JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 9 (2021), Issue 3 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Reviews

Original Papers

Review

Using Machine Learning Technologies in Pressure Injury Management: Systematic Review

Mengyao Jiang[1*], MSN; Yuxia Ma[1*], MSN; Siyi Guo[2], BSc; Liuqi Jin[2], BSc; Lin Lv[3], MSN; Lin Han[1,4], PhD; Ning An[2], PhD

[1]Evidence-based Nursing Center, School of Nursing, Lanzhou University, Lanzhou, China

[2]Key Laboratory of Knowledge Engineering with Big Data of the Ministry of Education, School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China

[3]Wound and Ostomy Center, Outpatient Department, Gansu Provincial Hospital, Lanzhou, China

[4]Department of Nursing, Gansu Provincial Hospital, Lanzhou, China

[*]these authors contributed equally

Corresponding Author:
Lin Han, PhD
Department of Nursing
Gansu Provincial Hospital
No 160, Donggang West Road, Chengguan District
Lanzhou, 730000
China
Phone: 86 0931 8281971
Email: LZU-hanlin@hotmail.com

Abstract

Background: Pressure injury (PI) is a common and preventable problem, yet it is a challenge for at least two reasons. First, the nurse shortage is a worldwide phenomenon. Second, the majority of nurses have insufficient PI-related knowledge. Machine learning (ML) technologies can contribute to lessening the burden on medical staff by improving the prognosis and diagnostic accuracy of PI. To the best of our knowledge, there is no existing systematic review that evaluates how the current ML technologies are being used in PI management.

Objective: The objective of this review was to synthesize and evaluate the literature regarding the use of ML technologies in PI management, and identify their strengths and weaknesses, as well as to identify improvement opportunities for future research and practice.

Methods: We conducted an extensive search on PubMed, EMBASE, Web of Science, Cumulative Index to Nursing and Allied Health Literature (CINAHL), Cochrane Library, China National Knowledge Infrastructure (CNKI), the Wanfang database, the VIP database, and the China Biomedical Literature Database (CBM) to identify relevant articles. Searches were performed in June 2020. Two independent investigators conducted study selection, data extraction, and quality appraisal. Risk of bias was assessed using the Prediction model Risk Of Bias ASsessment Tool (PROBAST).

Results: A total of 32 articles met the inclusion criteria. Twelve of those articles (38%) reported using ML technologies to develop predictive models to identify risk factors, 11 (34%) reported using them in posture detection and recognition, and 9 (28%) reported using them in image analysis for tissue classification and measurement of PI wounds. These articles presented various algorithms and measured outcomes. The overall risk of bias was judged as high.

Conclusions: There is an array of emerging ML technologies being used in PI management, and their results in the laboratory show great promise. Future research should apply these technologies on a large scale with clinical data to further verify and improve their effectiveness, as well as to improve the methodological quality.

(JMIR Med Inform 2021;9(3):e25704) doi:10.2196/25704

KEYWORDS

pressure injuries; pressure ulcer; pressure sore; pressure damage; decubitus ulcer; decubitus sore; bedsore; artificial intelligence; machine learning; neural network; support vector machine; natural language processing; Naive Bayes; bayesian learning; support

vector; random forest; boosting; deep learning; machine intelligence; computational intelligence; computer reasoning; management; systematic review

Introduction

Pressure injury (PI) is a significant indicator of the quality of care and a substantial burden on the public health system and the economy [1,2]. PI is a common but potentially preventable problem; however, current PI management is far from satisfactory. PI incidence and prevalence in the intensive care unit (ICU) were reported to be 10.0% to 25.9% and 16.9% to 23.8%, respectively [3]. The prevalence of PI in acute care settings ranged from 6% to 18.5% [4] and the hospital-acquired PI prevalence was 8.5% [5]. As for long-term care facilities, the PI prevalence was 27% in Italy [6] and 9.6% in Japan [7]. The overall prevalence of PI in the United States decreased from 13.5% in 2006 to 9.3% in 2015 [8]. Also, 95% of PIs are avoidable [9]. Nurses are primarily responsible for preventing PIs [10]. Several surveys have revealed that the majority of nurses, internationally, have insufficient knowledge of PI [11-14]. Besides, the global nursing shortage is a well-known fact [15]. Also, the most universally used PI risk assessment tool—the Braden scale—is subjective and inaccurate [16]. In a nutshell, medical practitioners need better PI management tools.

Artificial intelligence (AI) has been exerting a positive impact on daily living [17]. Moreover, machine learning (ML) is a way to achieve AI. Over the past two decades, ML has progressed from a laboratory curiosity to practical tools commonly applied in the medical field [18,19]. ML will continue to contribute to improving prognosis and diagnostic accuracies, even potentially taking on some of the work of medical practitioners' [20,21]. While researchers have developed various novel methods for PI management [22], there is no systematic review to our knowledge that evaluates current ML technologies used in PI management.

The objective of this paper was to synthesize and evaluate the nascent literature on the use of ML technologies in PI management, noting the strengths and weaknesses of the studies, and identify improvement opportunities for future research and practice.

Methods

Protocol

This review is reported according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [23].

Search Strategy

We conducted a systematic search of nine health science databases: PubMed, EMBASE, Web of Science, Cumulative Index to Nursing and Allied Health Literature (CINAHL), Cochrane Library, China National Knowledge Infrastructure (CNKI), the Wanfang database, the VIP database, and the China Biomedical Literature Database (CBM). We used Medical Subject Headings (MeSH) terms, Emtree terms, subject headings, and free text associated with the concepts of ML and PI. Searches were performed in June 2020. We also undertook a manual search of the reference list of all potentially eligible studies. Textbox 1 shows the search strategy that was used.

Textbox 1. Search strategy and search terms used.

- #1 pressure ulcer* OR pressure injur* OR pressure sore* OR pressure damage OR decubitus ulcer* OR decubitus sore* OR bedsore* OR bed sore*

AND

- #2 artificial intelligence OR machine learning OR neural network* OR support vector machine OR natural language processing OR Naive Bayes OR bayesian learning OR support vector* OR random forest* OR boosting OR deep learning OR machine intelligence OR computational intelligence OR computer reasoning

Inclusion and Exclusion Criteria

This review included studies that met the following criteria: (1) used a method related to ML technologies (including support vector machine, k-nearest neighbor [KNN], decision tree [DT], convolutional neural network, Bayesian network model, and logistic regression) in PI management, and (2) was published in English or Chinese. We excluded studies that met any of the following criteria: (1) review papers, opinion papers, editorials, discussion papers, dissertations, or conference abstracts; (2) papers on PI education; (3) papers about PI in animals; (4) papers lacking an outcome; and (5) papers without explicit algorithms.

Study Selection Methods

Two independent investigators screened titles and abstracts using the eligibility criteria. They then obtained full-text versions of all potential articles and scrutinized the full texts independently. Any discrepancies about study inclusion were resolved through discussion or by referral to a third investigator.

Data Extraction

Data were extracted from all identified studies using a predefined format. Variables included the first author, year of publication, country, aim, subject, algorithm used, study outcomes, performance of the algorithm, and findings. One investigator extracted the information into a standard data extraction sheet and a second investigator cross-checked the entries. Any disagreements were resolved via discussion.

Quality Appraisal

The methodological quality of the included studies was assessed independently by two investigators using the Prediction model Risk Of Bias ASsessment Tool (PROBAST) [24].

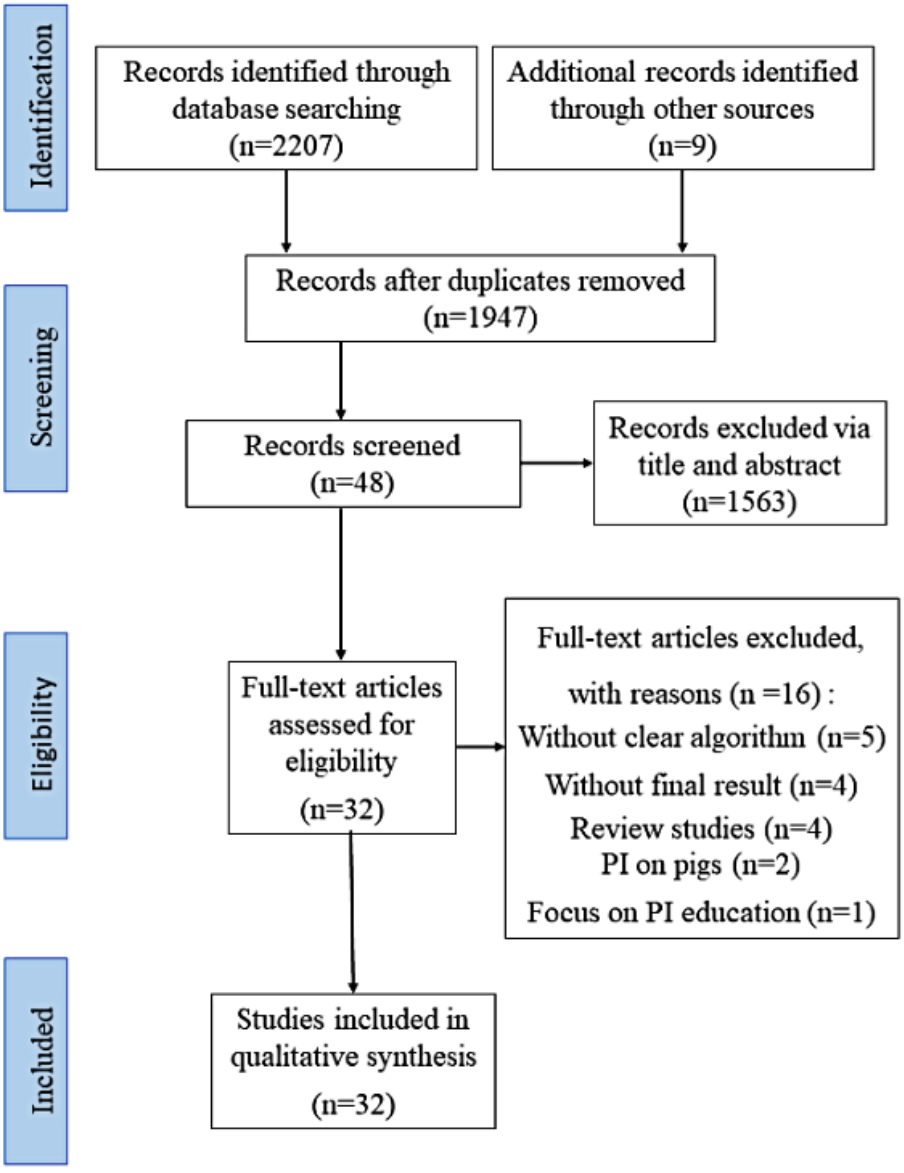
Disagreements were resolved by discussion. The PROBAST was designed to assess the risk of bias and applicability of diagnostic and prognostic prediction model studies, and it includes 20 signaling questions to judge the risk of bias from four domains (participants, predictors, outcome, and analysis). The risk of bias is judged as low, high, or unclear. If one domain is found to have a high risk of bias, the overall risk of bias is judged as high. Similarly, if one domain is assessed as unclear, the overall risk of bias is judged as unclear even if all other domains are assessed to have a low risk of bias.

Results

Study Process

Our initial search retrieved 2207 published articles, of which 269 were duplicates. After screening titles and abstracts, the full texts of 48 articles were obtained and assessed for potential eligibility. Of those 48 articles, 16 did not fulfill the inclusion criteria. The reasons for studies being ineligible were as follows: (1) lacking a clear algorithm (n=5); (2) lacking a result (n=4); (3) review studies (n=4); (4) studies in pigs (n=2); and (5) study on PI education (n=1). Finally, a total of 32 studies were eligible for our research (see Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the inclusion process. PI: pressure injury.



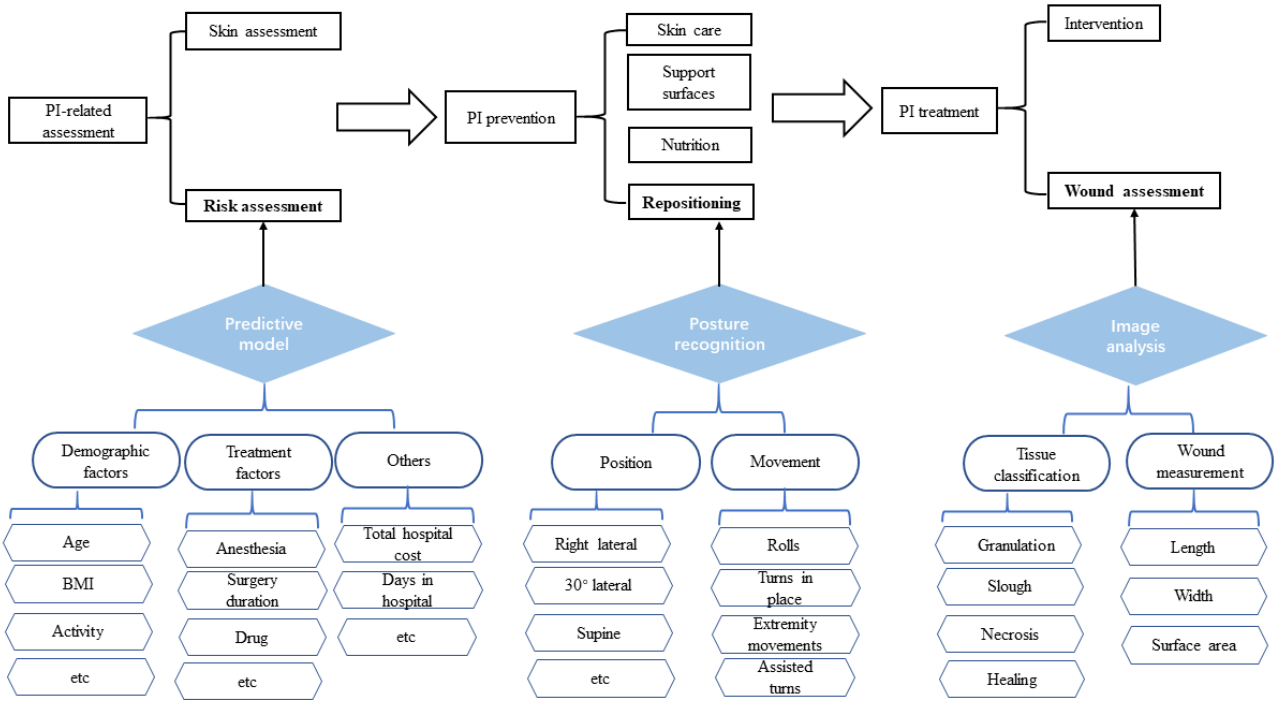
Characteristics of Included Studies

The articles that were included in our analysis were published between 2007 and 2020 and were undertaken in the United States [25-35], China [36-44], Spain [45-50], Japan [51,52], Italy [53,54], Korea [55], and Greece [56]. According to the applied area of the included studies, we divided the articles into three components: predictive model (12 studies), posture recognition (11 studies), and image analysis (9 studies). The characteristics of the included studies are presented in Multimedia Appendix 1.

Figure 2 shows the roles of the three components in the PI management process:

- Predictive model: when a patient is admitted into the hospital, a nurse needs to perform PI-related assessments—skin assessment and risk assessment. The predictive model is used to identify related risk factors.
- Posture recognition: when a patient is determined to be at risk, according to PI guidelines, proper measures such as repositioning, nutrition, support surfaces, and skin care need to be taken to prevent PI. The posture recognition can be used in the repositioning to help nurses to detect and classify the patient's position and movement.
- Image analysis: when a PI occurs, it is necessary to do wound assessment prior to treating the wounds. The image analysis can help to classify the wound tissue and measure the wound size.

Figure 2. The roles of machine learning technologies used in pressure injury (PI) management.



The performance indicators of ML algorithms include sensitivity, specificity, precision, accuracy, F score, positive predictive value, negative predictive value, geometric mean, false-positive rate, run time, and so on. Multimedia Appendix 2 shows the detailed results of the included studies.

Predictive Model

Twelve studies explored PI risk factors by data mining from the electronic health records (EHRs) of patients. The patients included in the studies were from a variety of settings: ICU (3 studies); operating room (2 studies); long-term care facilities (1 study); acute care hospital (1 study); orthopedic department (1 study); oncology department (1 study); end-of-life care (1 study); medical-surgical, critical care, and step-down units (1 study); and with mobility-related disabilities (1 study). The number of EHRs ranged from 147 to 125,213. The identified risk factors were different due to diverse input variables. In the majority of included studies, the PI percentage (the number of patients with PI/the number of total patients) of the data sets analyzed was imbalanced, and the minimum was 0.6% (51/8286). The accuracy ranged from 63.0% to 90.0%, the sensitivity ranged from 47.8% to 84.8%, and the specificity ranged from 70.3% to 94.7%. The DT algorithm was a typical data mining approach.

Posture Recognition

Eleven studies were concerned with posture identification by analyzing the pressure distribution of the body to achieve a robust assessment. Regarding the subjects of posture recognition, one study focused on wheelchair users [38], while the others looked at bed bound patients. The number of sensors was between 4 and 8192, and the number of subjects ranged from 2 to 58. Of the 11 studies, 10 studies detected and classified different postures or movements of a person and one study classified the bed inclination [31]. The common postures detected were supine, right lateral, and left lateral.

All articles reported on accuracy, which ranged from 49.1% to 100%. The difference in run times among different algorithms was quite large, from 0.04 seconds to 320.34 seconds. No articles reported on specificity. The sensitivity ranged from 62.0% to 100%, and the precision ranged from 65.0% to 100%. All eight studies applied the KNN algorithm in the processing of pressure sensor data.

Image Analysis

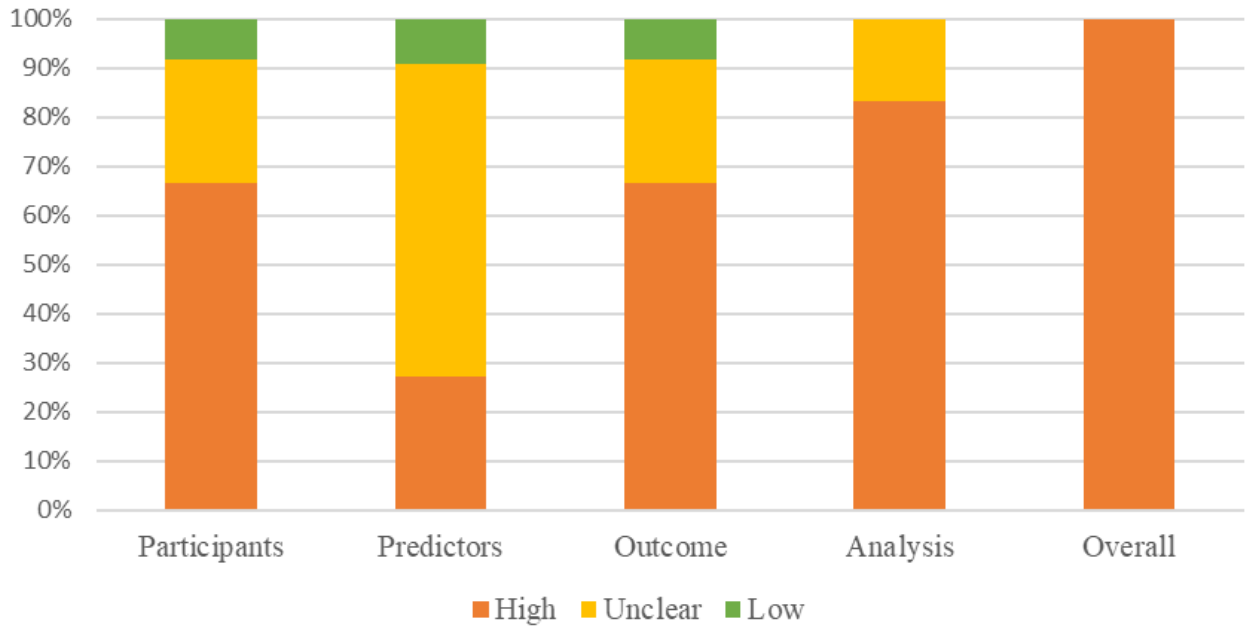
Nine studies conducted PI wounds' tissue segmentation and measurement using ML algorithms. We included studies that only analyzed PI images and excluded those involving the wound images of diabetes foot ulcers or venous leg ulcers. The number of digital images ranged from 14 to 193. Three articles were written by Veredas et al [46,48,49] using the same 113

color images to achieve tissue classification. Because different algorithms were used, we considered these three articles as independent research. Furthermore, the number of tissue segmentations ranged from 3 to 6. The most common PI wound tissue classifications were granulation, slough, and necrosis. One study developed an image processing algorithm that automatically measured the PI size [30]. The accuracy ranged from 78.3% to 92.0%, the sensitivity ranged from 61.7% to 99.9%, and the specificity ranged from 93.9% to 99.8%. Convolutional neural network algorithms, as deep learning architectures, were often used in medical image analysis in recent years.

Risk of Bias

The PROBAST was used to assess the risk of bias of the predictive model studies from four domains (participants, predictors, outcome, and analysis). However, the PROBAST was not suitable for the posture recognition and image analysis studies; to the best of our knowledge, there is still no appropriate tool to assess these engineering articles. The overall risk of bias of all of the predictive model studies was judged as high, and there was no low risk in the analysis domain (Figure 3).

Figure 3. Risk of bias assessment for the predictive model studies.



Discussion

Principal Findings

Our systematic review provided a broad overview of the ML technologies applied to PI management. After study selection, we were able to categorize these technologies into three components: predictive model, posture recognition, and image analysis. We discuss these different components in detail below.

Component 1: Predictive Model

The predictive model studies were all retrospective studies that analyzed the EHRs of patients to develop a prediction model via data mining techniques. The objective of the predictive model was to (1) identify the PI risk factors so that nurses could take customized preventive measures to arrest the PI progression, or (2) compare different algorithm performances and interpretability in constructing a predictive model. Even though the data sets were often imbalanced, Setoguchi et al [51] suggested that an alternating DT algorithm could effectively analyze highly imbalanced data. Shi et al [57] identified 22 empirically derived predictive models for PI risk using traditional statistical techniques. Compared with the previous predictive models, these advanced models can use the information available in EHRs rather than require investigators to input information into a questionnaire, and they can handle a large volume of various data at a faster velocity. Relative to the 2019 international guideline [1], we found a gap between the ML models and the empirical models. The risk factors mentioned in the guideline are mainly patient characteristics (eg, older age, spinal cord injuries, diabetes, incontinence, impaired sensory perception, etc) and treatment plan (eg, duration of surgery, anesthesia, use of vasopressors, etc). By employing ML models using data from patients' EHRs, Moon and Lee [55] found that the total hospital cost was associated with PIs, which had not been revealed by the guideline. However, it must be noted that these ML-based predictive models were lacking external validation. The results we got from one database had not been validated in temporal or spatial difference. Clearly, providing external validation for these models should be a focus of future research.

Component 2: Posture Recognition

PIs (also called bedsores) are common among bedridden older patients. However, the subjects in the included research studies were all healthy adults of different weights rather than patients at high risk for PIs. The research to test the ML technologies' performance was all conducted in the laboratory. In other words, these technologies are still in the development phase and have not transitioned from bench to bedside. The current research

focused simply on posture detection, and the majority of repositioning recommendations from the 2019 international guideline were based on expert opinion. Future research should combine posture recognition with the predictive model to develop the most effective repositioning schedules. For example, it is generally acknowledged that patients should be repositioned or mobilized every 2 hours. For a high-risk patient, it may be better to reposition every hour, while a low-risk patient may need to be repositioned every 3 hours. When it is time to change the patient's position, the related alarm will alert the nurse to help the patient to reposition, thus lightening the clinical nurse's workload.

Component 3: Image Analysis

It is worth mentioning that 6 of 9 (67%) studies were conducted in Spain. All three articles of Veredas et al (45,47,48) analyzed 113 digital images of PI of patients with home-care assistance, and we can assume that these were the same subjects; however, it is quite interesting to note that the images in the article published in 2010 were taken with a Canon digital camera, while the images in the 2015 article were taken with a Sony digital camera. In the real world, PI wounds are always irregular in shape, and it is inaccurate and unreliable to measure the size of the PI wound by multiplying length and width [58]. The computer-aided measurement system can offer an objective and efficient result. Using a photo of the PI wound, it is convenient and possible to analyze the characteristics of the lesion by the size and color of the ulcer, which helps clinicians monitor the developing and healing process of PI. Note that these subjects of image analysis are visible wounds, which are always stage IV—the severest PIs. Certainly, we do not want to see the most terrible situation happen, and thus future research is needed to optimize technologies so that we can assess PIs in their early stage via microclimate (eg, moisture, temperature, etc), not just via images. The current research is focused on classifying the wound tissue, and it is necessary to combine the percentage of the different tissue with the grading of PI to define the severity of PI. It is better to rely on objective indicators than to rely on human experience.

Future Research

PI management should be a holistic process, but the current research in these three components is separate. We'll use the case of a patient admitted to hospital to illustrate. First, according to the predictive model, we rated the patient as low risk. The repositioning schedule was implemented as the low risk required. Unfortunately, the patient developed PI, so we needed to assess the PI wound. The ML technologies on the predictive model and posture recognition need feedback from the PI wound image analysis to improve their performance. However, the research in these three components was conducted in different populations in different locations at different times. This point should be explored in future research.

The results on the risk of bias, surprisingly, were far from satisfactory. Similar to the research of Nagendran et al [59], the analysis domain was the major deficiency. More attention needs to be paid to the methodological quality of predictive model studies. The participants in posture recognition studies were healthy volunteers and the subjects in image analysis studies were images, so we could not judge these types of articles as medical research. There is a growing literature on interdisciplinary research such as in the fields of engineering and medicine. It is essential to develop a tool to assess the methodological quality of the relevant articles.

In summary, ML technologies furnish new alternatives to PI management. Given the global shortage of professional nurses and PI-related knowledge deficit, ML technologies will significantly reduce the burden on frontline clinicians and help to improve the quality of care, as Obermeyer and Emanuel [20] pointed out in 2016. However, because the current technologies only cover three components of PI management, there is a marked lack of novel technologies to assess potentially healthy skin, to achieve better skin care, to manage nutrition status, and to create intelligent support surfaces. Besides, IBM has discovered that its powerful technology is no match for the messy reality of today's health care system [60]. There is still a long way to go to integrate ML technologies into clinical care practices.

It is important to acknowledge some limitations. First, we only include articles published in English and Chinese. It will be better to include other language research for representing the current evidence. Second, due to the various aims and outcomes of the included studies, the quantitative synthesis has not been performed to obtain a direct result. Third, the aim of our review was to survey the current status of ML algorithms applied in PI management, so the eligibility criteria were defined broadly. After study selection, we found the related research can be divided into three components. We have no specific criteria for one component. Hence, under the guidance of our findings, future research can define detailed eligibility criteria.

Conclusions

The study results from various laboratory settings show an array of ML technologies with potential uses in PI management. Future research should apply these technologies on a large scale with clinical data to verify their effectiveness, enhance their performance, and improve methodological quality.

Acknowledgments

This work was partially supported by the National Nature Science Foundation of China (grants 71363004, 71663002, and 71704071), National Research Training Program of Gansu provincial hospital (19SYPYA-4), the Fundamental Research Funds for the Central Universities (lzujbky-2018-ct05 and lzujbky-2019-58), the National Key R&D Program of China (2018YFB1003204), the Anhui Provincial Key Technologies R&D Program (1804b06020378), and the Program of Introducing Talents of Discipline to Universities (111 program) (B14025).

Conflicts of Interest

None declared.

Multimedia Appendix 1
The characteristics of the included studies.
[DOCX File , 54 KB - medinform_v9i3e25704_app1.docx]

Multimedia Appendix 2
The detailed performance measurements of machine learning technologies in the included studies.
[DOCX File , 108 KB - medinform_v9i3e25704_app2.docx]

References

1. Prevention and Treatment of Pressure Ulcers: Clinical Practice Guideline, 3rd Edition (2019). European Pressure Ulcer Advisory Panel (EPUAP), National Pressure Injury Advisory Panel (NPIAP), and the Pan Pacific Pressure Injury Alliance (PPPIA). URL: http://internationalguideline.com/guideline [accessed 2020-08-08]
2. Sen CK, Gordillo GM, Roy S, Kirsner R, Lambert L, Hunt TK, et al. Human skin wounds: a major and snowballing threat to public health and the economy. Wound Repair Regen 2009;17(6):763-771 [FREE Full text] [doi: 10.1111/j.1524-475X.2009.00543.x] [Medline: 19903300]
3. Chaboyer WP, Thalib L, Harbeck EL, Coyer FM, Blot S, Bull CF, et al. Incidence and Prevalence of Pressure Injuries in Adult Intensive Care Patients: A Systematic Review and Meta-Analysis. Crit Care Med 2018 Nov;46(11):e1074-e1081. [doi: 10.1097/CCM.0000000000003366] [Medline: 30095501]
4. Tubaishat A, Papanikolaou P, Anthony D, Habiballah L. Pressure Ulcers Prevalence in the Acute Care Setting: A Systematic Review, 2000-2015. Clin Nurs Res 2018 Jul;27(6):643-659. [doi: 10.1177/1054773817705541] [Medline: 28447852]
5. Li Z, Lin F, Thalib L, Chaboyer W. Global prevalence and incidence of pressure injuries in hospitalised adult patients: A systematic review and meta-analysis. Int J Nurs Stud 2020 May;105:103546. [doi: 10.1016/j.ijnurstu.2020.103546] [Medline: 32113142]
6. Capon A, Pavoni N, Mastromattei A, Di Lallo D. Pressure ulcer risk in long-term units: prevalence and associated factors. J Adv Nurs 2007 May;58(3):263-272. [doi: 10.1111/j.1365-2648.2007.04232.x] [Medline: 17474915]
7. Igarashi A, Yamamoto-Mitani N, Gushiken Y, Takai Y, Tanaka M, Okamoto Y. Prevalence and incidence of pressure ulcers in Japanese long-term-care hospitals. Arch Gerontol Geriatr 2013;56(1):220-226. [doi: 10.1016/j.archger.2012.08.011] [Medline: 22974661]
8. VanGilder C, Lachenbruch C, Algrim-Boyle C, Meyer S. The International Pressure Ulcer Prevalence™ Survey: 2006-2015: A 10-Year Pressure Injury Prevalence and Demographic Trend Analysis by Care Setting. J Wound Ostomy Continence Nurs 2017;44(1):20-28. [doi: 10.1097/WON.0000000000000292] [Medline: 27977509]
9. Hibbs P. The past politics of pressure sores. J Tissue Viability 1998 Oct;8(4):14-15. [doi: 10.1016/s0965-206x(98)80029-6] [Medline: 10480966]
10. Pressure ulcers quality standard. National Institute for Health and Care Excellence (NICE). URL: https://www.nice.org.uk/guidance/qs89 [accessed 2020-08-08]
11. Ayello EA, Zulkowski K, Capezuti E, Jicman WH, Sibbald RG. Educating Nurses in the United States about Pressure Injuries. Adv Skin Wound Care 2017 Feb;30(2):83-94. [doi: 10.1097/01.ASW.0000511507.43366.a1] [Medline: 28106637]
12. Usher K, Woods C, Brown J, Power T, Lea J, Hutchinson M, et al. Australian nursing students' knowledge and attitudes towards pressure injury prevention: A cross-sectional study. Int J Nurs Stud 2018 May;81:14-20. [doi: 10.1016/j.ijnurstu.2018.01.015] [Medline: 29427831]
13. Tallier PC, Reineke PR, Asadoorian K, Choonoo JG, Campo M, Malmgreen-Wallen C. Perioperative registered nurses knowledge, attitudes, behaviors, and barriers regarding pressure ulcer prevention in perioperative patients. Appl Nurs Res 2017 Aug;36:106-110. [doi: 10.1016/j.apnr.2017.06.009] [Medline: 28720229]
14. Demarré L, Vanderwee K, Defloor T, Verhaeghe S, Schoonhoven L, Beeckman D. Pressure ulcers: knowledge and attitude of nurses and nursing assistants in Belgian nursing homes. J Clin Nurs 2012 May;21(9-10):1425-1434. [doi: 10.1111/j.1365-2702.2011.03878.x] [Medline: 22039896]
15. Drennan VM, Ross F. Global nurse shortages-the facts, the impact and action for change. Br Med Bull 2019 Jun 19;130(1):25-37. [doi: 10.1093/bmb/ldz014] [Medline: 31086957]
16. Hyun S, Vermillion B, Newton C, Fall M, Li X, Kaewprag P, et al. Predictive validity of the Braden scale for patients in intensive care units. Am J Crit Care 2013 Nov;22(6):514-520 [FREE Full text] [doi: 10.4037/ajcc2013991] [Medline: 24186823]
17. Duan Y, Edwards JS, Dwivedi YK. Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. Int J Inf Manage 2019 Oct;48:63-71. [doi: 10.1016/j.ijinfomgt.2019.01.021]
18. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science 2015 Jul 17;349(6245):255-260. [doi: 10.1126/science.aaa8415] [Medline: 26185243]

19. Deo RC. Machine Learning in Medicine. Circulation 2015 Nov 17;132(20):1920-1930 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.001593] [Medline: 26572668]
20. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219 [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
21. An N, Ding H, Yang J, Au R, Ang TFA. Deep ensemble learning for Alzheimer's disease classification. J Biomed Inform 2020 May;105:103411. [doi: 10.1016/j.jbi.2020.103411] [Medline: 32234546]
22. Brennan PF, Bakken S. Nursing Needs Big Data and Big Data Needs Nursing. J Nurs Scholarsh 2015 Sep;47(5):477-484. [doi: 10.1111/jnu.12159] [Medline: 26287646]
23. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
24. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med 2019 Jan 01;170(1):W1-W33 [FREE Full text] [doi: 10.7326/M18-1377] [Medline: 30596876]
25. Baran Pouyan M, Birjandtalab J, Nourani M, Matthew Pompeo M. Automatic limb identification and sleeping parameters assessment for pressure ulcer prevention. Comput Biol Med 2016 Aug 01;75:98-108. [doi: 10.1016/j.compbiomed.2016.05.017] [Medline: 27268736]
26. Duvall J, Karg P, Brienza D, Pearlman J. Detection and classification methodology for movements in the bed that supports continuous pressure injury risk assessment and repositioning compliance. J Tissue Viability 2019 Feb;28(1):7-13 [FREE Full text] [doi: 10.1016/j.jtv.2018.12.001] [Medline: 30598376]
27. Raju D, Su X, Patrician PA, Loan LA, McCarthy MS. Exploring factors associated with pressure ulcers: a data mining approach. Int J Nurs Stud 2015 Jan;52(1):102-111. [doi: 10.1016/j.ijnurstu.2014.08.002] [Medline: 25192963]
28. Kaewprag P, Newton C, Vermillion B, Hyun S, Huang K, Machiraju R. Predictive models for pressure ulcers from intensive care unit electronic health records using Bayesian networks. BMC Med Inform Decis Mak 2017 Jul 05;17(Suppl 2):65 [FREE Full text] [doi: 10.1186/s12911-017-0471-z] [Medline: 28699545]
29. Alderden J, Pepper GA, Wilson A, Whitney JD, Richardson S, Butcher R, et al. Predicting Pressure Injury in Critical Care Patients: A Machine-Learning Model. Am J Crit Care 2018 Nov;27(6):461-468 [FREE Full text] [doi: 10.4037/ajcc2018525] [Medline: 30385537]
30. Li D, Mathews C. Automated measurement of pressure injury through image processing. J Clin Nurs 2017 Nov;26(21-22):3564-3575. [doi: 10.1111/jocn.13726] [Medline: 28071843]
31. Baran PM, Ostadabbas S, Nourani M, Pompeo M. Classifying bed inclination using pressure images. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2014; 2014 Presented at: 36th Annual International Conference of the IEEE-Engineering-in-Medicine-and-Biology-Society (EMBC); AUG 26-30, 2014; Chicago, IL p. 4663-4666. [doi: 10.1109/embc.2014.6944664]
32. Heydarzadeh M, Nourani M, Ostadabbas S. In-bed posture classification using deep autoencoders. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2016 Aug; 2016 Presented at: 38th Annual International Conference of the IEEE-Engineering-in-Medicine-and-Biology-Society (EMBC); AUG 16-20, 2016; Orlando, FL p. 3839-3842. [doi: 10.1109/embc.2016.7591565]
33. Matar G, Lina J, Kaddoum G. Artificial Neural Network for in-Bed Posture Classification Using Bed-Sheet Pressure Sensors. IEEE J Biomed Health Inform 2020 Jan;24(1):101-110. [doi: 10.1109/JBHI.2019.2899070] [Medline: 30762571]
34. Enayati M, Skubic M, Keller J, Popescu M, Farahani N. Sleep Posture Classification Using Bed Sensor Data and Neural Networks. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2018 Jul; 2018 Presented at: 40th Annual International Conference of the IEEE-Engineering-in-Medicine-and-Biology-Society (EMBC); JUL 18-21, 2018; Honolulu, HI p. 461-465. [doi: 10.1109/embc.2018.8512436]
35. Sprigle S, McNair D, Sonenblum S. Pressure Ulcer Risk Factors in Persons with Mobility-Related Disabilities. Adv Skin Wound Care 2020 Mar;33(3):146-154. [doi: 10.1097/01.ASW.0000653152.36482.7d] [Medline: 32058440]
36. Xu X, Lin F, Wang A, Hu Y, Huang M, Xu W. Body-Earth Mover's Distance: A Matching-Based Approach for Sleep Posture Recognition. IEEE Trans Biomed Circuits Syst 2016 Oct;10(5):1023-1035. [doi: 10.1109/TBCAS.2016.2543686] [Medline: 27483475]
37. Hsiao R, Mi Z, Yang B, Kau L, Bitew MA, Li T. Body posture recognition and turning recording system for the care of bed bound patients. Technol Health Care 2015;24 Suppl 1:S307-S312. [doi: 10.3233/THC-151088] [Medline: 26444814]
38. Ma C, Li W, Gravina R, Fortino G. Posture Detection Based on Smart Cushion for Wheelchair Users. Sensors (Basel) 2017 Mar 29;17(4) [FREE Full text] [doi: 10.3390/s17040719] [Medline: 28353684]
39. Li H, Lin S, Hwang Y. Using Nursing Information and Data Mining to Explore the Factors That Predict Pressure Injuries for Patients at the End of Life. Comput Inform Nurs 2019 Mar;37(3):133-141. [doi: 10.1097/CIN.0000000000000489] [Medline: 30418245]

40. Su C, Wang P, Chen Y, Chen L. Data mining techniques for assisting the diagnosis of pressure ulcer development in surgical patients. J Med Syst 2012 Aug;36(4):2387-2399. [doi: 10.1007/s10916-011-9706-1] [Medline: 21503743]
41. Chen H, Yu S, Xu Y, Yu S, Zhang J, Zhao J, et al. Artificial Neural Network: A Method for Prediction of Surgery-Related Pressure Injury in Cardiovascular Surgical Patients. J Wound Ostomy Continence Nurs 2018;45(1):26-30. [doi: 10.1097/WON.0000000000000388] [Medline: 29189496]
42. Dai L, Xu Q, Gao J. Investigate on pressure ulcer risk factors of orthopedic patients. Chinese Journal of Modern Nursing 2012;18(31):3726-3729.
43. Deng X, Wang Q, Li M, Hu A. Predicting the risk of hospital-required pressure ulcers in intensive care unit patinets based on decision tree. Chinese Journal of Practical Nursing 2016;32(7):485-489.
44. Yang Q, Wang G, Jiang B, Zhang H, Lu X. Study on risk prediction model of unavoidable pressure ulcers in cancer patients based on decision tree. Journal of Nursing Science 2019;34(13):4-7.
45. García-Zapirain B, Elmogy M, El-Baz A, Elmaghraby AS. Classification of pressure ulcer tissues with 3D convolutional neural network. Med Biol Eng Comput 2018 Dec;56(12):2245-2258. [doi: 10.1007/s11517-018-1835-y] [Medline: 29949023]
46. Veredas FJ, Mesa H, Morente L. Efficient detection of wound-bed and peripheral skin with statistical colour models. Med Biol Eng Comput 2015 Apr;53(4):345-359. [doi: 10.1007/s11517-014-1240-0] [Medline: 25564183]
47. Zahia S, Sierra-Sosa D, Garcia-Zapirain B, Elmaghraby A. Tissue classification and segmentation of pressure injuries using convolutional neural networks. Comput Methods Programs Biomed 2018 Jun;159:51-58. [doi: 10.1016/j.cmpb.2018.02.018] [Medline: 29650318]
48. Veredas FJ, Luque-Baena RM, Martín-Santos FJ, Morilla-Herrera JC, Morente L. Wound image evaluation with machine learning. Neurocomputing 2015 Sep;164:112-122. [doi: 10.1016/j.neucom.2014.12.091]
49. Veredas F, Mesa H, Morente L. Binary tissue classification on wound images with neural networks and bayesian classifiers. IEEE Trans Med Imaging 2010 Feb;29(2):410-427. [doi: 10.1109/TMI.2009.2033595] [Medline: 19825516]
50. Zahia S, Garcia-Zapirain B, Elmaghraby A. Integrating 3D Model Representation for an Accurate Non-Invasive Assessment of Pressure Injuries with Deep Learning. Sensors (Basel) 2020 May 21;20(10) [FREE Full text] [doi: 10.3390/s20102933] [Medline: 32455753]
51. Setoguchi Y, Ghaibeh AA, Mitani K, Abe Y, Hashimoto I, Moriguchi H. Predictability of Pressure Ulcers Based on Operation Duration, Transfer Activity, and Body Mass Index Through the Use of an Alternating Decision Tree. J Med Invest 2016;63(3-4):248-255 [FREE Full text] [doi: 10.2152/jmi.63.248] [Medline: 27644567]
52. Noguchi H, Kitamura A, Yoshida M, Minematsu T, Mori T, Sanada H. Clustering and Classification of Local Image of Wound Blotting for Assessment of Pressure Ulcer. 2014 Presented at: World Automation Congress (WAC) on Emerging Technologies for a New Paradigm in System of Systems Engineering; AUG 03-07, 2014; Waikoloa Hilton, HI. [doi: 10.1109/wac.2014.6935984]
53. Barsocchi P. Position recognition to support bedsores prevention. IEEE J Biomed Health Inform 2013 Jan;17(1):53-59. [doi: 10.1109/TITB.2012.2220374] [Medline: 23014763]
54. Cicceri G, De Vita F, Bruneo D, Merlino G, Puliafito A. A deep learning approach for pressure ulcer prevention using wearable computing. Hum Cent Comput Inf Sci 2020 Feb 03;10(1). [doi: 10.1186/s13673-020-0211-8]
55. Moon M, Lee S. Applying of Decision Tree Analysis to Risk Factors Associated with Pressure Ulcers in Long-Term Care Facilities. Healthc Inform Res 2017 Jan;23(1):43-52 [FREE Full text] [doi: 10.4258/hir.2017.23.1.43] [Medline: 28261530]
56. Kosmopoulos DI, Tzevelekou FL. Automated pressure ulcer lesion diagnosis for telemedicine systems. IEEE Eng Med Biol Mag 2007;26(5):18-22. [doi: 10.1109/emb.2007.901786] [Medline: 17941318]
57. Shi C, Dumville JC, Cullum N. Evaluating the development and validation of empirically-derived prognostic models for pressure ulcer risk assessment: A systematic review. Int J Nurs Stud 2019 Jan;89:88-103. [doi: 10.1016/j.ijnurstu.2018.08.005] [Medline: 30352322]
58. Langemo D, Spahn J, Spahn T, Pinnamaneni VC. Comparison of standardized clinical evaluation of wounds using ruler length by width and Scout length by width measure and Scout perimeter trace. Adv Skin Wound Care 2015 Mar;28(3):116-121 [FREE Full text] [doi: 10.1097/01.ASW.0000461117.90346.0d] [Medline: 25679463]
59. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020 Mar 25;368:m689 [FREE Full text] [doi: 10.1136/bmj.m689] [Medline: 32213531]
60. Strickland E. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. IEEE Spectr 2019 Apr;56(4):24-31. [doi: 10.1109/MSPEC.2019.8678513]

Abbreviations

AI: artificial intelligence
CBM: China Biomedical Literature Database
CINAHL: Cumulative Index to Nursing and Allied Health Literature
CNKI: China National Knowledge Infrastructure
DT: decision tree

EHR: electronic health record
ICU: intensive care unit
KNN: k-nearest neighbor
MeSH: Medical Subject Headings
ML: machine learning
PI: pressure injury
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction model Risk Of Bias ASsessment Tool

Edited by C Lovis; submitted 12.11.20; peer-reviewed by L Ge, J Alderden; comments to author 05.12.20; revised version received 21.01.21; accepted 05.02.21; published 10.03.21.

Please cite as:
Jiang M, Ma Y, Guo S, Jin L, Lv L, Han L, An N
Using Machine Learning Technologies in Pressure Injury Management: Systematic Review
JMIR Med Inform 2021;9(3):e25704
URL: https://medinform.jmir.org/2021/3/e25704
doi:10.2196/25704
PMID:33688846

Original Paper

Electronic Health Record Use in Swiss Nursing Homes and Its Association With Implicit Rationing of Nursing Care Documentation: Multicenter Cross-sectional Survey Study

Dietmar Ausserhofer[1,2], PhD; Lauriane Favez[2], MA; Michael Simon[2,3], PhD; Franziska Zúñiga[2], PhD

[1]College of Health Care-Professions Claudiana, Bolzano-Bozen, Italy

[2]Nursing Science, Department of Public Health, University of Basel, Basel, Switzerland

[3]Nursing Research Unit, Inselspital Bern University Hospital, Bern, Switzerland

Corresponding Author:
Franziska Zúñiga, PhD
Nursing Science
Department of Public Health
University of Basel
Bernoullistrasse 28
Basel, 4056
Switzerland
Phone: 41 61 207 09 13
Email: franziska.zuniga@unibas.ch

Abstract

Background: Nursing homes (NHs) are increasingly implementing electronic health records (EHRs); however, little information is available on EHR use in NH settings. It remains unclear how care workers perceive its safety, quality, and efficiency, and whether EHR use might ease the burden of documentation, thereby reducing its implicit rationing.

Objective: This study aims to describe nurses' perceptions regarding the usefulness of the EHR system and whether sufficient numbers of computers are available in Swiss NHs, and to explore the system's association with implicit rationing of nursing care documentation.

Methods: This was a multicenter cross-sectional study using survey data from the Swiss Nursing Homes Human Resources Project 2018. It includes a convenience sample of 107 NHs, 302 care units, and 1975 care workers (ie, registered nurses and licensed practical nurses) from Switzerland's German- and French-speaking regions. Care workers completed questionnaires assessing the level of implicit rationing of nursing care documentation, their perceptions of the EHR system's usefulness and of how sufficient the number of available computers was, staffing and resource adequacy, leadership ability, and teamwork and safety climate. For analysis, we applied generalized linear mixed models, including individual-level nurse survey data and data on unit and facility characteristics.

Results: Overall, the care workers perceived the EHR systems as useful; ratings ranged from 69.42% (1362/1962; *guarantees safe care and treatment*) to 78.32% (1535/1960; *allows quick access to relevant information on the residents*). However, less than half (914/1961, 46.61%) of the care workers reported sufficient computers on their unit to allow timely documentation. Half of the care workers responded that they sometimes or often had to ration the documentation of care. After adjusting for work environment factors and safety and teamwork climate, both higher care worker ratings of the EHR system's usefulness (β=−.12; 95% CI −0.17 to −0.06) and sufficient numbers of computers (β=−.09; 95% CI −0.12 to −0.06) were consistently associated with lower implicit rationing of nursing care documentation.

Conclusions: Both the usefulness of the EHR system and the number of computers available were important explanatory factors for care workers leaving care activities (eg, developing or updating nursing care plans) unfinished. NH managers should carefully select and implement their information technology infrastructure with greater involvement and attention to the needs of their care workers and residents. Further research is needed to develop and implement user-friendly information technology infrastructure in NHs and to evaluate their impact on care processes as well as resident and care worker outcomes.

(JMIR Med Inform 2021;9(3):e22974) doi:10.2196/22974

KEYWORDS

electronic health records; nursing homes; nursing care; health care rationing; rationing of nursing care; unfinished care; documentation; patient care planning; mobile phone

Introduction

Background

Health care organizations worldwide are increasingly using electronic health records (EHRs) to improve health care safety, quality, and efficiency. EHRs are defined as an electronic version of a person's medical history, including key administrative clinical data relevant to that person's care [1]. Although digital transformation in acute care is progressing quickly, the implementation of EHR in long-term care is following at a slower pace. In the United States, less than 50% of nursing homes (NHs) have implemented EHRs, with nonprofit and government NHs, those with more than 100 beds, and those with higher staffing levels (ie, registered nurses [RNs] and certified nursing assistants) more likely to use EHRs [2-6]. Among the barriers identified for successful EHR implementation, NH settings were costs, the need for training, and the culture change required to embrace technology [6,7].

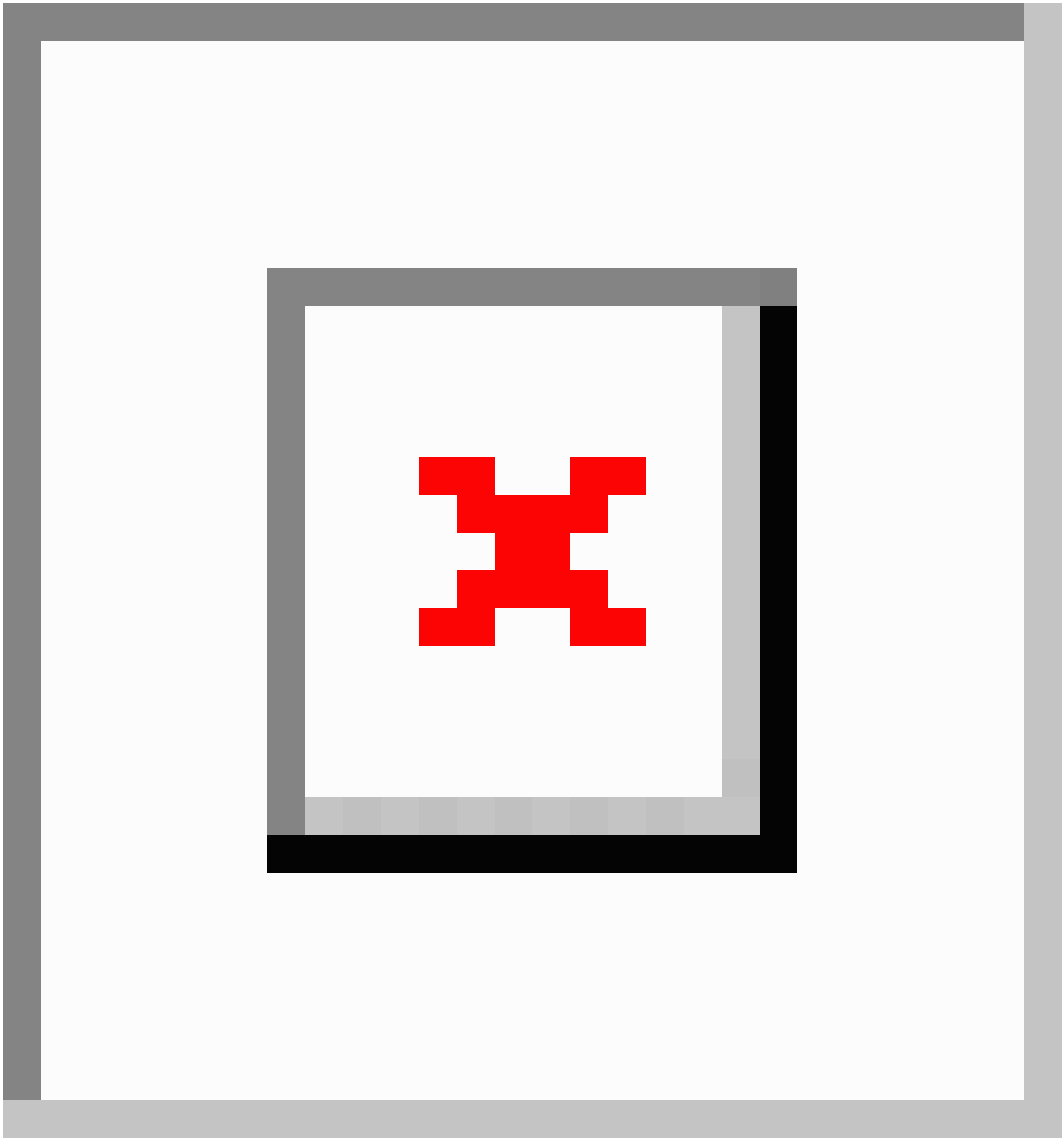
Although little is known regarding the impact of EHR adoption on the provision of NH care, positive effects on the processes and outcomes of acute care provision have been reported. These include increased adherence to guideline-based care, enhanced surveillance and monitoring, improved clinical decision making, and decreased medication errors [8-13]. Despite concerns that EHR implementation might negatively impact safety and quality of care during the transition period, acute care studies found no differences between pre- and postimplementation on short-term inpatient mortality, adverse events, or readmissions [14]. Some benefits of EHR use (eg, increased access to resident information, cost avoidance, and increased documentation accuracy) are increasingly recognized by health care professionals, including physicians [15] and nurses [16].

