2%</td>
<td>39.3 (SD 4.9)</td>
<td>26.7%</td>
</tr>
<tr>
<td>Lymphocyte proportion (%)</td>
<td>29.2 (SD 11.7)</td>
<td>27.6%</td>
<td>29.3 (SD 11.7)</td>
<td>27.1%</td>
</tr>
<tr>
<td>Platelet count (/μL)</td>
<td>236,697 (SD 82,897)</td>
<td>27.1%</td>
<td>236,776 (SD 82,534)</td>
<td>26.7%</td>
</tr>
<tr>
<td>White blood cell count (/μL)</td>
<td>6126 (SD 2824)</td>
<td>27.1%</td>
<td>6121 (SD 2841)</td>
<td>26.7%</td>
</tr>
<tr>
<td>WHO OSCI ≥6</td>
<td>271 (4.8%)</td>
<td>0.0%</td>
<td>242 (4.8%)</td>
<td>0.0%</td>
</tr>
<tr>
<td>Pregnancy</td>
<td>19 (0.3%)</td>
<td>0.4%</td>
<td>17 (0.3%)</td>
<td>0.3%</td>
</tr>
<tr>
<td>Pregnancy weeks</td>
<td>0.05 (SD 1.1)</td>
<td>0.4%</td>
<td>0.06 (SD 1.1)</td>
<td>0.4%</td>
</tr>
</tbody>
</table>

<sup>a</sup>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.
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 ($P=0.04$). 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 $P \leq .01$).

https://medinform.jmir.org/2021/11/e32726

JMIR Med Inform 2021 | vol. 9 | iss. 11 | e32726 | p.168
Table 3. Comparison of model performance.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of variables</th>
<th>AUROC(^a), value (95% CI)</th>
<th>Specificity, value (95% CI)</th>
<th>Sensitivity, value (95% CI)</th>
<th>Accuracy, value (95% CI)</th>
<th>PPV(^b), value (95% CI)</th>
<th>NPV(^c), value (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>37</td>
<td>0.965 (0.958-0.972)</td>
<td>0.900 (0.892-0.909)</td>
<td>0.905 (0.868-0.942)</td>
<td>0.900 (0.892-0.908)</td>
<td>0.314 (0.295-0.335)</td>
<td>0.995 (0.993-0.997)</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>0.946 (0.936-0.956)</td>
<td>0.900 (0.891-0.908)</td>
<td>0.839 (0.793-0.884)</td>
<td>0.897 (0.888-0.905)</td>
<td>0.297 (0.276-0.319)</td>
<td>0.991 (0.988-0.994)</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>0.963 (0.955-0.971)</td>
<td>0.900 (0.892-0.908)</td>
<td>0.884 (0.839-0.921)</td>
<td>0.899 (0.891-0.907)</td>
<td>0.309 (0.289-0.329)</td>
<td>0.994 (0.991-0.996)</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>0.942 (0.931-0.953)</td>
<td>0.901 (0.892-0.909)</td>
<td>0.810 (0.756-0.860)</td>
<td>0.896 (0.888-0.904)</td>
<td>0.291 (0.270-0.313)</td>
<td>0.989 (0.987-0.992)</td>
</tr>
</tbody>
</table>

\(^a\)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) with 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.