Even if the overall quality of documentation is not improved in the electronic system, for example, in cases where paper-based documentation standards were already extremely high [17], one expected benefit of EHR is increased time efficiency. In fact, at least during the implementation phase, the opposite has been reported, with documentation time increasing from 16% to 28% for physicians and from 9% to 23% for nurses [18]. Although EHRs should support health care professionals by reducing their documentation burden, thus allowing them more time for dedicated patient care, this initial impact on their workloads might prove a major barrier to their implementation and long-term use [18].

Nurses spend around one-fifth of their working time on documentation activities, such as developing or updating nursing care plans [19]. Although these activities are considered crucial to the provision of high-quality professional NH care [20], these *indirect care* activities performed away from residents are often either rationed or missed. Nurses place higher priority on direct care activities, that is, those that require interactions with the residents or their families, such as assisting with drinking and food intake [21,22]. A previous study reported that NH care workers who reported less rationing of direct care, rehabilitation, monitoring, and social care activities tended to perceive the overall quality of NH care as higher, whereas they actually associated *more* rationing of documentation with better self-perceived quality of NH care [23].

Implicit rationing of nursing care or missed care—recently summarized also under the umbrella term unfinished nursing care [24]—has become a global phenomenon of concern affecting the safety and quality of hospital and NH care [25,26]. NH studies indicate that up to 75% of nurses leave at least one necessary care activity unfinished on every shift [22,27]. Implicit rationing of nursing care has been defined as "the withholding of or failure to carry out all needed nursing interventions in the face of inadequate time, staffing or skill mix" [28]. Although this mainly refers to *direct care* activities with residents, failure to document nursing care is equally dangerous, as it hinders continuity of care. As this study's conceptual model describes (Figure 1), alongside perceived shortfalls in the information technology (IT) infrastructure (ie, EHRs and computers), care workers' perceptions of facility and unit characteristics, work environment, teamwork and safety climate, and even individual care worker characteristics can all impact NH care provision processes, meaning they can also result in implicit rationing of nursing care, including documentation. Evidence supports this conceptual underpinning, as lower levels of nurse staffing [29] and teamwork and safety climate [21] were all associated with higher amounts of missed or rationed care.

Figure 1. Conceptual framework: factors related to implicit rationing of nursing care documentation. EHR: electronic health records; FTE: full-time equivalent; RN: registered nurse.



Research Gap and Objectives

To date, little information is available on EHR use in NHs, for example, how nurses, as the main users, perceive their workplace system's quality and efficiency. Moreover, it remains unclear what roles EHRs' uses and characteristics might have on NH care processes, for example, whether more efficient EHRs might reduce care workers' documentation burden, thereby reducing the perceived need to implicitly ration it and allowing better continuity of care. As increasing numbers of NHs have implemented EHRs in recent years with the objective of increasing efficiency, in this study, we aim (1) to explore Swiss NH care workers' perceptions regarding their EHR systems' usefulness and the sufficiency of the number of computers and (2) to explore the association between the IT infrastructure and implicit rationing of nursing care documentation.

Methods

Study Design

This study is based on data from the 2018 Swiss Nursing Home Human Resources Project (SHURP), a cross-sectional, multicenter study.

Sample and Setting

A convenience sample of 107 NHs, housing 302 care units, and 1975 care workers (ie, RNs and licensed practical nurses) in Switzerland's German- and French-speaking regions were included in this study. The mean response rate to the care worker survey was 66.0%, ranging from 12.7% to 98.2% at the facility level. NHs who had participated in the first edition of the SHURP study (2013-2015) [30] were invited to participate in this new edition and were automatically included if they accepted. To increase the sample size, we sent waves of invitations to randomly selected NHs. In parallel, uninvited NHs that were willing to participate could contact the study team directly to be included. Finally, to further increase the inclusion rate, collaborations were set up with diverse NH associations. Additional NHs were included until March 2019. Inclusion criteria were that each NH was recognized by cantonal authorities and had a minimum of 20 beds.

Data Collection

The survey was administered, as appropriate, in two language versions, German and French, between September 2018 and October 2019. All directors of the participating NHs provided written consent to participate in the study. For care workers, sending back the voluntary care worker questionnaire was considered as informed consent.

Ethical Aspects

An ethics waiver was obtained from the responsible Swiss ethics committee (the Northwest and Central Switzerland ethics committee, BASEC Nr Req-2018-00420).

Variables and Measures

To measure *the rationing of nursing care documentation*, we used the 3-item subscale of the NH version of the Basel Extent of Rationing of Nursing Care instrument. Care workers were asked how often in the past 7 days they had been unable to study care plans at the beginning of their shift, set up or update residents' care plans, or document the care provided because of lack of time or high workload [31]. As lack of time or workload is a matter not only of resources (eg, staffing levels) but also of demand, EHR systems might increase the demand in terms of documentation.

The main explanatory variables were *care workers' perceptions of the EHR system's usefulness* (5 items) and *sufficiency of the number of computers on the units* (one item). These items were developed based on a literature review of EHR use in NHs [32,33]. The explanatory factor analyses of the internal structure of the 5 items on care workers' perceptions revealed a good fit, suggesting a one-dimensional solution (Tucker Lewis Index of factoring reliability=0.976; root mean square error of approximation index=0.079; 95% CI 0.063-0.096; Cronbach α=.88). Therefore, we calculated the scale's mean score. To facilitate further analyses, we kept the coding of the 5-point Likert scale of the single item assessing the sufficiency of computers on the units.

All potential confounding and control variables, including facility and unit characteristics, perceptions of work environment factors, teamwork and safety climate, and care worker characteristics, are described in Multimedia Appendix 1.

Data Analyses

Descriptive statistics (frequencies, percentages, means, and SDs) were calculated to describe the measured variables. To explore differences between care workers' professional backgrounds with regard to the EHR system's usefulness and whether a sufficient number of computers were available, we used chi-square tests. To explore the relationship between care workers' perceptions with regard to the EHR systems and whether sufficient computers were available and implicit rationing of nursing care documentation, 2-level generalized linear mixed models were used. On the basis of the intraclass correlation coefficient 1 (ICC1), which was >0.05, multilevel modeling was required [34]. Therefore, we computed ICC1 to assess the variability of the outcome variable (implicit rationing of nursing care documentation) between units and facilities. In this case, an ICC1 of 0.155 at the unit level and 0.118 at the facility level indicated a need to account for the clustering of care worker data within units and facilities.

We report unadjusted (*crude*) associations and 2 adjusted models: (1) not including staffing and resources adequacy and (2) including staffing and resources adequacy. To compare the models' relative fits, we used Akaike information criterion; a lower value indicates a better fit. Data analyses were performed with R (version 3.4.2; R Foundation for Statistical Computing, 2017) using the rptR package for the calculation of ICC1 [35] and the lme4 package for generalized linear mixed models [36]. Depending on the variable, between 0.1% and 8.3% of the data for unit and facility characteristics were missing. In the nurse survey, data missing varied between 0.1% (ie, educational background) and 3% (ie, professional experience). A *P* value of less than .05 was considered significant.

Results

Sample Description

This substudy used a sample of 1975 care workers. More than 90% were female; the majority were older than 41 years and had more than 5 years of professional experience. The majority worked part time, with employment levels between 51% and 90% and with regular changes in shifts. Of the 107 Swiss NHs included in the study, the majority were medium sized (between 50 and 100 beds) and private or privately subsidized. Table 1 summarizes the care worker, unit, and facility characteristics.

Table 1. Facility, unit, and care worker characteristics.

Facility and unit characteristics	Total (N=107 NHs[a], 302 units, 1975 care workers)	German-speaking region (n=88 NHs, 268 units, 1794 care workers)	French-speaking region (n=19 NHs, 34 units, 181 care workers)
NH size, n (%)			
Small (<50 beds)	24 (22.4)	20 (22.7)	4 (21.1)
Medium (50-100 beds)	55 (51.4)	42 (47.8)	13 (68.4)
Large (>100 beds)	28 (26.2)	26 (29.5)	2 (10.5)
NH profit status, n (%)			
Public	45 (42.1)	41 (46.6)	4 (21.1)
Privately subsidized or private	62 (57.9)	47 (53.4)	15 (78.9)
NH unit characteristics			
Clinical focus on dementia, n (%)	218 (74.4)	196 (75.1)	22 (68.8)
Bed capacity, median (IQR)	24 (12)	24 (12)	29 (19)
Full-time equivalent per 100 beds, median (IQR)	48.5 (23.2)	48.0 (23.4)	51.6 (16.8)
Skill mix level (% registered nurse), median (IQR)	26.5 (16.7)	27.8 (17.0)	20.3 (9.2)
Care worker characteristics			
Age (years), n (%)			
<21	127 (6.46)	120 (6.73)	7 (3.87)
21-30	408 (20.76)	361 (20.24)	47 (25.97)
31-40	336 (17.10)	295 (16.54)	41 (22.65)
41-50	396 (20.15)	360 (20.18)	36 (19.89)
51-60	556 (28.30)	519 (29.09)	37 (20.44)
>60	142 (7.23)	129 (7.23)	13 (7.18)
Gender: female, n (%)	1783 (91.25)	1613 (90.92)	170 (94.44)
Educational background, n (%)			
Registered nurse	944 (47.80)	861 (47.99)	83 (45.86)
Licensed practical nurse	1031 (52.20)	933 (52.01)	98 (54.14)
Tenure in current nursing home, n (%)			
Up to 5 years	921 (48.02)	836 (47.96)	85 (48.57)
5-10 years	387 (20.18)	348 (19.97)	39 (22.29)
≥10 years	610 (31.80)	559 (32.07)	51 (29.14)
Employment level, n (%)			
<51%	319 (16.32)	303 (17.07)	16 (8.89)
51%-90%	1105 (56.52)	982 (55.32)	123 (68.33)
91%-100%	531 (27.16)	490 (27.61)	41 (22.78)
Main shift, n (%)			
Regular change of shifts	1003 (50.99)	921 (51.57)	82 (45.30)
Day evening shift	783 (39.81)	702 (39.31)	81 (44.75)
Night shift	181 (9.20)	163 (9.12)	18 (9.95)

[a]NH: nursing home.

Variable Result Description

Care Workers' Perceptions of the EHR System's Usefulness and the Sufficiency of the Number of Computers on Their Unit

Overall, the care workers perceived their facilities' EHR systems as useful (Table 2). The percentage agreeing or strongly agreeing with the respective statements ranged from 69.42% (*guarantees safe care and treatment*) to 78.32% (*allows quick access to relevant information on the residents*). However, less than half (46.61%) of the care workers reported sufficient computers on their units to allow timely documentation.

As summarized in Table 2, we observed differences between RNs' and licensed practical nurses' perceptions as well as between language regions. For instance, compared with RNs, licensed practical nurses more often agreed that *the EHR system gives a good daily overview of all residents on the care unit*.

Table 2. Care workers' perception of the electronic health record system's usefulness and of whether the number of computers was sufficient (N=1975).

6 items on care workers' perceptions of the electronic health record system's usefulness and sufficiency of the number of computers on the units	Total (N=1975), n (%[a])	Educational background, n (%[a])		*P* value[b]
		Registered nurses (n=966)	Licensed practical nurses (n=1058)	
The electronic health record system allows timely communication between the nursing and therapy teams	1367 (69.96)	667 (71.11)	700 (68.90)	.29
The electronic health record system provides a good overview on the main focus of care and treatment for the individual residents	1507 (76.89)	710 (75.53)	797 (78.14)	.17
The electronic health record system gives a good daily overview on all residents on the care unit	1429 (72.98)	664 (70.78)	765 (75.00)	*.04*
The electronic health record system guarantees safe care and treatment	1362 (69.42)	645 (68.54)	717 (70.23)	.42
The electronic health record system allows quick access to relevant information on the residents	1535 (78.32)	720 (76.51)	815 (79.98)	.06
On our unit there are sufficient computers to allow timely documentation	914 (46.61)	451 (47.98)	463 (45.35)	.24

[a]Percentage agreement (agree and strongly agree).

[b]Chi-square test, *P*<.05 highlighted in italic.

Implicit Rationing of Nursing Care Documentation, Work Environment, and Teamwork and Safety Climate

Approximately half of the care workers responded that they sometimes or often had to ration care activities related to documentation (range: 46.02% [*studying care plans*] to 50.06% [*set up or update residents' care plans*]; Table 3). The mean rating for implicit rationing of nursing care documentation was 2.38 (SD 0.90; rarely to sometimes). As Table 4 shows, care workers rated adequate staffing and resources at the neutral midpoint (mean 2.67, SD 0.67) and strongly felt that they were supported by leadership (mean 3.18, SD 0.62). The mean teamwork and safety climate was rated as favorable (mean 3.89, SD 0.81). Furthermore, ICCs of the rationing of documentation items and whether sufficient numbers of computers were available ranged between 0.077 and 0.221, indicating substantial variation between units and between facilities (Table 4).

Table 3. Frequencies of implicit rationing of nursing care documentation (N=1975).

Care activities rationed by care workers in the last 7 days	Activity not necessary, n (%)	Never, n (%)	Seldom, n (%)	Sometimes, n (%)	Often, n (%)
Studying care plans at the beginning of the shift	13 (0.67)	478 (24.72)	553 (28.59)	480 (24.82)	410 (21.2)
Set up or update residents' care plans	110 (6.83)	270 (16.77)	424 (26.34)	481 (29.88)	325 (20.19)
Documentation of care	4 (0.21)	429 (22.26)	654 (33.94)	561 (29.11)	279 (14.48)

Table 4. Characteristics of variables under study (N=1975).

Variables	Mean (SD)	Facility level, ICC1[a] (95% CI)	Unit level, ICC1 (95% CI)
Rationing of nursing care documentation	2.38 (0.9)	0.118 (0.076-0.165)	0.155 (0.111-0.202)
Care workers' perception of the electronic health record system's usefulness	3.86 (0.77)	0.077 (0.043-0.112)	0.097 (0.064-0.135)
Care workers' perception of sufficient number of computers	3.13 (1.33)	0.116 (0.072-0.161)	0.221 (0.176-0.269)
Work environment			
Leadership	3.18 (0.62)	0.156 (0.104-0.205)	0.278 (0.228-0.326)
Staffing and resources adequacy	2.67 (0.67)	0.214 (0.151-0.271)	0.254 (0.207-0.302)
Teamwork and safety climate	3.89 (0.81)	0.111 (0.068-0.156)	0.196 (0.152-0.244)

[a]ICC1: intraclass correlation coefficient 1.

Factors Associated With Implicit Rationing of Nursing Care Documentation

In the crude models (Table 5), as well as models 1 and 2 (Table 6), care workers' perceptions of both the EHR system's usefulness and whether a sufficient number of computers were available were significantly associated with implicit rationing of nursing care documentation. More positive care workers' perceptions of the EHR system's usefulness (β=−.12; 95% CI −0.17 to −0.06) and of the sufficiency of the number of computers (β=−.09; 95% CI −0.12 to −0.06) were associated with lower implicit rationing of nursing care documentation (model 2).

Table 5. Implicit rationing of nursing care documentation regressed on care workers' perceptions of their electronic health record systems and the sufficiency of the number of computers, along with facility, unit and care worker characteristics and staffing variables, work environment, and teamwork and safety climate.

Variables	Crude models[a]	
	β (95% CI)	SE
Explanatory variables		
Care workers' perception of the electronic health record system's usefulness	−.31[b] (−0.36 to −0.26)	0.03
Care workers' perception of whether sufficient numbers of computers were available on their units	−.19[b] (−0.21 to −0.16)	0.02
Control variables		
Facility characteristics		
Language region	0.18 (−0.03 to 0.40)	0.11
Nursing home size	−.03 (−0.14 to 0.08)	0.05
Profit status	−.04 (−0.19 to 0.12)	0.08
Electronic health record system	0.01 (−0.01-to 0.04)	0.01
Unit characteristics		
Staffing levels	0 (−0.01 to 0.00)	0
Skill mix levels	0 (−0.01 to 0.00)	0
Work environment		
Leadership	−.37[b] (−0.44 to −0.31)	0.03
Staffing and resources adequacy	−.63[b] (−0.69 to −0.58)	0.03
Safety and teamwork climate	−.39[b] (−0.46 to −0.34)	0.03
Care workers' characteristics		
Gender	−.07 (−0.21 to 0.06)	0.07
Age	0.01 (−0.02 to 0.03)	0.01
Educational background	−.08[b] (−0.16 to −0.01)	0.04
Professional experience	0.04 (−0.01 to 0.08)	0.02
Employment level	−.04 (−0.09 to 0.03)	0.03
Fixed effects (intercept)	2.39[b] (2.32 to 2.47)	0.03

[a]Random effect: Facility-level variance (SD)=0.07 (0.27), Unit-level variance (SD)=0.06 (0.25).

[b]*P* value less than .05.

Higher ratings of leadership and safety teamwork climate were significantly associated with lower levels of implicit rationing of nursing care documentation only in model 1 (not accounting for staffing and resource adequacy). In model 2, care worker–perceived staffing and resources adequacy was the strongest explanatory factor, that is, higher ratings for staffing and resources adequacy were associated with lower levels of implicit rationing of nursing care documentation (β=−.52; 95% CI −0.58 to −0.45). Moreover, care workers' educational backgrounds were significantly associated with implicit rationing of nursing care documentation in both models (Table 6), with licensed practical nurses in both cases reporting lower levels of rationing of nursing care documentation than RNs (β=−.09; 95% CI −0.15 to −0.02).

Table 6. Implicit rationing of nursing care documentation regressed on care workers' perceptions of their electronic health record systems and the sufficiency of the number of computers, along with facility, unit and care worker characteristics and staffing variables, work environment, and teamwork and safety climate.

Variables	Multiple model 1[a] (without staffing and resources adequacy)		Multiple model 2[a] (with staffing and resources adequacy)	
	β (95% CI)	SE	β (95% CI)	SE
Explanatory variables				
Care workers' perception of the EHR[b] system's usefulness	−.14[c] (−0.20 to −0.09)	0.03	−.12[c] (−0.17 to −0.06)	0.03
Care workers' perception of whether sufficient numbers of computers were available on their units	−.12[c] (−0.15 to −0.09)	0.02	−.09[c] (−0.12 to −0.06)	0.01
Control variables				
Facility characteristics				
Language region	__[d]	—	—	—
Nursing home size	—	—	—	—
Profit status	—	—	—	—
EHR system	—	—	—	—
Unit characteristics				
Staffing levels	—	—	—	—
Skill mix levels	—	—	—	—
Work environment				
Leadership	−.12[c] (−0.21 to −0.04)	0.04	0.08 (−0.04 to 0.12)	0.04
Staffing and resources adequacy	—	—	−.52[c] (−0.58 to −0.45)	0.03
Safety and teamwork climate	−.20[c] (−0.27 to −0.12)	0.04	−.08[c] (−0.15 to −0.01)	0.04
Care workers' characteristics				
Gender	—	—	—	—
Age	—	—	—	—
Educational background	−.08[c] (−0.16 to −0.02)	0.04	−.09[c] (−0.15 to −0.02)	0.04
Professional experience	—	—	—	—
Employment level	—	—	—	—
Fixed effects (intercept)	4.67[c] (4.39 to 4.94)	0.14	4.80[c] (4.53 to 5.05)	0.13

[a]Random effects: Multiple model 1: Facility-level variance (SD)=0.05 (0.22), Unit-level variance (SD)=0.04 (0.21), Akaike information criterion=4598.8; Multiple model 2: Facility-level variance (SD)=0.03 (0.17), Unit-level variance (SD)=0.01 (0.12), Akaike information criterion=4405.8.

[b]EHR: electronic health record.

[c]*P* value <.05.

[d]Variable not included in the model.

Discussion

Principal Findings

In this study, we aimed to explore Swiss NH care workers' perceptions of their EHR systems' usefulness, whether their units had sufficient numbers of computers, and the association with rationing of nursing care documentation. Overall, the majority of care workers perceived the EHR systems as useful; however, fewer than half of the care workers reported having sufficient computers on their unit to allow timely documentation, and more than half of the care workers reported sometimes or often having to ration care activities related to documentation. Higher implicit rationing of nursing care documentation was reported by those who rated their EHR system's usefulness as low and the number of computers as insufficient.

Most care workers in our study sample perceived that the EHR was useful, for example, that it provided a good overview of the main focus of care and treatment and allowed quick access to relevant information on residents. Earlier studies have found that various advantages of EHR compared with traditional paper records were reported in long-term care settings. These included the structured collection of and accessibility to information

about residents' family histories, contact information, medications, information regarding current and previous care, medical treatments and procedures, and other relevant health-related information [37]. Likewise, Swiss care workers appreciated the various benefits of their EHR systems. Although EHRs are supposed to improve the safety and quality of care by offering tools (eg, alerts and reminders) to help avoid adverse events such as those related to medication errors [8-12], nearly one-third of our sample did not consider the EHR useful for guaranteeing safe care and treatment. We cannot explain this perception, but it could be based on the structure, accessibility, monitoring tools, usability, or other aspects of EHRs as well as on the handling and common understanding of a team about how to deal with the system.

It is clear, however, that EHR use does not automatically improve documentation, that is, its adoption does not necessarily mean that its users will provide timelier, more complete records; better continuity of care; or safer care or treatments [38]. Although safety concerns linked to EHR implementation, especially during the initial adjustment to digital documentation, have been reported elsewhere [39], once care workers are familiar with their particular systems [40], EHRs ultimately have a strong potential to improve the quality and safety of workflows. As with other systems that have delivered widespread improvements, the expected benefits of EHR can only be achieved in real-world settings through continuous feedback and improvement [41]. Improving our understanding of how EHRs contribute to safe care and how their use in NHs may actually lead to safety issues will require further qualitative research.

One less complicated matter is that half of our respondents reported not having sufficient computers on their units for the timely completion of their documentation. Care workers, especially RNs, spend a considerable amount of their working time on documentation activities, such as developing or updating nursing care plans [19]. A lack of computers on the unit (often there is only one) might impede timely care planning and documentation and increase the documentation burden. Therefore, NHs need to allow care workers timely access to EHRs and avoid waiting times. For example, to eliminate waiting time for computers, it may be practical to perform activities such as developing or updating nursing care plans or documenting nursing care in real time at the patient's bedside via mobile devices (eg, tablets or smartphones). Currently, however, no evidence is available on the effects or acceptability of such devices by NH care workers to either improve documentation or to reduce rationing of nursing care documentation. Further research on this topic is required.

More than half of our care worker sample responded that they sometimes or often had to ration documentation-related care activities. Tasks such as developing or updating nursing care plans or documenting nursing care are important parts of daily patient care; however, they are often perceived as keeping care workers away from the residents. However, it might be some time before EHR technology can meet care workers' initial expectations that EHR use will reduce their documentation time, allowing them more time for direct care activities. In fact, initial adjustment to EHR may even increase documentation time [18]. Although health care is a complex, adaptive system, the software is not. It is complex, but adaptation tends to result from incremental and iterative improvements. Initially, this limitation might be the heart of the problem for NH care workers: rather than following and lightening their daily workload, they might find that EHR largely determines and adds to it [42].

After adjusting for important factors, our analysis showed that rationing of nursing documentation is consistently related to care workers' perceptions of both their EHR systems' usefulness and the sufficiency of the number of computers available to them. This finding provides new insights on why these *indirect care* activities often remain unfinished [21,22]. Former evidence has shown that work environment factors such as leadership and staffing and resources adequacy as well as the safety and teamwork climate explain certain levels of NH care rationing [21,43]. In addition, we now see that both EHRs' general lack of user-friendliness and the general unit-level shortage of documentation workstations are important factors explaining care workers' tendency to leave *indirect* care activities, such as developing or updating nursing care plans or documenting nursing care, unfinished.

As this leaves information gaps in the EHR, documentation rationing is likely accompanied by work-arounds, such as exchanging vital daily information on paper and via oral handovers to provide continuity of care. In other situations, information may simply be lost. Apart from presenting obvious legal problems if documentation is lacking or untraceable, both options increase the risk of adverse events and reduce the quality of care.

In our study sample alone, we found 12 separate EHR systems, which might differ regarding key EHR domains (eg, data transfer, structured clinical documentation, medication use processes, and communication) [44]. EHRs target a large and growing global market; according to a recently published report from Fortune Business Insights, a compound annual growth rate of 5.4% is expected until 2026 [45]. As buyers in that market, NH management could more forcefully demand IT solutions that support care workers' documentation needs while increasing safety and quality of care. EHR providers can reasonably be called upon to develop and design their software with input from all stakeholders—especially their users—in real-world settings. Therefore, care workers should be actively involved in testing and implementing the proposed IT infrastructure to ensure that, from the moment of implementation, it actually reduces their documentation burden [40].

Limitations

First, the cross-sectional design of the study did not allow inference of causal relationships. Second, as both the outcome variable (rationing of nursing care documentation) and the main explanatory variables (both involving perceptions of IT infrastructure) were assessed via a care worker survey, this measure might have introduced common method bias. Third, we unfortunately did not measure when each NH implemented its EHR, what basic and/or continuous training care workers

receive to use the EHR, or to what extent staff managers encourage or monitor the care workers in using the EHR information, which could have helped explain the association between care workers' perceptions regarding IT infrastructure and implicit rationing of nursing care documentation.

Conclusions

Although the surveyed RNs' and licensed practical nurses' overall perception of EHR systems' usefulness in Swiss NHs was high, only half of the care workers reported having sufficient numbers of computers on their units. After adjusting for other main explanatory variables, our analyses indicated that more positive perceptions of both EHR systems' usefulness and the sufficiency of the number of computers on their units were associated with less rationing of nursing care documentation. Thus, both the EHR system and the number of available computers influence care workers' decision to leave *indirect* care activities, such as developing or updating nursing care plans or documenting nursing care, unfinished. Bearing this in mind, NH managers should carefully select and implement their IT infrastructure with full engagement and according to the needs of the end users, that is, their care workers, as well as their residents. Although EHRs are increasingly implemented in NHs, there is still little evidence on how their use influences the safety and quality of NH care, including as it relates to efficiency. Future challenges to the research concerning EHR use in NHs are (1) to identify user-friendly designs and successful implementations of related IT infrastructure in NHs (eg, EHR access via mobile devices) and (2) to evaluate the impact of EHR implementation in NH settings not only on both direct and indirect care processes but also on resident and care worker outcomes.

Authors' Contributions

FZ and LF developed the idea for this study. MS, LF, and FZ contributed to the concept, design, and data collection. DA, MS, and FZ contributed to data analysis and interpretation. DA contributed to drafting of the manuscript. All authors contributed to the critical revision of the manuscript and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study variables.
[DOC File , 55 KB - medinform_v9i3e22974_app1.doc]

References

1. Electronic Health Records. CMS. 2012. URL: https://www.cms.gov/Medicare/E-Health/EHealthRecords [accessed 2021-02-16]
2. Bjarnadottir RI, Herzig CT, Travers JL, Castle NG, Stone PW. Implementation of electronic health records in US nursing homes. Comput Inform Nurs 2017 Aug;35(8):417-424. [doi: 10.1097/CIN.0000000000000344] [Medline: 28800581]
3. Abramson EL, McGinnis S, Moore J, Kaushal R. A statewide assessment of electronic health record adoption and health information exchange among nursing homes. Health Serv Res 2014 Feb;49(1 Pt 2):361-372 [FREE Full text] [doi: 10.1111/1475-6773.12137] [Medline: 24359612]
4. Park-Lee E, Rome V, Caffrey C. Characteristics of residential care communities that use electronic health records. Am J Manag Care 2015 Dec 1;21(12):e669-e676 [FREE Full text] [Medline: 26760430]
5. Holup AA, Dobbs D, Temple A, Hyer K. Going digital: adoption of electronic health records in assisted living facilities. J Appl Gerontol 2014 Jun;33(4):494-504. [doi: 10.1177/0733464812454009] [Medline: 24781968]
6. Holup AA, Dobbs D, Meng H, Hyer K. Facility characteristics associated with the use of electronic health records in residential care facilities. J Am Med Inform Assoc 2013;20(4):787-791 [FREE Full text] [doi: 10.1136/amiajnl-2012-001564] [Medline: 23645538]
7. Cherry B, Carter M, Owen D, Lockhart C. Factors affecting electronic health record adoption in long-term care facilities. J Healthc Qual 2008;30(2):37-47. [Medline: 18411891]
8. McCarthy B, Fitzgerald S, O'Shea M, Condon C, Hartnett-Collins G, Clancy M, et al. Electronic nursing documentation interventions to promote or improve patient safety and quality care: A systematic review. J Nurs Manag 2019 Apr;27(3):491-501. [doi: 10.1111/jonm.12727] [Medline: 30387215]
9. Campanella P, Lovato E, Marone C, Fallacara L, Mancuso A, Ricciardi W, et al. The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. Eur J Public Health 2016 Feb;26(1):60-64. [doi: 10.1093/eurpub/ckv122] [Medline: 26136462]
10. Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med 2006 May 16;144(10):742-752. [Medline: 16702590]

11. Poissant L, Pereira J, Tamblyn R, Kawasumi Y. The impact of electronic health records on time efficiency of physicians and nurses: a systematic review. J Am Med Inform Assoc 2005 Oct;12(5):505-516 [FREE Full text] [doi: 10.1197/jamia.M1700] [Medline: 15905487]
12. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. J Am Med Assoc 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
13. Kruse CS, Mileski M, Vijaykumar AG, Viswanathan SV, Suskandla U, Chidambaram Y. Impact of electronic health records on long-term care facilities: systematic review. JMIR Med Inform 2017 Sep 29;5(3):e35 [FREE Full text] [doi: 10.2196/medinform.7958] [Medline: 28963091]
14. Barnett ML, Mehrotra A, Jena AB. Adverse inpatient outcomes during the transition to a new electronic health record system: observational study. Br Med J 2016 Jul 28;354:i3835 [FREE Full text] [Medline: 27471242]
15. King J, Patel V, Jamoom EW, Furukawa MF. Clinical benefits of electronic health record use: national findings. Health Serv Res 2014 Feb;49(1 Pt 2):392-404 [FREE Full text] [doi: 10.1111/1475-6773.12135] [Medline: 24359580]
16. Cherry BJ, Ford EW, Peterson LT. Experiences with electronic health records: early adopters in long-term care facilities. Health Care Manage Rev 2011;36(3):265-274. [doi: 10.1097/HMR.0b013e31820e110f] [Medline: 21646885]
17. Wang N, Yu P, Hailey D. The quality of paper-based versus electronic nursing care plan in Australian aged care homes: A documentation audit study. Int J Med Inform 2015 Aug;84(8):561-569. [doi: 10.1016/j.ijmedinf.2015.04.004] [Medline: 26004340]
18. Baumann LA, Baker J, Elshaug AG. The impact of electronic health record systems on clinical documentation times: a systematic review. Health Policy 2018 Dec;122(8):827-836. [doi: 10.1016/j.healthpol.2018.05.014] [Medline: 29895467]
19. Yee T, Needleman J, Pearson M, Parkerton P, Parkerton M, Wolstein J. The influence of integrated electronic medical records and computerized nursing notes on nurses' time spent in documentation. Comput Inform Nurs 2012 Jun;30(6):287-292. [doi: 10.1097/NXN.0b013e31824af835] [Medline: 22411414]
20. Johnson L, Edward K, Giandinoto J. A systematic literature review of accuracy in nursing care plans and using standardised nursing language. Collegian 2018 Jun;25(3):355-361. [doi: 10.1016/j.colegn.2017.09.006]
21. Zúñiga F, Ausserhofer D, Hamers JPH, Engberg S, Simon M, Schwendimann R. The relationship of staffing and work environment with implicit rationing of nursing care in Swiss nursing homes--A cross-sectional study. Int J Nurs Stud 2015 Sep;52(9):1463-1474. [doi: 10.1016/j.ijnurstu.2015.05.005] [Medline: 26032730]
22. Nelson ST, Flynn L. Relationship between missed care and urinary tract infections in nursing homes. Geriatr Nurs 2015;36(2):126-130. [doi: 10.1016/j.gerinurse.2014.12.009] [Medline: 25563066]
23. Zúñiga F, Ausserhofer D, Hamers JPH, Engberg S, Simon M, Schwendimann R. Are staffing, work environment, work stressors, and rationing of care related to care workers' perception of quality of care? A cross-sectional study. J Am Med Dir Assoc 2015 Oct 1;16(10):860-866. [doi: 10.1016/j.jamda.2015.04.012] [Medline: 26027721]
24. Jones TL, Hamilton P, Murry N. Unfinished nursing care, missed care, and implicitly rationed care: State of the science review. Int J Nurs Stud 2015 Jun;52(6):1121-1137. [doi: 10.1016/j.ijnurstu.2015.02.012] [Medline: 25794946]
25. Recio-Saucedo A, Dall'Ora C, Maruotti A, Ball J, Briggs J, Meredith P, et al. What impact does nursing care left undone have on patient outcomes? Review of the literature. J Clin Nurs 2018 Jun;27(11-12):2248-2259 [FREE Full text] [doi: 10.1111/jocn.14058] [Medline: 28859254]
26. Ausserhofer D, Zander B, Busse R, Schubert M, De GS, Rafferty AM, RN4CAST consortium. Prevalence, patterns and predictors of nursing care left undone in European hospitals: results from the multicountry cross-sectional RN4CAST study. BMJ Qual Saf 2014 Feb;23(2):126-135. [doi: 10.1136/bmjqs-2013-002318] [Medline: 24214796]
27. Knopp-Sihota JA, Niehaus L, Squires JE, Norton PG, Estabrooks CA. Factors associated with rushed and missed resident care in western Canadian nursing homes: a cross-sectional survey of health care aides. J Clin Nurs 2015 Oct;24(19-20):2815-2825. [doi: 10.1111/jocn.12887] [Medline: 26177787]
28. Schubert M, Glass TR, Clarke SP, Schaffert-Witvliet B, De Geest S. Validation of the basel extent of rationing of nursing care instrument. Nurs Res 2007;56(6):416-424. [doi: 10.1097/01.NNR.0000299853.52429.62] [Medline: 18004188]
29. Griffiths P, Recio-Saucedo A, Dall'Ora C, Briggs J, Maruotti A, Meredith P, Missed Care Study Group. The association between nurse staffing and omissions in nursing care: a systematic review. J Adv Nurs 2018 Jul;74(7):1474-1487. [doi: 10.1111/jan.13564] [Medline: 29517813]
30. Schwendimann R, Zúñiga F, Ausserhofer D, Schubert M, Engberg S, de Geest S. Swiss Nursing Homes Human Resources Project (SHURP): protocol of an observational study. J Adv Nurs 2014 Apr;70(4):915-926. [doi: 10.1111/jan.12253] [Medline: 24102650]
31. Zúñiga F, Schubert M, Hamers JPH, Simon M, Schwendimann R, Engberg S, et al. Evidence on the validity and reliability of the German, French and Italian nursing home version of the Basel Extent of Rationing of Nursing Care instrument. J Adv Nurs 2016 Aug;72(8):1948-1963. [doi: 10.1111/jan.12975] [Medline: 27062508]
32. Xiao Y, Montgomery DC, Philpot LM, Barnes SA, Compton J, Kennerly D. Development of a tool to measure user experience following electronic health record implementation. J Nurs Admin 2014;44(7/8):423-428. [doi: 10.1097/nna.0000000000000093]

33. Meehan R. Electronic health records in long-term care: staff perspectives. J Appl Gerontol 2017 Oct;36(10):1175-1196. [doi: 10.1177/0733464815608493] [Medline: 26464335]
34. LeBreton J, Senter JL. Answers to 20 questions about interrater reliability and interrater agreement. Organizational Research Methods 2008:815-852 [FREE Full text] [doi: 10.1177/1094428106296642]
35. Stoffel MA, Nakagawa S, Schielzeth H. rptR: repeatability estimation and variance decomposition by generalized linear mixed ‐ effects models. Methods Ecol Evol 2017 May 30;8(11):1639-1644. [doi: 10.1111/2041-210x.12797]
36. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using. J Stat Soft 2015;67(1):-. [doi: 10.18637/jss.v067.i01]
37. Kruse CS, Mileski M, Alaytsev V, Carol E, Williams A. Adoption factors associated with electronic health record among long-term care facilities: a systematic review. BMJ Open 2015 Jan 28;5(1):e006615 [FREE Full text] [doi: 10.1136/bmjopen-2014-006615] [Medline: 25631311]
38. Wang N, Yu P, Hailey D. Description and comparison of documentation of nursing assessment between paper-based and electronic systems in Australian aged care homes. Int J Med Inform 2013 Sep;82(9):789-797. [doi: 10.1016/j.ijmedinf.2013.05.002] [Medline: 23786709]
39. Harrington L, Kennerly D, Johnson C. Safety issues related to the electronic medical record (EMR): synthesis of the literature from the last decade, 2000-2009. J Healthc Manag 2011;56(1):31-43; discussion 43. [Medline: 21323026]
40. Ko M, Wagner L, Spetz J. Nursing home implementation of health information technology: review of the literature finds inadequate investment in preparation, infrastructure, and training. Inquiry 2018;55:46958018778902. [doi: 10.1177/0046958018778902] [Medline: 29888677]
41. Singh H, Sittig DF. Measuring and improving patient safety through health information technology: the Health IT Safety Framework. BMJ Qual Saf 2015 Sep 14;25(4):226-232. [doi: 10.1136/bmjqs-2015-004486]
42. De Groot K, De Veer AJE, Paans W, Francke AL. Use of electronic health records and standardized terminologies: a nationwide survey of nursing staff experiences. Int J Nurs Stud 2020 Apr;104:103523. [doi: 10.1016/j.ijnurstu.2020.103523] [Medline: 32086028]
43. Zúñiga F, Ausserhofer D, Hamers JP, Engberg S, Simon M, Schwendimann R. Are staffing, work environment, work stressors, and rationing of care related to care workers' perception of quality of care? A cross-sectional study. J Am Med Directors Assoc 2015 Oct;16(10):860-866. [doi: 10.1016/j.jamda.2015.04.012]
44. Degenholtz HB, Resnick A, Lin M, Handler S. Development of an applied framework for understanding health information technology in nursing homes. Journal of the American Medical Directors Association 2016 May;17(5):434-440. [doi: 10.1016/j.jamda.2016.02.002]
45. Market Research Report. Fortune Business Insights. 2020. URL: https://www.fortunebusinessinsights.com/electronic-health-records-ehr-market-102660 [accessed 2021-02-16]

Abbreviations

EHR: electronic health record
ICC1: intraclass correlation coefficient 1
IT: information technology
NH: nursing home
RN: registered nurse
SHURP: Swiss Nursing Home Human Resources Project

Edited by G Eysenbach; submitted 06.08.20; peer-reviewed by SB Ho, R Chan, Z Reis; comments to author 21.12.20; revised version received 30.12.20; accepted 17.01.21; published 02.03.21.

Please cite as:
Ausserhofer D, Favez L, Simon M, Zúñiga F
Electronic Health Record Use in Swiss Nursing Homes and Its Association With Implicit Rationing of Nursing Care Documentation: Multicenter Cross-sectional Survey Study
JMIR Med Inform 2021;9(3):e22974
URL: https://medinform.jmir.org/2021/3/e22974
doi:10.2196/22974
PMID:33650983

Original Paper

Commitment Levels of Health Care Providers in Using the District Health Information System and the Associated Factors for Decision Making in Resource-Limited Settings: Cross-sectional Survey Study

Shuma G Kanfe[1], MPH; Berhanu F Endehabtu[2], MPH; Mohammedjud H Ahmed[1], MPH; Nebyu D Mengestie[2], MPH; Binyam Tilahun[2], MSc, MPH, PhD

[1]Health Informatics, Mettu University, Metu Zuria, Ethiopia

[2]Department of Health Informatics, Institute of Public Health, University of Gondar, Gondar, Ethiopia

Corresponding Author:
Shuma G Kanfe, MPH
Health Informatics
Mettu University
Metu Zuria
Ethiopia
Phone: 251 0935054730
Email: shumagosha33@gmail.com

Abstract

Background: Changing the culture of information use, which is one of the transformation agendas of the Ministry of Health of Ethiopia, cannot become real unless health care providers are committed to using locally collected data for evidence-based decision making. The commitment of health care providers has paramount influence on district health information system 2 (DHIS2) data utilization for decision making. Evidence is limited on health care providers' level of commitment to using DHIS2 data in Ethiopia. Therefore, this study aims to fill this evidence gap.

Objective: This study aimed to assess the levels of commitment of health care providers and the factors influencing their commitment levels in using DHIS2 data for decision making at public health care facilities in the Ilu Aba Bora zone of the Oromia national regional state, Ethiopia in 2020.

Methods: The cross-sectional quantitative study supplemented by qualitative methods was conducted from February 26, 2020 to April 17, 2020. A total of 264 participants were approached. SPSS version 20 software was used for data entry and analysis. Descriptive and analytical statistics, including bivariable and multivariable analyses, were performed. Thematic analysis was conducted for the qualitative data.

Results: Of the 264 respondents, 121 (45.8%, 95% CI 40.0%-52.8%) respondents showed high commitment levels to use DHIS2 data. The variables associated with the level of commitment to use DHIS2 data were found to be provision of feedback for DHIS2 data use (adjusted odds ratio [AOR] 1.85, 95% CI 1.02-3.33), regular supervision and managerial support (AOR 2.84, 95% CI 1.50-5.37), information use culture (AOR 1.92, 95% CI 1.03-3.59), motivation to use DHIS2 data (AOR 1.80, 95% CI 1.00-3.25), health needs (AOR 3.96, 95% CI 2.11-7.41), and competency in DHIS2 tasks (AOR 2.41, 95% CI 1.27-4.55).

Conclusions: In general, less than half of the study participants showed high commitment levels to use DHIS2 data for decision making in health care. Providing regular supportive supervision and feedback and increasing the motivation and competency of the health care providers in performing DHIS2 data tasks will help in promoting their levels of commitment that can result in the cultural transformation of data use for evidence-based decision making in health care.

(JMIR Med Inform 2021;9(3):e23951) doi:10.2196/23951

KEYWORDS

commitment; district health information system; decision making; performance monitoring; health facilities; information use

Introduction

Health care providers, in particular, the performance monitoring team (PMT) is a team of multidisciplinary health workforce that is primarily responsible for improving data quality, using information regularly, monitoring the health progress, and improving the performance of health care delivery at all levels of the health care system. PMT members are selected health care providers who are involved in the collection, generation, and utilization of health information for decision making, and they serve as the focal persons in their departments/wards. PMT members in Ethiopia are prominent/selected health care providers who widely participate in using district health information system 2 (DHIS2) data for decision making. The commitment levels of the PMT members to their organizations and the use of health information for decision making is a topical issue that needs some attention in the delivery of quality health care.

Changing the culture of information use at each level of the health system is one of the transformation agendas of the Ministry of Health of Ethiopia. This cannot become real unless health care providers are committed to use locally collected data for evidence-based decision making. Health care providers' level of commitment to use DHIS2 data could provide comprehensive and dependable information, which is the basis for better decision making [1-3]. This is because DHIS2 data consist of global initiatives by Sustainable Development Goals and Countdown to 2030 that emphasize its contribution to monitoring of service delivery by health care providers [4]. The DHIS2 is used in more than 60 countries, and most global initiatives are interested in using DHIS2 data for monitoring the health performance [5-7]. Facility-based data (DHIS2 data) is one of the major identified strategies to achieve sustainable development goals—especially for maternal mortality and neonatal mortality to reach a global average of only 70 per 100,000 live births by 2030 [2,8].

As per the World Health Organization (WHO), health care providers' level of commitment has paramount influence on DHIS2 data utilization for decision making that will also be the basis for the provision of quality health service [9]. The WHO and the Institute for Health Metrics and Evaluation have stated that to improve the accuracy and utility of health information for decision making, the commitment levels of health care providers is the base [10]. This is because improving the quality of health service can be affected if health care providers are not responsible in using the highly generated medical data, which is but a mandatory step on the path to reaching the sustainable development goals and universal health coverage [11].

A study in Nigeria has identified that the commitment of health care providers to use health information should be taken into consideration, and currently, the level of commitment among health care providers in Nigeria is 60%-80% [1] A study conducted in Isfahan showed that the compliance of health care providers to use district health information was much lower than WHO standards (90%) and was limited to an average of 35.75% [12]. Another study conducted in Iran at hospitals proposed that health care providers' level of commitment toward the use of health information was a worthy path that every health care worker needs to be dedicated to in using and implementing routine health information for decision making. Currently, the average score of health care providers' level of commitment to use and implement routine health information, especially electronic medical records, has been achieved with an average of 74.7% [2]. Another study in Ghana indicated that health care providers are expected to have a sense of promoting responsibility to use health information and should feel committed to improving the health status of the target population. Factors such as punctuality at work, documentation of daily activities, and monitoring of data wisely have been associated with the level of commitment to use information. However, currently, the level of commitment among health care providers, specifically among senior managers, is 77.3% [13].

Studies have identified that health care providers need to be committed to using DHIS2 data, wherein over 90% of the available data have been generated within only 2 years [2,14]. In sub-Saharan Africa, the standard procedure for data use is poor and the measures of the health care performance are very low because of the inadequate commitment of health care providers [15]. The quality of health care depends on the dedication and commitment of health care service providers [2]. Being committed to using DHIS2 data will favor high-quality health systems, thereby ensuring relevant advancements toward achievement of sustainable development goals [2,14]. However, over the years, evidence for low commitment to use data have been less as decisions are not taken based on data [2]. Health information data show inconsistency and poor treatment responses because of the low levels of commitments of health care providers [16,17]. The government's health facilities are not committed to reporting data on a regular basis, data are not used for setting target programs, and these facilities are unresponsive to timely decision making [2,14]. The WHO has stated that factors that affect the commitment to use DHIS2 data are critical in affecting the quality of health service provision [9,10]. Thus, quality of health care will improve when health managers are committed to the use of health information for decision making because quality in health care is a production of cooperation between the patient and the health care provider in a supportive environment [1]. High-quality routine health system data are highly relevant for monitoring advancement toward achievement of the Millennium Development Goals 4 and 5, which are twins to the Sustainable Development Goal 3. However, the main determinant to reach this stage is the level of commitment toward the utilization of routine health information systems. The evidence for the provision of good-quality health service is lacking due to the low commitment of health care providers toward the utilization of health information systems [2]. Low commitment toward the utilization of DHIS2 data results in the production of late diagnosis and treatment reports and distorts the consistency within data space, making the overall utilization of district health information system for decision making to be low [3].

In Ethiopia, the PMT is one of the major platforms to review the performance, data quality, and information use of the health system at each level. The level of commitment of health care providers (especially PMT members) has direct influence on

DHIS2 data utilization for decision making [13,18]. Nevertheless, to the best of our knowledge, evidence is limited on PMT members' level of commitment to use DHIS2 data and the factors that determine the extent of their commitment levels. Therefore, this study aimed to fill the evidence gap on PMT members' level of commitment to use DHIS2 data for decision making and the factors that determine their commitment levels.

Methods

Study Design and Setting

A quantitative cross-sectional study design supplemented by a qualitative study design was conducted from February 26, 2020 to April 17, 2020. This study was conducted in public health facilities in the Ilu Aba Bora zone, Oromia, Ethiopia. The Ilu Aba Bora zone is one of the zones of the Oromia region of Ethiopia, which is 600 km away from Addis Ababa, Ethiopia. This study covered different types of health facilities, including referral hospitals, primary hospitals, and health care centers located in the southwest region of Ethiopia; 41 health centers and 2 hospitals (1 referral hospital and 1 primary hospital) were assessed as the areas for data collection.

Study Participants and Sample Size Determination

All selected health care providers who handle data, generate data, and use generated data for their decision making and those who serve as focal persons within their departments, collectively known as the PMT members according to the Ethiopian health system context, were the participants of this study. The total number of study participants within this zone was 264. Each study participant was approached and information was collected. For the qualitative study, purposive sampling techniques were used and the level of saturation was considered and saturated at the seventh participant.

Ethics Approval and Consent to Participate

This study protocol was reviewed and approved by the ethical review board of the University of Gondar and informed consent was obtained from each study participant. A permission letter was also obtained from each health facility. After the objective of this study was explained, verbal consent was obtained from each participant. The privacy and confidentiality of the information were strictly guaranteed by all data collectors and investigators. The information retrieved was used only for this study. Thus, the names of the participants and other personal identifiers were not included in the data collection tool.

Operational Definitions

PMT Members

The PMT members are the health care providers who serve as the focal persons in their respective departments (health management information system [HMIS] Officer, Medical Director, maternal and child health [MCH] Head, tuberculosis [TB] focal nurse, Triage Head nurse, primary health care unit manager, etc) according to Ethiopian health system contexts and the fact that they are responsible for the generation and utilization of data in addition to their clinical roles.

Commitment Level of PMT Members to Use DHIS2 Data

The commitment level of PMT members to use DHIS2 data was measured using 11 questions of the Likert scale, and respondents who scored the median score and higher were categorized as having high level of commitment to use DHIS2 data and those who scored less than the median score were categorized as having low level of commitment to use DHIS2 data.

Data Collection Tools and Procedures

For the quantitative approach, a self-administered English-version questionnaire was used. For qualitative data, in-depth interviews were conducted using an interview guide and a tape recorder. The maximum and minimum times for the in-depth interviews were 49 minutes and 31 minutes, respectively.

Data Quality Control

Data were collected by trained data collectors by using questionnaires. Before the actual data collection, a pretest was conducted among 5% of the samples at the Buno Bedele general hospital and health center in the Bedele town. The validity of the questionnaire was determined based on the views of experts and the reliability was obtained by calculating the Cronbach alpha value (α=.82). Qualitative data were collected by an investigator after debriefing an in-depth interview by arranging a favorable time and a place for the interviewee.

Data Processing and Analysis

The data entry and analysis were performed using SPSS version 20 (IBM Corp). To explain the study population in relation to relevant variables, descriptive statistics was used. Associations between dependent and independent variables were checked and their strengths were presented using odds ratios and 95% confidence intervals. Both bivariable and multivariable logistic regressions were used to assess the associations between the outcomes and explanatory variables. *P* values less than .05 were considered statistically significant in the multivariable logistic regression. The qualitative data were analyzed by thematic analysis methods.

Results

Sociodemographic Characteristics of the Study Participants

A total of 264 participants were approached with 100% response rate. About two-thirds of the study participants (186/264, 70.5%) were 30 years of age or younger. The majority of the study participants were from a health center (234/264, 88.6%). More than half of the participants were males (147/264, 55.7%). The majority (203/264, 76.9%) of the study participants had a work experience of 4 years and more. About 156 (59.1%) of the 264 study participants had a bachelor's degree, whereas only 23 (8.7%) had a master's degree (Table 1).

Table 1. Sociodemographic characteristics of the study participants at the health facilities of Ilu Aba Bora Zone in 2020 (N=264).

Variables, subcategories	Values, n (%)
Age	
≤30 years	186 (70.5)
>30 years	78 (29.5)
Sex	
Male	147 (55.7)
Female	117 (44.3)
Type of facility	
Referral hospitals	16 (6.1)
Primary hospitals	14 (5.3)
Health center	234 (88.6)
Educational level	
Master's degree	23 (8.7)
Bachelor's degree	156 (59.1)
Diploma	85 (32.2)
Work experience	
≤3 years	61 (23.1)
>4 years	203 (76.9)
Position at facility	
Head	101 (38.3)
Expert	163 (61.7)

Commitment Level of PMT Members to Use DHIS2 Data for Decision Making

Of the 264 respondents, 121 (45.8%, 95% CI 40.0%-52.8%) had high levels of commitment to use DHIS2 data for decision-making purposes.

Level of Commitment to Use DHIS2 Data for Decision Making by Sociodemographic Variables

Among 117 female respondents, only 50 (42.7%) had high levels of commitment to use DHIS2 data. Holders of master's degrees had higher levels of commitment than diploma and degree holders. Those who had more work experience had higher commitment levels to use DHIS2 data than those who had lesser work experience. Respondents serving in the Head positions (60/101, 59.4%) had higher levels of commitment than those serving in the expert positions. This detail is presented in Table 2.

Table 2. Commitment levels of the performance monitoring team members to use district health information system in accordance with the sociodemographic characteristics[a].

Variables	Commitment level to use district health information system 2 data (N=264)	
	Low commitment, n (%)	High commitment, n (%)
Sex		
Female (n=117)	67 (57.3)	50 (42.7)
Male (n=147)	76 (51.7)	71 (48.3)
Age		
≤30 years (n=186)	96 (51.6)	90 (48.4)
>30 years (n=78)	47(60.3)	31 (39.7)
Type of facilities		
Referral hospital (n=16)	10 (62.5)	6 (37.5)
Primary hospitals (n=14)	6 (42.9)	8 (57.1)
Health center (n=234)	127 (54.3)	107 (45.7)
Educational level		
Master's degree (n=23)	11 (47.8)	12 (52.2)
BSc degree (n=156)	87 (55.8)	69 (44.2)
Diploma (n=85)	45 (52.9)	40 (47.1)
Position at facility		
Expert position (n=163)	83 (50.9)	80 (66.1)
Head position (n=101)	60 (59.4)	41 (33.8)
Experience		
≤3 years (n=61)	27 (18.9)	34 (55.7)
>4 years (n=203)	116 (81.1)	87 (42.9)

[a]All the percentages were calculated for each sociodemographic category.

Factors Associated With the Commitment Levels to Use DHIS2 Data for Decision Making

PMT members who received feedback for their DHIS2 data use were 1.85 times (adjusted odds ratio [AOR] 1.85, 95% CI 1.02-3.33) more likely to have a higher commitment level to use DHIS2 data than those who did not receive feedback. PMT members who had regular supervision and managerial support on their daily use of DHIS2 data for decision making were 2.84 times (AOR 2.84, 95% CI 1.50-5.37) more likely to have higher levels of commitment to use DHIS2 data than those who had no supportive supervision. Respondents who were competent to use DHIS2 data for their decision making were 2.41 times (AOR 2.41, 95% CI 1.27-4.55) more likely to have higher levels of commitment to use DHIS2 data than those who were not competent in DHIS2 tasks. PMT members with good culture of information use were 1.92 times (AOR 1.92, 95% CI 1.03-3.59) more likely to have higher levels of commitment to use DHIS2 data for decision making than those who did not have good culture of information use. Similarly, PMT members who inquired for DHIS2 data for health management were 3.96 times (AOR 3.96, 95% CI 2.11-7.41) more likely committed to use DHIS2 data than those who did not need DHIS2 data for health management. PMT members having motivation to use DHIS2 data were 1.80 times (AOR 1.80, 95% CI 1.00-3.25) more likely committed to using DHIS2 data when compared to those who had low motivation to use DHIS2 data for their decision making. These data are presented in Table 3.

Table 3. Factors associated with the level of commitment to use district health information system 2 data among performance monitoring team members at health facilities in the Ilu Aba Bora zone, Oromia region in 2020[a].

Variable, category	Commitment level		Crude odds ratio	Adjusted odds ratio
	High commitment, n (%)	Low commitment, n (%)		
Culture of information use				
Good (n=164)	83 (50.6)	81 (49.4)	1.67 (1.00-2.77)*	1.92 (1.03-3.59) **
Poor (n=100)	38 (38.0)	62 (62.7)	1[b]	1
Health needs				
Yes (n=131)	76 (58.0)	55 (42.0)	2.70 (1.64-4.45)	3.96 (2.11-7.41)***
No (n=133)	45 (33.8)	88 (66.2)	1	1
Motivation				
High motivation (n=136)	71 (52.2)	65 (47.8)	1.70 (1.04-2.77) *	1.80 (1.00-3.25)**
Poor motivation (n=128)	50 (39.1)	78 (60.9)	1	1
Feedback				
Yes (n=140)	71 (50.7)	69 (49.3)	1.52 (0.93-2.48)	1.85 (1.02-3.33)**
No (n=124)	50 (40.3)	74 (59.7)	1	1
Supervision				
Yes (n=141)	84 (59.6)	57 (40.4)	3.42 (2.05-5.71)	2.84 (1.50-5.37)***
No (n=123)	37 (30.1)	86 (69.9)	1	1
Competency				
High (n=133)	76 (57.1)	57 (42.9)	2.54 (1.54-4.19)	2.41 (1.27-4.55)**
Low (n=131)	45 (34.4)	86 (65.6)	1	1

[a]All the percentages were calculated for each sociodemographic category.

[b]Reference.

*P<.05 for bivariable analysis.

**P<.05 for multivariable analysis.

***P≤.001.

Qualitative Results

Interview questions were expected to be directed toward 3 categories of investigation: level of commitment to use DHIS2 data for decision making, factors that could facilitate level of commitment, and challenges to use DHIS2 data for decision making. Analysis of the interview transcripts revealed key themes grouped into one of the above 3 categories. Most of the interviewees agreed that they were able to use DHIS2 data, that they were competent, and that they devoted their time, resources, and efforts to use DHIS2 data.

> *…Having taken training and also under supervision from my managers, I search DHIS2 data on where and when to do our activities. So I have confidence to say that I am familiar with effective utilization of DHIS2 data for decision making.* [HMIS Officer, 27 years old]

Respondents said that promoting the culture of information use would increase their confidence in using DHIS2 data.

…Thereis a good culture for using information. This enables us to carry out our attention to use effectively DHIS data. For this, we are able to compute with technology that inquires oneself to update himself with DHIS2 data used for decision making. [Medical Director, 29 years old]

Another respondent explained the members' commitment to use DHIS2 data as follows:

> *…The PMT members are those who raise why and how questions to make effective use of data for decision making. As a manager of the health center, I'm also playing a role even more than what is expected of me. We are always ready to cut off the problems encountered with using DHIS2 data for decision making. Even we are in need that always like to be guided by DHIS2 data.* [TB focal nurse, 30 years old]

In some areas, health care providers showed low responsibility toward using DHIS2 data for decision making.

> *…Some are unresponsive to what they are required to do, some are unaccountable to their duty. We are also facing a lack of budget to use DHIS2 data for decision making. On behalf of the facility, we do not have much materials like computers, internet*

> *connections, Wi-Fi, adequately trained human resources.* [Triage Head nurse, 26 years old]

To achieve a high level of commitment, respondents had problems as follows:

> *…On behalf of our facility, we have encountered numerous problems such as insufficient computers, no sufficient internet access, and no sufficient trained human power. All of the above use DHIS2 data for decision making at an optimum stage in our facility and we are expected to do more in future.* [HMIS officer, 31 years old]

> *…Sometimes there is incomplete data. Sometimes there is too late data. This is due to misunderstanding about using DHIS2 data. Resource is not provided at required stages. Example, we will be out of internet connection for three weeks, our computer may fail but may not be fixed until one month. We are asked to be supported but no response.* [MCH Head, 29 years old]

Discussion

This study focused on the level of commitment of health care providers to use DHIS2 data and the factors that affect their levels of commitment. We found that the 45.8% (121/264, 95% CI 40.0%-52.8%) of the PMT members used DHIS2 data for decision making, which was higher than that reported in a study conducted in Iran (35.75%) [19]. This finding may be attributed to the fact that the government of Ethiopia has given special attention to the utilization of health information systems for decision making and the internal commitment of health care providers in Ethiopia to use these data has increased [20]. However, the proportion of PMT members committed to using DHIS2 data in this study was lower than that reported in a study conducted in Ghana (77.3%) [21] and Iran (74.7%) [22]. This might be because infrastructures and advancements in technology in Ghana are more developed than those in Ethiopia. The proportion of the committed PMT members in this study was also lower than that of the PMT members in a study conducted in Nigeria, wherein the proportion of professionals committed to use the routine health information system was 60%-80% [23]; however, the target for this proportion in 2010 was 90% [24]. The possible explanations for this variation could be the size of the study participants, their scope of roles, availability of infrastructure, and availability of resources such as internet connection and other related electronic devices. This result was supported by qualitative findings as follows:

> *…We familiarized ourselves with DHIS2 data even more than expected from us. We are dedicated to accepting and using DHIS2 data, those who were taken by training everywhere else have given training to those who have not been taken. However we lack some requirements like sufficient internet connection and skills to amend our tools like computers, internet related materials.* [Primary health care unit manager, 31 years old]

> *…Almost by what we have, we sacrificed our efforts to use DHIS2 data for our decision making though we encounter some difficulties from the resources limitation.* [HMIS officer, 29 years old]

PMT members competent in DHIS2 data tasks were 2.41 times more likely to have a higher level of commitment to use DHIS2 data for decision making than those incompetent in DHIS2 data tasks (AOR 2.41, 95% CI 1.27-4.55). This finding was in line with those reported in studies conducted in Ethiopia [25], Ghana [2], Nairobi, Kenya [26], and another study conducted at the health facilities in Kenya (*P*=.03) (AOR 4.32, 95% CI 2.34-7.98) [27]. However, this finding was inconsistent with that of a study conducted in Kenya, which indicated that competency in DHIS2 task has no association with the performance of the health information systems [28]. This result was supported by qualitative finding as follows:

> *…We ought to have sufficient competency to use DHIS2 data, even we have a good competency in using DHIS2 data tasks though we don't have enough internet access and sufficient computer devices.* [TB focal nurse, 30 years old]

This study revealed that feedback on DHIS2 data use was positively associated with PMT members' commitment level to use DHIS2 data for their decision making in the Ilu Aba Bora zone health facilities (AOR 1.85, 95% CI 1.02-3.33), which was in line with the findings of the studies conducted in Ethiopia [12], Kenya [29], and Ghana (*P*=.04) [2]. However, this finding was inconsistent with the findings of a study conducted in Ghana [30].

The promotion of information use culture in health care providers would result in them being 1.92 times more likely to have higher levels of commitment to use DHIS2 data as compared to those who did not have a culture of information use (AOR 1.92, 95% CI 1.03-3.59). This result was supported by qualitative findings as follows:

> *…We need to use DHIS2 data for clinical decision making that it enables us to perform our duty more quickly and with full evidence.* [Psychiatry Head, 27 years old]

As this study revealed, commitment levels to use DHIS2 data for decision making were based on health needs (AOR 3.96, 95% CI 2.11-7.41). However, this finding was inconsistent with that reported in a cross-sectional study conducted in Ghana, which showed that the commitment to use DHIS2 data for decision making does not depend on the health needs [2]. This result was supported by a qualitative finding as follows:

> *…Applying and using of DHIS2 data for decision making could be tied to health needs, because it is when there is health needs that DHIS2 data will be put in to considerations that it helps us to deal with our focuses.* [Triage Head focal nurse, 32 years old]

Regarding study participants' motivation to use DHIS2 data, respondents with higher motivation were 1.80 times more likely to have higher levels of commitment when compared to those with lower motivation to use DHIS2 data for their decision making (AOR 1.80, 95% CI 1.00-3.25). This finding (*P*=.03) was in line with the findings of studies conducted in Ethiopia [25] and Ghana (*P*=.01) [2].

PMT members with regular supportive supervision visits were 2.84 times more likely to have a higher level of commitment than those who did not have regular supportive supervision (AOR 2.84, 95% CI 1.50-5.37). This result was similar to those reported in studies conducted in Ethiopia [12,25] and Ghana, which showed that the level of commitment to use DHIS2 data was directly associated with the daily managerial supervision (*P*=.04) [2].

This study attempted to reveal the commitment levels of health care providers to use DHIS2 data and the factors associated with their levels of commitment. The strength of this study lies in the attempt to cover the different types of health facilities such as health centers, primary hospitals, and referral hospitals. Moreover, our study used a mixed-methods approach and gives evidence on the commitment levels of PMT members to use DHIS2 data for decision making and the barriers in using it. However, our study has the following limitations. First, this study was a facility-based cross-sectional study; therefore, it could not provide the causal relationships with the factors. Second, this study was conducted at health facilities and might not be generalizable to all other administrative services in Ethiopia. In addition, this study did not include health care providers in private health care facilities.

In conclusion, less than half of the PMT members in this study were committed to using DHIS2 data for decision making. Based on WHO's criteria for commitment to use health information and other studies found in the literatures, our proportion was low. The culture of information use, motivation to use DHIS2 data, competency in DHIS2 tasks, health needs, managerial supervision, and feedback on DHIS2 data use were the most important factors determining the commitment of health care providers to use DHIS2 data for decision making. Thus, we found significant factors that affect PMT members' level of commitment to the use of DHIS2 data for their decision making. The findings of our study suggest that providing regular supportive supervision and feedback, increasing the motivation of health care providers, and changing their attitudes will help in bringing cultural transformation of data use for evidence-based decision making in health care.

Acknowledgments

The authors would like to thank the Institute of Public Health of the University of Gondar for the approval of ethical clearance, health facilities, data collectors, supervisors, and study participants. This work would not have been possible without the financial support of Doris Duke Charitable Foundation under grant number 2017187. The data sets generated and analyzed during this study will be available upon reasonable request from the corresponding author.

Authors' Contributions

SG, NB, and BF made significant contributions to the conception, design, data collection, supervision, data analysis, interpretation, and write-up of the manuscript. BT and MH contributed to extensive revision of the manuscript, analysis, and interpretation. SG, MH, and BF were involved in drafting the manuscript and revising it critically for important intellectual content. All authors have read and approved the final version of this manuscript. BT and BF were also involved in the conceptualization and guidance of the overall progress and correction of the manuscript.

Conflicts of Interest

None declared.

References

1. Cardoso I. How top-management commitment in information system implementation influences IS usage and benefits achievement? Atas da Cone da Asoka Port Sist Inf 2014;14:94. [doi: 10.18803/capsi.v14.174-194]
2. Effah F. Commitment among senior managers to the use of district health information management system 2 data for decision making in maternal and neonatal health in Greater Accra Region. University of Ghana Digital collections. 2019. URL: http://ugspace.ug.edu.gh/handle/123456789/30846 [accessed 2020-03-02]
3. Aqil A, Lippeveld T, Hozumi D. PRISM framework: a paradigm shift for designing, strengthening and evaluating routine health information systems. Health Policy Plan 2009 May;24(3):217-228 [FREE Full text] [doi: 10.1093/heapol/czp010] [Medline: 19304786]
4. Corbett J, Mellouli S. Winning the SDG battle in cities: how an integrated information ecosystem can contribute to the achievement of the 2030 sustainable development goals. Info Systems J 2017 Jan 27;27(4):427-461. [doi: 10.1111/isj.12138]
5. Bernadette A, Anthony K, et al. Enhancing health information system for evidence based decision making in the health sector. 2018. URL: https://www.health.go.ke/wp-content/uploads/2019/01/HIS-POLICY-BRIEF-.pdf [accessed 2020-03-05]
6. Building resilient and sustainable systems for health (RSSH) information. 2019. URL: https://www.theglobalfund.org/media/4759/core_resilientsustainablesystemsforhealth_infonote_en.pdf [accessed 2020-05-03]
7. Ogega P. Data use challenges and the potential of live data visualization tools: A case study of health data-use workshops in Zambia (Master's thesis). 2017 Nov. URL: https://www.duo.uio.no/handle/10852/60022 [accessed 2020-05-03]

8. Moran A, Jolivet R, Chou D, Dalglish S, Hill K, Ramsey K, et al. A common monitoring framework for ending preventable maternal mortality, 2015-2030: phase I of a multi-step process. BMC Pregnancy Childbirth 2016 Aug 26;16:250 [FREE Full text] [doi: 10.1186/s12884-016-1035-4] [Medline: 27565428]
9. Boerma JT. WHO: A commitment to improve global health information. 2015 Jun 13. URL: https://www.who.int/mediacentre/commentaries/improving-health-data/en/ [accessed 2020-03-01]
10. Memorandum of understanding between WHO and Institute for Health Metrics. URL: http://www.healthdata.org/sites/default/files/files/MOU_IHME_WHO_050615.pdf [accessed 2020-08-03]
11. Khan SI, Hoque A, Ullah M. National Health Data Warehouse Bangladesh for remote health monitoring: Features, problems and privacy issues. Remote Heal Monit Work. 2016. URL: https://www.researchgate.net/publication/303408294_National_Health_Data_Warehouse_Bangladesh_for_Remote_Health_Monitoring_Features_Problems_and_Privacy_Issues [accessed 2020-01-02]
12. Teklegiorgis K, Tadesse K, Mirutse G, Terefe W. Level of data quality from Health Management Information Systems in a resources limited setting and its associated factors, eastern Ethiopia. S. Afr. j. inf. manag 2016 Aug 10;18(1). [doi: 10.4102/sajim.v18i1.612]
13. Tull K. Designing and implementing health management information systems. 2018. URL: https://assets.publishing.service.gov.uk/media/5c7003a6e5274a0ec6ed95c5/376_Designing_and_Implementing_HMIS.pdf [accessed 2020-01-02]
14. Kayode GA, Amoakoh-Coleman M, Brown-Davies C, Grobbee DE, Agyepong IA, Ansah E, et al. Quantifying the validity of routine neonatal healthcare data in the Greater Accra Region, Ghana. PLoS One 2014;9(8):e104053 [FREE Full text] [doi: 10.1371/journal.pone.0104053] [Medline: 25144222]
15. Mutale W, Chintu N, Amoroso C, Awoonor-Williams K, Phillips J, Baynes C, et al. Improving health information systems for decision making across five sub-Saharan African countries: Implementation strategies from the African Health Initiative. BMC Health Serv Res 2013 May 31;13(S2). [doi: 10.1186/1472-6963-13-s2-s9]
16. Muhindo R, Joloba EN. Health Management Information System (HMIS); Whose Data is it Anyway? Contextual Challenges. Review Pub Administration Manag 2016;4(2). [doi: 10.4172/2315-7844.1000190]
17. Awoonor-Williams JK, Bawah A, Nyonator F, Asuru R, Oduro A, Ofosu A, et al. BMC Health Serv Res 2013 May 31;13(S2):a [FREE Full text] [doi: 10.1186/1472-6963-13-s2-s3]
18. Bhattacharyya S, Berhanu D, Taddesse N, Srivastava A, Wickremasinghe D, Schellenberg J, et al. District decision-making for health in low-income settings: a case study of the potential of public and private sector data in India and Ethiopia. Health Policy Plan 2016 Sep;31 Suppl 2:ii25-ii34 [FREE Full text] [doi: 10.1093/heapol/czw017] [Medline: 27591203]
19. Raeisi A, Saghaeiannejad S, Karimi S, Ehteshami A, Kasaei M. District health information system assessment: a case study in iran. Acta Inform Med 2013 Mar;21(1):30-35 [FREE Full text] [doi: 10.5455/aim.2012.21.30-35] [Medline: 23572859]
20. Information Revolution. 2016. URL: http://indepth-network.org/workshop/2016/presentations/ethiopia_evidence_workshop/information%20revolution%20moh%20ethiopia.pdf [accessed 2020-01-02]
21. Okyere Boadu R, Adzakpah G, Agyei-Baffour P. The Role of Quality Improvement Process in Improving the Culture of Information among Health Staff in Ghana. Advances in Public Health 2019 Oct 27;2019:1-9. [doi: 10.1155/2019/7579569]
22. Jahanbakhsh M, Karimi S, Hassanzadeh A, Beigi M. Hospital managers' attitude and commitment toward electronic medical records system in Isfahan hospitals 2014. J Educ Health Promot 2017;6:37 [FREE Full text] [doi: 10.4103/jehp.jehp_13_15] [Medline: 28584837]
23. Makinde OA, Umar C, et al. Assessment of the routine health management information system in Niger state, Federal republic of Nigeria. 2012 Sep. URL: https://www.hfgproject.org/assessment-routine-health-management-information-system-niger-state-federal-republic-nigeria/ [accessed 2020-01-02]
24. Europe WHORO. Support tool to assess health information systems and develop and strengthen health information strategies. 2015. URL: https://www.euro.who.int/__data/assets/pdf_file/0011/278741/Support-tool-assess-HIS-en.pdf [accessed 2020-01-02]
25. Wude H, Woldie M, Melese D, Lolaso T, Balcha B. Utilization of routine health information and associated factors among health workers in Hadiya Zone, Southern Ethiopia. PLoS One 2020;15(5):e0233092 [FREE Full text] [doi: 10.1371/journal.pone.0233092] [Medline: 32437466]
26. Peter MN. Factors influencing utilization of routine health data in evidence based decision making in HIV/AIDS services by public health facilities in Nakuru County. University of Nairobi Research Archive. 2015. URL: http://erepository.uonbi.ac.ke/handle/11295/90875 [accessed 2020-01-02]
27. Use of aggregate data for health decision making at district level?: case study of west municipality of the greater Accra region. University of Ghana data collections. 2019. URL: http://ugspace.ug.edu.gh/handle/123456789/33538 [accessed 2020-01-02]
28. Nicholas S. Factors influencing performance of routine health information system? the case of Garissa sub county, Kenya. University of Nairobi Research Initiative. 2017. URL: http://erepository.uonbi.ac.ke/handle/11295/101966 [accessed 2020-01-02]

29. Kuyo R. Use of district health information system data to facilitate decision making in Uasin-Gishu sub county hospitals, Kenya. KeMu Digital Repository. 2019. URL: http://repository.kemu.ac.ke/handle/123456789/742?show=full [accessed 2020-01-02]
30. Amaniampong R, Agyei-Baffour P, et al. Knowledge of health information for healthcare decision making? A cross sectional study of health staff in Kumasi. Journal of Health, Medicine, and Nursing. 2017. URL: https://iiste.org/Journals/index.php/JHMN/article/view/38008 [accessed 2020-01-02]

Abbreviations

AOR: adjusted odds ratio
DHIS2: district health information system 2
HMIS: health management information system
MCH: maternal and child health
PMT: performance monitoring team
TB: tuberculosis
WHO: World Health Organization

Edited by C Lovis; submitted 29.08.20; peer-reviewed by M Mengiste, K Said Abasse; comments to author 05.11.20; revised version received 28.11.20; accepted 28.12.20; published 04.03.21.

Please cite as:
Kanfe SG, Endehabtu BF, Ahmed MH, Mengestie ND, Tilahun B
Commitment Levels of Health Care Providers in Using the District Health Information System and the Associated Factors for Decision Making in Resource-Limited Settings: Cross-sectional Survey Study
JMIR Med Inform 2021;9(3):e23951
URL: https://medinform.jmir.org/2021/3/e23951
doi:10.2196/23951
PMID:33661133

Original Paper

The Effect of Innovation Capabilities of Health Care Organizations on the Quality of Health Information Technology: Model Development With Cross-sectional Data

Moritz Esdar[1], MA; Ursula Hübner[1], PhD; Johannes Thye[1], MA; Birgit Babitsch[2], MPH, PhD; Jan-David Liebe[1,3], PhD

[1]Health Informatics Research Group, Faculty of Business Management and Social Sciences, University of Applied Sciences Osnabrueck, Osnabrueck, Germany

[2]Institute of Health and Education, New Public Health, Osnabrück University, Osnabrueck, Germany

[3]Institute of Medical Informatics, UMIT - Private University for Health Sciences, Medical Informatics and Technology, Hall in Tyrol, Austria

Corresponding Author:
Ursula Hübner, PhD
Health Informatics Research Group
Faculty of Business Management and Social Sciences
University of Applied Sciences Osnabrueck
Caprivistr 30A
Osnabrueck, 49076
Germany
Phone: 49 541 969 2012
Email: u.huebner@hs-osnabrueck.de

Abstract

Background: Large health organizations often struggle to build complex health information technology (HIT) solutions and are faced with ever-growing pressure to continuously innovate their information systems. Limited research has been conducted that explores the relationship between organizations' innovative capabilities and HIT quality in the sense of achieving high-quality support for patient care processes.

Objective: The aim of this study is to explain how core constructs of organizational innovation capabilities are linked to HIT quality based on a conceptual sociotechnical model on innovation and quality of HIT, called the IQ_{HIT} model, to help determine how better information provision in health organizations can be achieved.

Methods: We designed a survey to assess various domains of HIT quality, innovation capabilities of health organizations, and context variables and administered it to hospital chief information officers across Austria, Germany, and Switzerland. Data from 232 hospitals were used to empirically fit the model using partial least squares structural equation modeling to reveal associations and mediating and moderating effects.

Results: The resulting empirical IQ_{HIT} model reveals several associations between the analyzed constructs, which can be summarized in 2 main insights. First, it illustrates the linkage between the constructs measuring HIT quality by showing that the *professionalism of information management* explains the degree of *HIT workflow support* (R^2=0.56), which in turn explains the *perceived HIT quality* (R^2=0.53). Second, the model shows that HIT quality was positively influenced by innovation capabilities related to the top management team, the information technology department, and the organization at large. The assessment of the model's statistical quality criteria indicated valid model specifications, including sufficient convergent and discriminant validity for measuring the latent constructs that underlie the measures of HIT quality and innovation capabilities.

Conclusions: The proposed sociotechnical IQ_{HIT} model points to the key role of professional information management for HIT workflow support in patient care and perceived HIT quality from the viewpoint of hospital chief information officers. Furthermore, it highlights that organizational innovation capabilities, particularly with respect to the top management team, facilitate HIT quality and suggests that health organizations establish this link by applying professional information management practices. The model may serve to stimulate further scientific work in the field of HIT adoption and diffusion and to provide practical guidance to managers, policy makers, and educators on how to achieve better patient care using HIT.

(JMIR Med Inform 2021;9(3):e23306) doi:10.2196/23306

KEYWORDS

organizational innovation; health information management; organizational culture; diffusion of innovation; hospital information systems; organizational change management

Introduction

Background

Discussions on health information technologies (HITs) in research and practice have increasingly shifted from dealing with the question of *if* digital solutions are worth investing in [1,2] to questions on *how* higher degrees of successful digitalization can be achieved [3-6] and how HIT improves processes and outcomes [7-9]. Although the term HIT has been used and defined in various ways, we understand it to encompass the organization's electronic information technologies that health care professionals use to support the care process [7]. These include, but are not limited to, electronic medical records, health information exchange systems, computerized provider order entry, clinical decision support systems, and the related hardware (excluding medical devices) and their integration with each other.

It has been repeatedly demonstrated that large health organizations often struggle to adopt high-quality and modern HIT solutions and are challenged with increasingly shorter innovation cycles of these technologies [10-15]. The fact that there is considerable variation in the adoption and quality of HIT between organizations within and across countries points to the importance of focusing on the organizations themselves in terms of their *inner* capabilities with regard to managerial skills, the promotion of HIT use, project execution, and innovation promotion [16-18]. Although a wide range of general facilitating factors of successful HIT adoption have been acknowledged in several theoretical frameworks [19-24] and various systematic literature reviews [3,12,25-28], little is known about the exact constituents of capabilities of health care organizations to innovate in particular and how they affect not only the adoption of HIT but also their quality. Insights about this relationship could prove valuable for guiding managers, policy makers, and educators toward promoting and developing organizational behavior that facilitates better HIT use, which in turn might lead to improved clinical outcomes [29].

HIT Quality and Innovation Capabilities

HIT adoption is most often understood as the implementation, that is, the introduction of an application, and its acceptance and use in an organization and many adoption studies focus on specific functionalities or applications [12,21,27]. However, the complexity of organization-wide HIT solutions is usually far greater and requires the incorporation of many different facets of the organization's information system [30-33]. In addition, when extending the scope from adoption to the quality of HIT, even more aspects need to be incorporated as quality requirements are typically considered to incorporate not only various technical layers (eg, data and information, functions, hardware, interoperability) to support clinical care processes but also features of information management and the perceived quality of the IT systems [17,23,34,35]. Thus, in our study, we focus on HIT quality rather than mere adoption and consider it to be composed of the following 3 principal domains: HIT information management, HIT workflow support, and perceived HIT quality:

- HIT information management encompasses the full spectrum of strategic, tactical, and operational management tasks to build and operate an organization's information system [34,36]. Management practices are deemed to be essential preconditions for realizing the potential of HIT [37], especially those executed by the information technology (IT) department [38,39] and those that involve systematic clinical user participation [40,41].
- HIT workflow support refers to the degree to which an organization has implemented the information technologies needed to support patient care processes. This encompasses the availability of electronic patient data across various care processes as discussed by Liebe et al [42], the availability of clinical applications (eg, electronic medical records, computerized provider order entry, and clinical decision support systems), their integration with one another, and accommodation of hardware solutions. This confluence of technical factors has been discussed as indicative of structural and process quality [17,43].
- Finally, HIT quality manifests itself not only in the technical quality of HIT but also in the subjective assessment of the implemented IT solutions that is hereinafter referred to as the perceived HIT quality.

In addition to HIT quality, there is also little understanding about the identification and effect of the organization's capabilities to innovate; however, as van Gemert-Pijnen et al [44] emphasize, many HIT innovations might fail as a result of disregarding the interdependencies between technology and its organizational and cultural environment. In our understanding, innovation capabilities (ICs) refer to the culture regarding HIT at various organizational levels that reflect its ability to innovate, that is, the ability to adopt new HIT solutions (or to renew the existing ones) that enhance the quality of information provision in clinical care processes. These capabilities refer to latent phenomena, that is, they are inherently difficult to capture, as they are expressions of a commonly shared attitude in social networks that leads to certain sets of corresponding behaviors [45,46]. In light of the semantic variations and inconsistent definitions of related phenomena, scholars have pointed to the need for further work to examine this construct and its measurement [47-49]. The lack of measurements also implies that there are few studies that provide empirically tested claims regarding the effect of an organization's ICs on HIT adoption or quality [49,50].

Conceptual Model and Study Objectives

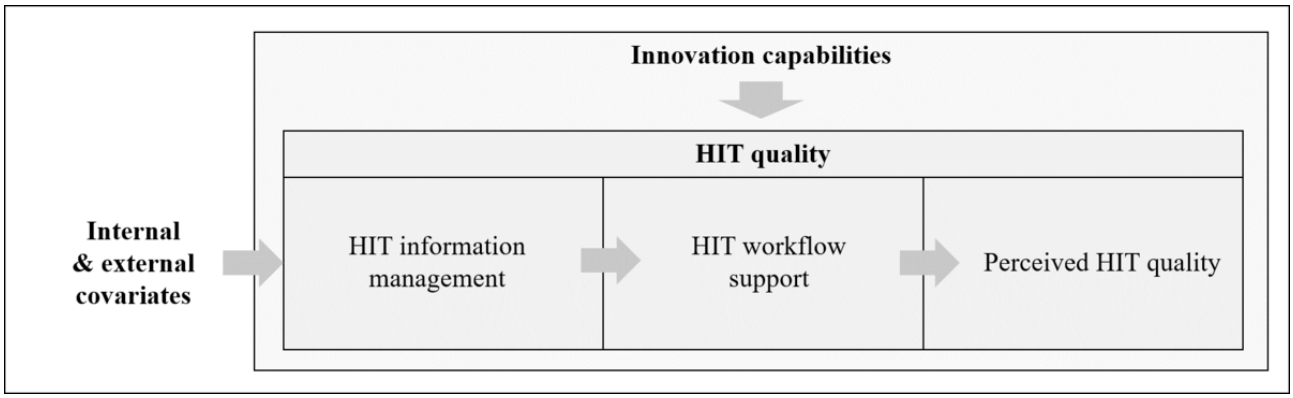
Only a few theoretical frameworks incorporate the peculiarities and complexity of organization-wide HIT solutions in a way that allows for an assessment of its quality and success [23,24,34]. Others acknowledge the facilitating role of domains comparable with ICs [19,20]; however, there is no framework

that puts the spotlight on the interrelationship between these 2 constructs and how they might enable better information provision in the care processes. Correspondingly, there is a need for validated measurement scales within such a framework to put its implicit hypotheses into the empirical test. Although some studies have begun to derive related scale sets [51-53], they are not yet ready to measure the full picture of the relationship between the 2 domains. In addition, the few that attempted to test more complex relationships between related constructs have limitations, particularly regarding small sample sizes and rather narrow outcome measures of HIT quality [54-56].

To investigate the sociotechnical interrelationships between ICs and HIT, we propose an initial conceptual model, that is, the IQ_{HIT} (innovation and quality of HIT) model (Figure 1). It rests on the underlying assumption of a directional process of antecedents and consequences of HIT as was similarly conceptualized in studies by Leidner et al [54] and Greenhalgh et al [57]. This is reflected in the assumption that HIT information management affects the degree of HIT workflow support that then determines the perceived HIT quality. Furthermore, these domains can be assumed to be influenced by an organization's ICs. In addition, internal structural characteristics such as the organization's size, teaching status, and ownership as well as external influences in terms of national health policies and legal regulations need to be accounted for as possible covariates in the model, as both have been shown to be significantly associated with HIT use [58-61].

On the basis of this model, the research goal was to empirically test and explain how health organizations' ICs are linked to HIT quality.

Figure 1. Initial conceptual innovation and quality of health information technology model of the layered relationship between innovation capabilities, health information technology quality, and covariates. HIT: health information technology.



Methods

Data Collection

Serving as empirical input for the model, data from chief information officers (CIOs) as hospital representatives were obtained. Hospitals are particularly interesting because of both the complexity of their IT and their organizational environment. We chose CIOs as our target group because they have the best oversight of the entirety of the IT landscape and top management issues [62,63]. We included Austrian, German, and Swiss hospitals in our target population to control for external influences in terms of different national health policies. The questionnaire and its constructs were based on the redevelopment and refinement of previous surveys and included a total of 188 question items (Multimedia Appendices 1 and 2) [64]. The final questionnaire was pretested by 5 hospital CIOs, 10 researchers (comprising health IT experts, statisticians, and 1 psychologist), and 1 clinician to evaluate whether the question items were understandable and answerable and whether they were sufficiently precise to measure the organization's information system. This led to some minor adjustments of item scales (response options), changes in the wording of items, and a few supplementary definitions.

Email addresses of 1669 CIOs were compiled through internet and telephone searches. The CIOs were responsible for 2324 hospitals (92% of all 2542 hospitals across Austria, Germany, and Switzerland). Data collection took place during the first half of 2017 via a web-based survey. Of the 1669 emails sent, 1499 had come through and 251 CIOs participated (17% response rate)—19 answers were discarded because of incompleteness (ie, the respondent did not finish the survey or sections were left out). The descriptive results were made available in 2018 [65], and as an incentive for participation, CIOs were offered access to a web-based benchmarking dashboard that allowed them to compare their hospital with peer groups [66].

Modeling and Data Analysis

We applied structural equation modeling (SEM) to test various interrelationships between constructs. Specifically, we chose partial least squares structural equation modeling as it is tolerant of the use of categorical data and allows for including reflective measurement models (ie, manifest indicators *reflect* the latent construct), formative measurement models (ie, manifest indicators *form* the latent construct), and single-item scales without identification problems [67].

Specification of the Measurement Models

We operationalized each of the 5 domains in the conceptual model (Figure 1), with a total of 10 constructs (Table 1). All items and scales associated with these constructs are detailed in Multimedia Appendices 1 and 2.

Table 1. Overview of the constructs used to operationalize the domains of the conceptual innovation and quality of health information technology model.

Domains	Constructs		
HIT[a] quality			
HIT information management	Professionalism of information management	Clinical IT[b] agents	N/A[c]
HIT workflow support	Workflow composite score including technical descriptors and care processes	N/A	N/A
Perceived HIT quality	Perceived HIT workflow support	Overall goodness of information provision	N/A
Innovation capabilities	Innovation capability: top management team support	Innovation capability of the information technology department	Organization-wide innovation capability
Covariates	Structural characteristics	Country	N/A

[a]HIT: health information technology.

[b]IT: information technology.

[c]N/A: not applicable.

HIT Quality

HIT information management was operationalized using 2 constructs. First, we applied a construct that captures the degree of professionalism of information management (PIM) in health care in terms of the regularity of 15 management key tasks and practices, as proposed by Thye et al [36]. As PIM consists of 3 latent and correlated subcomponents (strategic, tactical, and operational information management), we incorporated it as a reflective higher order model with PIM as the higher order construct and the 3 subcomponents as the lower order constructs using the repeated indicator approach [68]. Second, to reflect institutionalized user participation, we included the formal appointment of clinical IT agents as a reflective measurement model with 2 underlying items (one referring to physicians and the other one to nurses).

HIT workflow support can be theorized as being constituted by the descriptors data and information, IT functions, integration, and distribution of data and IT functions [17]. These 4 descriptors are the central building blocks of the Workflow Composite Score (WCS), an aggregated score that proved to be reliable and valid in measuring the degree of HIT supported patient care in core clinical processes [17,43,65]: ward rounds to reflect diagnostic and therapeutic decision making at the bedside, presurgery and postsurgery processes that reflect the information flow between departments, and admission and discharge as core interface processes between outpatient and inpatient care. The WCS comprises 146 items grouped along these 5 clinical processes and the 4 descriptors (Multimedia Appendix 2). We included it in the SEM analysis as a single-item scale, as its composite structure was largely predefined in previous studies [17,65].

Perceived HIT quality was measured using the 2 constructs. First, we asked the CIOs to grade the HIT workflow support (perceived HIT workflow support) in all 5 aforementioned clinical care processes separately and included the resulting indicators in a reflective measurement model. Second, we asked for a concluding assessment (single-item scale) of the overall goodness of information provision, that is, the organization's general ability to provide the right information, at the right time, at the right place, for the right persons, and in the right quality to support patient care processes. This indicator was applied in a previous study [38].

Innovation Capabilities

We investigated this domain and the underlying constructs across the 2 preceding surveys [38,52]. The initial exploratory study on this topic pointed to a latent construct, represented by 5 items that describe the top management team (TMT) support and the organization-wide innovation culture with regard to HIT [52]. A second study signified that the ICs relating to the IT department could be considered as another separate component [38]. To explore the emerging constructs in greater depth, we added 9 items to capture additional details on the TMT support and the organization-wide innovation culture and 6 additional items that refer to the IT department. An exploratory factor analysis using the unweighted least squares estimation and oblique factor rotation was computed, which resulted in a 3-factor structure that reflected ICs at the TMT level (IC TMT), ICs at the IT department level (IC ITD), and ICs at the organization-wide level (IC OW). For SEM, the underlying items were then included in 3 reflective measurement models. A total of 4 items with low outer loadings (<0.50) were removed to establish sufficient convergent and discriminant validity.

Covariates

A total of 2 covariates were included in the model. First, to control for well-known structural characteristics, we included a formative measurement model that was composed of the hospital size (bed count) and its teaching status. Second, the country was included as a single-item scale to account for external conditions. Austrian and Swiss hospitals were pooled to obtain more balanced group sizes.

Specification of the Structural Model

The specifications of the structural model resulted from a step-wise build-up of testing the direct and mediated effects along the components of the conceptual model. Each step was thereby rooted in findings from studies that suggest individual

linkages between the constructs, which we summarized into a set of 12 theoretical assumptions (Table 2). On the basis of these assumptions, we deduced one or more hypotheses for specifying the structural equation model paths.

Table 2. Theoretical assumptions and corresponding hypotheses guiding the structural model specification.

Assumption	Exemplary study
The PIM[a] might be linked to HIT[b] workflow support • H1: PIM has a positive effect on the WCS[c]	Ammenwerth et al (2006) [69], Avgar et al (2012) [70], Bradley et al (2012) [71], Winter et al (2011) [72]
Formal participation in terms of the appointment of clinical IT[d] agents might results from PIM practices and might lead to better HIT workflow support • H2: The effect of PIM on the WCS is partly mediated by clinical IT agents	Cresswell and Sheikh (2013) [12], Liebe et al (2018) [42], Potts et al (2011) [73], Sligo et al (2017) [26]
There likely is a direct link between the technical and the perceived quality of HIT workflow support • H3: The WCS has a positive effect on the perceived HIT workflow support • H4: The WCS has a positive effect on the overall goodness of information provision	Hadji and Degoulet (2016) [74], Hübner (2015) [75], Yusof et al (2008) [23]
The perceived quality of HIT is likely linked to the perceived goodness of information provision • H5: Perceived HIT workflow support has a positive effect on the overall goodness of information provision	Gorla et al (2010) [76], Suki (2012) [77]
A top management team that is capable and willing to innovate might facilitate an innovation-friendly culture throughout the organization, including the IT department • H6: Innovation capability: top management team support has a positive effect on organization-wide innovation capability • H7: Innovation capability: top management team support has a positive effect on the innovation capability of the IT department	Abdekhoda et al (2015) [78], Carpenter et al (2004) [79], Laukka et al [80]
The tasks and procedures that manifest PIM might also be facilitated by an innovation-friendly top management team • H8: Innovation capability: top management team support has a positive effect on PIM	Bradley et al (2012) [71], Liebe et al (2018) [81], Weintraub and McKee (2019) [82]
Innovation capabilities of the top management team and the IT department might determine the degree of HIT workflow support • H9: Innovation capability: top management team support has a positive effect on the WCS • H10: Innovation capability of the IT department has a positive effect on the WCS	Esdar et al (2018) [38], Paré et al (2020) [56], Leidner et al (2010) [54]
The ability of the IT department to innovate might be linked to information management practices • H11: Innovation capability of the IT department has a positive effect on PIM	Liebe et al (2017) [83], Watts and Henderson [84]
HIT quality might be a function of the organization-wide climate toward IT. Such climate might also facilitate a stronger effect of the technical HIT quality (ie, the WCS) on the perceived quality of information provision • H12: Organization-wide innovation capability has a positive effect on the WCS • H13: Organization-wide innovation capability has a positive effect on the perceived HIT workflow support • H14: Organization-wide innovation capability has a positive effect on the overall goodness of information provision • H15: Organization-wide innovation capability positively moderates the relationship between the WCS and the overall goodness of information provision	Caccia-Bava et al (2006) [45], Gagnon et al (2012) [85], Taylor et al (2015) [86], Vest et al (2019) [50]
Structural characteristics might be linked to HIT quality, possibly also to the TMT's capabilities to innovate • H16: Structural characteristics have a positive effect on the WCS • H17: Structural characteristics have a positive effect on PIM • H18: Structural characteristics have a positive effect on the innovation capability: top management team support	DesRoches et al (2012) [58], Fadol et al (2015) [87], Kruse et al (2014) [88], Troilo et al (2014) [89]
Compared with Germany, hospitals from Austria and Switzerland exhibit higher degrees of HIT workflow support and a more pronounced culture toward innovation • H19: Country has a positive effect on the WCS • H20: Country has a positive effect on organization-wide innovation capability • H21: Country has a positive effect on the innovation capability of the IT department	Esdar et al (2018) [38], Haux et al (2018) [90], Hübner et al (2010) [91], Hüsers et al (2017) [49], Naumann et al (2019) [11]

[a]PIM: professionalism of information management.

[b]HIT: health information technology.

[c]WCS: Workflow Composite Score.

[d]IT: information technology.

Parameter Estimations and Model Assessment

We applied partial least squares structural equation modeling using SmartPLS version 3 [92]. The measurement models were assessed for internal consistency using Cronbach α and composite reliability. Convergent and discriminant validity was evaluated according to the height of the outer loadings, the average variance extracted, the Fornell-Larcker criterion, and the Heterotrait-Monotrait ratio.

Inner variance inflation factor values were used to test for collinearity within the structural model. Path coefficients and mediation effects were evaluated based on the direct, total, and indirect effects as well as on f^2 effect sizes with *P* values and 95% CIs obtained from 10,000 bootstrap replications. Besides the R^2 values for the endogenous latent variables, we used blindfolding to obtain Stone-Geisser Q^2 values to determine the cross-validated predictive relevance of the exogenous constructs.

Results

Descriptive Statistics

The sample consisted of data from 232 hospitals, most of which were from Germany (Table 3), which corresponds to the higher baseline number of German hospitals. The participating hospitals were rather large, with an average size of 492 (SD 239) beds, and many (112/232, 48.3%) were in public ownership. Nevertheless, hospitals from all relevant demographic categories were included in the sample. The WCS, as the central measure of HIT workflow support in our model, showed an overall mean value of 56 (SD 14) points (ranging between 0 and 100 points; Multimedia Appendix 3). The mean values and SD of the remaining constructs are shown in Multimedia Appendix 1.

Table 3. Demographic characteristics of participating hospitals (N=232).

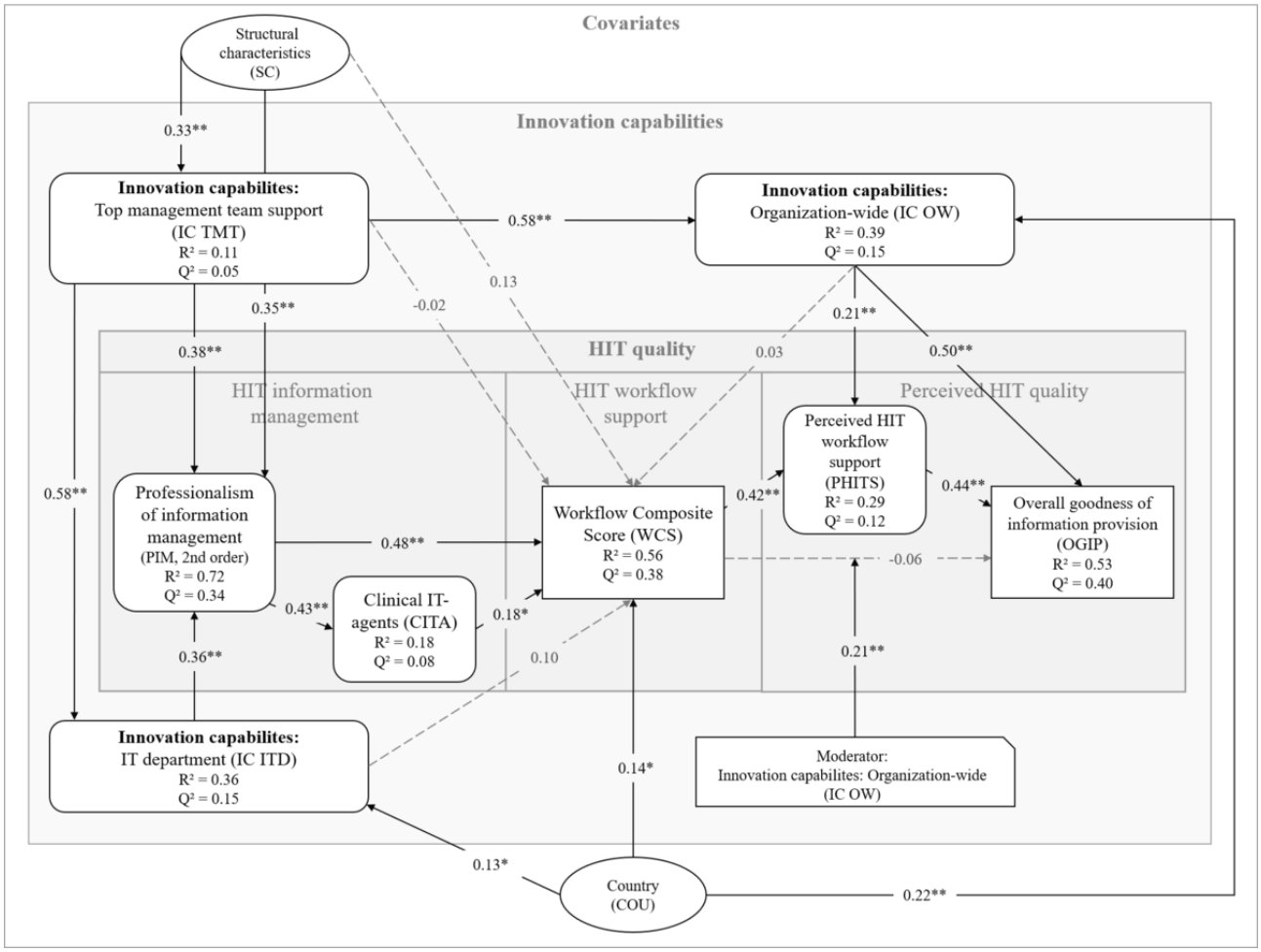
Characteristics	Value
Country, n (response rate in %)	
Austria	14 (8.8)
Germany	205 (18.3)
Switzerland	13 (11.3)
Ownership, n (% in sample)	
For-profit	42 (18.1)
Nonprofit	78 (33.6)
Public	112 (48.3)
Teaching status, n (% in sample)	
Major teaching hospital	22 (9.5)
Minor teaching hospital	101 (43.5)
Nonteaching hospital	109 (47.0)
Member of a hospital group, n (% in sample)	
Yes	140 (60.3)
No	92 (39.7)
Number of beds, mean (SD)	491.9 (238.5)

Structural Equation Model

The parameters assessing the measurement models pointed to valid specifications of the reflective models as well as the formative model in terms of convergent validity and internal consistency (Multimedia Appendices 4 and 5). In addition, sufficient discriminant validity was established according to the Fornell-Larcker criterion assessment, as indicated by the Heterotrait-Monotrait ratios of the correlations that were all below the recommended threshold value of 0.85 [93] (Multimedia Appendix 6). No collinearity was found in the structural model, as all the inner variance inflation factor values ranged within the limits of 0.20 and 4. Moreover, the Stone-Geisser Q^2 values of the endogenous variables indicate a good out-of-sample predictive power of the path model, especially with regard to the WCS (Q^2=0.38) and the overall goodness of information provision (Q^2=0.40).

The 21 hypotheses (Table 2) led to a variety of interrelationships in the structural model in terms of direct, mediated, and moderated effects. Approximately 50% of the variance in the key constructs for measuring HIT quality, the HIT workflow support (as measured by the WCS), and the perceived overall goodness of information provision (OGIP) could be explained by the model (Figure 2).

Figure 2. The structural model of innovation and quality of health information technology with path coefficients, explained variance (R^2), and predictive relevance measures (Q^2) of the endogenous constructs. Latent constructs are displayed with rounded edges, the exogenous covariates as ellipses and the moderator variable with a cut-off corner. *P<.05; **P<.01. HIT: health information technology.



Within the HIT quality domain, the results showed a strong effect of PIM on the WCS with a path coefficient estimate of 0.48 (P<.001). This association was partially mediated by the use of clinical IT agents to a small but significant extent (Multimedia Appendix 7). Furthermore, WCS was associated with OGIP via an indirect effect between the 2, which was mediated by the perceived HIT workflow support. The exact P values of the path coefficients are shown in Multimedia Appendix 8.

Within the innovation layer, the IC TMT exhibited a strong effect on IC ITD and IC OW.

Furthermore, the model revealed a strong association between innovation and quality at various levels (the total and indirect effects are given in Multimedia Appendix 7): the ICs of the TMT and of the IT department significantly and similarly affected PIM, whereas IC OW had a strong effect on the perceived HIT quality in terms of OGIP and a weaker but still significant effect related to perceived HIT workflow support. Contrary to some of our initial assumptions, as expressed in hypotheses H9, H10, and H12, there was no significant direct effect of any of the constructs representing IC on the WCS (Table 4). Instead, the results showed significant indirect effects of IC TMT and IC ITD on the WCS mediated by PIM (Multimedia Appendix 7). The effect of the WCS on OGIP, which did not become significant, was, however, significantly moderated by IC OW (hypothesis H15). In summary, ICs possessed many points of application at the HIT quality path, that is, at the beginning influencing PIM and later affecting the overall quality of information provision for patient care.

Table 4. Summarized results of the hypothesis tests in reference to *P* values <.05.

Hypothesis	Support by the model
H1: PIM[a] has a positive effect on the WCS[b]	Supported
H2: The effect of PIM on the WCS is partly mediated by clinical IT[c] agents	Supported
H3: The WCS has a positive effect on perceived HIT[d] workflow support	Supported
H4: The WCS has a positive effect on the overall goodness of information provision	Not supported
H5: Perceived HIT workflow support has a positive effect on the overall goodness of information provision	Supported
H6: Innovation capabilities: Top management team support has a positive effect on organization-wide innovation capability	Supported
H7: Innovation capabilities: Top management team support has a positive effect on the innovation capability of the IT department	Supported
H8: Innovation capabilities: Top management team support has a positive effect on PIM	Supported
H9: Innovation capabilities: Top management team support has a positive effect on the WCS	Not supported
H10: Innovation capability of the IT department has a positive effect on the WCS	Not supported
H11: Innovation capability of the IT department has a positive effect on PIM	Supported
H12: Organization-wide innovation capability has a positive effect on the WCS	Not supported
H13: Organization-wide innovation capability has a positive effect on perceived HIT workflow support	Supported
H14: Organization-wide innovation capability has a positive effect on the overall goodness of information provision	Supported
H15: Organization-wide innovation capability positively moderates the relationship between the WCS and the overall goodness of information provision	Supported
H16: Structural characteristics have a positive effect on the WCS	Not supported
H17: Structural characteristics have a positive effect on PIM	Supported
H18: Structural characteristics have a positive effect on innovation capabilities: top management team support	Supported
H19: Country has a positive effect on the WCS	Supported
H20: Country has a positive effect on the organization-wide innovation capability	Supported
H21: Country has a positive effect on the innovation capability of the IT department	Supported

[a]PIM: professionalism of information management.

[b]WCS: Workflow Composite Score.

[c]IT: information technology.

[d]HIT: health information technology.

With regard to the covariates, the country had a significant effect on the WCS and was also associated with higher degrees of IC ITD and IC TMT, albeit with rather small effect sizes f^2 (Multimedia Appendix 8). The organization's structural characteristics did not exhibit a direct effect on the WCS in our model but instead on the *preceding* latent variables in the model, namely, PIM and IC TMT.

Discussion

Principal Findings

On the basis of data from 232 hospitals in Austria, Germany, and Switzerland, a sociotechnical IQ_{HIT} model was developed and tested. To the best of our knowledge, this is the first model that investigates HIT quality in light of the organizations' ability to innovate. It does so in a strictly empirical manner using a validated instrument. The model sets out the internal composition of HIT quality in establishing a consecutive connection between HIT information management, HIT workflow support, and perceived quality. Furthermore, an organization's ICs were positively associated with HIT quality at various levels. Most notably, an innovation-friendly attitude on the TMT level appeared to strongly but indirectly facilitate HIT-based workflow support, mediated by professional information management practices.

The Inner Workings of HIT Quality

At the core of the IQ_{HIT} model, the WCS was used to measure HIT quality in terms of the workflow to support the IT solutions provided for improving patient care. The WCS is a multifaceted indicator that consists of a plethora of underlying items (Multimedia Appendix 2). By incorporating it, the model considers the complexity of interdepartmental and multifunctional health information systems.

According to the model, HIT workflow support depends on professional information management, that is, professionally conceptualized and performed activities at the strategic, tactical, and operational levels, as has been conjectured by Winter et al [34] and empirically conceptualized by Thye et al [36]. Only the HIT workflow support that is managed in an orderly and

professional manner by the IT department can work properly regarding data and information provision, IT functions in place, their integration with one another, and the ability to distribute the data and the information to the point of care. Part of this effect is mediated by the presence of clinical IT agents, confirming the importance of establishing a formal link between IT department information management and clinical end users. Interestingly, the structural characteristics (bed count and teaching status) did not affect the HIT workflow support directly but only via the mediating effect of professional information management. This is rather surprising, as most studies suggest a direct link, particularly between the size of an organization and its HIT use [25].