<table>
<thead>
<tr>
<th>Influx</th>
<th>Optimal threshold</th>
<th>Optimized mortality rate</th>
<th>Decreased mortality rate&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>H&lt;sub&gt;p1&lt;/sub&gt;</td>
<td>0.10</td>
<td>0.022</td>
<td>0.298</td>
</tr>
<tr>
<td>H2</td>
<td>0.01</td>
<td>0.015</td>
<td>0.047</td>
</tr>
<tr>
<td>H3</td>
<td>0.04</td>
<td>0.019</td>
<td>0.146</td>
</tr>
<tr>
<td>H4</td>
<td>0.24</td>
<td>0.031</td>
<td>0.209</td>
</tr>
<tr>
<td>R&lt;sub&gt;0&lt;/sub&gt;=1.5</td>
<td>0.02</td>
<td>0.017</td>
<td>0.061</td>
</tr>
<tr>
<td>R&lt;sub&gt;0&lt;/sub&gt;=2</td>
<td>0.16</td>
<td>0.025</td>
<td>0.179</td>
</tr>
<tr>
<td>R&lt;sub&gt;0&lt;/sub&gt;=4</td>
<td>0.39</td>
<td>0.032</td>
<td>0.181</td>
</tr>
<tr>
<td>R&lt;sub&gt;0&lt;/sub&gt;=6</td>
<td>0.43</td>
<td>0.041</td>
<td>0.068</td>
</tr>
<tr>
<td>R&lt;sub&gt;0&lt;/sub&gt;=8</td>
<td>0.62</td>
<td>0.042</td>
<td>0.071</td>
</tr>
<tr>
<td>R&lt;sub&gt;0&lt;/sub&gt;=10</td>
<td>0.66</td>
<td>0.042</td>
<td>0.069</td>
</tr>
</tbody>
</table>

<sup>a</sup>Decreased mortality rate: (J-index mortality rate – optimized mortality rate) / J-index mortality rate.

<sup>b</sup>H: historical epidemic patient influx scenario.

<sup>c</sup>R<sub>0</sub>: 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<sub>death</sub> (50.7%), was maximized in proportion to total death. For example, when R<sub>0</sub> 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. R<sub>0</sub>: 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 available online. [21].

https://medinform.jmir.org/2021/11/e32726
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 \( \text{P} \) 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

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
Stroke Outcome Measurements From Electronic Medical Records: Cross-sectional Study on the Effectiveness of Neural and Nonneural Classifiers

Bruna Stella Zanotto1,2*, MSc, PharmD; Ana Paula Beck da Silva Etges1,3*, Eng, MSc, PhD; Avner dal Bosco3*, MSc; Eduardo Gabriel Cortes4*, MSc; Renata Ruschel1*, PT; Ana Claudia De Souza5*, MD, PhD; Claudio M V Andrade6*, MSc; Felipe Viegas6*, MSc; Sergio Canuto6*, MSc, PhD; Washington Luiz6*, MSc; Sheila Ouriques Martins5*, MSc, MD, PhD; Renata Vieira7*, MSc, PhD; Carisi Polanczyk1,2*, MSc, MD, PhD; Marcos André Gonçalves6*, MSc, PhD

1National Institute of Health Technology Assessment - INCT/IATS (CNPQ 465518/2014-1), Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
2Graduate Program in Epidemiology, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
3School of Technology, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil
4Graduate Program of Computer Science, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
5Brazilian Stroke Network, Hospital Moinhos de Vento, Porto Alegre, Brazil
6Computer Science Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
7Centro 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

https://medinform.jmir.org/2021/11/e29120
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 (intragtask 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 approved by the hospital ethics committee (CAAE: 29694720000005330).
between them measured by $\kappa$, 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+Char-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:

$$\text{Macro-F1} = \frac{1}{\text{num classes}} \sum_{i=1}^{\text{num classes}} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. Precision (positive predictive value) $= \text{TP} / \text{TP} + \text{FP}$ is 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.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Number of classes</th>
<th>Supporting information for classes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Health care status achieved (tier 1)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rankin</td>
<td>8</td>
<td>• 0-6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>National Institutes of Health Stroke Scale</td>
<td>42</td>
<td>• 1-41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Death</td>
<td>3</td>
<td>• Absence of vital signs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Vital signs present</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td><strong>Process of recovery (tier 2)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobility level</td>
<td>16</td>
<td>• 1-15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Self-care</td>
<td>3</td>
<td>• Able</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Unable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Pain</td>
<td>4</td>
<td>• No pain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Low to intermediate pain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Intense pain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Strength</td>
<td>7</td>
<td>• 0-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Paresis</td>
<td>3</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Ability to feed orally</td>
<td>3</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Ability to communicate</td>
<td>4</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Poorly or symptomatic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Ability of understanding</td>
<td>4</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Poorly or symptomatic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Ability to ambulate</td>
<td>4</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Poorly or symptomatic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td><strong>Treatment or care related</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thrombolytic therapy</td>
<td>3</td>
<td>• No delta</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Thrombectomy</td>
<td>3</td>
<td>• No delta</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Tasks</td>
<td>Number of classes</td>
<td>Supporting information for classes</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------------------</td>
<td>-----------------------------------------------------</td>
</tr>
</tbody>
</table>
| Location                       | 4                 | • Emergency room  
• ICU  
• 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**
<table>
<thead>
<tr>
<th>Tasks</th>
<th>Number of classes</th>
<th>Supporting information for classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>High blood pressure</td>
<td>3</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Smoking status</td>
<td>4</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Former</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>3</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>3</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Diabetes</td>
<td>3</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Prior stroke</td>
<td>3</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Cancer</td>
<td>3</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Alcoholism</td>
<td>4</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Former</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Obesity</td>
<td>3</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
<tr>
<td>Dyslipidemia</td>
<td>3</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NI</td>
</tr>
</tbody>
</table>