HIT quality was conceptualized to encompass both, a technical component that bundles manifest, self-reported attributes about the information system, that is, the WCS, and a subjective judgment about its perceived quality. According to the CIOs' viewpoint, the very abstract judgment of the perceived goodness of information provision appears to not be directly linked to the WCS but requires some intermediate interpretation, that is, the perceived HIT workflow support, which refers to a more detailed perspective of admission, ward rounds, presurgery and postsurgery, and discharge processes. This also suggests that there is no strict automatism between a high degree of HIT quality in terms of its technical components and the perceived quality of information provision in an organization. This points to the need for good implementation practices of HIT interventions to successfully reap their benefits.

Innovation Capabilities in Health Care Organizations

The IQ_{HIT} model also specifies the inner fabric of organizational IC. The underlying scales yielded good psychometric properties and reflected an innovation-friendly attitude and behavior at different organizational levels: at the executive level (IC TMT), the items mirror the motivational and monetary support of the TMT for IT innovation and their proactive engagement with respective projects as part of the organization's vision. Similar to the views in the Upper Echelons Theory, which stresses the crucial role of senior leadership in fostering innovation, this factor had a strong predictive relevance across the model [94]. IC ITD reflect the kind of CIO leadership that facilitates creativity, communication, and participation of end users. On the third level (IC OW), openness and widespread flexibility for embracing new IT solutions that prevail throughout the organization at large were the defining elements. Most of these characteristics were suspected [47,95,96] and partly known [27,97,98] to facilitate innovation in a variety of contexts; however, the way they statistically cluster along different organizational levels and their different effects has not been specified before. Therefore, the innovative capacity of health organizations cannot be viewed as monolithic blocks or mere buzzwords. Its contents are woven throughout various organizational levels to varying degrees.

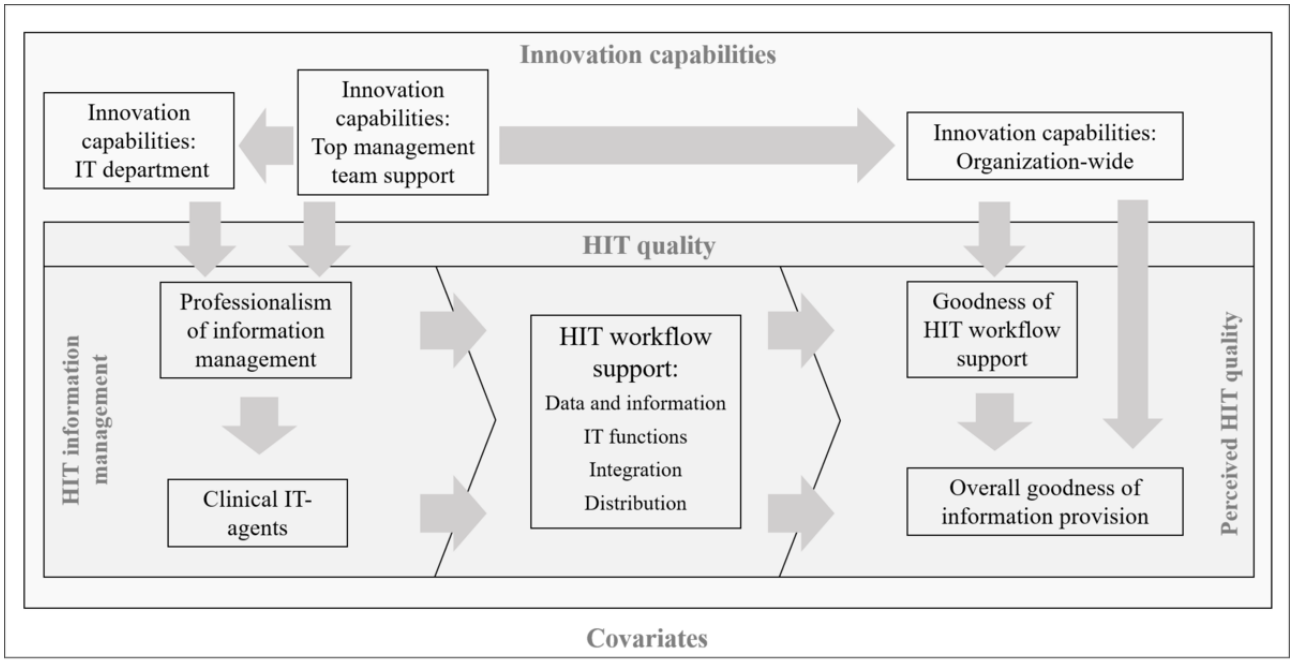
This study did not explicitly focus on how these capabilities can be built or how they are determined. However, when controlling for the covariates, we found that TMT support is a function of certain structural characteristics, namely, a higher bed count and teaching status, both of which can be interpreted as indicators of greater financial flexibility in terms of slack resources. However, ICs at the IT department and the organization at large depend on the respective country. More precisely, these 2 domains are more pronounced in Austria and Switzerland than in Germany, which corresponds well with previous findings on different samples [11,49].

HIT Implementation Between Innovation and Quality

Traditionally, empirical research conducted on HIT quality has frequently disregarded aspects of innovation, and both have often been discussed separately from one another [75]. Our model establishes a connection between the two by showing that attaining high levels of HIT quality is facilitated and mediated by an organization's ability to create space for creativity, agility, and communication in relation to IT-based innovation.

Overall, the structural model (Figure 2) can be translated into a more schematic model (Figure 3) based on the major findings. It shows that PIM mediates the effects of the 2 IC domains—IC TMT and IC ITD—on HIT workflow support, which illustrates the interplay between the *right attitude* toward innovation and formalized management practices for innovational strength. The attitude and intent to innovate play an important role in and of itself; however, professional information management is needed for the practical execution of this intention to improve HIT workflow support.

Figure 3. The innovation and quality of health information technology model. HIT: health information technology; IT: information technology.



Furthermore, we found IC OW to partly moderate the relationship between the HIT workflow support and the perceived OGIP, implying that there might actually be a direct effect between the 2 as long as the organization is agile, flexible, and open toward IT (equals high levels of IC OW). This could be interpreted as an indication that an organization-wide positive attitude toward using the IT in place, irrespective of how advanced it actually is, leads to better information provision in the clinical care processes, at least from the vantage point of CIOs. Overall, it becomes clear that ICs are not only needed at the TMT level but also at the IT department level and throughout the organization to establish high-quality HIT solutions. Executive managers and policy makers should therefore consider how to establish higher levels of these capabilities at various levels.

Limitations

Our study had several limitations. Most notably, this is an observational study, and despite the statistical specifications that might suggest otherwise, it cannot be inferred that the relationship between constructs is truly causal. For instance, there might be temporal displacements between the current beliefs of executives and higher degrees of HIT quality as implementation processes take time [99].

Furthermore, this sociotechnical model reflects the perspective of the CIOs and their points of view of the HIT cosmos and ICs. This is both a strength and a weakness. The strength is its consistency and authenticity regarding technical and organizational issues related to IT. Its weakness is the CIOs cannot accurately evaluate clinicians' view on the timely and correct provision of data and information (ie, the *right side* of the model), which requires a more detailed assessment in future research. Ultimately, the clinical outcome is the improvement or stabilization of the patient's condition. None of this is captured in this model, as it mirrors the vantage point of CIOs. The next step will be to develop a model that incorporates the views of physicians and nurses. This approach can also cope with potential common-method biases. The sample is also based on voluntary participation, which is why we cannot rule out a nonresponse bias in the data.

Finally, not all possibly relevant factors at play can be accurately accounted for in one model, which is reflected by the R^2 values that leave parts of the variance in the endogenous constructs unexplained. Given these limitations, further studies are needed to validate and differentiate the relationships between and within IC and HIT quality, and our model provides various access points to do so.

Conclusions

On the basis of survey data provided by the CIOs of 232 hospitals, we proposed a sociotechnical IQ_{HIT} model to explain how organizational innovation relates to various facets of HIT quality. Although some associations in the model could be presumed by the literature, it clearly and uniquely highlights the key role of ICs and information management for HIT-based workflow support. Thus, it demonstrates that innovation and quality do not contradict each other. In particular, an innovation-friendly attitude of TMT and the IT department determines the degree of HIT workflow support, albeit not directly, but by means of professional information management practices that eventually facilitate the perceived goodness of information provision for patient care. This suggests that managers of health organizations should strive for both a more pronounced culture toward innovation and professional information management to translate such a culture into HIT quality. Furthermore, the IQ_{HIT} model might be useful for studies on HIT adoption and diffusion and for the definition of HIT maturity models. To this end, it provides validated measurement scales that can be utilized in future research.

Acknowledgments

This study was funded by the State of Lower Saxony, Germany (grant number ZN 3103).

Authors' Contributions

UH initiated the study; ME, UH, and JL specified the theoretical framework, supported by BB, and conceptualized the research design. All authors were responsible for the scale development. ME, JL, and JT collected and prepared the data, supervised by UH, and assisted by several members of the Health Informatics Research Group of the University of Applied Sciences Osnabrück. ME specified the model and conducted the statistical analysis. ME and UH wrote the manuscript that was reviewed and approved by all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Measurement models and underlying items (questionnaire part A).
[DOCX File , 22 KB - medinform_v9i3e23306_app1.docx]

Multimedia Appendix 2
Workflow Composite Score (WCS): structure and underlying items (questionnaire part B).
[XLSX File (Microsoft Excel File), 22 KB - medinform_v9i3e23306_app2.xlsx]

Multimedia Appendix 3
Descriptive statistics of the Workflow Composite Score (WCS; n=232).
[DOCX File , 14 KB - medinform_v9i3e23306_app3.docx]

Multimedia Appendix 4
Convergent validity and internal consistency of the measurement models with bias corrected 95% CIs.
[DOCX File , 17 KB - medinform_v9i3e23306_app4.docx]

Multimedia Appendix 5
Convergent validity and internal consistency of lower order constructs reflecting the latent variable “Professionalism of Information Management (PIM)” with bias corrected 95% CIs.
[DOCX File , 16 KB - medinform_v9i3e23306_app5.docx]

Multimedia Appendix 6
Heterotrait-Monotrait (HTMT) ratios.
[DOCX File , 15 KB - medinform_v9i3e23306_app6.docx]

Multimedia Appendix 7
Total effects and total indirect effects of the structural model with bias corrected 95% CIs and significance tests of the path coefficients.
[DOCX File , 17 KB - medinform_v9i3e23306_app7.docx]

Multimedia Appendix 8
Direct path coefficients with bias corrected 95% CIs, significance tests of path coefficients, and their effect sizes.
[DOCX File , 16 KB - medinform_v9i3e23306_app8.docx]

References

1. Thouin MF, Hoffman JJ, Ford EW. The effect of information technology investment on firm-level performance in the health care industry. Health Care Manage Rev 2008;33(1):60-68. [doi: 10.1097/01.HMR.0000304491.03147.06] [Medline: 18091445]
2. Driessen J, Cioffi M, Alide N, Landis-Lewis Z, Gamadzi G, Gadabu OJ, et al. Modeling return on investment for an electronic medical record system in Lilongwe, Malawi. J Am Med Inform Assoc 2013;20(4):743-748 [FREE Full text] [doi: 10.1136/amiajnl-2012-001242] [Medline: 23144335]

3. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
4. Yen PY, McAlearney AS, Sieck CJ, Hefner JL, Huerta TR. Health information technology (HIT) adaptation: refocusing on the journey to successful HIT implementation. JMIR Med Inform 2017 Sep 07;5(3):e28 [FREE Full text] [doi: 10.2196/medinform.7476] [Medline: 28882812]
5. Cresswell KM, Sheikh A. Health information technology in hospitals: current issues and future trends. Future Hosp J 2015 Feb;2(1):50-56 [FREE Full text] [doi: 10.7861/futurehosp.2-1-50] [Medline: 31098079]
6. Desveaux L, Soobiah C, Bhatia RS, Shaw J. Identifying and overcoming policy-level barriers to the implementation of digital health innovation: qualitative study. J Med Internet Res 2019 Dec 20;21(12):e14994 [FREE Full text] [doi: 10.2196/14994] [Medline: 31859679]
7. Kruse CS, Beane A. Health information technology continues to show positive effect on medical outcomes: systematic review. J Med Internet Res 2018 Feb 05;20(2):e41 [FREE Full text] [doi: 10.2196/jmir.8793] [Medline: 29402759]
8. Lin SC, Jha AK, Adler-Milstein J. Electronic health records associated with lower hospital mortality after systems have time to mature. Health Aff (Millwood) 2018 Jul;37(7):1128-1135. [doi: 10.1377/hlthaff.2017.1658] [Medline: 29985687]
9. Plantier M, Havet N, Durand T, Caquot N, Amaz C, Philip I, et al. Does adoption of electronic health records improve organizational performances of hospital surgical units? Results from the French e-SI (PREPS-SIPS) study. Int J Med Inform 2017 Feb;98:47-55. [doi: 10.1016/j.ijmedinf.2016.12.002] [Medline: 28034412]
10. Asthana S, Jones R, Sheaff R. Why does the NHS struggle to adopt eHealth innovations? A review of macro, meso and micro factors. BMC Health Serv Res 2019 Dec 21;19(1):984 [FREE Full text] [doi: 10.1186/s12913-019-4790-x] [Medline: 31864370]
11. Naumann L, Esdar M, Ammenwerth E, Baumberger D, Hübner U. Same goals, yet different outcomes: analysing the current state of eHealth adoption and policies in Austria, Germany, and Switzerland using a mixed methods approach. Stud Health Technol Inform 2019 Aug 21;264:1012-1016. [doi: 10.3233/SHTI190377] [Medline: 31438077]
12. Cresswell K, Sheikh A. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. Int J Med Inform 2013 May;82(5):e73-e86. [doi: 10.1016/j.ijmedinf.2012.10.007] [Medline: 23146626]
13. Kim YG, Jung K, Park YT, Shin D, Cho SY, Yoon D, et al. Rate of electronic health record adoption in South Korea: a nation-wide survey. Int J Med Inform 2017 May;101:100-107. [doi: 10.1016/j.ijmedinf.2017.02.009] [Medline: 28347440]
14. Piening EP. Insights into the process dynamics of innovation implementation. Pub Manage Rev 2011 Jan;13(1):127-157. [doi: 10.1080/14719037.2010.501615]
15. Colicchio TK, Cimino JJ, Del Fiol G. Unintended consequences of nationwide electronic health record adoption: challenges and opportunities in the post-meaningful use era. J Med Internet Res 2019 Jun 03;21(6):e13313 [FREE Full text] [doi: 10.2196/13313] [Medline: 31162125]
16. Sabes-Figuera R, Maghiros I. European hospital survey: benchmarking deployment of e-Health services (2012-2013). Luxembourg: Publications Office of the European Union; 2013:978-992.
17. Liebe JD, Hübner U, Straede MC, Thye J. Developing a workflow composite score to measure clinical information logistics. A top-down approach. Methods Inf Med 2015;54(5):424-433. [doi: 10.3414/ME14-02-0025] [Medline: 26419492]
18. Martin G, Clarke J, Liew F, Arora S, King D, Aylin P, et al. Evaluating the impact of organisational digital maturity on clinical outcomes in secondary care in England. NPJ Digit Med 2019;2:41 [FREE Full text] [doi: 10.1038/s41746-019-0118-9] [Medline: 31304387]
19. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
20. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
21. Rogers EM. Diffusion of innovations. Fifth edition. New York: Free Press; 2003.
22. Tornatzky LG, Fleischer M. The processes of technological innovation. Maryland, United States: Lexington Books; 1990.
23. Yusof MM, Kuljis J, Papazafeiropoulou A, Stergioulas LK. An evaluation framework for Health Information Systems: human, organization and technology-fit factors (HOT-fit). Int J Med Inform 2008 Jun;77(6):386-398. [doi: 10.1016/j.ijmedinf.2007.08.011] [Medline: 17964851]
24. Lau F, Price M, Keshavjee K. From benefits evaluation to clinical adoption: making sense of health information system success in Canada. Healthc Q 2011;14(1):39-45. [doi: 10.12927/hcq.2011.22157] [Medline: 21301238]
25. Kruse CS, Kothman K, Anerobi K, Abanaka L. Adoption factors of the electronic health record: a systematic review. JMIR Med Inform 2016 Jun 01;4(2):e19 [FREE Full text] [doi: 10.2196/medinform.5525] [Medline: 27251559]
26. Sligo J, Gauld R, Roberts V, Villa L. A literature review for large-scale health information system project planning, implementation and evaluation. Int J Med Inform 2017 Jan;97:86-97. [doi: 10.1016/j.ijmedinf.2016.09.007] [Medline: 27919399]

27. Ben-Zion R, Pliskin N, Fink L. Critical success factors for adoption of electronic health record systems: literature review and prescriptive analysis. Info Sys Manage 2014 Oct 28;31(4):296-312. [doi: 10.1080/10580530.2014.958024]
28. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of eHealth services: systematic literature analysis. J Med Internet Res 2019 Nov 22;21(11):e14197 [FREE Full text] [doi: 10.2196/14197] [Medline: 31755869]
29. Brenner SK, Kaushal R, Grinspan Z, Joyce C, Kim I, Allard RJ, et al. Effects of health information technology on patient outcomes: a systematic review. J Am Med Inform Assoc 2016 Sep;23(5):1016-1036 [FREE Full text] [doi: 10.1093/jamia/ocv138] [Medline: 26568607]
30. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. Int J Med Inform 2008 May;77(5):291-304. [doi: 10.1016/j.ijmedinf.2007.09.001] [Medline: 17951106]
31. Everson J, Lee SD, Friedman CP. Reliability and validity of the American Hospital Association's national longitudinal survey of health information technology adoption. J Am Med Inform Assoc 2014 Oct;21(e2):e257-e263 [FREE Full text] [doi: 10.1136/amiajnl-2013-002449] [Medline: 24623194]
32. Pettit L. Understanding EMRAM and how it can be used by policy-makers, hospital CIOs and their IT teams. World Hosp Health Serv 2013;49(3):7-9. [Medline: 24377140]
33. Carvalho JV, Rocha A, Abreu A. Maturity models of healthcare information systems and technologies: a literature review. J Med Syst 2016 Jun;40(6):131. [doi: 10.1007/s10916-016-0486-5] [Medline: 27083575]
34. Winter A, Takabayashi K, Jahn F, Kimura E, Engelbrecht R, Haux R, et al. Quality requirements for electronic health record systems. A Japanese-German information management perspective. Methods Inf Med 2017 Aug 07;56(7):e92-e104 [FREE Full text] [doi: 10.3414/ME17-05-0002] [Medline: 28925415]
35. Ammenwerth E, Ehlers F, Hirsch B, Gratl G. HIS-Monitor: an approach to assess the quality of information processing in hospitals. Int J Med Inform 2007;76(2-3):216-225. [doi: 10.1016/j.ijmedinf.2006.05.004] [Medline: 16777476]
36. Thye J, Esdar M, Liebe J, Jahn F, Winter A, Hübner U. Professionalism of information management in health care: development and validation of the construct and its measurement. Methods Inf Med 2020 Jun;59(S 01):e1-e12 [FREE Full text] [doi: 10.1055/s-0040-1712465] [Medline: 32620017]
37. Adler-Milstein J, Woody Scott K, Jha AK. Leveraging EHRs to improve hospital performance: the role of management. Am J Manag Care 2014 Nov;20(11 Spec No. 17):SP511-SP519 [FREE Full text] [Medline: 25811825]
38. Esdar M, Liebe J, Babitsch B, Hübner U. Determinants of clinical information logistics: tracing socio-organisational factors and country differences from the perspective of clinical directors. Stud Health Technol Inform 2018;253:143-147. [Medline: 30147060]
39. Haux R. Health information systems - past, present, future. Int J Med Inform 2006;75(3-4):268-281. [doi: 10.1016/j.ijmedinf.2005.08.002] [Medline: 16169771]
40. Ingebrigtsen T, Georgiou A, Clay-Williams R, Magrabi F, Hordern A, Prgomet M, et al. The impact of clinical leadership on health information technology adoption: systematic review. Int J Med Inform 2014 Jun;83(6):393-405. [doi: 10.1016/j.ijmedinf.2014.02.005] [Medline: 24656180]
41. Pagliari C. Design and evaluation in eHealth: challenges and implications for an interdisciplinary field. J Med Internet Res 2007 May 27;9(2):e15 [FREE Full text] [doi: 10.2196/jmir.9.2.e15] [Medline: 17537718]
42. Liebe JD, Esdar M, Hübner U. Measuring the availability of electronic patient data across the hospital and throughout selected clinical workflows. Stud Health Technol Inform 2018;253:99-103. [Medline: 30147050]
43. Esdar M, Hübner U, Liebe J, Hüsers J, Thye J. Understanding latent structures of clinical information logistics: a bottom-up approach for model building and validating the workflow composite score. Int J Med Inform 2017 Jan;97:210-220. [doi: 10.1016/j.ijmedinf.2016.10.011] [Medline: 27919379]
44. van Gemert-Pijnen JEWC, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 05;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
45. Caccia-Bava MDC, Guimaraes T, Harrington SJ. Hospital organization culture, capacity to innovate and success in technology adoption. J Health Organ Manag 2006;20(2-3):194-217. [doi: 10.1108/14777260610662735] [Medline: 16869354]
46. Luu TT, Venkatesh S. Organizational culture and technological innovation adoption in private hospitals. Int Busi Res 2010 Jun 11;3(3):144. [doi: 10.5539/ibr.v3n3p144]
47. Wisdom JP, Chor KHB, Hoagwood KE, Horwitz SM. Innovation adoption: a review of theories and constructs. Adm Policy Ment Health 2014 Jul;41(4):480-502 [FREE Full text] [doi: 10.1007/s10488-013-0486-4] [Medline: 23549911]
48. Allen JD, Towne SD, Maxwell AE, DiMartino L, Leyva B, Bowen DJ, et al. Meausures of organizational characteristics associated with adoption and/or implementation of innovations: a systematic review. BMC Health Serv Res 2017 Aug 23;17(1):591 [FREE Full text] [doi: 10.1186/s12913-017-2459-x] [Medline: 28835273]
49. Hüsers J, Hübner U, Esdar M, Ammenwerth E, Hackl WO, Naumann L, et al. Innovative power of health care organisations affects IT adoption: a bi-national health IT benchmark comparing Austria and Germany. J Med Syst 2017 Feb;41(2):33. [doi: 10.1007/s10916-016-0671-6] [Medline: 28054195]

50. Vest JR, Jung H, Wiley K, Kooreman H, Pettit L, Unruh MA. Adoption of health information technology among US nursing facilities. J Am Med Dir Assoc 2019 Aug;20(8):995-1000 [FREE Full text] [doi: 10.1016/j.jamda.2018.11.002] [Medline: 30579920]
51. Fernandez ME, Walker TJ, Weiner BJ, Calo WA, Liang S, Risendal B, et al. Developing measures to assess constructs from the Inner Setting domain of the Consolidated Framework for Implementation Research. Implement Sci 2018 Mar 27;13(1):52 [FREE Full text] [doi: 10.1186/s13012-018-0736-7] [Medline: 29587804]
52. Esdar M, Liebe J, Weiß JP, Hübner U. Exploring innovation capabilities of hospital CIOs: an empirical assessment. Stud Health Technol Inform 2017;235:383-387. [Medline: 28423819]
53. Tomlins JC. Is it possible for the NHS to become fully digital? In: Cons Informatics and Digi Health. Amsterdam: Springer; 2019:359-374.
54. Leidner DE, Preston D, Chen D. An examination of the antecedents and consequences of organizational IT innovation in hospitals. J Strateg Info Sys 2010 Sep;19(3):154-170. [doi: 10.1016/j.jsis.2010.07.002]
55. Faber S, van Geenhuizen M, de Reuver M. eHealth adoption factors in medical hospitals: a focus on the Netherlands. Int J Med Inform 2017 Apr;100:77-89. [doi: 10.1016/j.ijmedinf.2017.01.009] [Medline: 28241940]
56. Paré G, Guillemette MG, Raymond L. IT centrality, IT management model, and contribution of the IT function to organizational performance: a study in Canadian hospitals. Info and Manage 2020 Apr;57(3). [doi: 10.1016/j.im.2019.103198]
57. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q 2004;82(4):581-629 [FREE Full text] [doi: 10.1111/j.0887-378X.2004.00325.x] [Medline: 15595944]
58. DesRoches CM, Worzala C, Joshi MS, Kralovec PD, Jha AK. Small, nonteaching, and rural hospitals continue to be slow in adopting electronic health record systems. Health Aff (Millwood) 2012 May;31(5):1092-1099. [doi: 10.1377/hlthaff.2012.0153] [Medline: 22535503]
59. Liebe JD, Egbert N, Frey A, Hübner U. Characteristics of German hospitals adopting health IT systems - results from an empirical study. Stud Health Technol Inform 2011;169:335-338. [Medline: 21893768]
60. Zhang NJ, Seblega B, Wan T, Unruh L, Agiro A, Miao L. Health information technology adoption in U.S. acute care hospitals. J Med Syst 2013 Apr;37(2):9907. [doi: 10.1007/s10916-012-9907-2] [Medline: 23340826]
61. Esdar M, Hüsers J, Weiß JP, Rauch J, Hübner U. Diffusion dynamics of electronic health records: a longitudinal observational study comparing data from hospitals in Germany and the United States. Int J Med Inform 2019 Nov;131:103952. [doi: 10.1016/j.ijmedinf.2019.103952] [Medline: 31557699]
62. Szydlowski S, Smith C. Perspectives from nurse leaders and chief information officers on health information technology implementation. Hosp Top 2009;87(1):3-9. [doi: 10.3200/HTPS.87.1.3-9] [Medline: 19103582]
63. Smaltz DH, Sambamurthy V, Agarwal R. The antecedents of CIO role effectiveness in Organizations: an empirical study in the healthcare sector. IEEE Trans Eng Manage 2006 May;53(2):207-222. [doi: 10.1109/tem.2006.872248]
64. IT-Report Healthcare 2020. Hübner U. URL: https://www.hs-osnabrueck.de/de/it-report-gesundheitswesen/ [accessed 2020-05-04]
65. Hübner U, Esdar M, Hüsers J, Liebe J, Rauch J, Thye J, et al. Status of digitization and the use of technology in German hospitals. In: Hospital Report 2019. Berlin, Heidelberg: Springer; 2019:33-48.
66. Weiß JP, Thye J, Rauch J, Tissen M, Esdar M, Teuteberg F, et al. IT benchmarking as an interaction between science and practice - a web portal for the dissemination of individual results for hospitals. In: Proceedings of the Multi-Conference of Information Systems 2018. 2018 Presented at: Multi-Conference of Information Systems 2018; March 2018; Lüneburg URL: https://tinyurl.com/9v4n4pxn
67. Hair J, Hult GTM, Ringle CM, Sarstedt M. A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). Thousand Oaks, California, United States: Sage Publications Inc; 2017.
68. Becker JM, Klein K, Wetzels M. Hierarchical latent variable models in PLS-SEM: guidelines for using reflective-formative type models. Long Range Planning 2012 Oct;45(5-6):359-394. [doi: 10.1016/j.lrp.2012.10.001]
69. Ammenwerth E, Iller C, Mahler C. IT-adoption and the interaction of task, technology and individuals: a fit framework and a case study. BMC Med Inform Decis Mak 2006 Jan 09;6:3 [FREE Full text] [doi: 10.1186/1472-6947-6-3] [Medline: 16401336]
70. Avgar AC, Litwin AS, Pronovost PJ. Drivers and barriers in health IT adoption: a proposed framework. Appl Clin Inform 2012;3(4):488-500 [FREE Full text] [doi: 10.4338/ACI-2012-07-R-0029] [Medline: 23646093]
71. Bradley RV, Byrd TA, Pridmore JL, Thrasher E, Pratt RM, Mbarika VW. An empirical examination of antecedents and consequences of IT governance in us hospitals. J Info Tech 2012 Jun 01;27(2):156-177. [doi: 10.1057/jit.2012.3]
72. Winter A, Gardner RM. Health information systems: architectures and strategies. 2nd ed. Switzerland: Springer; 2011.
73. Potts HW, Keen J, Denby T, Featherstone I, Patterson D, Anderson J, et al. Towards a better understanding of delivering e-health systems: a systematic review using the meta-narrative method and two case studies: final report 2011. URL: http://www.netscc.ac.uk/hsdr/files/project/SDO_FR_08-1602-131_V01.pdf [accessed 2021-02-17]
74. Hadji B, Degoulet P. Information system end-user satisfaction and continuance intention: a unified modeling approach. J Biomed Inform 2016 Jun;61:185-193 [FREE Full text] [doi: 10.1016/j.jbi.2016.03.021] [Medline: 27033175]

75. Hübner U. What are complex eHealth innovations and how do you measure them? Position paper. Methods Inf Med 2015;54(4):319-327. [doi: 10.3414/ME14-05-0001] [Medline: 25510406]
76. Gorla N, Somers TM, Wong B. Organizational impact of system quality, information quality, and service quality. J Strateg Info Sys 2010 Sep;19(3):207-228. [doi: 10.1016/j.jsis.2010.05.001]
77. Suki NM. Correlations of perceived flow, perceived system quality, perceived information quality, and perceived user trust on mobile social networking service (SNS) users' loyalty. J Inf Technol Res 2012;5(2):1-14. [doi: 10.4018/jitr.2012040101]
78. Abdekhoda M, Ahmadi M, Gohari M, Noruzi A. The effects of organizational contextual factors on physicians' attitude toward adoption of Electronic Medical Records. J Biomed Inform 2015 Feb;53:174-179 [FREE Full text] [doi: 10.1016/j.jbi.2014.10.008] [Medline: 25445481]
79. Carpenter MA, Geletkanycz MA, Sanders WG. Upper echelons research revisited: antecedents, elements, and consequences of top management team composition. J Manage 2016 Jun 23;30(6):749-778. [doi: 10.1016/j.jm.2004.06.001]
80. Laukka E, Huhtakangas M, Heponiemi T, Kanste O. Identifying the roles of healthcare leaders in hit implementation: a scoping review of the quantitative and qualitative evidence. Int J Environ Res Public Health 2020 Apr 21;17(8) [FREE Full text] [doi: 10.3390/ijerph17082865] [Medline: 32326300]
81. Liebe JD, Esdar M, Thye J, Hübner UH. Auf dem Weg zum digitalen Krankenhausine empirische Analyse über die gemeinsame Wirkung von Intrapreneurship und Informationsmanagement. In: Proceedings of the conference : Data driven X — Turning Data into Value. 2018 Presented at: Data driven X — Turning Data into Value; March 2018; Leuphana Universität Lüneburg.
82. Weintraub P, McKee M. Leadership for innovation in healthcare: an exploration. Int J Health Policy Manag 2019 Mar 01;8(3):138-144 [FREE Full text] [doi: 10.15171/ijhpm.2018.122] [Medline: 30980629]
83. Liebe JD, Esdar M, Thye J, Hübner U. Antecedents of CIOs' innovation capability in hospitals: results of an empirical study. Stud Health Technol Inform 2017;243:142-146. [Medline: 28883188]
84. Watts S, Henderson JC. Innovative IT climates: CIO perspectives. J Strateg Info Sys 2006 Jun;15(2):125-151. [doi: 10.1016/j.jsis.2005.08.001]
85. Gagnon MP, Desmartis M, Labrecque M, Car J, Pagliari C, Pluye P, et al. Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals. J Med Syst 2012 Feb;36(1):241-277 [FREE Full text] [doi: 10.1007/s10916-010-9473-4] [Medline: 20703721]
86. Taylor N, Clay-Williams R, Hogden E, Braithwaite J, Groene O. High performing hospitals: a qualitative systematic review of associated factors and practical strategies for improvement. BMC Health Serv Res 2015 Jun 24;15:244 [FREE Full text] [doi: 10.1186/s12913-015-0879-z] [Medline: 26104760]
87. Fadol Y, Barhem B, Elbanna S. The mediating role of the extensiveness of strategic planning on the relationship between slack resources and organizational performance. Manage Deci 2015 Jun 15;53(5):1023-1044. [doi: 10.1108/md-09-2014-0563]
88. Kruse CS, DeShazo J, Kim F, Fulton L. Factors associated with adoption of health information technology: a conceptual model based on a systematic review. JMIR Med Inform 2014 May 23;2(1):e9 [FREE Full text] [doi: 10.2196/medinform.3106] [Medline: 25599673]
89. Troilo G, De Luca LM, Atuahene-Gima K. More innovation with less? A strategic contingency view of slack resources, information search, and radical innovation. J Prod Innov Manag 2013 Oct 08;31(2):259-277. [doi: 10.1111/jpim.12094]
90. Haux R, Ammenwerth E, Koch S, Lehmann CU, Park H, Saranto K, et al. A brief survey on six basic and reduced eHealth indicators in seven countries in 2017. Appl Clin Inform 2018 Jul;9(3):704-713 [FREE Full text] [doi: 10.1055/s-0038-1669458] [Medline: 30184560]
91. Hübner U, Ammenwerth E, Flemming D, Schaubmayr C, Sellemann B. IT adoption of clinical information systems in Austrian and German hospitals: results of a comparative survey with a focus on nursing. BMC Med Inform Decis Mak 2010 Feb 02;10:8 [FREE Full text] [doi: 10.1186/1472-6947-10-8] [Medline: 20122275]
92. Ringle CM, Wende S, Becker JM. SmartPLS 3. URL: http://www.smartpls.com [accessed 2021-02-17]
93. Henseler J, Ringle CM, Sarstedt M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. J Acad Mark Sci 2014 Aug 22;43(1):115-135. [doi: 10.1007/s11747-014-0403-8]
94. Hambrick DC, Mason PA. Upper echelons: the organization as a reflection of its top managers. Aca of Manage Rev 1984 Apr;9(2):193-206. [doi: 10.5465/amr.1984.4277628]
95. Paulus RA, Davis K, Steele GD. Continuous innovation in health care: implications of the Geisinger experience. Health Aff (Millwood) 2008;27(5):1235-1245. [doi: 10.1377/hlthaff.27.5.1235] [Medline: 18780906]
96. Patterson F, Kerrin M, Gatto-Roissard G. Characteristics and behaviours of innovative people in organisations. Literature review prepared for the NESTA Policy & Research Unit. 2009. URL: https://media.nesta.org.uk/documents/characteristics_behaviours_of_innovative_people.pdf [accessed 2021-02-17]
97. Elenkov DS, Judge W, Wright P. Strategic leadership and executive innovation influence: an international multi-cluster comparative study. Strat Mgmt J 2005 Jul;26(7):665-682. [doi: 10.1002/smj.469]
98. Patel VM, Ashrafian H, Uzoho C, Nikiteas N, Panzarasa P, Sevdalis N, et al. Leadership behaviours and healthcare research performance: prospective correlational study. Postgrad Med J 2016 Nov;92(1093):663-669. [doi: 10.1136/postgradmedj-2016-134088] [Medline: 27190092]

99. McAlearney AS, Hefner JL, Sieck CJ, Huerta TR. The journey through grief: insights from a qualitative study of electronic health record implementation. Health Serv Res 2015 Apr;50(2):462-488 [FREE Full text] [doi: 10.1111/1475-6773.12227] [Medline: 25219627]

Abbreviations

CIO: chief information officer
HIT: health information technology
IT: information technology
IC: innovation capability
IC ITD: innovation capabilities at the information technology department level
IC TMT: innovation capabilities at the top management team level
IC OW: innovation capabilities at the organization-wide level
IQHIT: innovation and quality of HIT
OGIP: overall goodness of information provision
PIM: professionalism of information management
TMT: top management team
WCS: Workflow Composite Score

Edited by C Lovis; submitted 08.08.20; peer-reviewed by S Kujala, A Vagelatos, D Walker, H Oh; comments to author 18.10.20; revised version received 13.12.20; accepted 07.02.21; published 15.03.21.

Please cite as:
Esdar M, Hübner U, Thye J, Babitsch B, Liebe JD
The Effect of Innovation Capabilities of Health Care Organizations on the Quality of Health Information Technology: Model Development With Cross-sectional Data
JMIR Med Inform 2021;9(3):e23306
URL: https://medinform.jmir.org/2021/3/e23306
doi:10.2196/23306
PMID:33720029

Original Paper

Impact of Web-Based Self-Scheduling on Finalization of Well-Child Appointments in a Primary Care Setting: Retrospective Comparison Study

Frederick North[1], MD; Elissa M Nelson[2], MA; Rebecca J Majerus[2], BAS; Rebecca J Buss[2], MS; Matthew C Thompson[2], MBA; Brian A Crum[3], MD

[1]Division of Community Internal Medicine, Department of Medicine, Mayo Clinic, Rochester, MN, United States

[2]Enterprise Office of Access Management, Mayo Clinic, Rochester, MN, United States

[3]Department of Neurology, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Frederick North, MD
Division of Community Internal Medicine
Department of Medicine
Mayo Clinic
200 First Street SW
Rochester, MN, 55905
United States
Phone: 1 507 284 2511
Email: north.frederick@mayo.edu

Abstract

Background: Web-booking of flights, hotels, and sports events has become commonplace in the travel and entertainment industry, but self-scheduling of health care appointments on the web is not yet widely used. An electronic health record that integrates appointment scheduling and patient web-based access to medical records creates an opportunity for patient self-scheduling. The Mayo Clinic developed and implemented a feature in its Patient Online Services (POS) web and mobile platform that allows software-managed self-scheduling of well-child visits.

Objective: This study aims to examine the use of a new self-scheduling appointment feature within POS in both web and mobile formats and determine the use characteristics, outcomes, and efficiency of self-scheduling compared with staff scheduling.

Methods: Within a primary care setting, we collected 13 months of all appointment activity for the well-child visit for children aged 2-12 years. As these specific appointment types are for minors, self-scheduling is performed by parents or other proxies. We compared the appointment actions of scheduling and cancelling for both self-scheduled and staff-scheduled appointments. The frequency in which patients were using self-scheduling outside of normal business hours was quantified, and we compared no-show outcomes of finalized appointments.

Results: Of the 1099 patients who performed any self-scheduling actions, 73.1% (803/1099) exclusively used self-scheduling and self-cancelling software. For those with access to self-scheduling (patients registered with the Mayo Clinic POS), 4.92% (1201/24,417) of all well-child appointment-scheduling actions were self-scheduled. Staff scheduling required more than a single appointment step (eg, schedule, cancel, reschedule) in 28.32% (3729/13,168) compared with only 6.93% (53/765) of self-scheduled appointments (P<.001). Self-scheduling appointment actions took place outside of regular business hours 29.5% (354/1201) of the time. No-shows accounted for 3.07% (28/912) of the self-scheduled finalized appointments compared with 4.12% (693/16,828) of staff-scheduled appointments, which is a nonsignificant difference (P=.12). Staff-scheduled finalized appointments (that allowed for scheduling appointments for more than 12 weeks in the future) revealed a potential demand of 11.15% (1876/16,828) for appointments with longer lead times.

Conclusions: Self-scheduling can generate a significant number of finalized appointments, decreasing the need for staff scheduler time. We found that 29.5% (354/1201) of the self-scheduling activity took place outside of the usual staff scheduler hours, adding convenience value to the scheduling process. For exclusive self-schedulers, 93.1% (712/765) finalized the appointment in a single step. The no-show rates were not adversely affected by the self-scheduling.

(JMIR Med Inform 2021;9(3):e23450) doi:10.2196/23450

KEYWORDS

electronic health record; schedules; patient appointment; preventive health service; office visit; outpatient care; software tool; computer software application; mobile applications; child health; pediatric; preventive care; self

Introduction

The travel and entertainment industries have provided web booking of flights, hotel rooms, and sports and entertainment events for many years, whereas web-based scheduling of medical appointments is not widely available. Gupta and Denton [1] summarized many of the unique challenges in health care that make scheduling rules for medical appointments complex and difficult to code into software. Ahmadi-Javid et al [2] reviewed much of the extant literature on outpatient appointment scheduling and decision making involved in the appointment-scheduling processes. In addition to the complex rules needed for scheduling, there are patient barriers to web-based appointment scheduling. A survey in Australia showed that 89% of primary care patients with access to a web-based appointment-scheduling system were reluctant to adopt it [3]. Although all patients had access to the system, only 11% used the web-based appointment service at least once, and 74% were not inclined to use the web-based appointment service in the near future. In interviews, some of the patients preferred phone call appointments that they perceived "provided them with more opportunities to discuss the options for more complex situations than the online self-service" [3]. Others cited low computer literacy and lack of access to the internet at home [3].

Despite these barriers, independent vendors are filling some demand for medical appointments on the web. ZocDoc (TM), for example, has been offering web-based appointment scheduling for health care practices [4]. Kurtzman et al [5] assessed appointments from 4150 physicians available in 20 cities where ZocDoc was available and found a "substantial number of appointments available for patients on ZocDoc," with the conclusion that "ZocDoc is a promising method for obtaining reliable primary care appointments in the cities evaluated" [5]. Internationally, similar web-based platforms for self-scheduling health care appointments exist, such as Lybrate in India [6]. Zhao et al [7] reviewed the literature on web-based appointment systems and found support for associations between web-based appointment systems and improved no-show rates, decreased waiting time, improved patient satisfaction, and decreased staff labor.

Mayo Clinic implemented a web- and mobile-based self-scheduling option for a well-child visit in 2019. We examined the patient uptake and outcomes of the self-scheduling option to see if there were differences in use and appointment outcomes between self-scheduling and the use of Mayo staff appointment schedulers.

Methods

Setting

This study was conducted in the Mayo Clinic Health System in the Rochester, Minnesota, and Northwest Wisconsin regions for clinic well-child visits scheduled for the 13-month time interval from February 1, 2019, to February 28, 2020. Providers eligible to have well-child visits on their schedules were physicians, physician assistants, and nurse practitioners in family medicine and community pediatrics departments.

The Mayo Clinic uses Epic as its electronic health record (EHR) system. The Mayo Clinic has a patient portal, named Patient Online Services (POS), that patients can access via a mobile app or on the web. With the Mayo Clinic POS, patients can communicate with providers via secure messages, review their medical records, and view future appointment details. Patients of the Mayo Clinic have been increasingly engaged with POS, and portal registration has increased from 33% in 2013 to 62% in 2018 [8]. Although POS has been available for many years, self-scheduling of office visits through the POS has been made available only recently.

The self-scheduling process required POS, so portal registration was a prerequisite for self-scheduling. In this study, we use self-scheduling as a generic term for scheduling via software, without assistance from a staff scheduler. Owing to age limitations on the well-child appointment, self-scheduling and self-cancellations were accomplished by patient proxies such as parents or other adults who had portal access to the child's EHR.

Well-child appointment scheduling with a staff scheduler was generally limited to Monday through Friday, from 7 AM to 5 PM. For self-scheduling, there was 24/7 access to a web-based self-scheduling process and mobile self-scheduling app, except for rare occasions of a software outage.

Well-Child Appointment Type

The well-child examination is a periodic exam recommended by the American Academy of Pediatrics [9]. The Mayo-implemented self-scheduling feature is limited to exams for ages 2 to 12 years to decrease the complexity of rules associated with scheduling.

The well-child visit type is a good visit type for self-scheduling in primary care practice. By definition, the appointment is for a healthy child, so no symptoms require an urgent visit. Many appointments in primary care are symptom-based and can require some symptom assessment to determine the urgency of the visit. Self-scheduling symptom-based visits are a larger informatics challenge because of the patient safety issue around the urgency of visits that is not present in the well-child visit type.

The well-child visit is also a visit type that allows the provider some autonomy to schedule these visits in blocks or spread out, earlier or later in the day, or for a specific day of the week. The same visit type was used for self-scheduling as was used for staff scheduling, which reduced the software build needs. Prebuilt provider calendar templates with well-child visits were being used by children's providers at the Mayo Clinic long before self-scheduling was implemented. Thus, self-scheduling of this visit type required very little change management other

than communication of the change. Self-schedulers were able to see an open well-child visit they wanted and could book it. A subsequent informal provider survey confirmed that most providers (20/25, 80%) were unaware that a proxy had booked the appointment.

Well-child appointments also did not need a provider order to initiate scheduling. Many primary care visits involve lab and radiology procedures, which require orders for scheduling. This is needed to identify providers able to *close the loop* on imaging and lab test results. Preventive services such as screening mammography and chronic disease visits for diabetes, hypercholesterolemia, and hypertension are examples of visits requiring orders. Screening mammograms require radiology visit orders, and laboratory orders (eg, hemoglobin A_{1c}, lipids, etc) are often requested in advance of chronic disease appointments. Chronic disease visits also have some appointment length challenges because many patients have multiple chronic diseases, so appointment lengths often need to be individualized. The well-child visit was a good candidate for self-scheduling; it required no decision support for appointment urgency or appointment duration, and it did not require an order before scheduling.

Scheduling Rules in Software

The scheduling rules in the software for the well-child visit include the following:

1. Frequency limitations of appointments: the software looks back at the date of previous well-child appointments to ensure that frequency limitations are met.
2. Age: limited to ages 2-12 years.
3. Assigned primary care provider: children need an assigned primary care provider to be eligible. To ensure continuity of care, there is no option to schedule a well-child appointment with any provider other than the assigned primary care provider. The software automatically pulls the primary care provider scheduling template.
4. Appointment lead time: calendar availability was 12 weeks in the future. Provider templates were built for no more than 12 weeks in the future for the Rochester, Minnesota, site; therefore, specific appointment times beyond 12 weeks could not be self-scheduled at that site. Although the Northwest Wisconsin site had provider calendar templates available for more than 12 weeks, the self-scheduling rules for the initial implementation did not account for the expanded scheduling ability at the Northwest Wisconsin location.

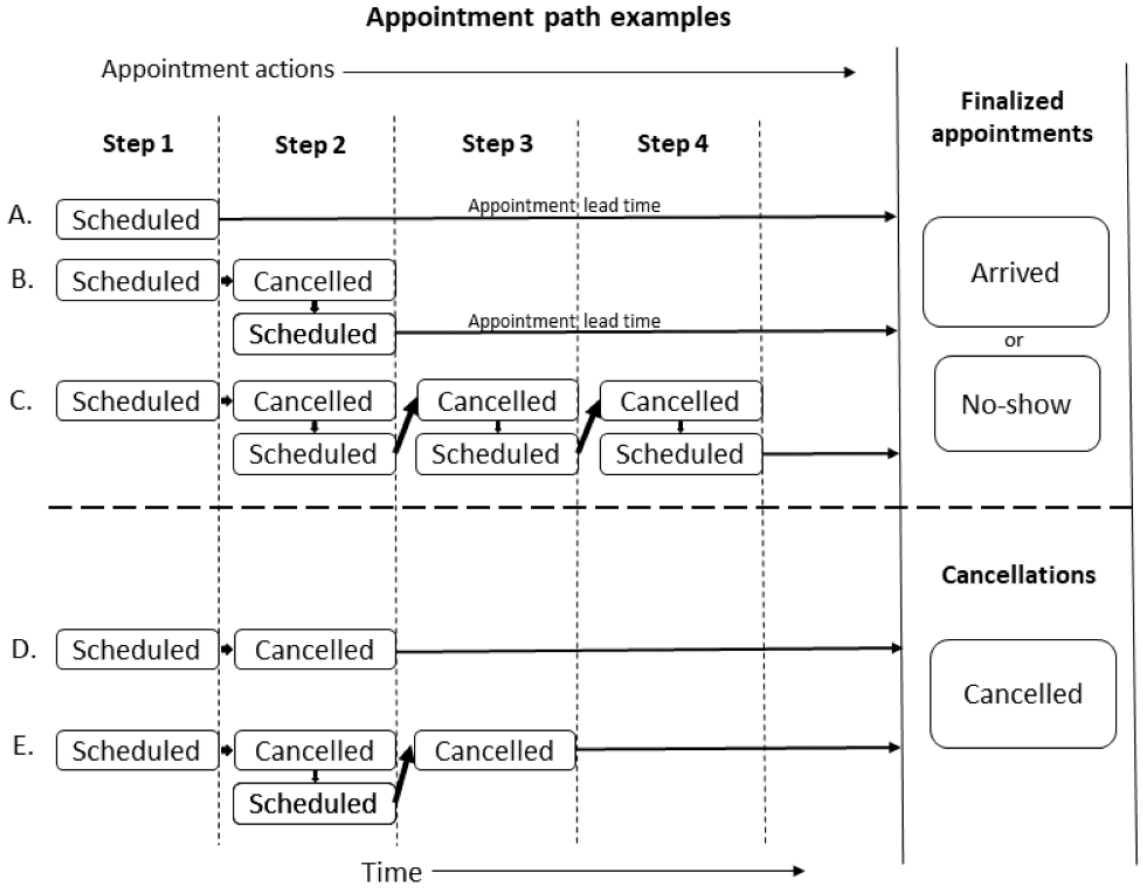
Appointment Definitions and Data

Individual appointment actions are dichotomously classified as a *schedule* or *cancel* action. A schedule action reserves a single appointment time; a cancel action opens a previously scheduled appointment time. Well-child visits are typically 30 minutes but could be scheduled for 45 minutes by staff schedulers.

Staff schedulers are clinic staff employees who schedule or cancel appointments for patients. Self-schedulers or self-cancellers were the parent or proxy who used the Mayo software interface (web or mobile) to self-schedule or self-cancel the child's appointments. It should be noted that we focused on self-scheduled actions in this paper. Some proxies did not use the self-scheduling feature but self-cancelled appointments made by staff schedulers. To be considered self-scheduled, a patient had to have at least one appointment action of self-scheduling (booking an appointment with the self-schedule software). The few patients who self-cancelled their staff-scheduled appointments were classified as staff scheduled.

An appointment path is the sequence of appointment actions leading to a finalized appointment or cancellation outcome. Finalized appointments were those scheduled appointments that were left scheduled up to the appointment date and time (not cancelled before the appointment time). Figure 1 shows examples of appointment paths and appointment outcomes. Our data start with a time-stamped appointment schedule action. We dichotomized appointment actions into those created by staff schedulers and those created by self-scheduling. As shown in Figure 1, each patient (whether self-scheduled or staff scheduled) begins with a scheduling action that we term appointment step 1. Patients can then go through several decision steps of whether to cancel or reschedule (a cancel and schedule pair). Some patients would cancel and reschedule multiple times before a finalized appointment. To quantify this activity, we counted the appointment steps, as shown in Figure 1. Example (A) within Figure 1 shows an appointment path to appointment finalization with just 1 step, the initial scheduling action. Examples (B) and (C) within Figure 1 show appointment paths for appointment finalization taking 2 and 4 steps, respectively. Appointment paths ending in a cancellation outcome may also take several appointment steps. Figure 1 shows examples (D) and (E) that take 2 and 3 appointment steps, respectively, to a cancellation.

Figure 1. Examples of different appointment paths showing the appointment actions and appointment steps leading to a finalized appointment or cancellation.



Appointment outcomes are dichotomously categorized as finalized appointments or cancellations. Finalized appointments are further dichotomously categorized as completed or no-show (never arrived at the scheduled appointment time). The well-child visit appointment was an in-person visit; therefore, this study did not include any telephonic or video appointments.

Figure 1 example (A) also shows the appointment lead time, which is the scheduled appointment date and time minus the date and time the appointment was made. This is the lead time that the patient has from the date of scheduling the appointment to the actual future reserved appointment date.

Data Collection and Study Metrics

We used appointment data sets from the Mayo Clinic Enterprise Office of Access Management in this study. The data set captured all appointment activity dichotomously as a scheduling or cancellation action, and whether the scheduling or cancellation action was done by the appointment staff or self. We obtained complete scheduling and cancellation actions for all well-child visits encompassing ages 2 to 12 years from February 1, 2019, through February 28, 2020. The time of the appointment action (scheduling or cancelling) was included in the data set and was categorized as weekend (Saturday or Sunday) and after-hours weekday (not occurring within 7 AM to 5 PM). There were mobile and web versions for self-scheduling, and we were able to capture which was used for each self-scheduled action. If a patient used the web version on a mobile device, it was captured as web use.

Demographic information was obtained from all children whose proxy or proxies either cancelled or made a well-child appointment for the 13 months of the study. Demographic data on the proxies were not collected for this study.

Finalized appointment outcomes were obtained from a final data set that contained only scheduled appointments still in the system on the day of the expected visits. Appointments cancelled any time up to the appointment date and hour were excluded from the finalized appointment outcome analysis to leave behind only those scheduled visits that providers expected to see face-to-face. The finalized appointment outcomes were dichotomously categorized as no-show or arrived.

Statistics and Ethics

We used JMP Pro 14.2 (SAS) for descriptive statistics and statistical analyses. For comparison between categorical variables, we used chi-square tests and odds ratios (ORs). This study met the criteria for institutional review board exemption (20-006809).

Results

Well-Child Visit Scheduling Counts and Provider Counts

There were 36,392 well-child scheduling actions for 399 providers. Pediatric providers accounted for 65.20% (23,727/36,392) and family medicine providers for 34.80% (12,665/36,392) of the well-child scheduled visits. We limited

this study to those who could access self-scheduling, so only those patients with proxy access to POS (portal registration) were included. This resulted in 24,417 scheduling actions for analysis, with 4.92% (1201/24,417) of all scheduling actions being self-scheduled.

Patient Characteristics

Figure 2 shows the unique patient counts of those who scheduled well-child appointments during the study. There were 32.83% (7898/24,059) of patients who could not self-schedule because they did not have POS registration. Of the 16,161 patients who had portal access, 6.80% (1099/16,161) used self-scheduling. Of the 1099 patients who used self-scheduling, 73.1% (803/1099) used self-scheduling and self-cancelling exclusively, and 26.9% (296/1099) had appointment actions that used both self-scheduling and staff scheduling.

Sex, race, and ethnicity of children were not statistically different between those who were self-scheduled and those who were not (Table 1). However, self-scheduled appointments were proportionately greater for those aged 6 to 12 years than staff-scheduled appointments (Table 1).

Figure 2. Patient counts by category of those who completed well-child visit scheduling actions during the 13 month study period.

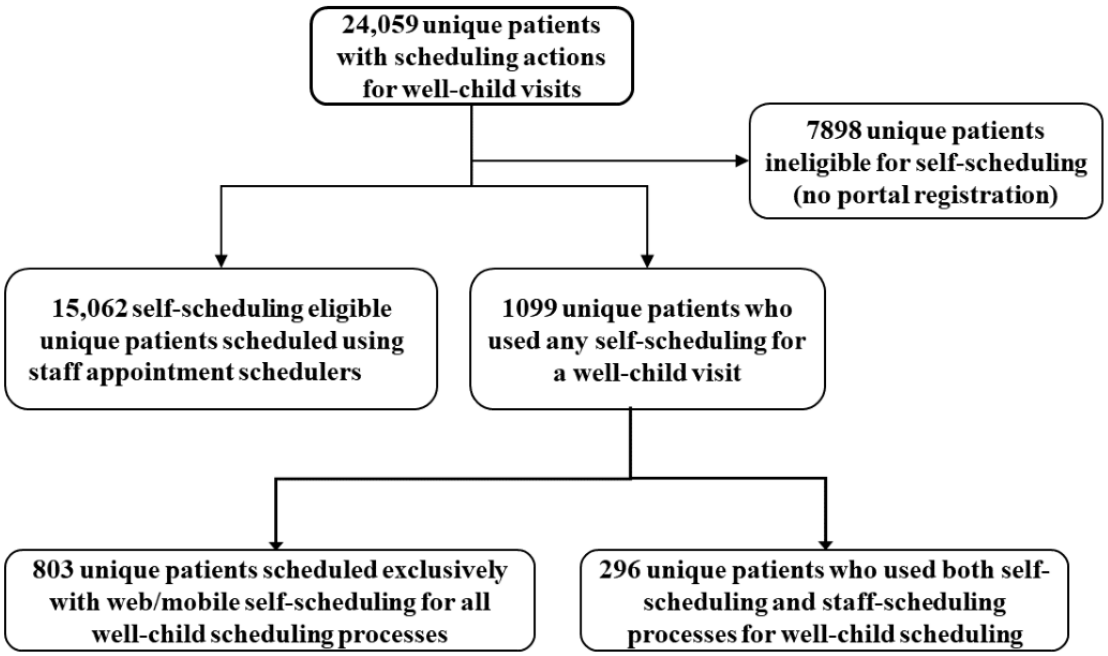


Table 1. Demographic comparison of patients with portal registration with well-child appointment actions (N=16,161). The self-scheduled group completed at least one self-scheduling action. Demographics compared are only those with access to self-scheduling (those without portal registration were not included).

Demographics	Self-scheduled (n=1099), n (%)	Staff scheduled (n=15,062), n (%)	*P* value[a]
Age (years)			<.001
2-5	480 (43.68)	8523(56.59)	
6-12	619 (56.32)	6539 (43.41)	
Female sex	524 (47.68)	7290 (48.40)	.64
Race			.27
White	963 (87.63)	12,987 (86.22)	
Black	20 (1.82)	374 (2.48)	
Asian	31 (2.82)	545 (3.62)	
Other or not disclosed	85 (7.73)	1156 (7.67)	
Ethnicity			.97
Hispanic	41 (3.73)	585 (3.88)	
Not Hispanic	1026 (93.36)	14,041 (93.22)	
Undisclosed or unknown	32 (2.91)	436 (2.89)	

[a]Null hypothesis (H^0): patient demographic proportions are equal.

After-Hours Scheduling and Appointment Lead Time

The 1099 patients who used the self-scheduling feature generated 1490 appointment actions (1201 self-scheduling actions and 289 cancelling actions), resulting in 912 finalized appointments. Similarly, 15,062 patients who used staff scheduling generated 29,604 appointment actions (23,216 scheduling actions and 6388 cancelling actions), resulting in 16,828 finalized appointments. Cancelling actions in the self-scheduled group accounted for 19.4% (289/1490) of all appointment actions in that group and 21.58% (6388/29,604) in the staff-scheduled group (*P*=.046). The differences in scheduling between self-scheduled and staff-scheduled actions are shown in Table 2. There were across-the-board differences when scheduling actions occurred on weekend days and weekdays after usual staff scheduler hours and when scheduling lead time was greater than 12 weeks. As noted in the *Methods* section, staff scheduler hours were mostly limited to usual outpatient weekday hours, so staff-scheduling actions on weekend days and after hours on weekdays were expected to be low; 12.99% (3015/23,216) of staff-scheduling actions had appointment lead times greater than 12 weeks. As noted in the *Methods* section, a software rule excluded self-scheduling with lead times over 12 weeks; thus, patients wanting a longer lead time had to be scheduled by staff.

Table 2. Comparison of self- versus staff-scheduling actions (does not include cancelling actions). Scheduling actions are limited to those who could access self-scheduling (those with portal registration).

Appointment metric	Self-scheduled (scheduling actions only), n (%)	Staff scheduled (scheduling actions only), n (%)	*P* value[a]
Scheduling appointment action count	1201 (100)	23,216 (100)	N/A[b]
Any appointment-scheduling action outside regular business hours (Monday-Friday, 7 AM to 5 PM)	354 (29.48)	199 (0.86)	<.001
Appointment-scheduling action on weekdays (Monday-Friday) outside of 7 AM to 5 PM	227 (18.90)	180 (0.78)	<.001
Appointment-scheduling action on weekend days (Saturday or Sunday)	127 (10.57)	19 (0.08)	<.001
Scheduling action lead time over 12 weeks	0 (0)	3015 (12.99)	<.001

[a]Null hypothesis (H^0): proportions are equal between self-scheduled and staff scheduled.

[b]N/A: not applicable.

Staff Scheduler Work Involved in Self-Scheduled Appointments

As indicated in Figure 1, appointment scheduling can go through many steps of scheduling and cancelling over the span of a finalized or cancelled appointment. A patient could use both self-scheduling and self-cancelling for parts of the appointment steps and staff scheduling and staff cancelling for other parts of the appointment steps, leading to a single finalized appointment. Self-scheduling would be inefficient if patients who self-scheduled also relied on staff schedulers to cancel or reschedule the appointment. To determine whether staff were involved in the rework of this type, we examined all finalized appointments to determine how much of the scheduling and cancelling work was being done by staff and how much was being done by the patients themselves. Table 3 shows that of the 912 finalized appointments with any self-scheduling activity, only 9.9% (147/1490) involved a staff scheduler. Thus, self-scheduling activity did not lead to large amounts of rework by staff schedulers to obtain a finalized appointment. Table 3 also shows that there were on average 1.63 appointment actions per finalized appointment for those with self-scheduling activity, with only 0.16 actions per finalized appointment attributable to staff schedulers. In contrast, on average, staff schedulers took 1.76 appointment actions for each staff-scheduled finalized appointment.

Table 3. Comparison of the average patient-performed appointment actions per finalized appointment for self-scheduled and staff scheduled.

Appointment metric	Self-scheduled (but staff could cancel)	Staff scheduled (but the patient could cancel on the web)
Total appointment actions (schedule and cancel), n	1490	29,604
Scheduling actions, n (%)	1201 (80.6)	23,216 (78.4)
Cancelling actions, n (%)	289 (19.4)	6388 (21.6)
Appointment actions (schedule or cancel action) performed by patient or proxy (web or mobile), n (%)	1343 (90.1)	540[a] (1.8)
Appointment actions performed by Mayo scheduler, n (%)	147[b] (9.9)	29,064 (98.2)
Appointments finalized, n (remaining on calendar up to visit date and time)	912	16,828
Average patient or proxy performed appointment actions per finalized appointment (total count of patient or proxy appointment actions divided by total count of finalized appointments)	1.47 (1343/912)	0.032 (540/16,828)
Average Mayo scheduler actions per finalized appointment (total count of Mayo scheduler actions divided by total count of finalized appointments)	0.16 (147/912)	1.73 (29,064/16,828)
Average appointment actions per finalized appointment (total count of appointment actions divided by total count of finalized appointments)	1.63 (1490/912)	1.76 (29,604/16,828)

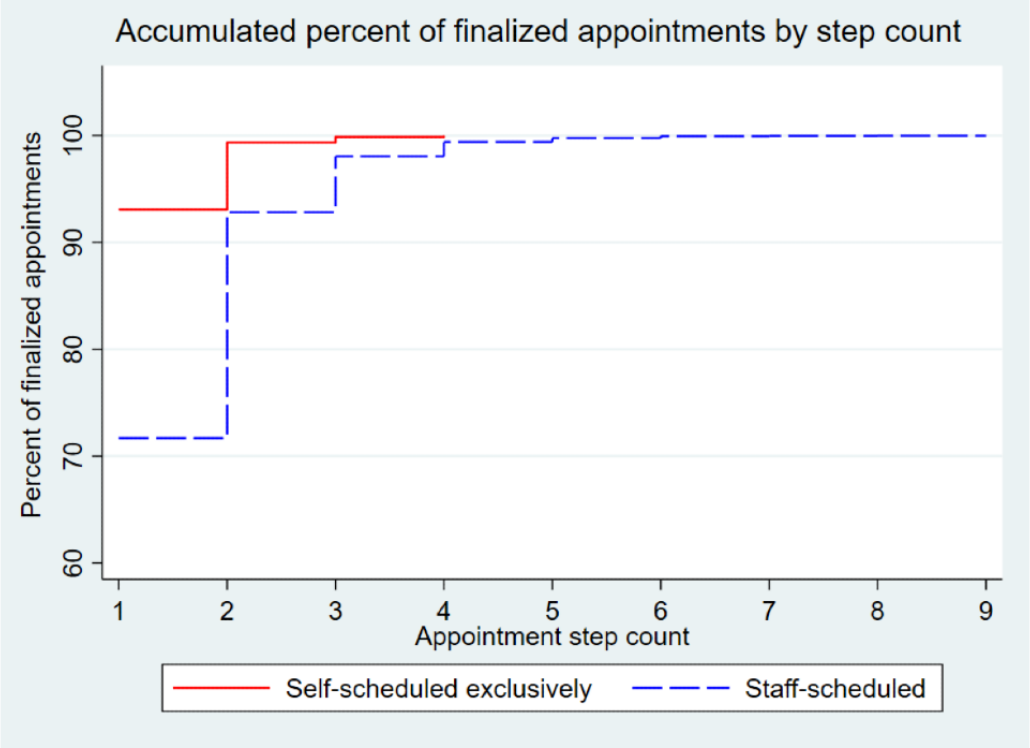
[a]Staff scheduled but was self-cancelled.

[b]Self-scheduled but had cancel actions taken by staff schedulers (however, 142 of the 289 cancels were self-cancelled).

Comparison of Appointment Steps With Finalized Appointment

As shown in Figure 1, the initial appointment step starts with scheduling a future appointment, which we term appointment step 1. In addition, as indicated in Figure 1, there may be multiple steps before a finalized appointment. Figure 3 shows a comparison between exclusively self-scheduled and staff-scheduled appointments on the accumulated appointment steps taken before appointments were finalized. To make our analysis comparable, we limited it to only those patients who had a single finalized appointment within the timeframe of the study. A total of 765 patients completed a single appointment exclusively using self-scheduled appointment actions. Of those, 93.07% (712/765) finalized the appointment in a single step; for the 13,168 patients who used staff schedulers for a single appointment, 71.68% (9439/13,168) finalized the appointment in a single step (P<.001). Thus, 28.32% (3729/13,168) of the staff-scheduled appointments required multiple appointment steps compared with 6.93% (53/765) requiring multiple steps for a self-scheduled appointment (OR 5.3, 95% CI 4.0-7.0).

Figure 3. Accumulated percent of finalized appointments by step count.



Comparison of Finalized Appointment Outcomes

Of the 1201 self-scheduling appointment actions, 912 became finalized appointments. Self-scheduling accounted for 5.14% (912/17,740) of finalized, well-child appointments. Table 4 shows the differences in appointment outcomes for those who had finalized appointments. No-shows for well-child appointments were not statistically different between self-scheduled and staff-scheduled appointments. As with scheduling actions, we found a significant number of staff-scheduled finalized appointments (1876/16,828, 11.15%) involved appointment lead times over 12 weeks.

Table 4. Comparison of appointment outcomes for finalized appointments (those remaining scheduled on the day of appointment).

Visit outcome	Self-scheduled	Staff scheduled	*P* value[a]
Finalized appointments (scheduled and not cancelled before the visit date), n (%)	912 (100)	16,828 (100)	N/A[b]
Arrived for appointment (percent of patients seen for visit day appointment), n (%)	884 (96.93)	16,135 (95.88)	.12
No-show (percent not arriving at the appointment), n (%)	28 (3.07)	693 (4.12)	.12
Appointment lead time greater than 12 weeks (84 days), n (%)	0 (0)	1876 (11.15)	<.001
Median appointment lead time (days)	27	32	N/A

[a]Null hypothesis (H^0): proportions are equal between self-scheduled and staff-scheduled appointments.

[b]N/A: not applicable.

Mobile-Based Versus Web-Based Self-Scheduling

For the 1201 self-scheduled appointment actions, 61.20% (735/1201) were completed through the patient web application and 38.80% (466/1201) were completed through the mobile app. Of the 912 appointments finalized, 60.2% (549/912) were through the web and 39.8% (363/912) were through the mobile app.

Discussion

Principal Findings

For those with the opportunity to self-schedule (registered with POS), 5.14% (912/17,740) of finalized well-child visits were self-scheduled. For those with portal registration, no-show rates were statistically similar to those who did not self-schedule. Self-scheduling occurred 29.5% (354/1201) of the time outside of the usual business hours. There was a significant demand for appointment lead times greater than the 12-week window allowed in self-scheduling, demonstrated by 11.15% (1876/16,828) of staff-scheduled appointments with lead times

over 12 weeks. Self-scheduling resulted in a significantly higher percentage of single-step appointment scheduling (*one and done*) than with staff scheduling.

Practice Implications

Although it was outside the scope of this study to examine cost, a significant number of patients (n=803) exclusively used self-scheduling, likely decreasing scheduling expenses for that group. For patients using any self-scheduling, there were only 9.87% (147/1490) of scheduling actions performed by staff schedulers, so there was little indication of unintended consequences leading to more staff work.

Paré et al [10] found that the flexibility of being able to book appointments when it was most convenient was one of the highest patient-perceived benefits of the e-booking system they studied. Ballantyne et al [11] also noted that parents of special needs children thought it was important to be able to self-schedule appointments with a computer or mobile device. With 29.5% (354/1201) of our self-scheduling occurring outside of normal business hours and 38.8% (466/1201) of self-scheduling by mobile devices, many parents or proxies were able to access and complete scheduling activities anytime and anywhere. With this increase in scheduling convenience, it is possible that self-scheduling can significantly improve patient satisfaction.

With the amount of self-scheduling that occurred after hours, our findings might help decide whether to have appointment schedulers on duty during evenings and on weekend days. This could be helpful for scheduling patients who do not have portal access and those who have access to self-schedule but may need additional help.

No-Shows Associated With Self-Scheduling

Although no-shows were less frequent in the self-scheduled group, this was not statistically significant. In other studies, self-scheduling has been associated with lower no-show rates [10,12-14]. Our finding that missed appointments in self-scheduled patients were not statistically less suggests that self-scheduling itself may not decrease missed appointments. It should be noted that at the Mayo Clinic, patients receive appointment reminders by text or phone for all appointments, whether self-scheduled or staff scheduled.

Appointment Lead Time

The median appointment lead times were about 1 month in both the self-scheduled and staff-scheduled groups (Table 4). Self-scheduled lead times greater than the 12-weeks were not allowed by the software. However, it is notable that 11.15% (1876/16,828) of staff-scheduled finalized appointments had an appointment lead time greater than 12 weeks. For the subgroup of Northwest Wisconsin, where appointment lead times up to 6 months were available using staff scheduling, 27.25% (1759/6455) had a lead time of over 12 weeks. This suggests that there might be increased uptake in self-scheduling if future software updates can accommodate longer lead times.

Comparison With Previous Studies

Our 5% uptake is similar to that observed by Zhang et al [15], who found an uptake of 4% after 29 months of using an e-appointment–scheduling system in an Australian primary health care clinic. In that study, they found that many patients did not see the value of the e-appointment system when they could easily call by phone to make appointments, and the patients noted limitations in the functionality of the self-scheduling system [15].

Lack of awareness of the self-scheduling feature has been an issue with implementation elsewhere. Cao et al [16] noted that 53% of patients were unaware of their web-based appointment system. Although we attempted to increase patient awareness of the ability to self-schedule this visit type, we do not know how many who needed appointments were actually aware of the self-scheduling option.

Lessons for Future Enhancements

As a large percentage of the staff-scheduled visits were made with more than a 12-week lead time, there is likely a significant demand for this to be incorporated into the self-scheduling of this visit type. Software enhancement could be made to set future visit requests greater than 12 weeks in a placeholder visit and then automatically generate a reminder to the patient as soon as appointment templates are opened for scheduling. For our Mayo Clinic Northwest Wisconsin site, where provider schedule templates are built out 6 months in advance, there may be a trade-off in provider satisfaction for those who do not want to be locked into a 6-month schedule. Another option could be to have a pool of providers available to give more scheduling options, but for the well-child visit, this may negate the advantage of continuity of care [17-19].

Limitations

Our study was limited to just 1 self-scheduled appointment, the well-child visit, limiting the potential generalizability of our results. There are numerous other types of visits, which may have different results. The study also took place in a majority White community, so there may be differences in communities with different demographics. To control for portal access, we included only those with portal registration; there would be a smaller percentage engaged in self-scheduling had we included those without access to self-scheduling. The scheduling platform we used (Mayo Clinic POS) is specific to the Mayo Clinic, but appointments and rules were managed with information from Epic, the Mayo Clinic's EHR, which has a wide user base across the United States.

The uptake of self-scheduling may also differ in other practices. Although there was some promotion of this new module, additional promotion may have resulted in a larger uptake. It is possible that the uptake of self-scheduling was influenced by the advantage of 24/7 availability compared to the more limited availability of Mayo Clinic schedulers (mostly 7 AM-5 PM on weekdays). Self-scheduling uptake could be lower if staff schedulers were available during the evening hours and on weekends when some of the self-scheduling occurred.

Future Research Implications

The impact of self-scheduling on patient satisfaction is unclear. At the Mayo Clinic, patient satisfaction with access significantly decreased, for a time, associated with an EHR switch [20]. It is

possible that the ability to self-schedule could also be associated with a measurable change in patient satisfaction. Future research is also needed on patient acceptance of self-scheduling, especially in view of some studies that have shown patients' reluctance to use a self-scheduling feature [3,15,21].

Our study showed that the impact of self-scheduling on no-show rates was not significant when limited to those with portal registration. In a systematic review by Dantas et al [22], longer appointment lead times were found to be a major driving factor for higher no-show rates. It deserves restatement that the well-child visit is a special visit type where, as we show in this study, proxies may find a long lead time desirable. Additional research is needed to clarify the confounders related to self-scheduling and no-show rates. Additional investigation is also needed for other types of self-scheduled appointments (such as acute care visits) for more generalizable conclusions on self-scheduling quality and safety issues.

Conclusions

Well-child appointments were successfully scheduled entirely within the appointment software, resulting in fewer interactions with appointment schedulers, frequently outside of the hours staff schedulers usually work. Self-scheduled appointments were more likely to be completed in a single appointment step than staff-scheduled appointments; self-schedulers were unlikely to need additional help from a scheduler to finalize an appointment. Self-scheduled appointment no-show rates were not statistically different from those of staff-scheduled appointments. Self-scheduling software may need to accommodate patients wanting to schedule appointments further in the future than their providers' appointment templates allow.

Authors' Contributions

FN conceptualized the study. FN, EN, RM, RB, and MT contributed to the study design and data analysis, interpretation, and statistics. FN drafted the first version of the manuscript. FN, EN, RM, RB, MT, and BC performed final manuscript editing and critical revisions and approved the manuscript's final version.

Conflicts of Interest

None declared.

References

1. Gupta D, Denton B. Appointment scheduling in health care: challenges and opportunities. IIE Transactions 2008 Jul 21;40(9):800-819. [doi: 10.1080/07408170802165880]
2. Ahmadi-Javid A, Jalali Z, Klassen KJ. Outpatient appointment systems in healthcare: a review of optimization studies. Eur J Oper Res 2017 Apr;258(1):3-34. [doi: 10.1016/j.ejor.2016.06.064]
3. Zhang X, Yu P, Yan J. Patients' adoption of the e-appointment scheduling service: a case study in primary healthcare. Stud Health Technol Inform 2014;204:176-181. [Medline: 25087546]
4. Oberoi A. How zocdoc works: business model and revenue streams. daffodil software. 2018 Jul 17. URL: https://insights.daffodilsw.com/blog/how-zocdoc-works-business-model-and-revenue-streams [accessed 2021-02-19]
5. Kurtzman GW, Keshav MA, Satish NP, Patel MS. Scheduling primary care appointments online: differences in availability based on health insurance. Healthc (Amst) 2018 Sep;6(3):186-190. [doi: 10.1016/j.hjdsi.2017.07.002] [Medline: 28757308]
6. Lybrate. URL: https://www.lybrate.com/ [accessed 2021-02-19]
7. Zhao P, Yoo I, Lavoie J, Lavoie BJ, Simoes E. Web-based medical appointment systems: a systematic review. J Med Internet Res 2017 Apr 26;19(4):e134 [FREE Full text] [doi: 10.2196/jmir.6747] [Medline: 28446422]
8. North F, Luhman KE, Mallmann EA, Mallmann TJ, Tulledge-Scheitel SM, North EJ, et al. A retrospective analysis of provider-to-patient secure messages: how much are they increasing, who is doing the work, and is the work happening after hours? JMIR Med Inform 2020 Jul 08;8(7):e16521 [FREE Full text] [doi: 10.2196/16521] [Medline: 32673238]
9. American academy of pediatrics schedule of well-child care visits. American Academy of Pediatrics. 2020. URL: https://www.healthychildren.org/English/family-life/health-management/Pages/Well-Child-Care-A-Check-Up-for-Success.aspx [accessed 2021-02-19]
10. Paré G, Trudel M, Forget P. Adoption, use, and impact of e-booking in private medical practices: mixed-methods evaluation of a two-year showcase project in Canada. JMIR Med Inform 2014 Sep 24;2(2):e24 [FREE Full text] [doi: 10.2196/medinform.3669] [Medline: 25600414]
11. Ballantyne M, Liscumb L, Brandon E, Jaffar J, Macdonald A, Beaune L. Mothers' perceived barriers to and recommendations for health care appointment keeping for children who have cerebral palsy. Glob Qual Nurs Res 2019 Aug 14;6 [FREE Full text] [doi: 10.1177/2333393619868979] [Medline: 31453266]
12. Marhefka KM. The Impact of Digital Self-Scheduling on No-Show Event Rates in Outpatient Clinics. 2020. URL: https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=9673&context=dissertations [accessed 2021-02-19]
13. Parmar V, Large A, Madden C, Das V. The online outpatient booking system 'Choose and Book' improves attendance rates at an audiology clinic: a comparative audit. Inform Prim Care 2009 Sep 01;17(3):183-186 [FREE Full text] [doi: 10.14236/jhi.v17i3.733] [Medline: 20074431]

14. Siddiqui Z, Rashid R. Cancellations and patient access to physicians: ZocDoc and the evolution of e-medicine. Dermatol Online J 2013 Apr 15;19(4):14. [Medline: 24021373]
15. Zhang X, Yu P, Yan J, Spil ITAM. Using diffusion of innovation theory to understand the factors impacting patient acceptance and use of consumer e-health innovations: a case study in a primary care clinic. BMC Health Serv Res 2015 Feb 21;15:71 [FREE Full text] [doi: 10.1186/s12913-015-0726-2] [Medline: 25885110]
16. Cao W, Wan Y, Tu H, Shang F, Liu D, Tan Z, et al. A web-based appointment system to reduce waiting for outpatients: a retrospective study. BMC Health Serv Res 2011 Nov 22;11(1):318 [FREE Full text] [doi: 10.1186/1472-6963-11-318] [Medline: 22108389]
17. Oliver D, Deal K, Howard M, Qian H, Agarwal G, Guenter D. Patient trade-offs between continuity and access in primary care interprofessional teaching clinics in Canada: a cross-sectional survey using discrete choice experiment. BMJ Open 2019 Mar 23;9(3):e023578 [FREE Full text] [doi: 10.1136/bmjopen-2018-023578] [Medline: 30904840]
18. Rubin G, Bate A, George A, Shackley P, Hall N. Preferences for access to the GP: a discrete choice experiment. Br J Gen Pract 2006 Oct;56(531):743-748 [FREE Full text] [Medline: 17007703]
19. Turner D, Tarrant C, Windridge K, Bryan S, Boulton M, Freeman G, et al. Do patients value continuity of care in general practice? An investigation using stated preference discrete choice experiments. J Health Serv Res Policy 2007 Jul 24;12(3):132-137. [doi: 10.1258/135581907781543021] [Medline: 17716414]
20. North F, Pecina J, Tulledge-Scheitel S, Chaudhry R, Matulis J, Ebbert J. Is a switch to a different electronic health record associated with a change in patient satisfaction? J Am Med Inform Assoc 2020 Jun 01;27(6):867-876 [FREE Full text] [doi: 10.1093/jamia/ocaa026] [Medline: 32357370]
21. Zhang X, Yu P, Yan J, Hu H, Goureia N. Patients' perceptions of web self-service applications in primary healthcare. Stud Health Technol Inform 2012;178:242-249. [Medline: 22797048]
22. Dantas LF, Fleck JL, Cyrino OFL, Hamacher S. No-shows in appointment scheduling? A systematic literature review. Health Policy. 2018. URL: https://doi.org/10.1016/j.healthpol.2018.02.002 [accessed 2021-02-11]

Abbreviations

EHR: electronic health record
OR: odds ratio
POS: Patient Online Services

Edited by C Lovis; submitted 18.10.20; peer-reviewed by KM Kuo; comments to author 11.01.21; revised version received 02.02.21; accepted 03.02.21; published 18.03.21.

Please cite as:
North F, Nelson EM, Majerus RJ, Buss RJ, Thompson MC, Crum BA
Impact of Web-Based Self-Scheduling on Finalization of Well-Child Appointments in a Primary Care Setting: Retrospective Comparison Study
JMIR Med Inform 2021;9(3):e23450
URL: https://medinform.jmir.org/2021/3/e23450
doi:10.2196/23450
PMID:33734095

Original Paper

A Chatbot for Perinatal Women's and Partners' Obstetric and Mental Health Care: Development and Usability Evaluation Study

Kyungmi Chung[1,2,3], PhD; Hee Young Cho[4], MD, PhD; Jin Young Park[1,2,3], MD, PhD

[1]Department of Psychiatry, Yonsei University College of Medicine, Yongin Severance Hospital, Yonsei University Health System, Yongin-si, Republic of Korea

[2]Center for Digital Health, Yongin Severance Hospital, Yonsei University Health System, Yongin-si, Republic of Korea

[3]Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Yonsei University Health System, Seoul, Republic of Korea

[4]Department of Obstetrics and Gynecology, CHA Gangnam Medical Center, CHA University, Seoul, Republic of Korea

Corresponding Author:
Jin Young Park, MD, PhD
Department of Psychiatry, Yonsei University College of Medicine
Yongin Severance Hospital
Yonsei University Health System
363, Dongbaekjukjeon-daero, Giheung-gu
Yongin-si
Republic of Korea
Phone: 82 31 5189 8148
Email: empathy@yuhs.ac

Abstract

Background: To motivate people to adopt medical chatbots, the establishment of a specialized medical knowledge database that fits their personal interests is of great importance in developing a chatbot for perinatal care, particularly with the help of health professionals.

Objective: The objectives of this study are to develop and evaluate a user-friendly question-and-answer (Q&A) knowledge database–based chatbot (Dr. Joy) for perinatal women's and their partners' obstetric and mental health care by applying a text-mining technique and implementing contextual usability testing (UT), respectively, thus determining whether this medical chatbot built on mobile instant messenger (KakaoTalk) can provide its male and female users with good user experience.

Methods: Two men aged 38 and 40 years and 13 women aged 27 to 43 years in pregnancy preparation or different pregnancy stages were enrolled. All participants completed the 7-day-long UT, during which they were given the daily tasks of asking Dr. Joy at least 3 questions at any time and place and then giving the chatbot either positive or negative feedback with emoji, using at least one feature of the chatbot, and finally, sending a facilitator all screenshots for the history of the day's use via KakaoTalk before midnight. One day after the UT completion, all participants were asked to fill out a questionnaire on the evaluation of usability, perceived benefits and risks, intention to seek and share health information on the chatbot, and strengths and weaknesses of its use, as well as demographic characteristics.

Results: Despite the relatively higher score of ease of learning (EOL), the results of the Spearman correlation indicated that EOL was not significantly associated with usefulness (ρ=0.26; *P*=.36), ease of use (ρ=0.19; *P*=.51), satisfaction (ρ=0.21; *P*=.46), or total usability scores (ρ=0.32; *P*=.24). Unlike EOL, all 3 subfactors and the total usability had significant positive associations with each other (all ρ>0.80; *P*<.001). Furthermore, perceived risks exhibited no significant negative associations with perceived benefits (ρ=−0.29; *P*=.30) or intention to seek (SEE; ρ=−0.28; *P*=.32) or share (SHA; ρ=−0.24; *P*=.40) health information on the chatbot via KakaoTalk, whereas perceived benefits exhibited significant positive associations with both SEE and SHA. Perceived benefits were more strongly associated with SEE (ρ=0.94; *P*<.001) than with SHA (ρ=0.70; *P*=.004).

Conclusions: This study provides the potential for the uptake of this newly developed Q&A knowledge database–based KakaoTalk chatbot for obstetric and mental health care. As Dr. Joy had quality contents with both utilitarian and hedonic value, its male and female users could be encouraged to use medical chatbots in a convenient, easy-to-use, and enjoyable manner. To boost their continued usage intention for Dr. Joy, its Q&A sets need to be periodically updated to satisfy user intent by monitoring both male and female user utterances.

(JMIR Med Inform 2021;9(3):e18607) doi:10.2196/18607

KEYWORDS

chatbot; mobile phone; instant messaging; mobile health; perinatal care; usability; user experience; usability testing

Introduction

Background

With a growing interest in chatbots based on various digital platforms such as websites, social channels, and mobile apps, a wide range of gratifications have been suggested as motivators of chatbot use. In general, productivity is considered to be a key factor in driving chatbot use, which means that the ease, speed, and convenience of using chatbots can help their users, who seek instant gratification via quick and consistent feedback and dialogue to obtain information or assistance in a timely and efficient manner [1]. Particularly, medical chatbots as a virtual doctor or educator have been built to reduce the burden of health care costs, improve the accessibility of medical knowledge, and empower patients with their medical decision-making process [2-8]. When it comes to developing medical chatbots using artificial intelligence (AI), a number of previous studies have focused on not only accurate prediction, diagnosis, or personalized management and treatment of diseases based on their symptoms [3,4,6-9], but also conversational agent role in social and emotional support and mental health interventions [10-16]. However, the major challenge perceived by more than 70% of the medical physicians in one study is the inability of health care chatbots to address the full extent of a patient's needs and understand or display the emotional state of humans [17]. Furthermore, common concerns on inaccurate and inflexible information that chatbots provided have been raised [3,5,17-20]. Despite these continuous attempts to provide patients with better user experience (UX) on informational and emotional support, both costs and benefits are still associated with the use of medical chatbots.

In addition to productivity, entertainment, and social or relational benefits, there are other main motivations to use chatbots, which are considered to be more humanlike than other interactive systems designed to support enjoyable social interactions [1]. As patients with lower health literacy are more likely to use and trust informal health information sources, such as television, social media, friends, blogs, celebrity webpages, and pharmaceutical companies, than formal ones such as doctors and health professionals [21], medical chatbots are required to provide their users with evidence-based health information as answers to questions from them. Given that the majority of pregnant women tend to use multiple information sources for their antenatal and postnatal care [22,23], obtaining conflicting information can increase anxiety levels or add uncertainty on whether or not to use a medication [24]. In fact, the attention of prenatal women seeking informal information or multiple information from multiple sources can be readily directed to social and emotional support from other experienced mothers and friends who have been in a similar situation, but they can experience stigma and receive inappropriate support due to their lack of related knowledge [25]. As more and more online communities have formed with huge numbers of female members who have undergone many different situations during pregnancy and childbirth, maintaining social interaction with their peers can encourage perinatal women to satisfy their curiosity and interests in specific information and content, which is thus perceived as an immediate and enjoyable daily activity. In turn, it means that medical chatbots with the characteristics of these peers, as well as a valid, accurate, and credible medical knowledge database, can be more likely to capture perinatal women's attention when encountering medical problems.

To encourage people to adopt and use medical chatbots, both content quality and expertise of the chatbots should be first considered in the development process. From the perspective of utilitarian and hedonic value, content quality has strongly positive effects on perceived usefulness and enjoyment, both of which influence users' usage intention [26]. Perceived expertise of the medical chatbots can increase the users' trust in the chatbots, which in turn affects their continuance intention to use the service agents [27]. In addition to the effort to improve a chatbot's content quality and expertise, it is also important to iteratively evaluate its usability and UX, both in the development process and after the completion of its development. According to Lund, who developed the Usefulness, Satisfaction, and Ease of Use (USE) Questionnaire [28], ease of use and usefulness influence each other and drive satisfaction strongly related to predicted and actual usage; ease of use can be separated into two factors, ease of use (EOU) and ease of learning (EOL), if the systems to be assessed are internal systems that its users are required to use. However, it is less likely that the two factors will be highly correlated for this chatbot based on a mobile instant messenger (MIM), as it is a flexible system used in different contexts and for different needs of individuals. Furthermore, a wide range of satisfaction dimensions (ie, productivity, entertainment, social or relational benefit, etc) can serve as motivators of chatbot use [1], and therefore, there is a need to identify these motivations or any other barriers associated with the users' intention to seek and share health information on the medical chatbot via MIM.

From the findings of a previous study based on a net valence model [29], perceived benefits were positively related to the intention to seek and share health information in social media in both Chinese and Italian samples, but only the Chinese sample showed a negative relationship between perceived risk and the intention to share health information. Until recently, little was known about the relationship between the variables in MIM-based medical chatbot use in a Korean sample. Considering that a MIM app such as KakaoTalk, which is the most popular in South Korea, is more private than other social media platforms such as YouTube, Facebook, and Twitter, it is expected that the negative relationship between the variables will not be observed in this study sample. However, it is challenging to explore the motivators and barriers to chatbot use in everyday life, not in experimental contexts, and its associations with different intention behaviors by applying a single quantitative or qualitative method, particularly in contextual usability testing (UT) without the intervention of a facilitator.

Objectives

Taken together, the primary purpose of this study is to develop a user-centered question-and-answer (Q&A) knowledge database–based chatbot for perinatal women's and their partners' obstetric and mental health care by applying a text-mining technique. The secondary purpose is to evaluate it by conducting contextual UT, thereby measuring the perception of usability and UX and their associations with motivators and barriers to chatbot use and different intention behaviors and obtaining theoretical and practical implications to supplement the weaknesses of this chatbot. Based on relevant literature, we hypothesize that this chatbot will produce both utilitarian and hedonic value during the 7-day contextual UT period.

Methods

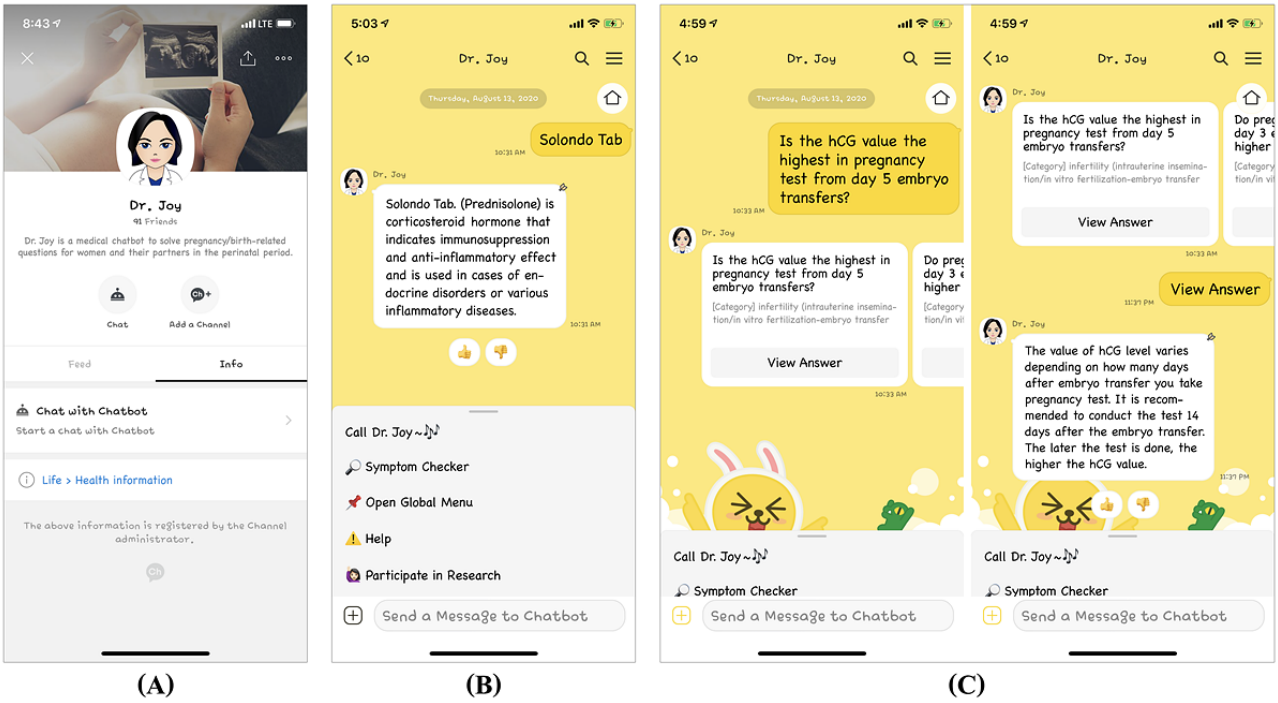
Chatbot Development

Dr. Joy was developed with the "kakao i" open builder, which allows businesses and users to create custom AI services provided in "KakaoTalk," the most widely used web- and mobile-based instant messaging application in South Korea. As kakao's AI platform could support two main features to develop a Q&A chatbot, (1) by uploading a structured Q&A Excel data file to its "knowledge+" menu or (2) by creating dialog blocks to add the users' text input and the chatbot's output to each scenario and linking these blocks in the "scenario" menu, both features were applied in this chatbot development. As this chatbot was only available in Korean, Multimedia Appendices 1-3 are provided to enhance Korean readers' understanding of all figures translated to English from Korean.

Persona

Dr. Joy was named after the second author, whose name is pronounced similarly to "Joy," as this chatbot was designed to lead users to perceive enjoyment when seeking health information and medical help for their prenatal and postnatal care. In order to look more professional to users, we provided Dr. Joy with a character of a "humanlike" female medical doctor (Figure 1 and Multimedia Appendix 1) and a formal, firm tone of voice, particularly when answering the questions. However, Dr. Joy demonstrated warmth in its informal, pleasant voice tone, manner, and emoji use when treating users in other scenarios.

Figure 1. Screenshots of (A) Dr. Joy's persona introduced on the KakaoTalk Channel and examples of (B) 1 Q&A pair and (C) 3 Q&A pairs that can be triggered when user intent matches most closely with them in Dr. Joy's obstetric and mental health–related Q&A knowledge database.



Perinatal and Postnatal Care Knowledge Database

By employing the data-mining technique, the user-friendly obstetric and mental health–related Q&A knowledge database was built. A list of 3524 refined Q&A sets was created by the following procedure. To build a data set of medical questions and their terms that are of real interest and concern among Korean perinatal women, we first developed a web crawler in Python. From one of South Korea's largest online communities for prenatal, postnatal, and maternal care, message boards with 6 different topics (ie, infertility: intrauterine insemination, in vitro fertilization, embryo transfer; pregnancy diagnosis: pregnancy test kit, ultrasound scan; pregnancy preparation; pregnancy; labor and delivery; and postpartum recovery) were chosen to be crawled, and then all posts during a 1-year period, from August 1, 2017, to August 31, 2018, were automatically collected. Contents retrieved by the web crawler were parsed into each Excel spreadsheet file by topics and stored into the following column headings: nickname (ID), timestamp (date and time), URL, title, body content (text only), and up to 3 replies.

As some contents were involved in more than one topic, all 6 topics were redefined to remove overlap. First, the topic of

"pregnancy preparation" was redefined to address all posts excluding the posts on "infertility" and "pregnancy diagnosis," and any posts irrelevant to the redefined topic were moved to the topic of either "infertility" or "pregnancy diagnosis" to effectively search questions or statements to be updated. Second, the topic of "pregnancy" was redefined to address the posts from the first to the ninth month of pregnancy because the message board about "pregnancy" covered all posts from the first to the last month of pregnancy and that about "labor and delivery" partially included posts at the tenth month of pregnancy.

From the title and body content of the posts, we extracted medical questions whose context and intent could be generally understood by both medical doctors and peer users and eliminated personal questions that were beyond a medical scope to satisfy one's own curiosity. After that, excessively long, complex questions or statements about medical and obstetric problems were refined as simple, conversational questions or statements that one might ask a MIM-based chatbot, particularly at medium length. In the next step, to establish the data set of user-friendly question and professional answer pairs on these particular topics, a total of 11 medical doctors, who were specialized in infertility (3/11, 27%), obstetrics and gynecology (6/11, 55%), and psychiatry (2/11, 18%), were recruited; 6 (55%) and 5 (45%) of these were recruited from local hospitals and university hospitals, respectively, by using a snowball sampling method. They first identified and revised inappropriate questions or statements with false terms or without user intent and contextual information, answered all 3524 questions with a consistent tone and manner, and finally cross-checked the Q&A pairs involved in their specialty. The 3524 Q&A sets were categorized as follows: (1) infertility (intrauterine insemination, in vitro fertilization, embryo transfer: 609 items), (2) pregnancy diagnosis (pregnancy test kit, ultrasound scan, blood test: 381 items), (3) pregnancy preparation (303 items), (4) pregnancy (1-36 weeks [1-9 months]: 1154 items), (5) labor and delivery (37-40 weeks [final months]: 446 items), and (6) postpartum recovery (631 items).

Following the aforementioned procedure, we filled in the Excel spreadsheet template that the chatbot builder provided, particularly with the following data: number, category, question, and answers. In addition to the Q&A knowledge database, we built a dictionary of synonyms to improve the accuracy of providing the Q&A pairs that match well with user intent (ie, search intent), as perinatal women tend to use a wide variety of abbreviations for medical terms and neologisms in the online community. This dictionary was also organized within the given Excel template and registered into the "my entity" menu.

Main Features and User Interface

As a Q&A chatbot, Dr. Joy had the main feature as a bot to answer user queries and frequently asked questions. The main feature, which was developed by the Knowledge+ feature of kakao's chatbot builder, worked by searching for questions similar to users' dialog input in the stored Q&A knowledge database and then outputting answers linked to those questions. As shown in Figure 1 (Multimedia Appendix 1), Dr. Joy, employing an AI engine called kakao i sympson (a similarity inference engine for evaluating semantic similarity between sentences), could answer all questions by offering either (1) only 1 Q&A pair that matches the best with the user intent or (2) the 3 Q&A pairs that match most closely. Even if the given 3 Q&A pairs did not completely meet users' intentions in asking a question to the chatbot, the users could come to know other peer mothers' current interests and concerns from the questions and the 11 aforementioned medical doctors' accurate, professional answers to the questions consisting of relevant medical knowledge and advice. To use this feature, users could type their questions into an input box directly or do so after calling Dr. Joy by dragging the generic menu up to open it and then tapping the button to call the chatbot. The input box and the generic menu were located at the bottom of the chatbot. Otherwise, users could also call the chatbot after accessing the graphical user interface (UI)–based global menu via the generic menu (Figure 1 and Multimedia Appendix 1).

With a particular focus on managing perinatal women's mental and physical health, other main features were developed based on predefined conversational design and rule- and choice-based dialogues, which only performed and worked within scenarios. To handle unexpected responses from the users and their unwanted escape from a prearranged conversational UI flow, Dr. Joy provided the users with dialog buttons to choose as their responses to call the linked dialog blocks, particularly motivating them to follow the given UI flow. The scenario-based additional features were designed to lead the users to learn about the importance of (1) early detection of physical and obstetric problems (if users experienced specific physical symptoms, they could check up on their current health status by answering symptom-related questions that Dr. Joy asked; this chatbot-assisted medical examination was the same as a medical doctor–administered medical examination), (2) preventative mental health care, such as a depression screening test and cognitive behavioral therapy (ie, sleep hygiene education and mindfulness-based intervention; Figure 2 and Multimedia Appendix 2), and (3) social supports from their male partners, such as fetal education and various useful tips for physical and mental health care (Figure 3 and Multimedia Appendix 3).

These needs for preventative mobile health care and social supports from the perinatal women's partners in everyday life were identified through in-depth interviews with 11 patients, 10 women and 1 man in the perinatal period, and a focus group interview with two obstetrician-gynecologist (ob/gyn) groups: (1) 3 ob/gyns at local hospitals and (2) 3 ob/gyns at university hospitals. According to the reports of the interviews, both patients and medical doctors highlighted the importance of the relationship between perinatal women and their partners on the women's mental health during the prenatal, pregnancy, and postnatal periods. Particularly, the female interviewees and the doctors' female patients who had experienced depressed symptoms expressed that they had a lack of opportunity to spend time with their partners in common; otherwise, a few women's partners had cheated on them during pregnancy. By contrast, it was reported that the male interviewee, whose wife had no specific mental problems throughout pregnancy and after birth but who experienced depressed symptoms instead of her, tried to help his wife to overcome postpartum blues by sharing house

chores, having a talk with her as much as possible, and ventilating her feelings of physical and emotional distress related to the double burden of childcare and housework. However, without their partners' support, most pregnant women and mothers had difficulty in going out to refresh themselves or to attend a variety of mental health care programs held in local community health centers, local or university hospitals, and postpartum care centers. Although the male partners were also susceptible to the women's mood fluctuations in the long-term period, both found it difficult to consult with health professionals and others (eg, family members and friends) about emotional or psychiatric problems and to consider using appropriate psychotropic medication about which a concern that it might negatively affect their fetuses might be raised. Furthermore, there has been a limitation in that the accessibility of useful information for effectively treating the women and even their partners was not significantly improved, particularly from the men's point of view.

On the basis of these findings from the interviews, the same sample of medical doctors who had participated in the development of Dr. Joy's Q&A knowledge contents as the main feature to answer questions regarding obstetric and mental health concerns in both perinatal women and their partners guided the development of additional features to enable them to manage these health-related concerns by themselves by using a medical examination, a depression screening test, alternative therapies, and more useful male partner–oriented tips and dialogues. Particularly, Dr. Joy had a male partner–friendly UI access point for use in paternal fetal education features: (1) know-how in fetal education and (2) fathers can do it (Figure 3 and Multimedia Appendix 3). Following Dr. Joy's instruction, would-be fathers or current fathers who were inexperienced in fetal education with their partners could perform step-by-step prenatal care. To promote male partners' involvement during routine prenatal care for a positive outcome in labor and delivery, Dr. Joy explained the need for partner support in a friendly tone and delivered practical strategies with relevant images in which a man actively supported his partner, showing empathic concerns and sympathetic responses to the men's difficult situation related to their pregnant partners and their social life.

Figure 2. Screenshots of user interface workflow for a depression screening test using the 10-item Edinburgh Postnatal Depression Scale that can be administered in the prenatal period, followed by the screening test result and therapy suggestions.

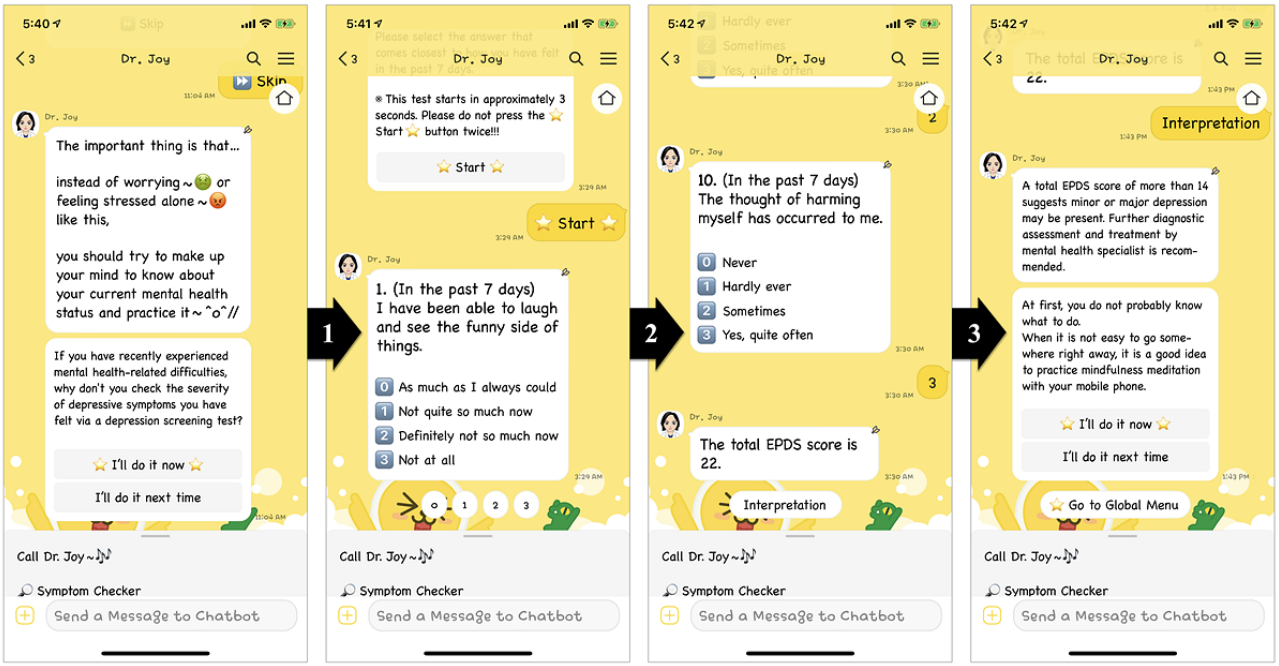
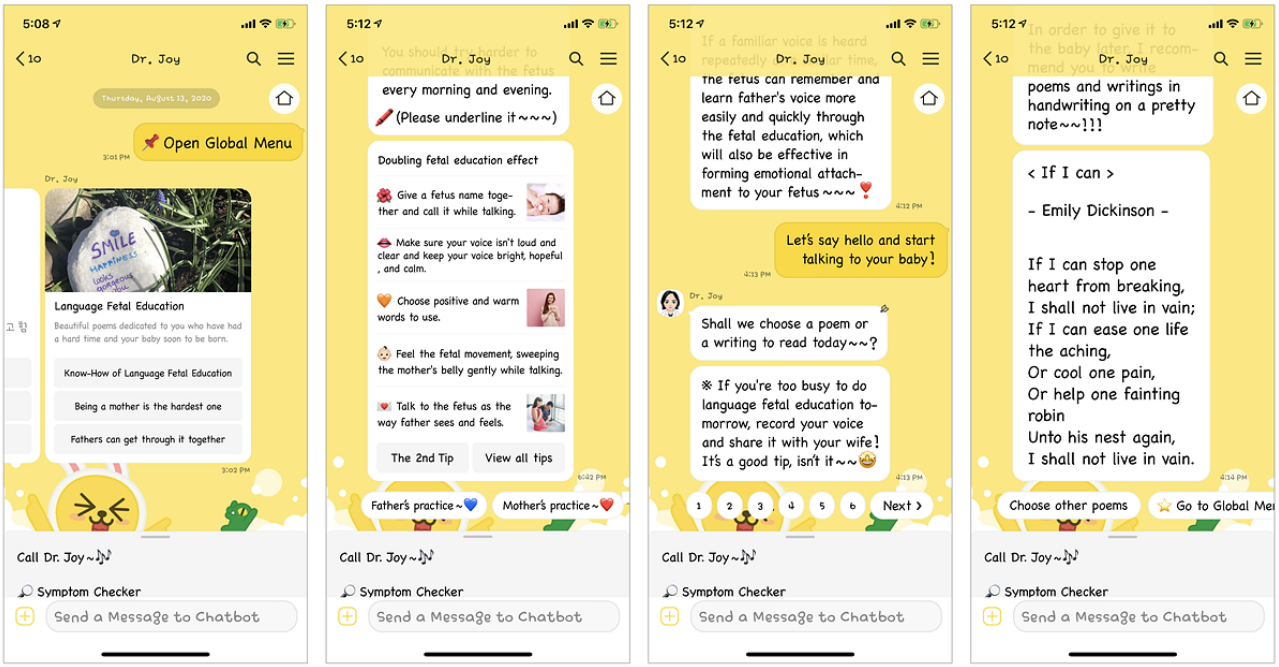


Figure 3. Screenshots of additional features with which male partners can provide their pregnant partners with social support that is needed for physical and mental health care, or women can take care of themselves.