aNI: noninformative.
bICU: 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.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Patients with ischemic stroke evaluated (n=188)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Values, median (range)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>79 (68-87)</td>
</tr>
<tr>
<td>LOS&lt;sup&gt;b&lt;/sup&gt; (days)</td>
<td>6 (4-12)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>N/A</td>
</tr>
<tr>
<td>Male</td>
<td>N/A</td>
</tr>
<tr>
<td>Comorbidities</td>
<td></td>
</tr>
<tr>
<td>Previous stroke</td>
<td>N/A</td>
</tr>
<tr>
<td>Previous coronary artery disease</td>
<td>N/A</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>N/A</td>
</tr>
<tr>
<td>Diabetes</td>
<td>N/A</td>
</tr>
<tr>
<td>Hypertension</td>
<td>N/A</td>
</tr>
<tr>
<td>Smoking status</td>
<td>N/A</td>
</tr>
<tr>
<td>Alcoholism</td>
<td>N/A</td>
</tr>
<tr>
<td>Treatment and care related</td>
<td></td>
</tr>
<tr>
<td>Antithrombotic therapy</td>
<td>N/A</td>
</tr>
<tr>
<td>Thrombolysis with rtPA&lt;sup&gt;c&lt;/sup&gt;</td>
<td>N/A</td>
</tr>
<tr>
<td>Thrombectomy</td>
<td>N/A</td>
</tr>
<tr>
<td>Thrombolysis and thrombectomy</td>
<td>N/A</td>
</tr>
<tr>
<td>NIHSS&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;8</td>
<td>N/A</td>
</tr>
<tr>
<td>&gt;8 and &lt;15</td>
<td>N/A</td>
</tr>
<tr>
<td>&gt;15</td>
<td>N/A</td>
</tr>
</tbody>
</table>

<sup>a</sup>N/A: not applicable.
<sup>b</sup>LOS: length of stay.
<sup>c</sup>rtPA: alteplase.
<sup>d</sup>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.
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 \cite{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.