Study Design

To measure perinatal women's and their partners' perceptions of the utilitarian and hedonic value of a medical chatbot experience, we conducted a 7-day contextual UT after completing the development of a Q&A knowledge database-based chatbot on KakaoTalk, named "Dr. Joy," for solving their obstetric and mental health problems. This study was approved by the institutional review board of CHA Bundang Medical Center, CHA University.

Recruitment

In this study, two different convenience sampling methods were used to prevent this study sample from being biased and to collect samples from the population of interest. According to the result of previous research by Nielsen and his colleagues [30], 5 users has found 85% of the usability problems, and at least 15 users were needed to discover all the usability problems. As the aim of UT was to improve the chatbot design based on the usability problems, a total of 15 participants were recruited. Of 15 participants, 6 (40%) were patients who were recruited from the outpatient clinic in the Department of Obstetrics and Gynecology, CHA Bundang Women's Hospital, CHA University. The rest (9/15, 60%) were recruited using the snowball sampling method, and therefore, 1 out of the 9 participants was asked for further potential participants who were patients at local hospitals. As Dr. Joy's medical knowledge database could cover perinatal women's questions ranging from antenatal care to postpartum care, women in pregnancy preparation and different pregnancy stages (ie, first [1-3 months: 1-12 weeks], second [4-7 months: 13-28 weeks], and third [8-10 months: 29-40 weeks] trimester and birth [puerperium: within 6 weeks after childbirth]) and their spouses were enrolled to complement the answers to both female and male partners' questions in this study. Particularly, 2 married couples, who were in first and second trimester, achieved pregnancy through infertility treatments.

Following the inclusion and exclusion criteria for recruitment, the women who gave birth but were not in the 6-week puerperal period were not eligible to participate in the study. However, if the ineligible women had a plan on pregnancy immediately after puerperium, their participation was allowed as women in pregnancy preparation.

Usability Testing: Task and Procedure

All enrolled participants completed the 7-day long UT during the entire study period, from September 30, 2019, to October 11, 2019. All the participants were given the daily tasks of asking Dr. Joy at least 3 questions at any time and place and then giving the chatbot either positive or negative feedback with emoji (Figure 2 and Multimedia Appendix 2), using at least one feature of the obstetrics chatbot, and finally sending a facilitator all screenshots for the history of the day's use via KakaoTalk before midnight. To make Dr. Joy available on their mobile phones, the participants were first required to search its name on the KakaoTalk Channel and add it as a friend, in order to readily access the chatbot service whenever they wanted to use it. One day after the UT completion, all participants were asked to fill out a questionnaire containing demographic characteristics, closed-ended questions about usability, perceived benefits and risks, and intention to seek and share health information on the chatbot, and open-ended questions about the strengths and weaknesses of its use.

Measurements

To measure the subjective usability of our newly developed chatbot service, the USE Questionnaire [28] was employed. The 30-item USE questionnaire examined the 4 subfactors of usability: usefulness (8 items), EOU (11 items), EOL (4 items), and satisfaction (7 items). All the items were anchored

from 1 (strongly disagree) to 7 (strongly agree), and these 4 mean scores were averaged across all participants and sex groups to calculate a total usability score. In addition to usability, perceived benefits (2 items) and risks (2 items), and intention to seek (SEE, 6 items) and share (SHA, 4 items) health information on the chatbot using KakaoTalk were measured on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree), and all items were adapted from Li and colleagues' net valence model [29]. Each mean score of these factors was computed for all participants and both male and female groups. Finally, the participants responded to open-ended questions about Dr. Joy's strengths and weaknesses, which could determine whether the chatbot led them to perceive utilitarian and hedonic value from using the chatbot.

Apart from the self-reported measures of chatbot UX, a list of users' utterances was collected from the reports in the analysis menu of the chatbot builder and the screenshots for the history of asking Dr. Joy at least 3 questions per day during the 7-day UT period. Based on the data on the specific questions or statements that triggered fallback messages as well as the users' positive or negative feedback on given Q&A sets extracted from the obstetric and mental health–related Q&A knowledge database, we could gain insight into the practical implications of what the questions related to real interests and concerns of male and female users were.

Statistical Analysis

To determine whether to use a nonparametric or parametric statistical analysis for the small-size data sets (N<50), a Shapiro-Wilk normality test was performed to check the normal distribution of the data. As the normality of EOL (W_{15}=0.84; *P*=.01) and perceived risks (W_{15}=0.88; *P*=.04) was violated, the Spearman correlation was chosen for the final analysis.

Results

Participant Characteristics

As presented in Table 1, 2 men, aged 38 and 40 years (mean 39.00 years, SD 1.41 years), and 13 women, aged 27 to 43 years (mean 34.31 years, SD 3.95 years), in pregnancy preparation or different pregnancy stages were enrolled in this study: (1) men: first trimester (1/2, 50%) and second trimester (1/2, 50%); (2) women: planned natural pregnancy (4/13, 31%), first trimester (2/13, 15%), second trimester (4/13, 31%), third trimester (1/13, 8%), and puerperium (2/13, 15%). All participants (15/15, 100%) reported KakaoTalk as the most frequently used instant messenger in everyday life.

When seeking health information on pregnancy or delivery to solve medical problems, all men referred to information sourced from books (2/2, 100%). However, women reported that they referred to multiple information sources, and the main source was acquaintances (7/13, 54%), followed by the internet (4/13, 31%), books (1/13, 8%), and health professionals (1/13, 8%). Particularly when using their personal computers or mobile phones to obtain online information on pregnancy or delivery, the 2 men employed different information search strategies: keyword search (1/2, 50%) and sentence search (1/2, 50%). A majority of women employed keyword search (11/13, 85%), and the others employed sentence search (1/13, 8%) and real-time search (1/13, 8%).

Table 1. Demographic information on the contextual UT participants (N=15).

ID[a]	Age (years)	Sex[b]	Pregnancy stage[c]	Pregnancy/delivery information source	Web-based information search strategy via computer or mobile phone
UTI-01	34	F	PP	Internet[d]	Keyword search[e]
UTI-02	35	F	PP	Acquaintances[f]	Keyword search
UTI-03	35	F	PP	Books[g]	Sentence search[h]
UTI-04	34	F	PP	Acquaintances	Real-time search[i]
UTI-05	31	F	FT (8 weeks)	Internet	Keyword search
UTC-06A	36	F	FT (8 weeks)	Acquaintances	Keyword search
UTC-07A	38	M	FT (8 weeks)	Books	Sentence search
UTC-08B	36	F	ST (15 weeks)	Acquaintances	Keyword search
UTC-09B	40	M	ST (15 weeks)	Books	Keyword search
UTI-10	43	F	ST (17 weeks)	Internet	Keyword search
UTI-11	33	F	ST (23 weeks)	Internet	Keyword search
UTI-12	39	F	ST (24 weeks)	Health professionals[j]	Keyword search
UTI-13	27	F	TT (32 weeks)	Acquaintances	Keyword search
UTI-14	31	F	P (3 weeks after birth)	Acquaintances	Keyword search
UTI-15	32	F	P (3 weeks after birth)	Acquaintances	Keyword search

[a]Two different ID labels were assigned to differentiate couples (UTC) from individuals (UTI) [31], and those with the same uppercase letters (A or B) are a married couple.

[b]F: female; M: male.

[c]PP: pregnancy preparation (planned natural pregnancy); FT: first trimester; ST: second trimester; TT: third trimester; P: puerperium.

[d]Internet includes portal/search engines, online communities, blogs, vlogs, etc.

[e]Keyword search with simple words, search operators, hashtags, etc.

[f]Acquaintances include friends, colleagues, online community members, experienced mothers in the same postnatal care center, etc.

[g]Books include encyclopedias of pregnancy and birth, essays and articles written by medical doctors, magazines, etc.

[h]Sentence search with a single statement/question or multiple statements/questions.

[i]Real-time search means choosing and looking for attention-capturing content published in real time on the internet.

[j]Health professionals include medical doctors, nurses, etc. An acquaintance who was a medical doctor was included in health professionals.

Quantitative Data Analysis

The results from the USE questionnaire are shown in Table 2. Among the psychometric aspects of usability, the mean score of EOL was the highest, followed by the EOU, satisfaction, and usefulness scores in this sample. Even though the number of participants was insufficient to determine statistical significance of the difference in all 4 subfactors and total usability scores across sex, male participants showed higher mean scores than female ones. Both men and women had a tendency to rate the scores of usefulness and satisfaction lower than those of EOU and EOL; these trends were also identified within the total scores of usability and its subfactors.

Despite the higher mean score of EOL in the entire participant group, the results of the Spearman correlation indicated that there were no significant associations with usefulness, EOU, satisfaction, or total usability scores (Table 3). Unlike EOL, the total usability and other 3 subfactors had significant positive associations with each other (all $\rho>0.80$; $P<.001$).

Table 2. Descriptive statistics for sex difference in responses to USE questionnaire on the medical chatbot via KakaoTalk (N=15).

Sex	Usability subfactors[a], mean (SD)				
	USE[b]	EOU[c]	EOL[d]	SAT[e]	Total
Men (n=2)	5.43 (1.21)	6.05 (0.96)	7.00 (0.00)	5.57 (2.02)	6.01 (2.02)
Women (n=13)	4.78 (1.12)	5.23 (0.67)	6.25 (0.71)	4.80 (1.20)	5.27 (0.82)
Total (N=15)	4.87 (1.11)	5.34 (0.73)	6.35 (0.71)	4.90 (1.26)	5.37 (0.85)

[a]Usability was measured by the average score of 4 subfactors, which is presented as the "Total" score in this table. All scales were rated from 1 (strongly disagree) to 7 (strongly agree).

[b]USE: usefulness.

[c]EOU: ease of use.

[d]EOL: ease of learning.

[e]SAT: satisfaction.

Table 3. Spearman rank correlation analysis of associations among individual and total usability scores from USE questionnaire on the medical chatbot via KakaoTalk (N=15).[a]

Subfactors	USE[b]	EOU[c]	EOL[d]	SAT[e]	Total
USE					
Correlation coefficient (ρ)	1.00	0.82	0.26	0.98	0.97
P value (2-tailed)	—[f]	<.001	.36	<.001	<.001
EOU					
Correlation coefficient (ρ)	0.82	1.00	0.19	0.81	0.89
P value (2-tailed)	<.001	—	.51	<.001	<.001
EOL					
Correlation coefficient (ρ)	0.26	0.19	1.00	0.21	0.32
P value (2-tailed)	.36	.51	—	.46	.24
SAT					
Correlation coefficient (ρ)	0.98	0.81	0.21	1.00	0.95
P value (2-tailed)	<.001	<.001	.46	—	<.001
Total					
Correlation coefficient (ρ)	0.97	0.89	0.32	0.95	1.00
P value (2-tailed)	<.001	<.001	.24	<.001	—

[a]Usability was measured by the average score of 4 subfactors, which is presented as the "Total" score in this table.

[b]USE: usefulness.

[c]EOU: ease of use.

[d]EOL: ease of learning.

[e]SAT: satisfaction.

[f]Not applicable.

Regardless of sex, the total mean score for SEE showed a similar trend to that for SHA. Compared to women, who rated the SEE score similar to the SHA score, men had a tendency to rate the mean score for SEE higher than that for SHA. Apart from the rating on SHA, the ratings on perceived benefits, SEE, and even perceived risks were higher in men than in women (Table 4).

According to the results of the Spearman correlation analysis, perceived risks exhibited no significant negative associations with perceived benefits, SEE, or SHA, whereas perceived benefits exhibited significant positive associations with both SEE and SHA. As can be seen in Table 5, perceived benefits were more strongly associated with SEE (ρ=0.94; *P*<.001) than with SHA (ρ=0.70; *P*=.004).

Table 4. Descriptive statistics for sex difference in responses to perceived benefits and risks and intention to seek and share health information on the medical chatbot via KakaoTalk (N=15).[a]

Sex	Factors, mean (SD)			
	PB[b]	PR[c]	SEE[d]	SHA[e]
Men (n=2)	6.25 (1.06)	3.00 (0.71)	6.17 (0.47)	5.00 (1.41)
Women (n=13)	5.19 (1.03)	2.54 (1.64)	5.01 (1.21)	4.98 (1.30)
Total (N=15)	5.33 (1.06)	2.60 (1.54)	5.17 (1.20)	4.98 (1.26)

[a]All scales were rated from 1 (strongly disagree) to 7 (strongly agree).

[b]PB: perceived benefits.

[c]PR: perceived risks.

[d]SEE: intention to seek health information.

[e]SHA: intention to seek health information.

Table 5. Spearman rank correlation analysis of associations among scores on perceived benefits and risks and intention to seek and share health information on the medical chatbot via KakaoTalk (N=15).

Factor	PB[a]	PR[b]	SEE[c]	SHA[d]
PB				
Correlation coefficient (ρ)	1.00	−0.29	0.94	0.70
P value (2-tailed)	—[e]	.30	<.001	.004
PR				
Correlation coefficient (ρ)	−0.29	1.00	−0.28	−0.24
P value (2-tailed)	.30	—	.32	.40
SEE				
Correlation coefficient (ρ)	0.94	−0.28	1.00	0.73
P value (2-tailed)	<.001	.32	—	.002
SHA				
Correlation coefficient (ρ)	0.70	−0.24	0.73	1.00
P value (2-tailed)	.004	.40	.002	—

[a]PB: perceived benefits.

[b]PR: perceived risks.

[c]SEE: Intention to seek health information

[d]SHA: Intention to seek health information

[e]Not applicable.

Qualitative Data Analysis

For the qualitative data analysis, thematic analysis [32,33] was conducted on user utterance data collected via two different sources, (1) kakao i open builder and (2) usability testers, in order to complement missing data and monitor the users' responses to a single answer or 3 Q&A sets that Dr. Joy provided. The raw data of user utterances during the 7-day UT period were extracted from the reports in the analysis menu of the chatbot builder and downloaded as separate text files, and then the files were combined into two different data sets: (1) default fallback intent (ie, users' questions or statements that triggered error messages from Dr. Joy) and (2) predefined user intent (ie, those which triggered users' positive or negative feedback on given Q&A sets from Dr. Joy's knowledge database).

From the data sets, initial major themes and the chatbot's identity, strengths, and weaknesses were produced from 316 user utterances (310 questions or statements and 6 responses to chatbot's answers in UT) and 30 open-ended responses to the post-test questionnaire after the UT was completed (15 strengths and 15 weaknesses) by the first author. More detailed descriptions of the major themes were then generated, compared, and revised by three coders (the first author and bachelors- and masters-level research assistants) before agreement between appropriate coding categories for the 5 refined minor themes and memorable quotes was reached (Textbox 1). To ensure intercoder reliability for all 5 themes, the coded transcripts on which all coders agreed were included based on an examination of coding disagreement.

Textbox 1. Illustrative quotes from user utterance data by theme.

Theme 1-1: Chatbot Identity as a Social Agent

- (1) These days, I tend to fall asleep easily at night. But...I wake up in the middle of the night, feel restless for more than two hours, and then...fall asleep again. It wasn't like this in the early first trimester of pregnancy, but since the 15th week, sleep quality has dramatically decreased. How can I improve the quality of my sleep?

  [UTI-10]

- (2) I'm 39 and pregnant with my third child. I'm so worried that my belly at 23 weeks pregnant is much bigger than that at the same week of my previous pregnancy. I'm also worried about the deep stretch marks on my belly. Anyway...my PCP said to me...that my baby and amniotic fluid volume were normal at 23 weeks of pregnancy. Is it all right if I don't have to worry about my belly size?

  [UTI-12]

- (3) Since I was a patient with an early cervical cancer, I have eaten turmeric powder with a teaspoon three times a day after each meal. After I found I was pregnant, I didn't eat it for 2 months. I reached a stable period of pregnancy, so I wonder if I can eat it once a day by reducing my turmeric powder intake.

  [UTI-10]

Theme 2-1: Strengths in Chatbot's Utilitarian and Hedonic Values

- (4) It was user-friendly to use and easy to understand how to ask questions.

  [UTC-08B]

- (5) Convenience, Speed, and Usefulness!

  [UTI-11]

- (6) A wide variety of information was provided by entering only a simple keyword.

  [UTI-13]

- (7) This chatbot was easy to access, and I could ask questions at any time.

  [UTI-12]

- (8) It was fun to see more answers to others' frequently asked questions as well as an answer to my question.

  [UTI-4]

- (9) It was so unique and enjoyable...that I could make more than one choice from other three Q&As.

  [UTI-10]

Theme 2-2: Strengths in Chatbot's Informational Support

- (10) For me, it was a good opportunity to know basic information more accurately.

  [UTI-5]

- (11) The strengthen was that I could look forward to more reliable responses from medical doctors, not incredible information from the Internet or online communities.

  [UTI-14]

- (12) While using this chatbot, I realized that I've had a lot of questions since I got pregnant and that I needed a mobile application like chatbot to solve them.

  [UTC-08B]

Theme 3: Weaknesses in Chatbot's Content Coverage

- (13) I had to keep asking questions to get the answers that I expected.

  [UTI-05]

- (14) Blunt answers to my pointed questions...

  [UTI-02]

- (15) Sometimes...this chatbot could not recognize all abbreviations commonly used. It left a lot to be desired.

  [UTI-01]

- (16) I think its database range was too narrow. It was impossible to check the information on the government policies to boost birthrate. If it has a dictionary-style user interface where I can see each of the Q&As whenever I want, I'll spend my spare time reading them.

  [UTI-11]

- (17) How can I have a child of the desired sex?

  [UTC-06A]
- (18) What is the chance of having a girl after a boy?

  [UTC-06A]
- (19) Can I tell the sex of my baby by my belly shape?

  [UTI-03]
- (20) What is the possibility that the baby's sex will change after the ultrasound scan?

  [UTI-02]
- (21) Although nightmares during pregnancy are a common symptom of pregnancy, it remains a little disappointing that I have not received a professional answer to that.

  [UTI-12]

Theme 1-1: Chatbot Identity as a Social Agent

Although Dr. Joy was a text-based Q&A chatbot whose weakness was the lack of ability to understand what users were saying and to interact with them in a natural manner, it was found that our participants tended to consider Dr. Joy as a social actor as follows:

When asking a question, excessively detailed, personal information or their stories were included in their questions as if they talked to a close friend or acquaintance (Textbox 1, quotes 1-3).

Humanlike responses to Dr. Joy's answers were yielded appreciating her valuable recommendations and professional medical knowledge. Our participants said the following: "Thank you."; "Yes!"; "I got it."; "OK, I see it."; "Sure, I will."; "I need to keep it properly!"

Theme 1-2: Chatbot Identity as a Male-Friendly Agent

Even though the facilitator gave them no specific instruction on what to ask, male participants raised questions about themselves as well as their wives, and female participants also did so about their husbands as well as themselves. They asked the following: "Can men have morning sickness?"; "Should men take folic acid?"; "Is there postpartum depression for fathers?"; "Does fathers' medication affect pregnancy?"; "Husband is really having a hard time"; "What age is considered advanced paternal age?"

Theme 2-1: Strengths in Chatbot's Utilitarian and Hedonic Values

According to the reports of all user utterance data, participants tried to view other given Q&A sets rather than press the negative feedback button or produce negative utterances on all Q&A sets. Regarding the strengths of this newly developed chatbot, a response that participants had in common was that Dr. Joy had both utilitarian and hedonic values (Textbox 1, quotes 4-9).

Theme 2-2: Strengths in Chatbot's Informational Support

In addition to these strong points, some participants mentioned the benefits from health-related information sourced from Dr. Joy (Textbox 1, quotes 10-12).

Theme 3: Weaknesses in Chatbot's Content Coverage

The most frequently reported weak point was that Dr. Joy failed to meet all user intents and to cover a much broader range of content domains because we focused more on helping perinatal women to prevent and solve their own mental and physical problems than on offering them answers to baby-oriented questions (Textbox 1, quotes 13-16).

In particular, routine, nonmedical questions, which were difficult for health professionals to answer, were quite often asked. For example, 2 couples at 8 and 15 weeks of pregnancy wondered about the sex of a child, so they hoped that plenty of relevant content would be supplemented in the next update. Other asked questions are listed in Textbox 1 (quotes 17-21).

Discussion

Principal Findings

In this study, we aimed to develop and evaluate a user-friendly Q&A chatbot with quality content and expertise for perinatal women's and their partners' obstetric and mental health care. This study could add to the literature by comparing the developed system or the approach of other existing chatbots with that of Dr. Joy, highlighting its technical and design contributions, and providing theoretical and empirical evidence for the perception of its UX values in the field of application addressed.

As productivity has been considered as the main motivation for chatbot use [1], this "always-on" Q&A chatbot for offering 24/7 digital support to perinatal women and their partners can be an easier and more efficient way to obtain credible information and be more intuitive to the target users than conventional means (eg, books, internet search, acquaintances, and health professionals). In line with this study, previous studies tried to expand the Q&A chatbots' own knowledge databases to ensure content quality and improve response capability. Chung and his colleagues [19,34] applied (1) the expert-based approach to create rules for the provision of the medical information and (2) the data-based approach to provide customized information based on the already established medical knowledge database for the chatbot-based health care service, thus increasing reliability. In order to create a domain-specific or context-based chatbot to provide optimal, up-to-date answers immediately,

the high-quality chatbot knowledge was extracted from social networking services such as Twitter [20], online discussion forums as web communities [35], and messengers [19]. Similar to our approach, Jeong and Seo [20] proposed a keyword matching–based answer retrieval technique based on the collection of Q&A sets from Twitter by utilizing the tweet-and-reply and the tweet-and-mention pairs and the refinement of the newly collected pairs by adding them to the existing Q&A knowledge database. As these related works focused on developing the Q&A chatbots' answer retrieval technique to provide more accurate and flexible answers to their users, the response appropriateness of each chatbot based on quantitative data such as self-report questionnaire [20] or recall and precision measurement [18,35] was evaluated. While these proposed knowledge databases and answer retrieval techniques for Q&A chatbots were appropriate to be applied to general health care or lifecare services whose target users and content coverage were not specified, there have also been a variety of informative chatbots designed for the specific purposes of supporting pregnant women and mothers or families with young children in emergency situations [6] and providing low-cost accessible fertility and preconception health education for perinatal women [36] or breastfeeding education for community health workers and mothers in under-developed areas [37].

As entertainment and social or relational benefits have been regarded as other main motivations for chatbot use [1], the chatbot can make the process of seeking medical help enjoyable and improve the relationships between couples who need social support from their partners or care for their mental state when undergoing a stressful situation. Particularly in this study, the recommendation of evidence-based digital therapeutics, fetal education, and useful tips applicable in their daily life, as well as the establishment of a specialized medical knowledge database which fits the personal interests of women and their partners, was of great importance in developing a medical chatbot to promote their physical and mental health in the perinatal period. In the development of the first version of Dr. Joy, we focused more on enhancing and assessing the utilitarian and hedonic quality of the KakaoTalk-based Q&A chatbot as follows: (1) by building and expanding its own Q&A knowledge database with questions that were collected from peer pregnant women's and mothers' posts in an online community for prenatal, postnatal, and maternal care via the text-mining technique and were answered by medical specialists in the field of infertility, obstetrics and gynecology, and psychiatry; (2) by suggesting 3 optional Q&A pairs in response to the question queries of women and their partners in the perinatal period via kakao's similarity inference engine for assessing semantic similarity between the new query and the existing Q&A sets; (3) by providing them with dialogue-based procedural recommendations and helping easily apply the knowledge to either themselves or their partners; and (4) by defining the chatbot's identity as a medical doctor and maintaining a differentiated tone, manner, and UI when responding directly to the query and when dealing with social support– and mental health–related issues. Unlike the developed chatbots and their approaches in the aforementioned studies, this study took into account three user motivations (ie, productivity, entertainment, and social or relational benefits) and two UX values (ie, utilitarian and hedonic values) at the same level in the process of developing and assessing this medical chatbot, respectively.

The main finding of this study was that both utilitarian and hedonic value could be produced by this newly developed Q&A knowledge database–based chatbot for perinatal women's and their partners' obstetric and mental health care during the 7-day contextual UT period. According to the results of the USE questionnaire, it was found that Dr. Joy was very easy to learn and quick to apply, while achieving a high level of usefulness, EOU, satisfaction, and total usability was not guaranteed by its high learnability. However, given the strong associations among these 3 usability subfactors and total usability scores, it can be expected that an increase in the level of one or more usability subfactors will ensure good usability. The weak association between EOL and other subfactors also reflects that this KakaoTalk-based chatbot is a flexible system used in different contexts and for different needs of individuals [28]. As perceived usefulness, as well as perceived enjoyment, can be strongly affected by content quality as one of influential determinants of usage intention [26], Dr. Joy could provide its users with more intriguing content in its multiple Q&A responses based on the Q&A knowledge database to motivate them to acquire credible knowledge, even if the response outcomes might be a little out of line with what they expected. As reflected in the responses to the open-ended question about the strengths of Dr. Joy, participants highlighted not only the hedonic value as represented by fun, pleasure, and enjoyment, but also the utilitarian value as represented by usefulness, speed and ease to use, and convenience. In terms of its weaknesses, participants who asked questions beyond the coverage of our Q&A knowledge database pointed out that Dr. Joy with medical expertise had to suggest the right set of answers that successfully aligned with user intent, thereby enhancing its users' trust in and their continued usage intention for the chatbot [27]. In this respect, the improvement in the quality of its Q&A set contents is of utmost importance.

Another finding was that the negative association between the perceived benefits and risks of using Dr. Joy was not significantly strong enough to influence behavioral intention in a negative direction. Furthermore, Dr. Joy led its users to perceive a low level of risks that discussing health-related information on this medical chatbot via KakaoTalk would confront them with unwanted problems or that the expected benefits of doing so would not materialize. With a low possibility of trade-off between benefit and risk, the different intentions to seek and share health information on Dr. Joy were significantly associated only with the perceived benefits, not with the perceived risks. The more its users think Dr. Joy can benefit them, the more likely they are to seek and share information from it. Compared to women, who scored SEE and SHA at similar level, the men had more intention to seek health information on medical chatbot via KakaoTalk than the women. This might be because the male partners have comparatively less opportunity to access information sources than perinatal women, who have tended to seek medical help from multiple informal and formal sources [25]. As pregnant women's partners, our male participants, whose main source of pregnancy or delivery information was books such as encyclopedias of

pregnancy and birth or essays written by medical doctors, were less likely to show the tendency to double-check information from other sources by sharing Dr. Joy's relatively more credible information verified by health professionals. In line with the findings of our previous study [23], it can be explained that female participants, who reported relying more on multiple word-of-mouth sources of information and less on health professionals, were highly likely to share many concerns that they were reluctant to discuss with their doctors in the outpatient clinic, particularly with this KakaoTalk chatbot with a humanlike medical doctor persona.

In addition to these theoretical implications, the qualitative data suggested empirical implications for developing the next version of Dr. Joy. The main Q&A feature of this version of the informative medical chatbot was based on the response selection for a single-turn conversation, thereby intending to elicit no specific conversational responses to the given Q&A sets from the users. Nevertheless, 6 (40%) out of 15 participants showed a positive, polite attitude toward the chatbot's answers, as if the participants had asked private questions with more personal information and responded to their doctors to show that they would follow their answers in reality (Thank you; Yes!; I got it; OK, I see it; Sure, I will; I need to keep it properly). Surprisingly, none of the participants left any negative feedback or rude, abusive utterances (eg, curses or insults) to the Q&A sets that might not meet their real intent in asking questions. This might be because the participants could not completely rule out the possibility that all their utterances would be monitored by the facilitator or researchers for the purpose of the data analysis. Despite the concern of the Hawthorne effect, these behaviors might also reflect that some users perceived Dr. Joy to be a social agent to maintain a doctor-patient–like relationship with the chatbot. As the greatest advantage of this mobile chatbot is that chatbot designers and developers can readily collect the users' dialog inputs that were not added to the dialog blocks in advance, it can be expected to update the users' utterance data for machine learning purposes and the chatbot's dialog outputs and conversational UI, as well as the content values that reside in the knowledge base on a regular cycle. Particularly in terms of regular updates of the contents of Q&A sets, nonmedical but pregnancy-related subjects (eg, pronatalist policies for increasing fertility and birth rate) extracted from active users' dialog input logs should be included to increase user retention and engagement and to decrease anxiety levels by clarifying the uncertainty of conflicting information from multiple sources, based on previous studies [22-24].

Last but not least, this study found that the male partners had needs for emotional support and information in the period of pregnancy, birth, and early fatherhood, indicating that the possibility of their needs might have been implicitly disregarded, as revealed by other studies [38-40]. Most importantly, given that pregnant women's psychological well-being and positive pregnancy experience are closely related to better partner relationships [41,42], it is important to support male partners by adding men-oriented Q&A sets from male partners' perspectives into this new chatbot's knowledge database, thus helping them to understand and manage the challenges of pregnancy, birth, and the postpartum period.

Limitations and Future Direction

As this study introduced an early-stage outcome of a government-funded research and development (R&D) project whose milestone was to investigate at least 10 perinatal women's uptake of this initial version of Dr. Joy, the sample size of the study (N=15) was too small and its sex ratio was too unbalanced to generalize the findings to a larger population and guarantee the effectiveness of the medical Q&A chatbot, in spite of both male and female participants' positive perceptions of the chatbot. Even though it is well-known that this sample size is enough to find out the practical implications for improving the UX of this chatbot based on its end users' real voice and log data [30], this study has further limitations as follows:

First, the user utterance data from the small sample might be insufficient to accumulate Q&A data sets of a wide variety of pregnant women's and their partners' questions and concerns differently expressed with their own terms and in their own problematic situations, because Dr. Joy was designed to cover a wide range of pregnancy- and delivery-related information that was classified into 6 subjects. After the update of the Q&A sets via this usability study, the aim of this R&D project is to increase the number of active chatbot users by at least 100, collect more utterance data, and keep the Q&A knowledge base up to date. Comparison between the perception of Dr. Joy before it was initially released and that after being updated will be drawn to examine its robust uptake and the favorable perception of its utilitarian and hedonic value.

Second, Dr. Joy is geared toward encouraging perinatal women relying on multiple informal information sources to obtain evidence-based information for decision support. For this reason, we only recruited a small number of targeted participants by adopting two different convenience sampling methods to refrain from recruiting only the patients who established a good rapport with the medical doctors involved in the development of Dr. Joy, or those whose main information source was solely their doctors. However, a relatively small sample is potentially biased given the nonprobability sampling method where the sample can be taken from the units of the population that are easily accessible, thus failing to accurately reflect the responses of a large population. To deal with this potential bias of the study sample, the right probability sampling methods such as simple random sampling or clustering sampling will be used with a large sample size in future studies.

Finally, considering that a full-term pregnancy lasts 38 weeks or longer, a 7-day study period is insufficient to assess whether Dr. Joy can improve the participants' knowledge, answer their questions effectively, or be useful for certain tasks, even if the participants provided positive usability and UX ratings in this study. To answer these research questions, which remain open for future studies, a perinatal and mental health–related variable should be directly adopted in the short-term study period, or a more longitudinal evaluation should be performed. Taken together, future studies will benefit from addressing these limitations.

Conclusions

In sum, this study provides the potential for the uptake of this newly developed Q&A knowledge database–based KakaoTalk chatbot for perinatal women's and their partners' obstetric and mental health care. As Dr. Joy has quality contents, which are positively linked with both utilitarian and hedonic value, its male and female users can be encouraged to adopt and use medical chatbots in a convenient, easy-to-use, and pleasant manner. To boost their intention to continue use of Dr. Joy, its Q&A sets should be periodically updated to satisfy more user intent by monitoring both male and female user utterances.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C0911).

Conflicts of Interest

None declared.

Multimedia Appendix 1
[PNG File , 299 KB - medinform_v9i3e18607_app1.png]

Multimedia Appendix 2
Original Korean screenshots of user interface workflow for a depression screening test using the 10-item Edinburgh Postnatal Depression Scale that can be administered in the prenatal period, followed by the screening test result and therapy suggestions.
[PNG File , 283 KB - medinform_v9i3e18607_app2.png]

Multimedia Appendix 3
Original Korean screenshots of additional features with which male partners can provide their pregnant partners with social support that is needed for physical and mental health care, or women can take care of themselves.
[PNG File , 340 KB - medinform_v9i3e18607_app3.png]

References

1. Brandtzaeg P, Følstad A. Why people use chatbots. 2017 Presented at: International Conference on Internet Science; 2017; Greece. [doi: 10.1007/978-3-319-70284-1_30]
2. Madhu D, Jain C, Sebastain E, Shaji S, Ajayakumar A. A novel approach for medical assistance using trained chatbot. 2017 Presented at: International Conference on Inventive Communication and Computational Technologies (ICICCT); 2017; New Delhi. [doi: 10.1109/icicct.2017.7975195]
3. Divya S, Indumathi V, Ishwarya S, Priyasankari M, Devi S. A self-diagnosis medical chatbot using artificial intelligence. Journal of Web Development and Web Designing 2018:1-7 [FREE Full text] [doi: 10.46610/jowdwd]
4. Ghosh S, Bhatia S, Bhatia A. Quro: Facilitating User Symptom Check Using a Personalised Chatbot-Oriented Dialogue System. Stud Health Technol Inform 2018;252:51-56. [Medline: 30040682]
5. Mishra S, Bharti D, Mishra N. Dr. Vdoc: A Medical Chatbot that Acts as a Virtual Doctor. Research & Reviews: Journal of Medical Science and Technology 2018 [FREE Full text]
6. Vaira L, Bochicchio M, Conte M, Casaluci F, Melpignano A. MamaBot: a System based on ML and NLP for supporting Women and Families during Pregnancy. 2018 Presented at: Proceedings of the 22nd International Database Engineering & Applications Symposium; 2018; Provincia di Reggio Calabria. [doi: 10.1145/3216122.3216173]
7. Chaix B, Bibault J, Pienkowski A, Delamon G, Guillemassé A, Nectoux P, et al. When Chatbots Meet Patients: One-Year Prospective Study of Conversations Between Patients With Breast Cancer and a Chatbot. JMIR Cancer 2019 May 02;5(1):e12856 [FREE Full text] [doi: 10.2196/12856] [Medline: 31045505]
8. Mugoye K, Okoyo H, Mcoyowo S. Smart-bot Technology: Conversational Agents Role in Maternal Healthcare Support. 2019 Presented at: IST-Africa Week Conference (IST-Africa); 2019; Nairobi. [doi: 10.23919/istafrica.2019.8764817]
9. Jungmann SM, Klan T, Kuhn S, Jungmann F. Accuracy of a Chatbot (Ada) in the Diagnosis of Mental Disorders: Comparative Case Study With Lay and Expert Users. JMIR Form Res 2019 Oct 29;3(4):e13863 [FREE Full text] [doi: 10.2196/13863] [Medline: 31663858]
10. Oh KJ, Lee D, Ko B, Choi HJ. A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. 2017 Presented at: 18th IEEE International Conference on Mobile Data Management (MDM); 2017; Daejeon. [doi: 10.1109/mdm.2017.64]
11. Hoermann S, McCabe KL, Milne DN, Calvo RA. Application of Synchronous Text-Based Dialogue Systems in Mental Health Interventions: Systematic Review. J Med Internet Res 2017 Jul 21;19(8):e267 [FREE Full text] [doi: 10.2196/jmir.7023] [Medline: 28784594]

12. Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
13. Inkster B, Sarda S, Subramanian V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. JMIR Mhealth Uhealth 2018 Nov 23;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
14. Liu B, Sundar SS. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot. Cyberpsychol Behav Soc Netw 2018 Oct;21(10):625-636. [doi: 10.1089/cyber.2018.0110] [Medline: 30334655]
15. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an Artificially Empathic Conversational Agent for Mental Health Applications: System Design and User Perceptions. J Med Internet Res 2018 Jun 26;20(6):e10148 [FREE Full text] [doi: 10.2196/10148] [Medline: 29945856]
16. Kretzschmar K, Tyroll H, Pavarini G, Manzini A, Singh I, NeurOx Young People's Advisory Group. Can Your Phone Be Your Therapist? Young People's Ethical Perspectives on the Use of Fully Automated Conversational Agents (Chatbots) in Mental Health Support. Biomed Inform Insights 2019;11:1178222619829083 [FREE Full text] [doi: 10.1177/1178222619829083] [Medline: 30858710]
17. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' Perceptions of Chatbots in Health Care: Cross-Sectional Web-Based Survey. J Med Internet Res 2019 Apr 05;21(4):e12887. [doi: 10.2196/12887] [Medline: 30950796]
18. Afifi Mohamad Safee M, Mohd Saudi M, Pitchay SA, Ridzuan F, Basir N, Saadan K, et al. Hybrid Search Approach for Retrieving Medical and Health Science Knowledge from Quran. IJET 2018 Oct 07;7(4.15):69. [doi: 10.14419/ijet.v7i4.15.21374]
19. Chung K, Park RC. Chatbot-based heathcare service with a knowledge base for cloud computing. Cluster Comput 2019;22(1):1925-1937. [doi: 10.1007/s10586-018-2334-5]
20. Jeong SS, Seo YS. Improving response capability of chatbot using twitter. J Ambient Intell Human Comput 2019 Jun 14. [doi: 10.1007/s12652-019-01347-6]
21. Chen X, Hay JL, Waters EA, Kiviniemi MT, Biddle C, Schofield E. Health Literacy and Health Information Technology Adoption: The Potential for a New Digital Divide. J Health commun 2018;23(8):724-734. [Medline: 27702738]
22. Hämeen-Anttila K, Jyrkkä J, Enlund H, Nordeng H, Lupattelli A, Kokki E. Medicines information needs during pregnancy: a multinational comparison. BMJ Open 2013 Apr;3(4) [FREE Full text] [doi: 10.1136/bmjopen-2013-002594] [Medline: 23624989]
23. Chung K, Cho HY, Kim YR, Jhung K, Koo HS, Park JY. Medical Help-Seeking Strategies for Perinatal Women With Obstetric and Mental Health Problems and Changes in Medical Decision Making Based on Online Health Information: Path Analysis. J Med Internet Res 2020 Mar 04;22(3):e14095 [FREE Full text] [doi: 10.2196/14095] [Medline: 32130139]
24. Hämeen-Anttila K, Nordeng H, Kokki E, Jyrkkä J, Lupattelli A, Vainio K, et al. Multiple information sources and consequences of conflicting information about medicine use during pregnancy: a multinational Internet-based survey. J Med Internet Res 2014;16(2):e60 [FREE Full text] [doi: 10.2196/jmir.2939] [Medline: 24565696]
25. Griffiths KM, Crisp DA, Barney L, Reid R. Seeking help for depression from family and friends: a qualitative analysis of perceived advantages and disadvantages. BMC Psychiatry 2011 Dec 15;11:196 [FREE Full text] [doi: 10.1186/1471-244X-11-196] [Medline: 22171567]
26. Yang H, Lee H. Understanding user behavior of virtual personal assistant devices. Inf Syst E-Bus Manage 2019;17(1):65-87. [doi: 10.1007/s10257-018-0375-1]
27. Molinillo-Jimenez S, Viglia G, Domínguez Gómez J, Ekinci Y. This chatbot is a smart one! Does perceived expertise increase willingness to interact with chatbots? 2019. Repositorio Institucional de la Universidad de Málaga. 2019. URL: https://hdl.handle.net/10630/18185 [accessed 2019-12-10]
28. Lund AM. Measuring usability with the use questionnaire 12. Usability interface 2001;8(2):3-6.
29. Li Y, Wang X, Lin X, Hajli M. Seeking and sharing health information on social media: A net valence model and cross-cultural comparison. Technological Forecasting and Social Change 2018 Jan;126:28-40. [doi: 10.1016/j.techfore.2016.07.021]
30. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. 1993 Presented at: INTERACT'93 and CHI'93 conference on Human factors in computing systems; 1993; Amsterdam. [doi: 10.1145/169059.169166]
31. Mizuno K, Tanaka M, Yamaguti K, Kajimoto O, Kuratsune H, Watanabe Y. Mental fatigue caused by prolonged cognitive load associated with sympathetic hyperactivity. Behav Brain Funct 2011 May 23;7:17 [FREE Full text] [doi: 10.1186/1744-9081-7-17] [Medline: 21605411]
32. Clarke V, Braun V, Hayfield N. Thematic analysis. In: Qualitative psychology: A practical guide to research methods. Thousand Oaks, CA: SAGE Publications; 2015:222-248.
33. Suffoletto B, Kristan J, Person ML, Chung T, Clark DB. Optimizing a Text Message Intervention to Reduce Heavy Drinking in Young Adults: Focus Group Findings. JMIR Mhealth Uhealth 2016 Jun 22;4(2):e73 [FREE Full text] [doi: 10.2196/mhealth.5330] [Medline: 27335099]

34. Jung H, Chung K. Sequential pattern profiling based bio-detection for smart health service. Cluster Comput 2015;18(1):209-219. [doi: 10.1007/s10586-014-0370-3]
35. Huang J, Zhou M, Yang D. Extracting Chatbot Knowledge from Online Discussion Forums. 2007 Presented at: IJCAI; 2007; Hyderabad. [doi: 10.5555/1625275.1625342]
36. Maeda E, Miyata A, Boivin J, Nomura K, Kumazawa Y, Shirasawa H, et al. Promoting fertility awareness and preconception health using a chatbot: a randomized controlled trial. Reprod Biomed Online 2020 Dec;41(6):1133-1143. [doi: 10.1016/j.rbmo.2020.09.006] [Medline: 33039321]
37. Yadav D, Malik P, Dabas K, Singh P. Understanding opportunities for chatbots in breastfeeding education of women in india. PACM 2019 Nov 07;3(CSCW):1-30. [doi: 10.1145/3359272]
38. Donovan J. The process of analysis during a grounded theory study of men during their partners' pregnancies. J Adv Nurs 1995 Apr;21(4):708-715. [doi: 10.1046/j.1365-2648.1995.21040708.x] [Medline: 7797707]
39. Chalmers B, Meyer D. What men say about pregnancy, birth and parenthood. J Psychosom Obstet Gynaecol 1996 Mar;17(1):47-52. [doi: 10.3109/01674829609025663] [Medline: 8860886]
40. Draper J. Men's passage to fatherhood: an analysis of the contemporary relevance of transition theory. Nurs Inq 2003 Mar;10(1):66-77. [doi: 10.1046/j.1440-1800.2003.00157.x] [Medline: 12622806]
41. Carter FA, Carter JD, Luty SE, Wilson DA, Frampton CMA, Joyce PR. Screening and treatment for depression during pregnancy: a cautionary note. Aust N Z J Psychiatry 2005 Apr;39(4):255-261. [doi: 10.1080/j.1440-1614.2005.01562.x] [Medline: 15777362]
42. Milgrom J, Gemmill AW, Bilszta JL, Hayes B, Barnett B, Brooks J, et al. Antenatal risk factors for postnatal depression: a large prospective study. J Affect Disord 2008 May;108(1-2):147-157. [doi: 10.1016/j.jad.2007.10.014] [Medline: 18067974]

Abbreviations

AI: artificial intelligence
EOL: ease of learning
EOU: ease of use
MIM: mobile instant messenger
ob/gyn: obstetrician-gynecologist
Q&A: question-and-answer
R&D: research and development
SEE: intention to seek health information
SHA: intention to share health information
UI: user interface
USE: Usefulness, Satisfaction, and Ease of Use
UT: usability testing
UX: user experience

Edited by G Eysenbach; submitted 30.03.20; peer-reviewed by G Pavarini, E Bellei, M Nitsch, JT te Gussinklo; comments to author 21.04.20; revised version received 15.08.20; accepted 09.12.20; published 03.03.21.

Please cite as:
Chung K, Cho HY, Park JY
A Chatbot for Perinatal Women's and Partners' Obstetric and Mental Health Care: Development and Usability Evaluation Study
JMIR Med Inform 2021;9(3):e18607
URL: https://medinform.jmir.org/2021/3/e18607
doi:10.2196/18607
PMID:33656442

Original Paper

Natural Language Processing of Clinical Notes to Identify Mental Illness and Substance Use Among People Living with HIV: Retrospective Cohort Study

Jessica P Ridgway[1], MD, MSc; Arno Uvin[1], BSc; Jessica Schmitt[1], LCSW; Tomasz Oliwa[2], PhD; Ellen Almirol[1], MSc; Samantha Devlin[1], MSc; John Schneider[1], MD, MPH

[1]Department of Medicine, University of Chicago, Chicago, IL, United States

[2]Center for Research Informatics, University of Chicago, Chicago, IL, United States

Corresponding Author:
Jessica P Ridgway, MD, MSc
Department of Medicine
University of Chicago
5841 S Maryland Ave
MC 5065
Chicago, IL, 60637
United States
Phone: 1 7737029185
Email: jessica.ridgway@uchospitals.edu

Abstract

Background: Mental illness and substance use are prevalent among people living with HIV and often lead to poor health outcomes. Electronic medical record (EMR) data are increasingly being utilized for HIV-related clinical research and care, but mental illness and substance use are often underdocumented in structured EMR fields. Natural language processing (NLP) of unstructured text of clinical notes in the EMR may more accurately identify mental illness and substance use among people living with HIV than structured EMR fields alone.

Objective: The aim of this study was to utilize NLP of clinical notes to detect mental illness and substance use among people living with HIV and to determine how often these factors are documented in structured EMR fields.

Methods: We collected both structured EMR data (diagnosis codes, social history, Problem List) as well as the unstructured text of clinical HIV care notes for adults living with HIV. We developed NLP algorithms to identify words and phrases associated with mental illness and substance use in the clinical notes. The algorithms were validated based on chart review. We compared numbers of patients with documentation of mental illness or substance use identified by structured EMR fields with those identified by the NLP algorithms.

Results: The NLP algorithm for detecting mental illness had a positive predictive value (PPV) of 98% and a negative predictive value (NPV) of 98%. The NLP algorithm for detecting substance use had a PPV of 92% and an NPV of 98%. The NLP algorithm for mental illness identified 54.0% (420/778) of patients as having documentation of mental illness in the text of clinical notes. Among the patients with mental illness detected by NLP, 58.6% (246/420) had documentation of mental illness in at least one structured EMR field. Sixty-three patients had documentation of mental illness in structured EMR fields that was not detected by NLP of clinical notes. The NLP algorithm for substance use detected substance use in the text of clinical notes in 18.1% (141/778) of patients. Among patients with substance use detected by NLP, 73.8% (104/141) had documentation of substance use in at least one structured EMR field. Seventy-six patients had documentation of substance use in structured EMR fields that was not detected by NLP of clinical notes.

Conclusions: Among patients in an urban HIV care clinic, NLP of clinical notes identified high rates of mental illness and substance use that were often not documented in structured EMR fields. This finding has important implications for epidemiologic research and clinical care for people living with HIV.

(JMIR Med Inform 2021;9(3):e23456) doi:10.2196/23456

KEYWORDS

natural language processing; HIV; substance use; mental illness; electronic medical records

Introduction

Behavioral health disorders are highly prevalent among people living with HIV [1,2], who have a 2 to 4-fold higher risk of depression than the general population, with prevalence rates ranging from 24% to 63% [3-9]. A recent study among over 10,000 people living with HIV at seven HIV care sites across the United States found the prevalence of substance use disorder to be 48%, with 20% of patients having polysubstance use disorder [10]. This is higher than the rate of the general US population, in which the prevalence of substance use disorder is 7.7% [11].

In addition to being common among people living with HIV, mental illness and substance use often lead to poor health outcomes for this population. People living with HIV who have mental illness and substance use disorder have lower rates of engagement in HIV care and are less likely to adhere to antiretroviral therapy than those without behavioral health disorders [12-18]. Depression has been independently associated with mortality among several large cohorts of people living with HIV [12,19-21]. Besides poor individual health outcomes, people living with HIV with mental illness or substance use disorder are more likely to transmit HIV to others, because behavioral health disorders are associated with elevated HIV viral loads and behaviors that increase the risk of HIV transmission [22-24]. Many people living with HIV have co-occurring mental health disorders and substance use disorders, further exacerbating these adverse health outcomes [5,14].

To improve understanding of mental illness and substance use among people living with HIV, electronic medical record (EMR)-based behavioral health data are increasingly being utilized in HIV-related clinical research and medical care [25-27]. For example, Tolson et al [25] used an electronic reporting tool within the EMR to identify people living with HIV with substance use disorders to determine the association of substance use with hospitalization and virologic suppression. Other researchers used electronic billing codes to identify risk factors for suicidal ideation among people living with HIV [27]. However, mental illness and substance use are often underdocumented in structured EMR fields (eg, diagnosis codes, Problem List) [26,28,29], potentially leading to the exclusion of people living with HIV with behavioral health disorders from these studies if only discrete EMR data are used.

Natural language processing (NLP) of unstructured text of clinical notes in the EMR may identify behavioral health disorders beyond those identified using structured EMR fields alone [30,31]. Afshar et al [31] used NLP of clinical notes to identify patients with alcohol misuse, demonstrating greater accuracy than EMR-based billing codes; however, this study was performed among hospitalized trauma patients rather than with outpatients living with HIV. Oliwa et al [32] used NLP of clinical notes to identify phrases associated with improved engagement in HIV care. Their study identified NLP phrases related to substance use and mental health among people living with HIV, but they did not compare their findings with documentation in structured EMR fields.

To fill these gaps, the aim of this study was to utilize NLP of clinical notes to detect mental illness and substance use among people living with HIV, and to determine how often these factors were documented in structured EMR fields.

Methods

We performed a retrospective cohort study among people living with HIV at the University of Chicago Medicine (UCM) in Chicago, Illinois. Participants were included in the study if they were HIV-positive, 18 years of age or older, and attended at least one outpatient HIV care encounter at UCM between May 1, 2011 and May 30, 2016. This study was approved by the University of Chicago Institutional Review Board.

For eligible participants, we collected both structured EMR data as well as the unstructured text of clinical HIV care notes during the study time period. Structured EMR data collected included demographics, diagnosis codes (International Classification of Disease [ICD]-9 and ICD-10), Problem List (a list of physician-assigned diagnoses in the EMR), and social history. Unstructured data included the text of notes written by physicians, advanced practice providers, nurses, and social workers in the Department of Infectious Diseases. Data were extracted from the University of Chicago Clinical Research Data Warehouse, which stores data from the EMR (EPIC, Verona, WI) as well as administrative databases.

To develop the NLP algorithms for detecting mental illness and substance use, subject matter experts (physicians at the Department of Infectious Diseases and HIV care social workers) defined potential indicative words and crafted regular expressions to search for these key words and phrases related to mental illness and substance use (see Textbox 1). NegEx with augmented negation terms was applied to the key words and phrases found in clinical notes [33]. Those that were identified as negated occurrences by NegEx were excluded for the subsequent NLP steps. The Lucene Porter stemmer was used as a stemming algorithm to provide matching generalization between the tokens and the words/phrases from Textbox 1 [34]. Stanford CoreNLP with additional domain-specific split patterns was employed as a tokenizer and sentence splitter to provide the NegEx input sentences [35].

Textbox 1. Words and phrases detected by natural language processing algorithms.

- Words/phrases for mental illness

  Depression, Depressed, Bipolar, Anxiety, Panic, Psychiatry, Schizophrenia, Bipolar, Psychosis, Care2Prevent (mental health program), Anxious, Therapist (excluding physical therapist), Behavioral health, C2P, Psychotic

  Note: Stemmed forms, regular expression word boundaries, and a negative lookbehind in the case of "therapist" are excluded from this list for readability purposes.

- Words/phrases for substance abuse

  IVDU (intravenous drug user), Cocaine, Heroin, Crack, Alcohol abuse, AA (Alcoholics Anonymous) meeting, Haymarket (drug treatment program), NA (Narcotics Anonymous) meeting, Drug treatment program

Textboxes 2-4 list the diagnosis codes and Problem List phrases used to identify mental illness and substance use. Structured data from the Social History EMR section was considered to identify substance use if there was documentation of any illegal drug use (with the exception of marijuana) or if there was specific documentation of abuse of substances, including both legal and illegal substances.

To validate the NLP algorithm for mental illness, a random sample of 100 clinical notes flagged as positive for mental illness and 100 clinical notes not flagged for mental illness were manually reviewed to determine if the note documented that the patient had a mental illness. Two reviewers examined each note, and any discrepancies were resolved based on discussion and mutual agreement between reviewers. Using the determination from the manual chart review as the gold standard, we calculated the positive predictive value (PPV) of the algorithm (ie, the number of notes in which mental illness was present based on chart review divided by the number of reviewed notes that were flagged as positive for mental illness). We also calculated the negative predictive value (NPV) of the algorithm (ie, the number of notes in which mental illness was not present based on chart review divided by the number of reviewed notes not flagged by the mental illness algorithm). Similarly, to validate the NLP algorithm for substance use, a random sample of 100 clinical notes in which the algorithm detected substance use and 100 clinical notes where substance use was not detected were manually reviewed. Subsequently, the PPV and NPV for the substance use algorithm were also calculated.

We compared numbers of patients with mental illness or substance use identified by structured EMR fields with those identified by the NLP algorithms.

Textbox 2. International Classification of Diseases (ICD) diagnosis codes used to identify mental illness.

- ICD-9 codes

  291.9, 293.81, 293.82, 293.83, 293.84, 294.9, 295.3, 295.31, 295.32, 295.33, 295.34, 295.35, 295.42, 295.44, 295.6, 295.7, 295.71, 295.72, 295.75, 295.8, 295.9, 295.92, 296, 296.01, 296.02, 296.1, 296.15, 296.2, 296.21, 296.22, 296.23, 296.24, 296.25, 296.26, 296.3, 296.31, 296.32, 296.33, 296.34, 296.35, 296.36, 296.4, 296.41, 296.42, 296.44, 296.5, 296.51, 296.52, 296.53, 296.54, 296.55, 296.6, 296.64, 296.7, 296.8, 296.9, 297.1, 297.9, 298.9, 300, 300.01, 300.21, 300.3, 300.4, 300.81, 301.7, 301.82, 301.83, 301.9, 309, 309.24, 309.28, 309.3, 309.4, 309.81, 310.8, 311, 312.81, 312.82, 313.81, 314, 314.01, 648.41, 648.44, E950.0, E950.2, E950.3, E950.4, E950.9, E953.0, V11.0, V40.0, V40.9

- ICD-10 codes

  F0630, F09, F203, F2089, F209, F23, F250, F251, F259, F3110, F3111, F312, F3130, F3132, F3170, F3181, F319, F321, F323, F329, F330, F331, F332, F333, F3340, F3341, F3342, F339, F4001, F4010, F411, F419, F4323, F458, F509, F603, F900, F901, F913, F919, Z915

Textbox 3. International Classification of Diseases (ICD) diagnosis codes used to identify substance use.

- ICD-9 codes

  291, 291.2, 291.3, 291.81, 291.9, 304, 304.01, 304.02, 304.2, 304.22, 304.23, 304.3, 304.31, 304.7, 304.71, 304.72, 304.8, 304.83, 305, 305.01, 305.02, 305.03, 305.2, 305.21, 305.22, 305.23, 305.4, 305.5, 305.51, 305.52, 305.53, 305.6, 305.61, 305.62, 305.63, 305.7, 305.91, 305.93, 425.5, 535.3, 571, 571.2, 648.33, 965.01, 970.81, E850.0, E850.1, E850.2, E854.8, E860.0, E860.9, E935.0

- ICD-10 codes

  F1010, F10120, F10129, F10188, F1020, F1021, F10239, F1029, F1099, F1110, F1120, F1121, F1123, F1190, F11959, F1210, F1220, F1290, F12929, F1410, F14129, F1414, F14188, F1420, F1421, F14259, F1490, F14929, F1494, F1510, F1520, I426, K7031, K852, K860, O99313

Textbox 4. Problem List phrases.

- Mental illness

  ADD (attention deficit disorder); ADHD (attention deficit hyperactivity disorder); ADHD (attention deficit hyperactivity disorder), inattentive type; ADHD, predominantly inattentive type; Adjustment disorder with depressed mood; Adjustment disorder with mixed anxiety and depressed mood; Agoraphobia with panic disorder; Anxiety; Anxiety and depression; Anxiety disorder; Anxiety disorder in conditions classified elsewhere; Anxiety state, unspecified; Anxiety, mild; Attention deficit disorder without mention of hyperactivity; Bipolar 1 disorder; Bipolar 2 disorder; Bipolar affective; Bipolar affective disorder; Bipolar affective disorder, currently depressed, moderate; Bipolar depression; Bipolar disorder; Bipolar disorder, currently in remission of unspecified degree, most recent episode type unspecified; Bipolar disorder, unspecified; Bipolar I disorder, most recent episode depressed; Bipolar I disorder, most recent episode depressed, severe with psychosis; Bipolar I disorder, most recent episode manic; Bipolar I disorder, most recent episode manic, mild; Borderline personality disorder; Bulimia nervosa; Depressed mood; Depression; Depression (disease); Depression with anxiety; Depression, major; Depression, major, recurrent, severe with psychosis; Depression, recurrent; Depressive disorder; Depressive disorder, not elsewhere classified; Depressive episode; H/O attempted suicide; History of depression; Major depression; Major depression, recurrent; Major depression, recurrent, chronic; Major depressive disorder; Major depressive disorder, recurrent episode, in full remission; Major depressive disorder, recurrent episode, in partial or unspecified remission; Major depressive disorder, recurrent episode, mild; Major depressive disorder, recurrent episode, moderate; Major depressive disorder, recurrent episode, severe, without mention of psychotic behavior; Major depressive disorder, recurrent episode, unspecified; Major depressive disorder, severe; Major depressive disorder, single episode in full remission; Major depressive disorder, single episode, moderate; Manic depression; MDD (major depressive disorder), recurrent episode; MDD, recurrent episode, moderate; Mechanical complication of other vascular device, implant, and graft; Mood disorder; Mood disorder due to known physiological condition; Mood disorder in conditions classified elsewhere; Panic attack; Panic attacks; Panic disorder without agoraphobia; Paranoia (psychosis); Paranoid schizophrenia, chronic condition; Paranoid schizophrenia, chronic condition with acute exacerbation; Paranoid schizophrenia, unspecified condition; Postpartum depression; Posttraumatic stress; Posttraumatic stress disorder; Psychiatric illness; Psychiatric pseudoseizure; Psychosis; Psychosis, organic; Psychotic disorder with delusions in conditions classified elsewhere; PTSD (posttraumatic stress disorder); Schizoaffective disorder; Schizoaffective disorder, unspecified condition; Schizophrenia; Schizophrenia, disorganized, chronic with acute exacerbation; Schizophrenia, paranoid type; Schizophrenia, unspecified type; Suicidal ideation; Suicide attempt by drug ingestion; Suicide ideation; Unspecified schizophrenia, unspecified condition

- Substance use disorders

  Addiction, marijuana; Alcohol abuse; Alcohol abuse, continuous drinking behavior; Alcohol abuse, daily use; Alcohol abuse, episodic; Alcohol dependence; Alcohol dependence in remission; Alcohol dependence with acute alcoholic intoxication; Alcohol use; Alcohol withdrawal; Alcoholic cirrhosis of liver; Alcoholism with alcohol dependence; Cannabis use disorder, mild, abuse; Cocaine abuse; Cocaine abuse, in remission; Cocaine addiction; Cocaine dependence, continuous; Cocaine substance abuse; Cocaine use; Cocaine withdrawal; Dementia associated with alcoholism; ETOH abuse; Excessive blood level of alcohol; H/O alcohol abuse; H/O drug abuse; H/O substance abuse; Habitual alcohol use; History of alcohol abuse; History of alcohol use; History of cocaine abuse; History of cocaine use; History of drug abuse; History of heroin abuse; History of opioid abuse; Hx of cocaine abuse; IV (intravenous) drug abuse; IVDU (intravenous drug user); Marijuana abuse; Methadone dependence; Methadone use; Methamphetamine abuse; Opioid abuse, unspecified; Pancreatitis, alcoholic, acute; Polysubstance abuse; Psychoactive substance-induced organic mood disorder

Results

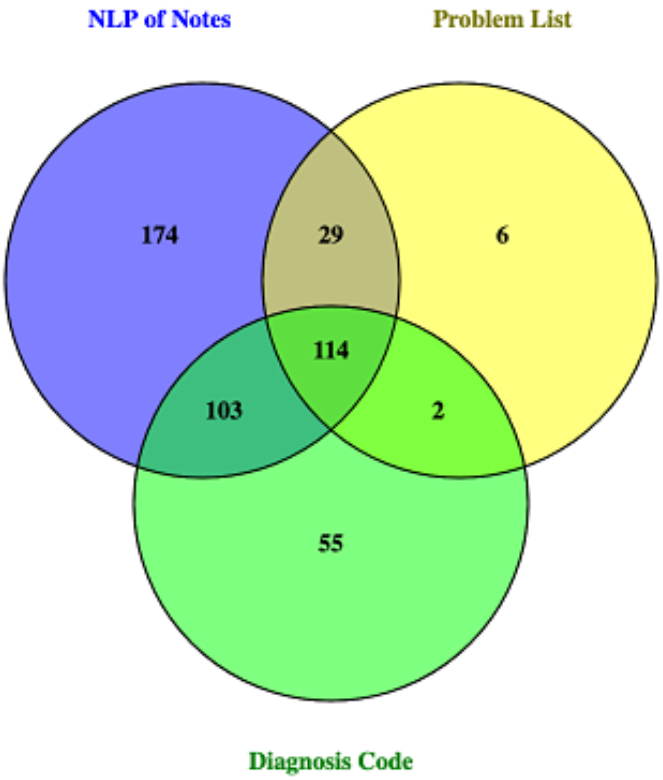
During the study period, 778 people living with HIV attended at least one HIV care appointment (Table 1). A total of 13,905 clinical notes were included, with a mean of 13 notes per patient (range 1-109). Based on manual review of clinical notes as described above, the NLP algorithm for detecting mental illness had a PPV of 98% and an NPV of 98%. The NLP algorithm for detecting substance use had a PPV of 92% and an NPV of 98%.

The NLP algorithm for mental illness identified 54.0% (420/778) of patients as having documentation of mental illness in the text of clinical notes (Figure 1). With the PPV of the algorithm of 98%, this would suggest that 412 patients truly had mental illness. Among the patients with mental illness detected by NLP, 58.6% (246/420) had documentation of mental illness in at least one structured EMR field (ie, Problem List or diagnosis code), including 34.0% (143/420) with a mental illness listed in the Problem List and 51.7% (217/420) with a diagnosis code related to mental illness. Sixty-three patients had documentation of mental illness in structured EMR fields that was not detected by NLP of clinical notes.

Table 1. Demographic characteristics of participants (N=778).

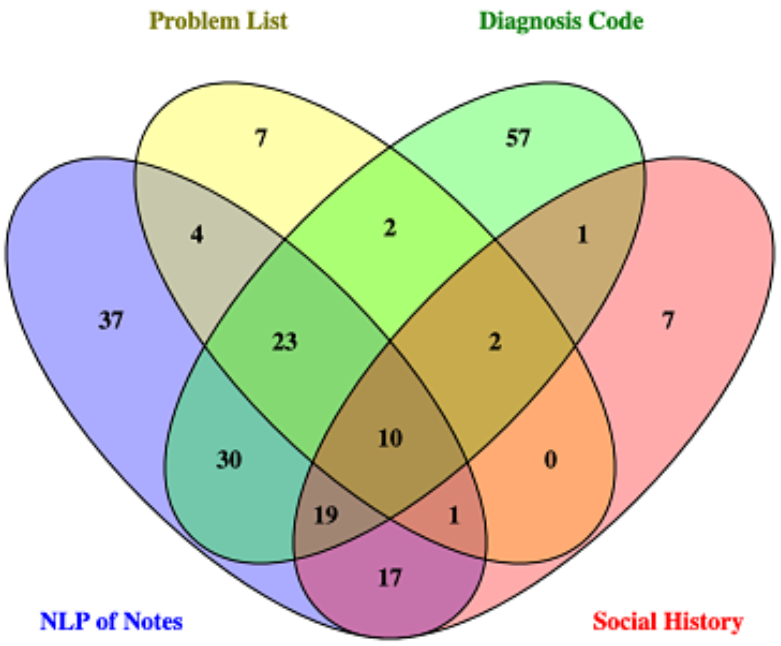
Characteristic	Value
Age (years), mean (SD)	43.1 (13.5)
Female, n (%)	287 (36.9)
Race/ethnicity, n (%)	
Black	620 (79.7)
White	107 (13.8)
Latinx	27 (3.5)
Asian	8 (1.0)
Other	16 (2.1)
Insurance, n (%)	
Medicaid	272 (35.0)
Medicare	228 (29.3)
Private	257 (33.0)
Other/self-pay	21 (2.7)

Figure 1. Electronic medical record documentation of mental Illness among people living with HIV. NLP: natural language processing.



The NLP algorithm for substance use detected substance use in the text of clinical notes in 18.1% (141/778) of participants (Figure 2). Based on the PPV of the algorithm of 92%, it is likely that 130 patients truly had substance use. Among patients with substance use detected by NLP, 73.8% (104/141) had documentation of substance use in at least one structured EMR field, including 27.0% (38/141) with documentation of substance use in the Problem List, 58.2% (82/141) with a diagnosis code related to substance use, and 33.3% (47/141) with substance use documented in the Social History section of the EMR. Seventy-six patients had documentation of substance use in structured EMR fields that was not detected by NLP of clinical notes.

Figure 2. Electronic medical record documentation of substance use among people living with HIV. NLP: natural language processing.



Discussion

Among patients in an urban HIV care clinic, NLP of clinical notes identified high rates of mental illness and substance use that were often not documented in structured EMR fields. This finding has important implications for clinical care and epidemiologic research among people living with HIV. Namely, relying on structured EMR fields alone to identify people living with HIV with behavioral health disorders may miss a substantial number of patients. Given the high PPV of our algorithms, addition of such NLP algorithms to current tools for identifying behavioral health disorders could augment detection of these disorders among people living with HIV.

To our knowledge, this is the first study to utilize NLP of EMR notes to detect mental illness and substance use among people living with HIV. Other studies have used NLP to detect depression and substance misuse in non-HIV care settings [30,31,36]. Adekkanattu et al [36] used NLP to identify depression from EMR notes among patients prescribed antidepressants, and found that 31% of patients with depression detected by NLP were missing a diagnosis code for depression. Zhou et al [37] similarly used NLP of hospital discharge summaries to identify depression among hospitalized patients, and found that 20% of patients with depression detected by NLP did not have a depression diagnosis code. These rates of discordant documentation are lower than that obtained in this study, in which nearly half of patients with mental illness detected by NLP did not have a diagnosis code for mental illness. This discrepancy may be explained by differences in the patient populations studied. Our patients are from a general HIV clinic, rather than inpatients or outpatients already prescribed antidepressants, populations in which medical providers may be more likely to enter a diagnosis code for mental illness.

Our NLP algorithm for mental illness identified 54% of people living with HIV in our clinical population as having mental illness. This is similar to other studies among people living with HIV, which have shown prevalence rates as high as 63% based on validated depression screening tools (eg, Patient Health Questionnaire-9) [3-9].The NLP algorithm detected substance use in 18% of our clinical population. This rate is within the lower end of what has previously been reported. Hartzler et al [10] found that the prevalence of substance use disorders among people living with HIV at 7 HIV care sites ranged from 21% to 71% based on substance use disorder screening tools. Of note, for both mental illness and substance use, the NLP algorithms failed to flag a substantial number of patients who had mental illness or substance use documented in structured EMR fields, suggesting that NLP algorithms should be used in combination with structured fields rather than as a replacement for structured fields for detecting these characteristics.

As EMR data are increasingly being used for clinical care and research among people living with HIV, extracting accurate behavioral health data from the EMR is essential. EMRs have been used to provide electronic feedback to providers to alert them that patients may have untreated depression [38,39]. Results from NLP of clinical notes could potentially augment such electronic alerts. Recent studies have used structured EMR fields, including documentation of substance use and mental illness, to create predictive models of HIV appointment

adherence [12,40]. However, if mental illness and substance use are not adequately documented in structured EMR fields, inclusion of NLP of clinician notes may improve such predictive models by identifying additional risk factors for appointment nonadherence.

Our study has several limitations. We did not review all clinical notes for the presence or absence of behavioral health disorder documentation, and some of the NLP-detected cases may be false positives. Although we adjusted for negation in the text, we may have falsely detected mental illness in some instances where providers wrote in a nonstandard format that patients did *not* have mental illness or where they documented that a family member and not the patient themselves had a behavioral health disorder. In addition, certain phrases (eg, Alcoholics Anonymous meeting) may have detected patients with past substance use disorder rather than active substance use disorder. However, in the review of a random sample of 400 notes, we found a high PPV for the NLP algorithms. The NLP algorithms may have also failed to flag notes that documented behavioral health disorders (ie, false negatives). Moreover, the NLP algorithms do not necessarily detect patients with mental illness or substance use, but only detect documentation in the clinical notes of mental illness or substance use. If providers did not ask patients about these topics or did not document regarding their conversations, then people living with HIV with behavioral health disorders may have been missed by our algorithms. Inclusion of validated behavioral health screening tools within the EMR would likely improve detection of mental illness and substance use. These screening tools were not routinely in place in our clinic at the time of the study, and therefore we were unable to assess how they would have affected the results.

In conclusion, we performed the first study of NLP of unstructured clinical notes for mental illness and substance use among people living with HIV. Although these behavioral health disorders were commonly detected by NLP, they were often undocumented in structured fields of the EMR. More research is needed to understand how to best utilize both structured and unstructured EMR data for clinical and epidemiologic research among people living with HIV.

Acknowledgments

This work was supported by National Institutes of Health (NIH; 1K23MH121190-01) and the NIH-funded Third Coast Center for AIDS Research (CFAR; P30 AI117943). Data from this study were provided by the Clinical Research Data Warehouse maintained by the Center for Research Informatics at University of Chicago. The Center for Research Informatics is funded by the Biological Sciences Division, Institute for Translational Medicine/CTSA (NIH UL1 TR000430) at the University of Chicago. The funders had no role in review or approval of the manuscript for publication.

Authors' Contributions

JR and JAS conceived of and designed the study. TO, EA, AU, and SD collected and analyzed the data. JR, JAS, and JS interpreted the data. JR drafted the manuscript, and JAS, TO, EA, AU, SD, and JS critically revised the manuscript. JR obtained funding and supervised the study.

Conflicts of Interest

None declared.

References

1. Remien R, Stirratt M, Nguyen N, Robbins R, Pala A, Mellins C. Mental health and HIV/AIDS: the need for an integrated response. AIDS 2019 Jul 15;33(9):1411-1420 [FREE Full text] [doi: 10.1097/QAD.0000000000002227] [Medline: 30950883]
2. Rezaei S, Ahmadi S, Rahmati J, Hosseinifard H, Dehnad A, Aryankhesal A, et al. Global prevalence of depression in HIV/AIDS: a systematic review and meta-analysis. BMJ Support Palliat Care 2019 Dec 19;9(4):404-412. [doi: 10.1136/bmjspcare-2019-001952] [Medline: 31537580]
3. Bing EG, Burnam MA, Longshore D, Fleishman JA, Sherbourne CD, London AS, et al. Psychiatric disorders and drug use among human immunodeficiency virus-infected adults in the United States. Arch Gen Psychiatry 2001 Aug 01;58(8):721-728. [doi: 10.1001/archpsyc.58.8.721] [Medline: 11483137]
4. Schumacher JE, McCullumsmith C, Mugavero MJ, Ingle-Pang PE, Raper JL, Willig JH, et al. Routine depression screening in an HIV clinic cohort identifies patients with complex psychiatric co-morbidities who show significant response to treatment. AIDS Behav 2013 Oct 20;17(8):2781-2791 [FREE Full text] [doi: 10.1007/s10461-012-0342-7] [Medline: 23086427]
5. Tegger MK, Crane HM, Tapia KA, Uldall KK, Holte SE, Kitahata MM. The effect of mental illness, substance use, and treatment for depression on the initiation of highly active antiretroviral therapy among HIV-infected individuals. AIDS Patient Care STDs 2008 Mar;22(3):233-243. [doi: 10.1089/apc.2007.0092] [Medline: 18290749]
6. Ciesla JA, Roberts JE. Meta-analysis of the relationship between HIV infection and risk for depressive disorders. Am J Psychiatry 2001 May;158(5):725-730. [doi: 10.1176/appi.ajp.158.5.725] [Medline: 11329393]

7. Chenneville T, Gabbidon K, Drake H, Rodriguez C. Comparison of the utility of the PHQ and CES-D for depression screening among youth with HIV in an integrated care setting. J Affect Disord 2019 May 01;250:140-144. [doi: 10.1016/j.jad.2019.03.023] [Medline: 30852366]
8. Shacham E, Nurutdinova D, Satyanarayana V, Stamm K, Overton E. Routine screening for depression: identifying a challenge for successful HIV care. AIDS Patient Care STDs 2009 Nov;23(11):949-955 [FREE Full text] [doi: 10.1089/apc.2009.0064] [Medline: 19925308]
9. Brandt C, Zvolensky MJ, Woods SP, Gonzalez A, Safren SA, O'Cleirigh CM. Anxiety symptoms and disorders among adults living with HIV and AIDS: A critical review and integrative synthesis of the empirical literature. Clin Psychol Rev 2017 Feb;51:164-184 [FREE Full text] [doi: 10.1016/j.cpr.2016.11.005] [Medline: 27939443]
10. Hartzler B, Dombrowski JC, Crane HM, Eron JJ, Geng EH, Christopher Mathews W, et al. Prevalence and predictors of substance use disorders among HIV care enrollees in the United States. AIDS Behav 2017 Apr 13;21(4):1138-1148 [FREE Full text] [doi: 10.1007/s10461-016-1584-6] [Medline: 27738780]
11. McCance-Katz EF. The National Survey on Druge Use and Health: 2019. Substance Abuse and Mental Health Services Administration. 2020 Sep. URL: https://www.samhsa.gov/data/sites/default/files/reports/rpt29392/Assistant-Secretary-nsduh2019_presentation/Assistant-Secretary-nsduh2019_presentation.pdf [accessed 2021-01-24]
12. Pence BW, Mills JC, Bengtson AM, Gaynes BN, Breger TL, Cook RL, et al. Association of increased chronicity of depression with HIV appointment attendance, treatment failure, and mortality among HIV-infected adults in the United States. JAMA Psychiatry 2018 Apr 01;75(4):379-385 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.4726] [Medline: 29466531]
13. Gonzalez J, Batchelder A, Psaros C, Safren S. Depression and HIV/AIDS treatment nonadherence: a review and meta-analysis. J Acquir Immune Defic Syndr 2011 Oct 01;58(2):181-187 [FREE Full text] [doi: 10.1097/QAI.0b013e31822d490a] [Medline: 21857529]
14. Chander G, Himelhoch S, Moore RD. Substance abuse and psychiatric disorders in HIV-positive patients: epidemiology and impact on antiretroviral therapy. Drugs 2006;66(6):769-789. [doi: 10.2165/00003495-200666060-00004] [Medline: 16706551]
15. Nanni MG, Caruso R, Mitchell AJ, Meggiolaro E, Grassi L. Depression in HIV infected patients: a review. Curr Psychiatry Rep 2015 Jan 21;17(1):530. [doi: 10.1007/s11920-014-0530-4] [Medline: 25413636]
16. Tucker JS, Burnam M, Sherbourne CD, Kung F, Gifford AL. Substance use and mental health correlates of nonadherence to antiretroviral medications in a sample of patients with human immunodeficiency virus infection. Am J Med 2003 May;114(7):573-580. [doi: 10.1016/s0002-9343(03)00093-7] [Medline: 12753881]
17. Azar MM, Springer SA, Meyer JP, Altice FL. A systematic review of the impact of alcohol use disorders on HIV treatment outcomes, adherence to antiretroviral therapy and health care utilization. Drug Alcohol Depend 2010 Dec 01;112(3):178-193 [FREE Full text] [doi: 10.1016/j.drugalcdep.2010.06.014] [Medline: 20705402]
18. Zuniga JA, Yoo-Jeong M, Dai T, Guo Y, Waldrop-Valverde D. The role of depression in retention in care for persons living with HIV. AIDS Patient Care STDs 2016 Jan;30(1):34-38 [FREE Full text] [doi: 10.1089/apc.2015.0214] [Medline: 26544915]
19. Ickovics JR, Hamburger ME, Vlahov D, Schoenbaum EE, Schuman P, Boland RJ, HIV Epidemiology Research Study Group. Mortality, CD4 cell count decline, and depressive symptoms among HIV-seropositive women: longitudinal analysis from the HIV Epidemiology Research Study. JAMA 2001 Mar 21;285(11):1466-1474. [doi: 10.1001/jama.285.11.1466] [Medline: 11255423]
20. Cook JA, Grey D, Burke J, Cohen MH, Gurtman AC, Richardson JL, et al. Depressive symptoms and AIDS-related mortality among a multisite cohort of HIV-positive women. Am J Public Health 2004 Jul;94(7):1133-1140. [doi: 10.2105/ajph.94.7.1133] [Medline: 15226133]
21. Todd J, Cole S, Pence B, Lesko CR, Bacchetti P, Cohen MH, et al. Effects of antiretroviral therapy and depressive symptoms on all-cause mortality among HIV-infected women. Am J Epidemiol 2017 May 15;185(10):869-878 [FREE Full text] [doi: 10.1093/aje/kww192] [Medline: 28430844]
22. Sikkema KJ, Watt MH, Drabkin AS, Meade CS, Hansen NB, Pence BW. Mental health treatment to reduce HIV transmission risk behavior: a positive prevention model. AIDS Behav 2010 Apr 15;14(2):252-262 [FREE Full text] [doi: 10.1007/s10461-009-9650-y] [Medline: 20013043]
23. Hutton HE, Lyketsos CG, Zenilman JM, Thompson RE, Erbelding EJ. Depression and HIV risk behaviors among patients in a sexually transmitted disease clinic. Am J Psychiatry 2004 May;161(5):912-914. [doi: 10.1176/appi.ajp.161.5.912] [Medline: 15121659]
24. Liang J, Nosova E, Reddon H, Nolan S, Socías E, Barrios R, et al. Longitudinal patterns of illicit drug use, antiretroviral therapy exposure and plasma HIV-1 RNA viral load among HIV-positive people who use illicit drugs. AIDS 2020 Jul 15;34(9):1389-1396. [doi: 10.1097/QAD.0000000000002551] [Medline: 32590435]
25. Tolson C, Richey LE, Zhao Y, Korte JE, Brady K, Haynes L, et al. Association of substance use with hospitalization and virologic suppression in a southern academic HIV clinic. Am J Med Sci 2018 Jun;355(6):553-558 [FREE Full text] [doi: 10.1016/j.amjms.2018.03.002] [Medline: 29891038]

26. O'Cleirigh C, Magidson JF, Skeer MR, Mayer KH, Safren SA. Prevalence of psychiatric and substance abuse symptomatology among HIV-infected gay and bisexual men in HIV primary care. Psychosomatics 2015 Sep;56(5):470-478 [FREE Full text] [doi: 10.1016/j.psym.2014.08.004] [Medline: 25656425]
27. Brown LA, Majeed I, Mu W, McCann J, Durborow S, Chen S, et al. Suicide risk among persons living with HIV. AIDS Care 2020 Aug 03:online ahead of print. [doi: 10.1080/09540121.2020.1801982] [Medline: 32741212]
28. Machado IK, Luz PM, Lake JE, Castro R, Velasque L, Clark JL, et al. Substance use among HIV-infected patients in Rio de Janeiro, Brazil: Agreement between medical records and the ASSIST questionnaire. Drug Alcohol Depend 2017 Sep 01;178:115-118 [FREE Full text] [doi: 10.1016/j.drugalcdep.2017.04.033] [Medline: 28646713]
29. Brown LA, Mu W, McCann J, Durborow S, Blank MB. Under-documentation of psychiatric diagnoses among persons living with HIV in electronic medical records. AIDS Care 2021 Mar 13;33(3):311-315. [doi: 10.1080/09540121.2020.1713974] [Medline: 31931621]
30. Topaz M, Murga L, Bar-Bachar O, Cato K, Collins S. Extracting alcohol and substance abuse status from clinical notes: the added value of nursing data. Stud Health Technol Inform 2019 Aug 21;264:1056-1060. [doi: 10.3233/SHTI190386] [Medline: 31438086]
31. Afshar M, Phillips A, Karnik N, Mueller J, To D, Gonzalez R, et al. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. J Am Med Inform Assoc 2019 Mar 01;26(3):254-261 [FREE Full text] [doi: 10.1093/jamia/ocy166] [Medline: 30602031]
32. Oliwa T, Furner B, Schmitt J, Schneider J, Ridgway J. Development of a predictive model for retention in HIV care using natural language processing of clinical notes. J Am Med Inform Assoc 2021 Jan 15;28(1):104-112. [doi: 10.1093/jamia/ocaa220] [Medline: 33150369]
33. Chapman W, Hilert D, Velupillai S. Extending the NegEx Lexicon for Multiple Languages. 2013 Presented at: 14th World Congress on Medical & Health Informatics; 2013; Copenhagen.
34. Apache Lucene. URL: https://lucene.apache.org/ [accessed 2021-01-26]
35. Manning C, Sureadnu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. 2014 Presented at: 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations; 2014; Baltimore p. 55-60. [doi: 10.3115/v1/p14-5010]
36. Adekkanattu P, Sholle ET, DeFerio J, Pathak J, Johnson SB, Campion TR. Ascertaining Depression Severity by Extracting Patient Health Questionnaire-9 (PHQ-9) Scores from Clinical Notes. 2018 Nov Presented at: AMIA Annual Symposium; 2018; San Francisco p. 147-156.
37. Zhou L, Baughman AW, Lei VJ, Lai KH, Navathe AS, Chang F, et al. Identifying patients with depression using free-text clinical documents. Stud Health Technol Inform 2015;216:629-633. [Medline: 26262127]
38. Rollman BL, Hanusa BH, Gilbert T, Lowe HJ, Kapoor WN, Schulberg HC. The electronic medical record. A randomized trial of its impact on primary care physicians' initial management of major depression [corrected]. Arch Intern Med 2001 Jan 22;161(2):189-197. [doi: 10.1001/archinte.161.2.189] [Medline: 11176732]
39. Frame A, LaMantia M, Reddy Bynagari B, Dexter P, Boustani M. Development and implementation of an electronic decision support to manage the health of a high-risk Population: the enhanced Electronic Medical Record Aging Brain Care software (eMR-ABC). EGEMS (Wash DC) 2013 Mar 11;1(1):1009 [FREE Full text] [doi: 10.13063/2327-9214.1009] [Medline: 25848560]
40. Ramachandran A, Kumar A, Koenig H, De Unanue A, Sung C, Walsh J, et al. Predictive analytics for retention in care in an urban HIV clinic. Sci Rep 2020 Apr 14;10(1):6421. [doi: 10.1038/s41598-020-62729-x] [Medline: 32286333]

Abbreviations

EMR: electronic medical record
ICD: International Classification of Diseases
NLP: natural language processing
NPV: negative predictive value
PPV: positive predictive value
UCM: University of Chicago Medicine

Edited by G Eysenbach; submitted 12.08.20; peer-reviewed by J Jain, J Walsh, M Torii; comments to author 07.12.20; revised version received 30.01.21; accepted 31.01.21; published 10.03.21.

Please cite as:
Ridgway JP, Uvin A, Schmitt J, Oliwa T, Almirol E, Devlin S, Schneider J
Natural Language Processing of Clinical Notes to Identify Mental Illness and Substance Use Among People Living with HIV: Retrospective Cohort Study
JMIR Med Inform 2021;9(3):e23456
URL: https://medinform.jmir.org/2021/3/e23456
doi:10.2196/23456
PMID:33688848

Original Paper

Hybrid Deep Learning for Medication-Related Information Extraction From Clinical Texts in French: MedExt Algorithm Development Study

Jordan Jouffroy[1,2], MD, MSc; Sarah F Feldman[1,2], MD; Ivan Lerner[1,2], MD; Bastien Rance[2,3], PhD; Anita Burgun[1,2], MD, PhD; Antoine Neuraz[1,2], MD, PhD

[1]Department of Biomedical Informatics, Necker-Enfants malades Hospital, Assistance Publique–Hôpitaux de Paris, Paris, France

[2]UMRS 1138 team 22, Institut National de la Santé et de la Recherche Médicale, Université de Paris, Paris, France

[3]Department of Biomedical Informatics, Georges Pompidou European Hospital, Assistance Publique–Hôpitaux de Paris, Paris, France

Corresponding Author:
Antoine Neuraz, MD, PhD
Department of Biomedical Informatics
Necker-Enfants malades Hospital
Assistance Publique–Hôpitaux de Paris
Bâtiment Imagine - Bureau 145
149 rue de Sèvres
Paris,
France
Phone: 33 171396585
Email: antoine.neuraz@aphp.fr

Abstract

Background: Information related to patient medication is crucial for health care; however, up to 80% of the information resides solely in unstructured text. Manual extraction is difficult and time-consuming, and there is not a lot of research on natural language processing extracting medical information from unstructured text from French corpora.

Objective: We aimed to develop a system to extract medication-related information from clinical text written in French.

Methods: We developed a hybrid system combining an expert rule–based system, contextual word embedding (embedding for language model) trained on clinical notes, and a deep recurrent neural network (bidirectional long short term memory–conditional random field). The task consisted of extracting drug mentions and their related information (eg, dosage, frequency, duration, route, condition). We manually annotated 320 clinical notes from a French clinical data warehouse to train and evaluate the model. We compared the performance of our approach to those of standard approaches: rule-based or machine learning only and classic word embeddings. We evaluated the models using token-level recall, precision, and F-measure.

Results: The overall F-measure was 89.9% (precision 90.8; recall: 89.2) when combining expert rules and contextualized embeddings, compared to 88.1% (precision 89.5; recall 87.2) without expert rules or contextualized embeddings. The F-measures for each category were 95.3% for medication name, 64.4% for drug class mentions, 95.3% for dosage, 92.2% for frequency, 78.8% for duration, and 62.2% for condition of the intake.

Conclusions: Associating expert rules, deep contextualized embedding, and deep neural networks improved medication information extraction. Our results revealed a synergy when associating expert knowledge and latent knowledge.

(JMIR Med Inform 2021;9(3):e17934) doi:10.2196/17934

KEYWORDS

medication information; natural language processing; electronic health records; deep learning; rule-based system, recurrent neural network; hybrid system

Introduction

In 2017, medication consumption in France represented €37.8 billion (approximately US $45.5 billion) in spending and 16% of the French health budget [1]. Adverse drug reactions are an important public health problem, representing a major cause of mortality (0.15% in France); one-third of admissions caused by adverse drug reactions are preventable, associated with a poorly reported drug history or rare adverse events [2,3].

Furthermore, electronic health records contain rich information about drug history that would be valuable to the care of patients (eg, to prevent interaction with another medication and to track side effects), for epidemiology, or pharmaco-vigilance [4]. A major hurdle in the use of electronic health records is the format of the data. Up to 80% of relevant clinical information is present solely in the form of unstructured text, which represents a major barrier to the secondary use of this type of information [5,6].

To overcome this issue, natural language processing techniques can be used to extract, normalize, and restructure drug-related information from clinical texts [6,7] and increase the information available for research and health care. Three approaches have been described for this task: expert knowledge modeling, machine learning, and hybrid methods (combining both).

The first approach relies on modeling expert knowledge using dictionaries or rules (ie, expert rules) such as MedEx, MedXN, or MedLEE based on lexicons or regular expressions [8-12]. Dictionary-based approaches allow for direct or approximate matching of terms from a dictionary or terminology. These approaches may offer poor results when the mentions used in texts deviate from the terms in the dictionary. Rule-based approaches allow for specific extractions but usually lack sensitivity and do not perform well on new data sets. Rule-based approaches also require domain experts to design and build the rules and are particularly time-consuming. In addition, expertise is rare and costly, which constitutes a severe bottleneck for the use of this type of method.

The second approach, using machine learning, has been developed in addition to expert approaches to extract medication name, dosage, frequency, duration, mode, reason for the intake and to detect adverse drug reactions [13,14]. Most systems included a conditional random field or a support vector machine for medication-related information extraction [15-18], 2 studies introduced bidirectional long short-term memory associated with conditional random field for named entity recognition and medication information extraction [19,20], and another used a semisupervised model [21].

For the 2018 N2C2 shared task on medication extraction in electronic health records [22], several systems were proposed. The data set used in the challenge consisted of 505 discharge summaries extracted from the MIMIC-III database [23]. This data set contained 16,225 drug mentions in the training set and a total of 50,951 entity annotations again in the training set. Among the best-performing algorithms, bidirectional long short term memory and bidirectional long short term memory with conditional random field architectures were popular [24-27]. Some systems combined attention mechanisms [28] or convolutional neural networks [27]. Others combined classic entity extraction systems such as cTakes with classifiers such as support vector machines [29]. Ensemble approaches, combining multiple classifiers were also proposed [24-26,30].

At the conjunction of machine learning and expert rules, hybrid approaches can leverage the frugality of expert rules (in terms of data needs) and the flexibility and generalizability of machine learning. Examples include identifying medication heading using a conditional random field for named entity identification and a support vector machine to classify relations combined with a rule-based context engine [31]; a conditional random field and 2 bidirectional long short term memory–conditional random field models to extract handcrafted features [25]; and using expert rules and a knowledge base to enrich text, then using a bidirectional long short term memory with attention to perform the medication extraction in electronic health records [28]. These approaches were designed for text written in English. To the best of our knowledge, there are only a few studies [32,33] on French corpora: Deleger et al [32] used a rule-based system, and Lerner et al [33] developed a hybrid system that associated expert rules using terminology and bidirectional gated recurrent units with a conditional random field.

In recent years, the adoption of word embedding methods has led to a significant increase in the level of performance achievable by many natural language processing tasks [34]. Word embeddings use dense vector representation of the vocabulary. Interestingly, word embeddings are computed using large amounts of unannotated data (eg, Wikipedia). In static word embeddings, a token is represented by a static numeric vector. Recently, contextual word embedding methods have appeared, such as embedding for language model [35]. Contextual word embeddings provide a varying representation of the tokens with regard to the context in the text. Contextual word embeddings lead to richer representations and help to improve the performance in clinical concept extraction tasks [36]; results further improve when semantic information is incorporated [37].

In this work, we aimed to extract medication-related information from clinical narratives written in French in a real-world setting (ie, with documents directly extracted from a clinical data warehouse). Once extracted, such information can be restructured to be used for different purposes (eg, clinical epidemiology, monitoring, pharmaco-epidemiology, adverse drug reaction detection). Our purpose was two-fold: (1) We aimed to develop a gold standard data set of annotated clinical documents in French, along with an annotation guide, and (2) we aimed to develop a hybrid approach combining an association of knowledge base and expert rules, contextualized word embeddings training on clinical text, and a deep learning model based on bidirectional long short term memory–conditional random field.

Methods

Data

Source

We leveraged the clinical data warehouse of the *Assistance Publique–Hôpitaux de Paris* (AP-HP), grouping data collected from 39 hospitals to build a data set of 1 million documents [38]. These clinical reports were medical prescriptions, discharge reports, examinations, observation reports, and emergency visits randomly selected from the clinical data warehouse.

Annotated Data Set

We created an annotated data set for training and evaluation. We iteratively developed an annotation guide during the first phase of annotation. A small portion of the extracted data set (320 documents) was manually annotated by 3 medical doctors using an annotation tool [39]. The annotations were converted to the inside, outside, beginning (IOB) standard. Tokens that refer to an entity were labeled *B-entity_type* for the first token and then *I-entity_type*, tokens outside entities mention are labeled O. We split the 320 annotated clinical notes in a training set (n=216), a development set (n=24), and a test set (n=80).

Knowledge Base for Drug Names

We relied upon 2 French databases—*Base de données publique des medicaments* (a publicly accessible, *National drug database*)[40] and OpenMedic, a database from the national medical insurance agency [41]. These 2 databases contain all the drugs distributed in France during a given year. They were mapped to the Anatomical Therapeutic Chemical classification system. We extracted data from 2015 to 2019 and created a curated and unified dictionary of drug mentions.

The corpus can be made available on the condition that a research project is accepted by the scientific and ethics committee of the AP-HP health data warehouse.

After preprocessing, the text was preannotated using a set of expert handcrafted rules, then the texts were embedded using contextual word embeddings trained on a large corpus of clinical texts. The preannotations and the embedded texts were input into a bidirectional long short term memory–conditional random field to produce the final annotations (Figure 1; Figure 2).

Figure 1. General architecture of the model. BiLSTM: bidirectional long short term memory; CRF: conditional random field.

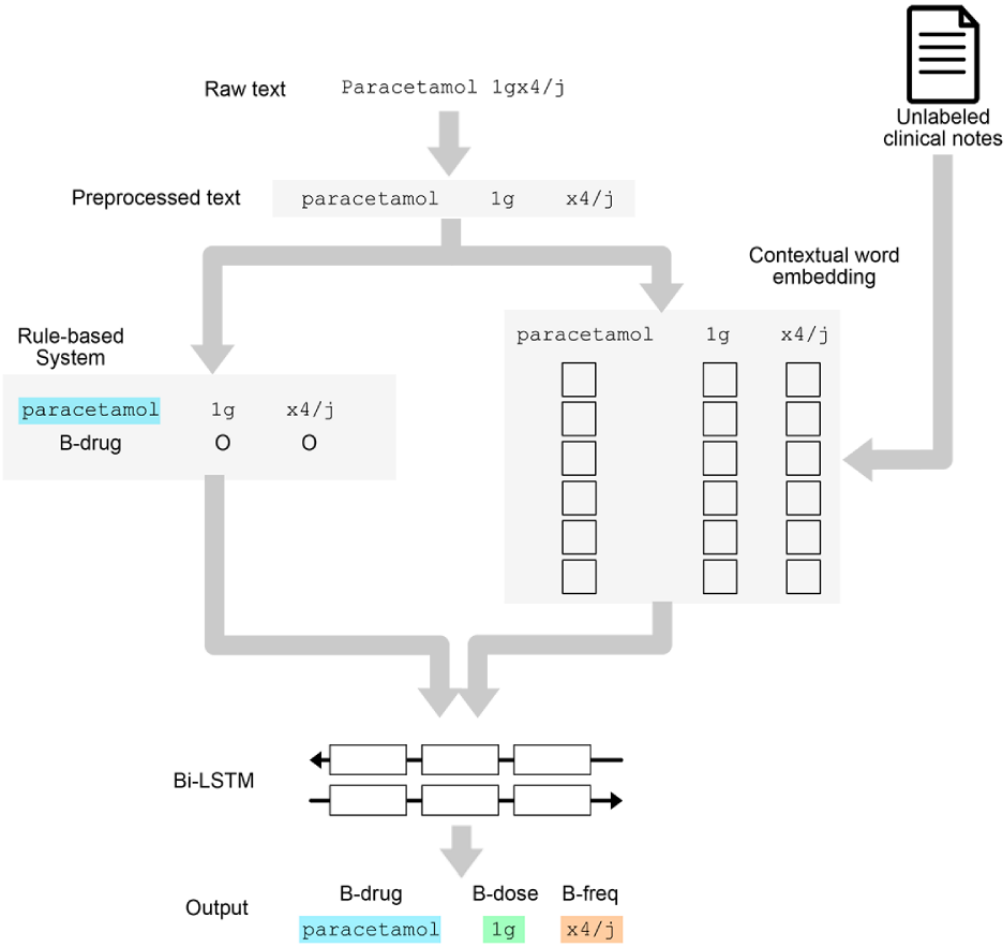
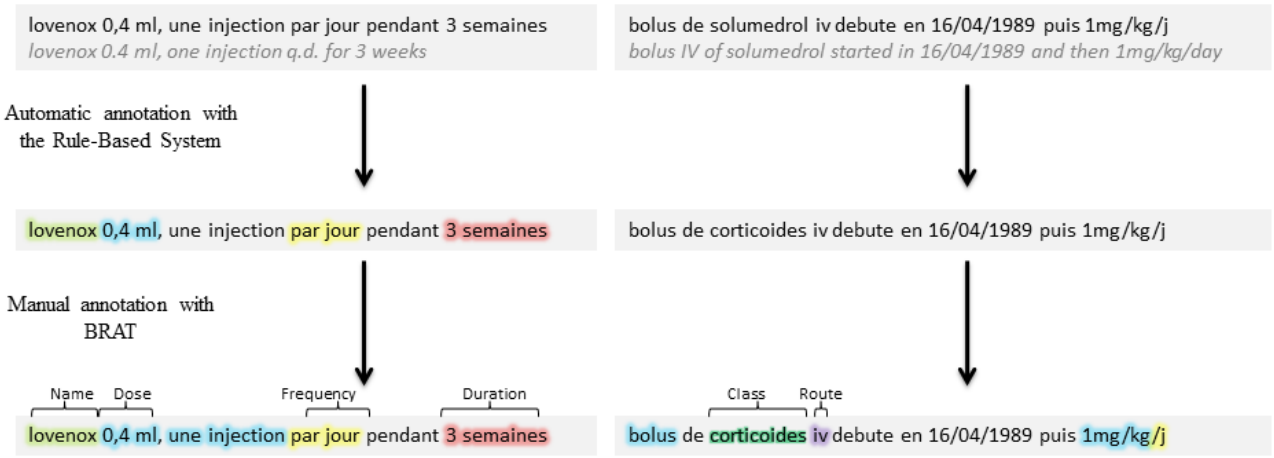


Figure 2. Annotation process with automatic annotation and completion with manual annotation.



Task Definition

We aimed to identify medication-related information in clinical documents in French. We were interested in drug names and a set of attributes related to the drug mentions: dosage, frequency, duration, route, and condition of administration. A detailed description of the types of entities is provided in Table 1.

Table 1. Description of the task.

Type	Description	Examples
Medication name	Descriptions that denote any medication, active molecule, association or protocol	doliprane, paracetamol, augmentin
Medication class	Descriptions that denote any Anatomical Therapeutic Chemical class or common therapy	ß-Lactam, antibiotherapy
Dosage	Dose or concentration of medication in prescription	3 mg, 2 tablets
Frequency	Frequency of medication administration	3 per day, every morning
Duration	Time range for the administration	3 weeks, until the surgery
Route	Medication administration mode	intravenous, per os
Condition	The event which provokes the administration	if pain, if infection

Preprocessing

We preprocessed the input texts as described in Textbox 1.

Textbox 1. Text preprocessing.

Steps

- Removing acronym points and replacing decimal points by comma
- Removing break lines added during documents conversion to text
- Removing accents
- Replacing apostrophes by spaces
- Detecting sentence boundaries: remaining points or break lines without transitive verbs, preposition or coordinating conjunctions.
- Detecting word boundaries and tokenization: sequence of alphanumeric characters or a repetition of a unique nonalphanumeric characters

Rule-Based Module

The overall approach was organized as follows: we first identified a drug mention or a drug-class mention with the knowledge-based dictionary using exact matching. The choice of exact matching for this step was driven by maximizing the precision of the annotations in this preannotation step. Then, using the identified mention as an anchor, we extended the search to the attributes of this mention (ie, frequency, dosage, duration, mode of administration, and condition of administration) in the area surrounding the seed mention. The attributes were detected using a set of handcrafted rules using

regular expressions. Examples of the rules are described in Table S1 of Multimedia Appendix 1. At this stage, the annotated entities were identified by their position and length relative to the beginning of the document. For the next steps, the annotations were converted to the IOB standard. The output of the rule-based system was used for preannotating the documents before the manual annotation step to speed up the annotation process of the gold standard data set and to serve as extra features to the input of the deep-learning module.

Deep Learning Module

Overview

We designed an approach leveraging deep neural networks. We tested 3 types of word embeddings—skip-gram [42], FastText embeddings [43], and embedding for language model [35]—and 2 neural network architectures—bidirectional long short term memory and bidirectional long short term memory–conditional random field.

Embeddings

We evaluated the impact of the word embeddings on the performance of the model. Our baseline was created using a skip-gram embedding trained on the training set only. We also considered FastText embedding (skip-gram model augmented with sub–word information) trained on a corpus of 1 million documents. Finally, we used embedding for language model embeddings, trained on 100,000 clinical notes that were contextualized embeddings computed through the internal states of a large bidirectional language model. The embeddings were kept fixed during model training.

Combination of the Rule-Based System Output

The output of the rule-based system was converted to the IOB standard. Then, this information was added as features to the input of the deep-learning module by concatenation with the word-embedding vectors.

Models

We used a deep recurrent neural network composed of long short-term memory units [44]. Specifically, we used bidirectional long short term memory composed of 2 concatenated long short term memory layers—one reading the input sequence forward, and another one reading the input sequence backward—allowing the model to take advantage of the context on the left and the right of a token when computing the latent states. The final prediction layer was either a standard dense layer with softmax or a conditional random field such as that in [19].

Implementation and Optimization of Hyperparameters

We implemented all the models using Keras and Keras-contrib [45] libraries using Python (version 3) with a TensorFlow backend [46]. We trained our models for 50 epochs, using an ADAM optimizer [47] with a learning rate of 0.001 and early stopping with a patience of 8 epochs. We applied a decrease of learning rate on plateau using a factor of 0.1. For models with a final dense layer, we used categorical cross-entropy loss and softmax activation. For the models with conditional random field, we used marginal optimization and categorical cross-entropy loss. We tuned (using Hyperas version 0.4) the following hyperparameters using a random search with 15 iterations on the parameter space: batch size: 64, 128; long short term memory size: 128, 256, 512; dropout before and after long short term memory; and recurrent dropout: 0.0, 0.1, 0.2, 0.3, 0.5, 0.6, 0.7 (Table S2, Multimedia Appendix 1). All models were trained using NVIDIA P40 GPUs (3840 CUDA cores, 24 GB of DDRAM).

Evaluation

Models

We compared the performance of the rule-based system only, bidirectional long short term memory only, and rule-based system plus bidirectional long short term memory (with and without conditional random field). For bidirectional long short term memory with and without conditional random field models, we tested the impact of adding FastText embeddings or embedding for language model embeddings.

Metrics

We considered an extracted token to be a true positive if it was annotated with the correct category, a false positive if it was falsely annotated with respect to the evaluated class, and a false negative if it was not annotated or if it was annotated with an incorrect class. We computed the precision, recall, and F-measure to evaluate each model, microaveraging over all entries (Multimedia Appendix 2)

We also used the slot error rate metric. A slot corresponded to a mention of an entity (ie, a sequence of B and I tokens of the same class), a deletion was a missing slot, an addition was a slot that had been incorrectly added, a substitution or type error was a class that had been replaced by another class, and a frontier error was a token that had been added or removed at the end or the start of the slot [48].

Results

Annotated Data Set

The labeled data set contained 320 clinical notes and 19,957 sentences with 173,796 words. Training, development, and test sets included 216, 24, and 80 clinical notes with 13,737, 1373, and 4847 sentences, respectively. Table 2 summarizes the number of tokens and slots for each class in each data set.

Table 2. Number of slots and tokens for each class per data set.

Label	Train		Development		Test	
	Tokens	Slots	Tokens	Slots	Tokens	Slots
Medication name	1385	1227	146	143	450	398
Medication class	309	228	38	30	97	76
Dosage	1366	761	115	62	606	311
Frequency	1604	600	142	46	468	184
Duration	161	70	26	13	68	37
Route	95	85	8	8	69	55
Condition	192	61	9	3	89	28

Overall Comparison of the Models

Table 3 summarizes the results of the different models. Overall, the best models were the hybrid models combining rule-based system, text embedding with embedding for language model, and bidirectional long short term memory (F-measure: 89.86). It had the lowest slot error rate (0.19) with a minimal deletion rate (0.05).

The bidirectional long short term memory with baseline embedding had the worst results (F-measure: 73.93). Adding FastText and embedding for language model trained on external data sets increased the F-measure by 14.15 and 9.81 points respectively. Combining rule-based system and bidirectional long short term memory increased the F-measure by 14.1 points.

The rule-based system alone had the highest precision (94.67) with the lowest insertion (0.03) and frontier (0.04) error rates. It had the second-lowest type error rate (0.02) but one of the highest deletion error rates (0.23). Adding bidirectional long short term memory and embedding for language model to the rule-based system increased the F-measure by 10.45 points.

Table 3. Overall medication component information predictions metrics by models.

Model[a]	F-measure	Precision	Recall	Slot error rate	Insertion error rate	Deletion error rate	Type error rate	Frontier error rate
RBS[b]	79.41	94.67	72.28	0.29	0.03	0.23	0.02	0.04
BiLSTM[c]	73.93	83.89	67.57	0.45	0.09	0.25	0.07	0.15
BiLSTM + FT[d]	88.08	89.48	87.17	0.21	0.07	0.08	0.03	0.09
BiLSTM + ELMo[e]	88.03	88.81	87.38	0.24	0.1	0.08	0.03	0.1
BiLSTM + RBS	83.74	88.46	80.24	0.27	0.08	0.13	0.03	0.09
BiLSTM + FT + RBS	88.18	91.73	85.54	0.21	0.07	0.09	0.01	0.07
BiLSTM + ELMo + RBS	89.86	90.83	89.17	0.19	0.09	0.05	0.03	0.08
BiLSTM-CRF[f]	70.12	79.04	65.57	0.53	0.11	0.26	0.11	0.21
BiLSTM-CRF + FT	87.16	88.58	86.41	0.25	0.09	0.08	0.03	0.12
BiLSTM-CRF + ELMo	88.66	87.95	89.44	0.23	0.11	0.06	0.02	0.11
BiLSTM-CRF + RBS	84.16	88.56	80.73	0.27	0.09	0.13	0.03	0.09
BiLSTM-CRF + FT + RBS	87.74	89.72	86.25	0.22	0.08	0.08	0.02	0.09
BiLSTM-CRF + ELMo + RBS	89.3	90.4	88.31	0.20	0.08	0.06	0.02	0.09

[a]Models are described according to their components; if neither ELMo nor FT is mentioned, then we used skip-gram embedding.

[b]RBS: rule-based system (ie, the outputs are added as extra features to the input of the deep learning module).

[c]BiLSTM: bidirectional long short term memory.

[d]FT: FastText embedding.

[e]ELMo: embedding for language model.

[f]CRF: conditional random field.

Comparison by Type of Annotation

Table 4 summarizes the metrics of the different models by type of entities. The rule-based system alone had the lowest F-measure for every class due to a very low recall (medication class: 7.22), but it had the highest precision for all classes with the exception of medication name and duration. Associating the rule-based system to a bidirectional long short term memory

increased medication name, medication class, dosage, and condition metrics (F-measures: 3.13, 3.12, 2.06, and 6.26, respectively) but decreased the F-measure for frequency, duration, and route (F-measures: –1, –3.38, and –2.66, respectively).

Table 4. Medication information predictions metrics results by models.