<table>
<thead>
<tr>
<th>Tier</th>
<th>Task</th>
<th>Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline characteristics</td>
<td>Dyslipidemia</td>
<td>96.3</td>
</tr>
<tr>
<td>Baseline characteristics</td>
<td>Diabetes</td>
<td>93.8</td>
</tr>
<tr>
<td>Treatment and care related</td>
<td>Pressure ulcer risk</td>
<td>92.2</td>
</tr>
<tr>
<td>Baseline characteristics</td>
<td>Obesity</td>
<td>91.3</td>
</tr>
<tr>
<td>Treatment and care related</td>
<td>Location</td>
<td>88.4</td>
</tr>
<tr>
<td>Healthcare status achieved (Tier 1)</td>
<td>Rankin (mRS)</td>
<td>87.1</td>
</tr>
<tr>
<td>Treatment and care related</td>
<td>Thrombolytic therapy</td>
<td>87.0</td>
</tr>
<tr>
<td>Baseline characteristics</td>
<td>Atrial fibrillation</td>
<td>84.6</td>
</tr>
<tr>
<td>Baseline characteristics</td>
<td>Alcoholism</td>
<td>83.4</td>
</tr>
<tr>
<td>Baseline characteristics</td>
<td>High blood pressure</td>
<td>81.7</td>
</tr>
<tr>
<td>Baseline characteristics</td>
<td>Coronary artery disease</td>
<td>81.7</td>
</tr>
<tr>
<td>Treatment and care related</td>
<td>Thrombectomy</td>
<td>68.7</td>
</tr>
<tr>
<td>Process of recovery (Tier 2)</td>
<td>Ability to ambulate</td>
<td>68.4</td>
</tr>
<tr>
<td>Process of recovery (Tier 2)</td>
<td>Ability to communicate</td>
<td>67.2</td>
</tr>
<tr>
<td>Process of recovery (Tier 2)</td>
<td>Paresis</td>
<td>64.1</td>
</tr>
<tr>
<td>Healthcare status achieved (Tier 1)</td>
<td>NIHSS</td>
<td>61.2</td>
</tr>
<tr>
<td>Process of recovery (Tier 2)</td>
<td>Ability to feed orally</td>
<td>60.9</td>
</tr>
<tr>
<td>Baseline characteristics</td>
<td>Smoking diabets</td>
<td>60.4</td>
</tr>
<tr>
<td>Treatment and care related</td>
<td>Fall Risk</td>
<td>52.9</td>
</tr>
<tr>
<td>Process of recovery (Tier 2)</td>
<td>Mobility level</td>
<td>38.1</td>
</tr>
<tr>
<td>Treatment and care related</td>
<td>Infection indication</td>
<td>34.9</td>
</tr>
<tr>
<td>Baseline characteristics</td>
<td>Prior stroke</td>
<td>16.9</td>
</tr>
<tr>
<td>Process of recovery (Tier 2)</td>
<td>Pain</td>
<td>13.2</td>
</tr>
<tr>
<td>Healthcare status achieved (Tier 1)</td>
<td>Death</td>
<td>0.6</td>
</tr>
</tbody>
</table>

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.

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>
<thead>
<tr>
<th>Class</th>
<th>Undersampling</th>
<th>Original sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Infection indicative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>−1</td>
<td>1</td>
<td>0.96</td>
</tr>
<tr>
<td>0</td>
<td>0.39</td>
<td>0.89</td>
</tr>
<tr>
<td>1</td>
<td>0.28</td>
<td>0.82</td>
</tr>
<tr>
<td>Thrombolytic therapy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>−1</td>
<td>1</td>
<td>0.98</td>
</tr>
<tr>
<td>0</td>
<td>0.32</td>
<td>0.62</td>
</tr>
<tr>
<td>1</td>
<td>0.31</td>
<td>0.91</td>
</tr>
<tr>
<td>Ability to communicate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>−1</td>
<td>1</td>
<td>0.96</td>
</tr>
<tr>
<td>0</td>
<td>0.34</td>
<td>0.63</td>
</tr>
<tr>
<td>1</td>
<td>0.35</td>
<td>0.81</td>
</tr>
<tr>
<td>2</td>
<td>0.32</td>
<td>0.93</td>
</tr>
</tbody>
</table>

<sup>a</sup>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.
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 (fémur)
• 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
• subocclusiva (subocclusive)
• perg
• mecania (mechanic 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)
- óbito (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)
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 textual context 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 (e.g., 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.
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.

References
5. Zanotto et al. JMIR MEDICAL INFORMATICS 2021 | vol. 9 | iss. 11 | e29120 | p.196https://medinform.jmir.org/2021/11/e29120


https://jmir.org/2021/11/e29120


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]


77. Healthcare Improvement - Patient-Reported Outcomes. ICHOM. URL: https://www.ichom.org/ [accessed 2021-09-07]


86. Zanotto et al. JMIR MEDICAL INFORMATICS 2021 | vol. 9 | iss. 11 | e29120 | p.199https://medinform.jmir.org/2021/11/e29120


**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


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