Label	RBS			BiLSTM + ELMo			BiLSTM + ELMo + RBS		
	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall
Medication name	90.31	96.46	84.89	92.2	93.79	90.67	95.33	95.33	95.33
Medication class	13.33	87.5	7.22	62.3	66.28	58.76	64.36	61.9	67.01
Dosage	90.43	96.62	84.98	92.17	91.13	93.23	95.29	95.52	95.05
Frequency	86.13	98.89	76.28	92.8	93.3	92.31	92.24	93.04	91.45
Duration	48.89	49.25	48.53	82.17	86.89	77.94	78.79	81.25	76.47
Route	47.92	85.19	33.33	75.52	72.97	78.26	72.86	71.83	73.91
Condition	33.64	100	20.22	55.9	62.5	50.56	62.16	77.97	51.69

[a]RBS: rule-based system

[b]BiLSTM: bidirectional long short term memory.

[c]ELMo: embedding for language models.

Discussion

Principal Findings

Our system achieved state-of-the-art performance for the task—an F-measure of 95.33 for medication names and an F-measure of 95.29 for dosage detection. Interestingly, these results were obtained using a data set representing only 10% of the size of similar data sets (N2C2 2018 shared task [22]). Combining expert knowledge (rule-based system) with a deep learning system increased the global F-measure, increased precision, increased recall, and decreased the slot error rate, having the most significant impacts on medication name, medication class, and dosage. While the rule-based system alone achieved the best precision and the worst recall, its association with the deep learning models helped to increase recall (for all information except condition) and increase precision (only for medication name, dosage, and condition of the intake). Adding a deep learning system with the embedding for language model on top of the rule-based system increased F-measures and recall for all categories. Adding a conditional random field layer increased the performance for the most frequent categories (ie, medication name, dosage, frequency). For other entities (ie, duration, route, condition), models with a conditional random field layer did not improve results (Multimedia Appendix 1). These results are consistent with those in the literature [18].

Technical Significance

It is interesting to note that leveraging the synergy between expert knowledge and deep learning allowed us to achieve performance comparable to state-of-the-art with only 10% of the data. Infusing knowledge into deep neural networks will probably be a key element in the future progress of the field. The use of externally trained embeddings is a first step in this direction given that they allow the incorporation of latent knowledge from large corpora into the models. The impact of contextualized embeddings proves that a more accurate representation is even more important. We can expect improved performance with more recent language representation approaches such as BERT [49] or XLNET [50]; however, the cost for fitting these types of models, in terms of computation, time, and data, will be a challenge for languages other than English, for which resources (ie, data) are less available. Therefore, it will be valuable to leverage other types of representations (such as ontologies) to infuse knowledge into neural networks. A possible path could be through specific embedding techniques such as Poincaré embeddings [51].

Our approach is highly versatile. It can be transposed to any language, as long as writing expert rules is feasible. We used regular expressions to this end, but any rule based can be used. Our approach is also transposable to other information extraction use cases (or even text classification).

Clinical Significance

The performance achieved by the system opens the way toward a large-scale use in real-life settings. We are currently developing an implementation to perform the medication information extraction at the scale of our institution. The versatility of the approach will enable its transposition to other types of clinical entities and information.

Related Works

Compared with systems developed on the I2B2 2009 medication data set, the performance of our system is competitive [31]. Regarding token metrics, we showed better performance (medication name, dosage, frequencies, and duration token-level F-measures: +5.03, +4.49, +4.54, +28.89, respectively). However, a direct comparison is difficult given that the data sets are different. First, we trained and evaluated our models on a different corpus of French clinical notes. Also, because of language differences, the annotation guidelines were not strictly identical.

In our corpus, the vast majority of medication name slots contained only one token (48.7% of the medication names in the dictionary contain only one token), therefore, we can approximate a phrase-level F-measure using the token-level

F-measure for medication names to compare with those in recent studies: Tao et al in 2018 reported a medication F-measure of 90.7 on the I2B2 corpus, and we achieved an F-measure of 95.3 [21]. However, regarding the mode of administration, our result was lower (token-level F-measures: 72.9 vs 93.3).

In French-language clinical data sets, mode of administration mentions are less structured and more variable than those in English-language clinical text. Therefore, it is logical to see lower results in this field, and our findings were consistent with the findings from a previous study [32]. Moreover, we took the condition of the intake, and not the reason for the intake, into consideration (which is more specific), and we added a tag regarding the class name; therefore, overall F-measures cannot be compared. Compared with results from a study [33] using a different French-language corpus that obtained a token-level F-measure of 90.4, our system's raw results were higher. Comparisons should be made with caution because the corpus used in [33], though in the same language, was from a different source and contained only 147 documents.

The rule-based system offered the highest precision in most classes. The combination of deep learning and rule-based system could not maintain this high level of precision. One explanation could be that the performance of the rule-based system on the training set led the deep learning module to rely heavily on it. But when the rule-based system failed to generalize on the evaluation set, it caused a drop in accuracy in the hybrid system. This issue could be overcome by forcing the machine learning system to not exclusively rely on one source of information, contextual embedding or rule-based system features, by adding dropouts to the inputs.

Using a rule-based system associated with a deep learning model had two major benefits: the synergy between the rules and the machine learning increased the performance and the preannotation of the documents with the rules decreased the annotation time. Even if hybrid systems had already proved to be efficient [16,21,31,33,52], combining expert knowledge (rules) and latent knowledge (neural network), demonstrated a synergistic effect by increasing the performance in all metrics. It will be interesting to also evaluate approaches combining rules and deep learning in a reverse manner—first using a deep-learning model and refining the results using rules.

Limitations and Perspectives

We have several perspectives from which to continue this work. First, we did not reproduce our study on a standard corpus such as that of the I2B2 challenge. We would, therefore, have to redevelop all the expert rules for this English corpus. Second, the embedding for language model was trained on a set of 100,000 French clinical notes from a single hospital [53]. However, even with these limits, using the embedding for language model proved to be efficient. We can anticipate even better results with an embedding for language model trained on a larger and more diverse corpus. Finally, our study focused on recognizing medication information entities without extracting the relationships among them. Tao et al [21] described a way to model the relationships by predicting boundaries of utterances that contain related medication entities. We plan to extend this to all types of sentences in our corpus, independently of the number of medications mentions. To this end, we will build a multitask model to predict medication fields and relations. We will also predict medication event markers such as start, stop, increase, decrease, switch, or unique intake of medication. Moreover, we could also predict meta-attribute markers that would provide information on the experiencer (patient, family, other), temporality (in the past, present, or for the future), and certainty (eg, factual, suggested, hypothetical, conditional, negated, or contraindicated [54]).

Conclusion

The combination of expert rules, deep contextualized embedding (embedding for language model), and deep neural networks improved medication information extraction. This association achieved high performance on a heterogeneous corpus of French-language clinical reports, despite the data set's small size.

Acknowledgments

The authors thank the AP-HP health data warehouse for supporting this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[DOCX File , 82 KB - medinform_v9i3e17934_app1.docx]

Multimedia Appendix 2
Formulas.
[DOCX File , 79 KB - medinform_v9i3e17934_app2.docx]

References

1. Les dépenses de santé en 2017 - résultats des comptes de la santé - édition 2018. Direction de la recherche, des études, de l'évaluation et des statistiques. 2018. URL: https://drees.solidarites-sante.gouv.fr/publications/panoramas-de-la-drees/les-depenses-de-sante-en-2017-resultats-des-comptes-de-la-sante [accessed 2019-10-01]
2. Olivier P, Boulbés O, Tubery M, Lauque D, Montastruc J, Lapeyre-Mestre M. Assessing the feasibility of using an adverse drug reaction preventability scale in clinical practice: a study in a French emergency department. Drug Saf 2002;25(14):1035-1044. [doi: 10.2165/00002018-200225140-00005] [Medline: 12408734]
3. Pirmohamed M, James S, Meakin S, Green C, Scott AK, Walley TJ, et al. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. BMJ 2004 Jul 03;329(7456):15-19 [FREE Full text] [doi: 10.1136/bmj.329.7456.15] [Medline: 15231615]
4. Zhou L, Mahoney LM, Shakurova A, Goss F, Chang FY, Bates DW, et al. How many medication orders are entered through free-text in EHRs?--a study on hypoglycemic agents. AMIA Annu Symp Proc 2012;2012:1079-1088 [FREE Full text] [Medline: 23304384]
5. Escudié JB, Jannot AS, Zapletal E, Cohen S, Malamut G, Burgun A, et al. Reviewing 741 patients records in two hours with FASTVISU. AMIA Annu Symp Proc 2015;2015:553-559 [FREE Full text] [Medline: 26958189]
6. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform 2017 Sep;73:14-29 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.012] [Medline: 28729030]
7. Gonzalez-Hernandez G, Sarker A, O'Connor K, Savova G. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. Yearb Med Inform 2017 Aug;26(1):214-227 [FREE Full text] [doi: 10.15265/IY-2017-029] [Medline: 29063568]
8. Sirohi E, Peissig P. Study of effect of drug lexicons on medication extraction from electronic medical records. Pac Symp Biocomput 2005:308-318 [FREE Full text] [doi: 10.1142/9789812702456_0029] [Medline: 15759636]
9. Jagannathan V, Mullett CJ, Arbogast JG, Halbritter KA, Yellapragada D, Regulapati S, et al. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. Int J Med Inform 2009 Apr;78(4):284-291. [doi: 10.1016/j.ijmedinf.2008.08.006] [Medline: 18838293]
10. Hyun S, Johnson SB, Bakken S. Exploring the ability of natural language processing to extract data from nursing narratives. Comput Inform Nurs 2009;27(4):215-23; quiz 224 [FREE Full text] [doi: 10.1097/NCN.0b013e3181a91b58] [Medline: 19574746]
11. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 2010 Jan 01;17(1):19-24. [doi: 10.1197/jamia.m3378]
12. Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. J Am Med Inform Assoc 2014;21(5):858-865 [FREE Full text] [doi: 10.1136/amiajnl-2013-002190] [Medline: 24637954]
13. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
14. Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf 2019 Jan;42(1):99-111 [FREE Full text] [doi: 10.1007/s40264-018-0762-z] [Medline: 30649735]
15. Doan S, Collier N, Xu H, Pham HD, Tu MP. Recognition of medication information from discharge summaries using ensembles of classifiers. BMC Med Inform Decis Mak 2012 May 07;12:36 [FREE Full text] [doi: 10.1186/1472-6947-12-36] [Medline: 22564405]
16. Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. BMC Med Inform Decis Mak 2015 May 06;15:37 [FREE Full text] [doi: 10.1186/s12911-015-0160-8] [Medline: 25943550]
17. Zhang Y, Xu J, Chen H, Wang J, Wu Y, Prakasam M, et al. Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning. Database (Oxford) 2016;2016 [FREE Full text] [doi: 10.1093/database/baw049] [Medline: 27087307]
18. Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning.: Morgan Kaufmann Publishers Inc; 2001 Presented at: Eighteenth International Conference on Machine Learning; June 28-July 1; Williamstown, Massachusetts, USA p. 282-289. [doi: 10.5555/645530.655813]
19. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. 2016 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 12-17; San Diego, California p. 260-270. [doi: 10.18653/v1/N16-1030]
20. Sadikin M, Fanany MI, Basaruddin T. A new data representation based on training data characteristics to extract drug name entity in medical text. Comput Intell Neurosci 2016;2016:3483528. [doi: 10.1155/2016/3483528] [Medline: 27843447]
21. Tao C, Filannino M, Uzuner Ö. FABLE: a semi-supervised prescription information extraction system. AMIA Annu Symp Proc 2018;2018:1534-1543 [FREE Full text] [Medline: 30815199]

22. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc 2020 Jan 01;27(1):3-12 [FREE Full text] [doi: 10.1093/jamia/ocz166] [Medline: 31584655]
23. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
24. Christopoulou F, Tran TT, Sahu SK, Miwa M, Ananiadou S. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. J Am Med Inform Assoc 2020 Jan 01;27(1):39-46 [FREE Full text] [doi: 10.1093/jamia/ocz101] [Medline: 31390003]
25. Dai H, Su C, Wu C. Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings. J Am Med Inform Assoc 2020 Jan 01;27(1):47-55 [FREE Full text] [doi: 10.1093/jamia/ocz120] [Medline: 31334805]
26. Kim Y, Meystre SM. Ensemble method-based extraction of medication and related information from clinical texts. J Am Med Inform Assoc 2020 Jan 01;27(1):31-38 [FREE Full text] [doi: 10.1093/jamia/ocz100] [Medline: 31282932]
27. Yang X, Bian J, Fang R, Bjarnadottir R, Hogan W, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. J Am Med Inform Assoc 2020 Jan 01;27(1):65-72 [FREE Full text] [doi: 10.1093/jamia/ocz144] [Medline: 31504605]
28. Chen L, Gu Y, Ji X, Sun Z, Li H, Gao Y, et al. Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning. J Am Med Inform Assoc 2020 Jan 01;27(1):56-64 [FREE Full text] [doi: 10.1093/jamia/ocz141] [Medline: 31591641]
29. Miller T, Geva A, Dligach D. Extracting adverse drug event information with minimal engineering. 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June 6-7; Minneapolis, Minnesota, USA p. 22-27. [doi: 10.18653/v1/W19-1903]
30. Xu J, Lee HJ, Ji Z, Wang J, Wei Q, Xu H. UTH_CCB system for adverse drug reaction extraction from drug labels track. 2017 Presented at: 2017 Text Analysis Conference; November 13-14; Gaithersburg, Maryland, USA.
31. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc 2010;17(5):524-527 [FREE Full text] [doi: 10.1136/jamia.2010.003939] [Medline: 20819856]
32. Deléger L, Grouin C, Zweigenbaum P. Extracting medication information from French clinical texts. Stud Health Technol Inform 2010;160(Pt 2):949-953. [Medline: 20841824]
33. Lerner I, Paris N, Tannier X. Terminologies augmented recurrent neural network model for clinical named entity recognition. J Biomed Inform 2020 Feb;102:103356 [FREE Full text] [doi: 10.1016/j.jbi.2019.103356] [Medline: 31837473]
34. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013 Presented at: 1st International Conference on Learning Representations; May 2-4; Scottsdale, Arizona, USA.
35. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018 Presented at: 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 1-6; New Orleans, Louisiana, USA. [doi: 10.18653/v1/n18-1202]
36. Zhu H, Paschalidis I, Tahmasebi A. Clinical concept extraction with contextual word embedding. arXiv. Preprint posted online on October 24, 2018 [FREE Full text]
37. Jiang M, Sanger T, Liu X. Combining contextualized embeddings and prior knowledge for clinical named entity recognition: evaluation study. JMIR Med Inform 2019 Nov 13;7(4):e14850 [FREE Full text] [doi: 10.2196/14850] [Medline: 31719024]
38. Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. the AP-HP experience. Comput Methods Programs Biomed 2019 Nov;181:104804. [doi: 10.1016/j.cmpb.2018.10.016] [Medline: 30497872]
39. Stenetorp P, Pyysalo S, Topic G, Ohta T, Ananiadou S, Tsujii J. BRAT: a web-based tool for NLP-assisted text annotation. 2012 Presented at: 13th Conference of the European Chapter of the Association for Computational Linguistics; April 23-27; Avignon, France p. 102-107 URL: http://dl.acm.org/citation.cfm?id=2380921.2380942
40. Base de données publique des médicaments. Ministère des Affaires Sociales et de la Santé. URL: http://base-donnees-publique.medicaments.gouv.fr [accessed 2019-10-01]
41. Médicaments remboursés par l'Assurance Maladie. Plateforme Ouverte des Données Publiques Françaises. URL: https://www.data.gouv.fr/fr/datasets/medicaments-rembourses-par-lassurance-maladie/ [accessed 2019-10-01]
42. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. 2013 Presented at: 26th International Conference on Neural Information Processing System; 2013; Lake Tahoe, Nevada, USA p. 3111-3119 URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf [doi: 10.5555/2999792.2999959]
43. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017 Dec;5:135-146. [doi: 10.1162/tacl_a_00051]

44. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
45. Chollet F. Keras: the python deep learning library. Astrophysics Source Code Library. 2018. URL: https://ui.adsabs.harvard.edu/abs/2018ascl.soft06022C [accessed 2019-10-01]
46. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation. 2016 Presented at: 12th USENIX Symposium on Operating Systems Design and Implementation; November 2-4; Savannah, Georgia, USA.
47. Kingma D, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint posted online on January 30, 2017 [FREE Full text]
48. Makhoul J, Kubala F, Schwartz R, Weischedel R. Performance measures for information extraction. In: Proceedings of DARPA Broadcast News Workshop. 1999 Presented at: DARPA Broadcast News Workshop; June 29; Herndon, Virginia, USA p. 249-252.
49. Devlin J, Chang M, Lee K, Toutanova K. BERT: pretraining of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on May 24, 2019 [FREE Full text]
50. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: generalized autoregressive pretraining for language understanding. 2019 Presented at: Advances in Neural Information Processing Systems 32; December 8-14; Vancouver, British Columbia, Canada.
51. Agarwal K, Eftimov T, Addanki R, Choudhury S, Tamang S, Rallo R. Snomed2Vec: random walk and Poincaré embeddings of a clinical knowledge base for health care analytics. arXiv. Preprint posted online on July 19, 2019 [FREE Full text]
52. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514-518 [FREE Full text] [doi: 10.1136/jamia.2010.003947] [Medline: 20819854]
53. Neuraz A, Llanos L, Burgun A, Rosset S. Natural language understanding for task oriented dialog in the biomedical domain in a low resources context. arXiv. Preprint posted online on November 29, 2018 [FREE Full text]
54. Jouffroy J, Feldman S, Neuraz A. Medication extraction annotation guide for french clinical texts. Équipe 22 GitHub. 2019 Jun 06. URL: https://equipe22.github.io/medExtAnnotation/ [accessed 2019-10-01]

Abbreviations

AP-HP: Assistance Publique–Hôpitaux de Paris
IOB: inside, outside, beginning

Edited by C Lovis; submitted 23.01.20; peer-reviewed by S Cossin, L Wang, S Kim; comments to author 06.09.20; revised version received 29.12.20; accepted 20.01.21; published 16.03.21.

Please cite as:
Jouffroy J, Feldman SF, Lerner I, Rance B, Burgun A, Neuraz A
Hybrid Deep Learning for Medication-Related Information Extraction From Clinical Texts in French: MedExt Algorithm Development Study
JMIR Med Inform 2021;9(3):e17934
URL: https://medinform.jmir.org/2021/3/e17934
doi:10.2196/17934
PMID:33724196

Original Paper

A Framework (SOCRATex) for Hierarchical Annotation of Unstructured Electronic Health Records and Integration Into a Standardized Medical Database: Development and Usability Study

Jimyung Park[1*], BS; Seng Chan You[2*], MD, PhD; Eugene Jeong[3], MS; Chunhua Weng[4], PhD; Dongsu Park[5], BS; Jin Roh[6], MD, PhD; Dong Yun Lee[5], MD; Jae Youn Cheong[7], MD, PhD; Jin Wook Choi[8], MD, PhD; Mira Kang[9], MD, PhD; Rae Woong Park[1,5], MD, PhD

[1]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea
[2]Department of Preventive Medicine and Public Health, Yonsei University College of Medicine, Seoul, Republic of Korea
[3]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, United States
[4]Department of Biomedical Informatics, Columbia University, New York, NY, United States
[5]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea
[6]Department of Pathology, Ajou University Hospital, Suwon, Republic of Korea
[7]Department of Gastroenterology, Ajou University School of Medicine, Suwon, Republic of Korea
[8]Department of Radiology, Ajou University School of Medicine, Suwon, Republic of Korea
[9]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Rae Woong Park, MD, PhD
Department of Biomedical Informatics
Ajou University School of Medicine
164, World cup-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do
Suwon, 16499
Republic of Korea
Phone: 82 31 219 4471
Fax: 82 31 219 4472
Email: veritas@ajou.ac.kr

Abstract

Background: Although electronic health records (EHRs) have been widely used in secondary assessments, clinical documents are relatively less utilized owing to the lack of standardized clinical text frameworks across different institutions.

Objective: This study aimed to develop a framework for processing unstructured clinical documents of EHRs and integration with standardized structured data.

Methods: We developed a framework known as Staged Optimization of Curation, Regularization, and Annotation of clinical text (SOCRATex). SOCRATex has the following four aspects: (1) extracting clinical notes for the target population and preprocessing the data, (2) defining the annotation schema with a hierarchical structure, (3) performing document-level hierarchical annotation using the annotation schema, and (4) indexing annotations for a search engine system. To test the usability of the proposed framework, proof-of-concept studies were performed on EHRs. We defined three distinctive patient groups and extracted their clinical documents (ie, pathology reports, radiology reports, and admission notes). The documents were annotated and integrated into the Observational Medical Outcomes Partnership (OMOP)-common data model (CDM) database. The annotations were used for creating Cox proportional hazard models with different settings of clinical analyses to measure (1) all-cause mortality, (2) thyroid cancer recurrence, and (3) 30-day hospital readmission.

Results: Overall, 1055 clinical documents of 953 patients were extracted and annotated using the defined annotation schemas. The generated annotations were indexed into an unstructured textual data repository. Using the annotations of pathology reports, we identified that node metastasis and lymphovascular tumor invasion were associated with all-cause mortality among colon and rectum cancer patients (both P=.02). The other analyses involving measuring thyroid cancer recurrence using radiology reports

and 30-day hospital readmission using admission notes in depressive disorder patients also showed results consistent with previous findings.

Conclusions: We propose a framework for hierarchical annotation of textual data and integration into a standardized OMOP-CDM medical database. The proof-of-concept studies demonstrated that our framework can effectively process and integrate diverse clinical documents with standardized structured data for clinical research.

(JMIR Med Inform 2021;9(3):e23983) doi:10.2196/23983

KEYWORDS

natural language processing; search engine; data curation; data management; common data model

Introduction

Background

With the universal adoption of electronic health records (EHRs), the secondary use of EHRs becomes important for translational research and improvement of the quality of health care [1-3]. EHRs comprise structured (ie, diagnoses, medications, procedures, laboratory tests, and medical device use) and unstructured records, such as clinical notes with diverse formats. Structured data have been widely utilized owing to their processable and standardized codes. In an international open science initiative, Observational Health Data Sciences and Informatics (OHDSI), the structured data of more than 200 hospitals worldwide were mapped into a standardized vocabulary and data structure referred to as the Observational Medical Outcomes Partnership (OMOP)-common data model (CDM) [4]. OHDSI is an open collaborative research community, and researchers from each country have collaborated for discovering medical knowledge. OMOP-CDM version 6.0 consists of 15 clinical data tables, four health system data tables, two health economics data tables, three derived tables, and 10 vocabulary tables. All of the tables are represented with standardized medical terminologies. Using the OMOP-CDM, OHDSI has generated medical evidence through large-scale observational research [5], which can be achieved by the software and user interface to facilitate standardized phenotyping [6], statistical analysis [7], and machine-learning application [8].

Clinical notes with natural language are keeping invaluable information that is not in available structured data, such as clinician's thoughts and medical profiles [9,10]. Although textual data can complement structured data and provide reliable clinical evidence, consistently processing textual data across multiple hospitals has been profoundly restricted. To process unstructured textual data, natural language processing (NLP) technology, an area of computer science for transforming human linguistics into a machine-readable form, is required [11-13]. Clinical documents in the OMOP-CDM have not been actively used for research in OHDSI because of difficulties in consistently processing the textual data and lack of standardized text mining pipelines. Therefore, a standardized clinical text framework for extracting, processing, and annotating unstructured clinical documents is essential to maximize the usefulness of the large body of clinical data in the OMOP-CDM format around the world.

One of the primary streams of clinical NLP is named entity recognition (NER), which extracts information of interest based on annotation schemas [14]. However, most NER studies have used a relatively narrow schema that permits restricted relationships and categories of medical concepts. The restricted medical concepts indicate that only limited information can be extracted from the narratives [15,16]. Conversely, hierarchical annotation leverages a multilevel data structure to extract a wide range of information. Users can richly annotate clinical notes and facilitate the annotations for various purposes. For example, the multilevel structure can contain the hierarchy and relations of the observed tumor, differentiation, gross type, invasion, size, and other characteristics, while the narrow schema cannot include this information. This rich information can be extracted through the hierarchical schema and can be facilitated for answering a variety of research questions. Therefore, a hierarchical annotation schema is more desirable for clinical research [17-20].

Comparison With Prior Work

One of the attempts to standardize diverse EHR formats into CDM is the Sentinel project. Sentinel and its component (ie, Mini-Sentinel) have been developed by the United States Food and Drug Administration (FDA), with the aim to create an active surveillance system for monitoring the safety of medical products [21]. Sentinel is a US domestic data model, and the OMOP-CDM was used in this study because of its international research network and wide coverage of standardized medical terminology [22].

In the aspects of NLP frameworks, many NLP information extraction and retrieval systems have been developed to process documents in EHRs for use in clinical practice or research. EMERSE is a clinical note searching system developed using Apache Lucene to increase the availability of EHRs and to help clinicians and researchers effectively retrieve information [23]. SemEHR provides a biomedical information extraction and semantic search system for clinical notes, and several case studies have proven the system's usability [24]. SemEHR facilitates Fast Healthcare Interoperability Resources (FHIR) to represent the clinical semantic concepts extracted from free text. cTAKES and CLAMP are widely used NLP systems that provide serial components for information extraction [25,26]. CREATE is an information retrieval system based on the OMOP-CDM for executing textual cohort selection queries on structured and unstructured data [27]. On the other hand, Sharma et al proposed a phenotyping system with NLP algorithms to extract features from the clinical documents of the OMOP-CDM database [28].

Despite well-performing systems, using the systems is still difficult since the systems require high optimization for the local environment and extensive domain knowledge [29]. Moreover, clinical note extraction and preprocessing are needed separately from the systems. The lack of a user interface limits the systems' usability and portability. Hence, clinical NLP systems that can be applied to standardized medical databases and provide serial NLP components to enhance research continuity are required for users. In this study, we chose the OMOP-CDM owing to its wide coverage of standardized medical terminology and worldwide distributed research networks.

Objectives

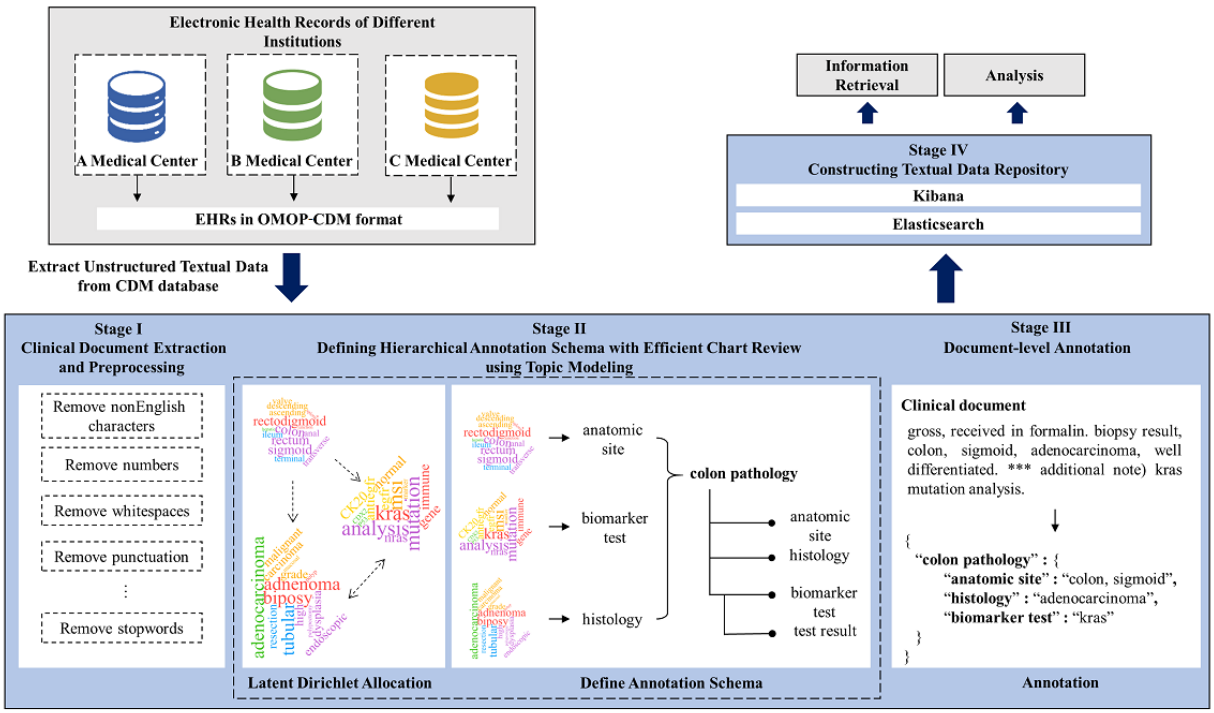
This study aimed to integrate unstructured clinical textual data with structured data through the framework referred to as Staged Optimization of Curation, Regularization, and Annotation of clinical text (SOCRATex). The proposed framework was designed (1) to define a flexible hierarchical annotation schema containing complex clinical information through efficient chart review, (2) to generate reusable annotations based on user-configurable JavaScript object notation (JSON) architecture, and (3) to construct a clinical text data repository that can be integrated with the standardized structured data.

Methods

System Architecture

SOCRATex follows a pipeline-based architecture with the following four stages: (1) extracting clinical notes for the target population and preprocessing the data, (2) defining the annotation schema with a hierarchical structure by referring clustered topics from the clinical notes; (3) performing document-level hierarchical annotation; and (4) constructing a textual data repository with a search engine (Figure 1). All source codes are available online [30].

Figure 1. The overall system architecture of Staged Optimization of Curation, Regularization, and Annotation of clinical text (SOCRATex). The system has the following four stages: (1) extracting clinical notes for the target population and preprocessing the data, (2) defining the annotation schema with a hierarchical structure, (3) performing document-level hierarchical annotation using the annotation schema, and (4) indexing annotations for a search engine system. CDM: common data model; EHR: electronic health record; OMOP: Observational Medical Outcomes Partnership.



Stage 1: Data Extraction and Preprocessing

In the first stage of SOCRATex, the user defines the target population. OHDSI provides an open-source software stack known as ATLAS, which enables users to define complex and transferrable phenotypes of interest based on structured data (ie, diagnosis, medication prescription, medical device use, and laboratory measurements) [31]. The documents in the OMOP-CDM are fully connected to other structured data through patient identifiers. Information regarding note type, language, and encoding system are stored with a fully standardized vocabulary [32].

In the NOTE table, foreign keys that can be connected with other tables exist in the CDM (Figures S1 and S2 in Multimedia Appendix 1). NOTE_EVENT_ID is a foreign key identifier of the event (ie, drug exposure, visit, and procedure) during which the note was recorded. NOTE_EVENT_FIELD_CONCEPT_ID is a standardized vocabulary showing which NOTE_EVENT_ID is being referred to. NOTE_TYPE_CONCEPT_ID represents the type, origin, or provenance of the recorded clinical notes. SOCRATex extracts a certain type of clinical document for the target population by using NOTE_TYPE_CONCEPT_ID.

The developed framework provides conventional preprocessing functions, such as eliminating stop words, white spaces, numbers, and punctuations; changing text to lowercase;

stemming; and generating a document-term matrix. SOCRATex users can add specific regular expressions or terms to the stop words list.

Stage 2: Defining the Annotation Schema With a Hierarchical Structure

To define an annotation schema for organizing hierarchical entities of medical documents, researchers with domain knowledge need to review the overall documents of interest thoroughly. By leveraging latent Dirichlet allocation (LDA), which clusters similar words based on the word distributions over documents, SOCRATex automatically identifies topic clusters among documents of interest and provides samples of each cluster to researchers [33,34]. It is assumed that the sampled documents can represent the semantic characteristics of the extracted documents because the topic clusters represent the latent semantics of the documents. Therefore, reviewing the samples can suggest an efficient chart review process rather than reviewing the documents. This reduces redundant labor for reviewing charts of similar content to understand the documents of interest comprehensively.

To calculate the optimal number of topics in LDA, we used perplexity scores, a statistical measure for probabilistic models. Users can decide the best hyperparameters for LDA performance based on the perplexity scores [35-39]. For the interpretation of LDA results, SOCRATex shows both words and documents from their associated topics (Figures S1 and S2 in Multimedia Appendix 2). Based on LDA topics, users can define the annotation schema using the JSON architecture, a machine readable and hierarchical architecture consisting of entity-value pairs.

Stage 3: Document-Level Annotation With a Defined Schema

Manual annotation is notorious for being an error-prone process. To limit the errors and ensure annotation quality, we applied the JSON schema that can restrict the values and data types of annotation entities [40]. Users need to specify the allowed values of annotation entities using the JSON format. For instance, diameters of observed tumors can be restricted to numeric values. The annotation schema can be distributed to other institutions for generating homogeneous annotations.

Stage 4: Constructing a Textual Data Repository for Data Exploration and Retrieval

Elastic Stack, a group of open-source products specialized in textual data exploration and retrieval, is used for constructing a textual data repository for the annotations. Elastic Stack is composed of Elasticsearch and Kibana. Elasticsearch is a full-text search and analytics engine for textual data, and Kibana is its visualization dashboard [41]. SOCRATex can index the generated annotations into Elasticsearch, and users can explore their data using Kibana (Figure S1 in Multimedia Appendix 3).

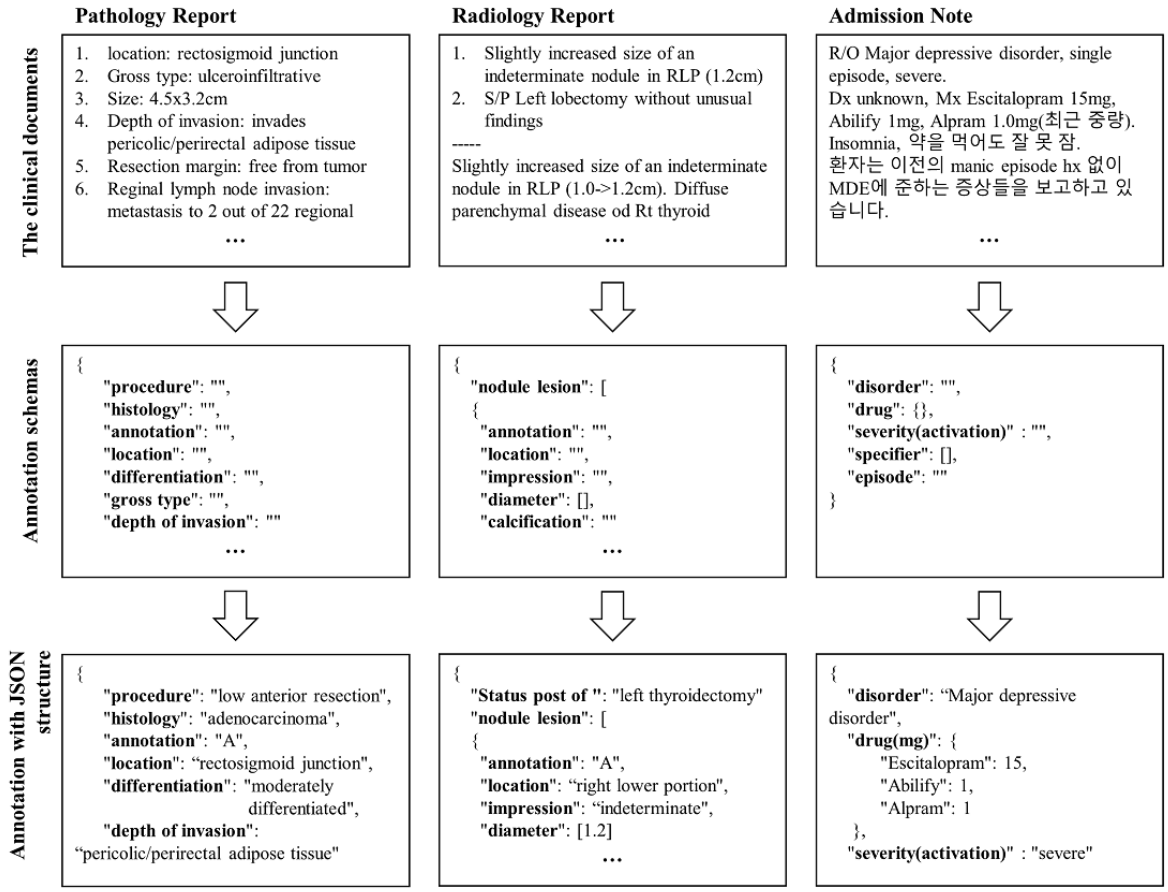
Validation Using EHRs

We applied SOCRATex against hospital data to validate the usability of the framework. The following three distinctive groups were defined using the OMOP-CDM database of Ajou University School of Medicine [42]: (1) patients who were diagnosed with malignant neoplasms of the colon and rectum between 2014 and 2017, (2) patients who were diagnosed with malignant neoplasms of the thyroid gland and who underwent thyroidectomy between 2014 and 2016, and (3) patients who were diagnosed with major depressive disorder and hospitalized via the emergency department between 2012 and 2018.

From each group of patients, we extracted a specific type of clinical note. Among the patients with colorectal cancer, we extracted their pathology reports with the statement of cancerous lesions of the colon and rectum. Radiology reports of postoperative thyroid ultrasonography were extracted for the patients who underwent thyroidectomy owing to thyroid cancer. Among the patients with major depressive disorder, admission notes were selected and identified with a description of the reason for hospitalization.

Each note type was selected because of its different characteristics (Figure 2). Pathology reports have a semistructured format that is similar to the synoptic pathology reporting form and are primarily written in English [43]. Radiology reports feature a semistructured data format and narrative sentences. Admission notes have narrative descriptions of medical history, disease diagnosis, and medication prescription of the patients. Korean characters were removed and only English characters were included for topic modeling analysis. During the annotation process, we used both languages for accurate annotation. To evaluate the accuracy and efficiency of the SOCRATex annotation process, we compared the annotation process of our system and traditional manual chart review.

Figure 2. Examples of annotating certain types of clinical documents and their annotation process. Pathology reports have a semistructured format, and radiology reports have a semistructured format with narrative sentences. Admission notes have narrative descriptions in both Korean and English.



Both structured and unstructured textual data were deidentified to protect patient data. The OMOP-CDM per se is a pseudonymized data model that does not allow identifying specific individuals with the data. Hence, it is compliant with pseudonymization of the EU General Data Protection Regulation and Health Insurance Portability and Accountability Act of 1996 (HIPAA) regulations [44,45]. Moreover, the deidentification process in Ajou University Hospital was applied to the data sets to ensure privacy protection (Figure S1 in Multimedia Appendix 4). With the process, patient IDs are encrypted and only the researcher with IRB approval is allowed to receive decryption keys. However, the unstructured textual data can still contain private information. Therefore, a rule-based algorithm was applied to eliminate HIPAA-defined protected health information (PHI) and Korean PHI from the narratives (Tables S1 and S2 in Multimedia Appendix 4). We applied the algorithm by Shin et al that was developed on bilingual clinical documents (ie, Korean and English), was validated on 5000 notes of 33 types, and showed 99.87% precision [46]. The rules for the data set of this study were then optimized and updated.

As proof-of-concept studies, we performed survival analyses to measure mortality rates, cancer recurrence, and hospital readmission using information from both structured clinical data and medical narratives. All-cause mortality, thyroid cancer diagnoses, and hospital readmission information were extracted from structured coded data and defined as outcomes of the analyses. From the annotations, we extracted the following clinical features that were not in structured data: node metastasis, lymphovascular tumor invasion, echogenicity of thyroid nodules, and episodes and specifiers of major depressive disorder. The episodes and specifiers were measured using the Diagnostics and Statistical Manual of Mental Disorder (DSM-5) [47]. Furthermore, we calculated the Korean Thyroid Imaging Reporting and Data System (K-TIRADS) score, a risk stratification of thyroid nodules using the extracted covariates (ie, size, content, and echogenicity of thyroid nodules) [48]. A high K-TIRADS score indicates that the observed thyroid lesions are suspected to be malignant.

In patients diagnosed with colon and rectum cancer, we measured all-cause mortality stratified by node metastasis and lymphovascular invasion. Thyroid cancer recurrence in patients who underwent thyroidectomy was measured with the K-TIRADS score and echogenicity on ultrasonography. Among the patients with major depressive disorder, hospital readmission was measured with specifiers and episodes of major depressive disorder. The *P* value of the log-rank test with Kaplan-Meier curves was measured on each annotation body. We used Cox proportional hazard models to assess and calculate the hazard ratio (HR) between the defined groups. HRs are presented with 95% CIs and *P* values. All *P* values <.05 were considered statistically significant.

To demonstrate external feasibility, we applied SOCRATex to pathology reports from another tertiary hospital's OMOP-CDM database. This study was approved by the Institutional Review Board at Ajou University Hospital (IRB approval number: AJIRB-MED-MDB-19-579).

Results

Stage 1: Defining Patient Groups and Extracting Clinical Documents

Overall, 600 pathology reports from 588 patients with colon and rectum cancer, 308 radiology reports from 220 patients who underwent thyroidectomy, and 147 admission notes from 145 patients with major depressive disorder were included in the study. The characteristics of the patients are shown in Table 1. To compare the cohorts, medical history of the patients was extracted using structured coded data.

Moreover, the information loss and accuracy of clinical note extraction were investigated (Tables S1, S2, and S3 in Multimedia Appendix 1). It showed that data sparsity dropped less than 1% in pathology and radiology reports and 4% in admission notes despite eliminating non-English character removal. The most frequent tokens in documents usually consisted of English characters and a few Korean characters, such as "환자는 (the patient)," "하였다 (did)," and "정신과 (department of psychiatry)."

Table 1. Baseline characteristics of the patient groups.

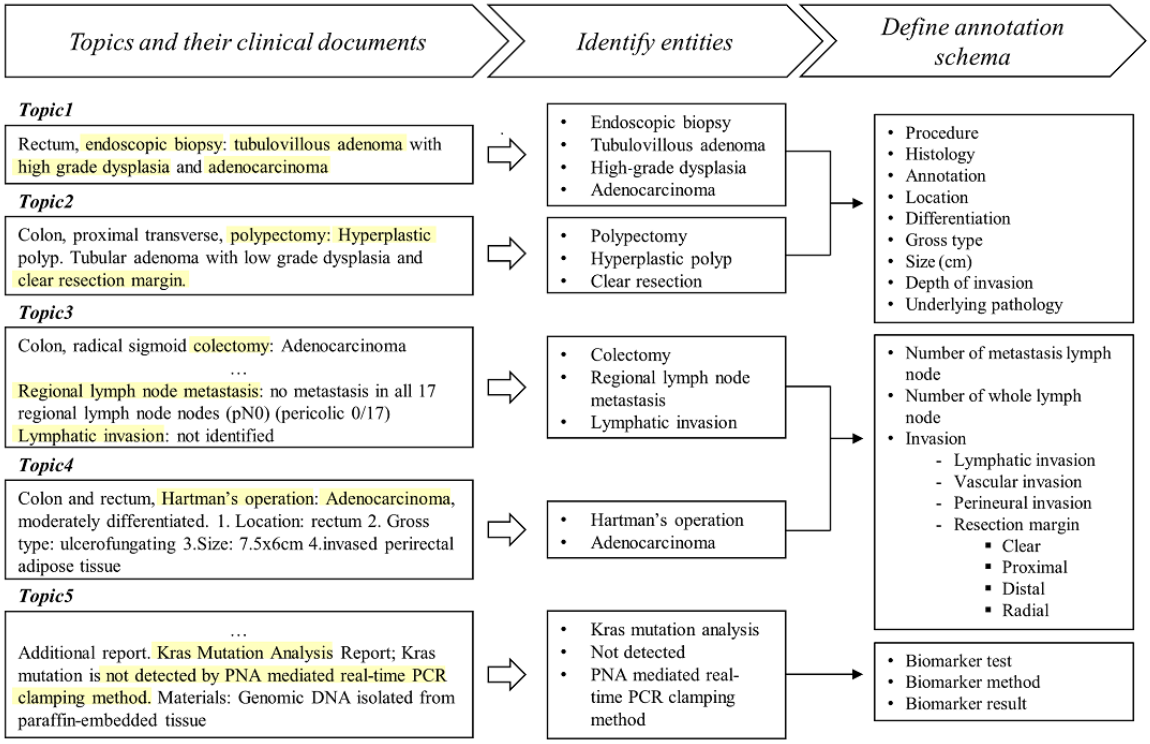
Characteristic	Patients with pathology reports (n=588)	Patients with radiology reports (n=220)	Patients with psychiatric admission notes (n=145)	*P* value
Age (years), mean (SD)	62.65 (12.58)	46.52 (18.69)	49.12 (19.59)	<.001
Female, n (%)	229 (38.9)	176 (80.0)	107 (73.8)	<.001
General medical history, n (%)				
Dementia	6 (1.1)	0 (0.0)	0 (0.0)	.23
Gastroesophageal reflux disease	9 (1.5)	8 (3.6)	0 (0.0)	.03
Gastrointestinal hemorrhage	31 (5.3)	1 (0.5)	0 (0.0)	<.001
Hyperlipidemia	9 (1.5)	11 (5.0)	3 (2.1)	<.001
Hypertensive disorder	165 (28.1)	15 (6.8)	2 (1.4)	<.001
Diabetes mellitus	84 (14.3)	18 (8.2)	0 (0.0)	<.001
Renal impairment	22 (3.7)	3 (1.4)	0 (0.0)	.01
Liver lesion	30 (5.1)	1 (0.5)	0 (0.0)	<.001
Cardiovascular disease, n (%)				
Atrial fibrillation	11 (1.9)	0 (0.0)	1 (0.7)	.08
Cerebrovascular disease	6 (1.0)	1 (0.5)	0 (0.0)	.64
Coronary arteriosclerosis	10 (1.7)	3 (1.4)	0 (0.0)	.34
Heart disease	39 (6.6)	8 (3.6)	1 (0.7)	.008
Heart failure	7 (1.2)	2 (0.9)	0 (0.0)	.45
Ischemic heart disease	16 (2.7)	2 (0.9)	0 (0.0)	.048
Peripheral vascular disease	10 (1.7)	3 (1.4)	1 (0.7)	.86

Stage 2: Defining the Annotation Schema With a Hierarchical Structure

The optimal number of topics for pathology reports was determined to be 5, whereas the optimal number of both radiology reports and admission notes was 4 (Table S1 in Multimedia Appendix 2).

We defined a hierarchical schema of pathology reports based on the topics and sample documents (Figure 3). The entities of pathology reports were classified into the following three groups: lesions, lymph nodes, and biomarker tests. Each entity has a multilevel structure, especially the invasion entity, which showed a deep multilevel structure containing the hierarchical information of lymphatic, vascular, and perineural invasion, and resection margin. The annotation schemas of radiology reports and admission notes are shown in Figures S3 and S4 in Multimedia Appendix 2. Overall, 23 entities were defined for pathology reports, 20 entities for radiology reports, and 5 entities for admission notes.

Figure 3. Defining a hierarchical annotation schema of pathology reports, which describes lesions of colon and rectum cancer. The process had the following three steps: (1) classifying documents using clustered topics from the latent Dirichlet allocation model, (2) identifying medical entities of interest, and (3) designing the annotation schema. PCR: polymerase chain reaction; PNA: peptide nucleic acid.



Stage 3: Document-Level Annotation With a Defined Schema

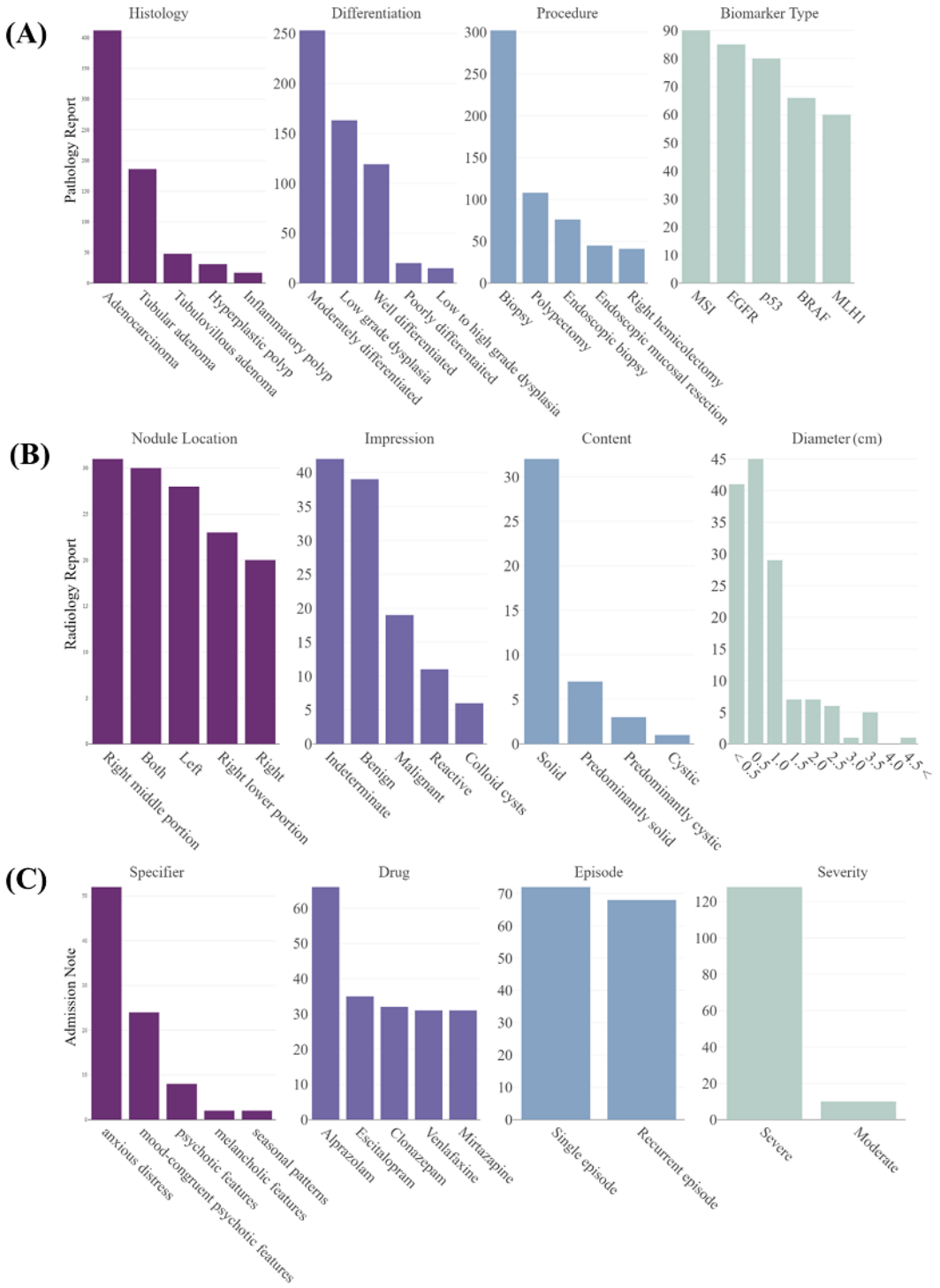
Document-level annotation was applied on the extracted documents, resulting in the annotation of 1055 clinical documents with the defined schema. A total of 1000 colonoscopy pathology reports from another tertiary hospital database were annotated with the distributed annotation schema (Multimedia Appendix 5). The comparison between SOCRATex annotation and traditional chart review is described in Multimedia Appendix 6. It shows that the mean accuracy of traditional chart review was 0.917 and its mean annotation time was 548 minutes. On the other hand, the mean accuracy of SOCRATex annotation was 0.937 and its mean annotation time was 360 minutes.

Stage 4: Constructing a Textual Data Repository for Data Exploration and Retrieval

The generated annotations were indexed into Elasticsearch to construct a textual data repository. Table S1 in Multimedia Appendix 3 demonstrates that the admission notes were identified as having the largest tokens (24,319 tokens) and that the radiology reports were identified as having 1006 tokens. The tokens of pathology reports were 3561.

Using the constructed textual data repository, we explored the entity distributions of the annotations using the Kibana interface (Figure 4). Figure 4A shows the distributions of pathology entities. It shows that adenocarcinoma was the most frequent tumor, which was observed in 412 of 600 documents (68.7%). Tubular and tubulovillous adenomas were the second most frequent tumors, which were observed in 186 (31.0%) and 48 (8.0%) documents, respectively. Among the biomarker tests, the microsatellite instability test was identified as the most frequent biomarker test with 90 (50.3%) occurrences, followed by epidermal growth factor receptor with 85 (47.5%) occurrences. The distributions of radiology entities showed that solid or predominantly solid thyroid nodules were observed in 34 of 148 documents, in which 209 (16.2%) nodules were observed via thyroid ultrasonography (Figure 4B). There were only 4 (2.70%) documents describing cystic or predominantly cystic nodules. Of 144 observed lesions with nodule size, 20 (14.1%) nodules were larger than 2.0 cm and the other 122 (85.9%) nodules were less than 2.0 cm. Using the DSM-5, we identified the severity, episode, and specifier of major depressive disorder from admission notes (Figure 4C). As a result, 52 (35.4%) hospitalized cases and 33 (22.5%) cases were identified as having anxious distress of major depressive disorder and psychotic or mood-congruent psychotic features, respectively. In addition, we identified the medication usage patterns of patients. The most frequently prescribed medication was alprazolam with 66 (22.6%) prescriptions, followed by escitalopram with 35 (11.3%) prescriptions.

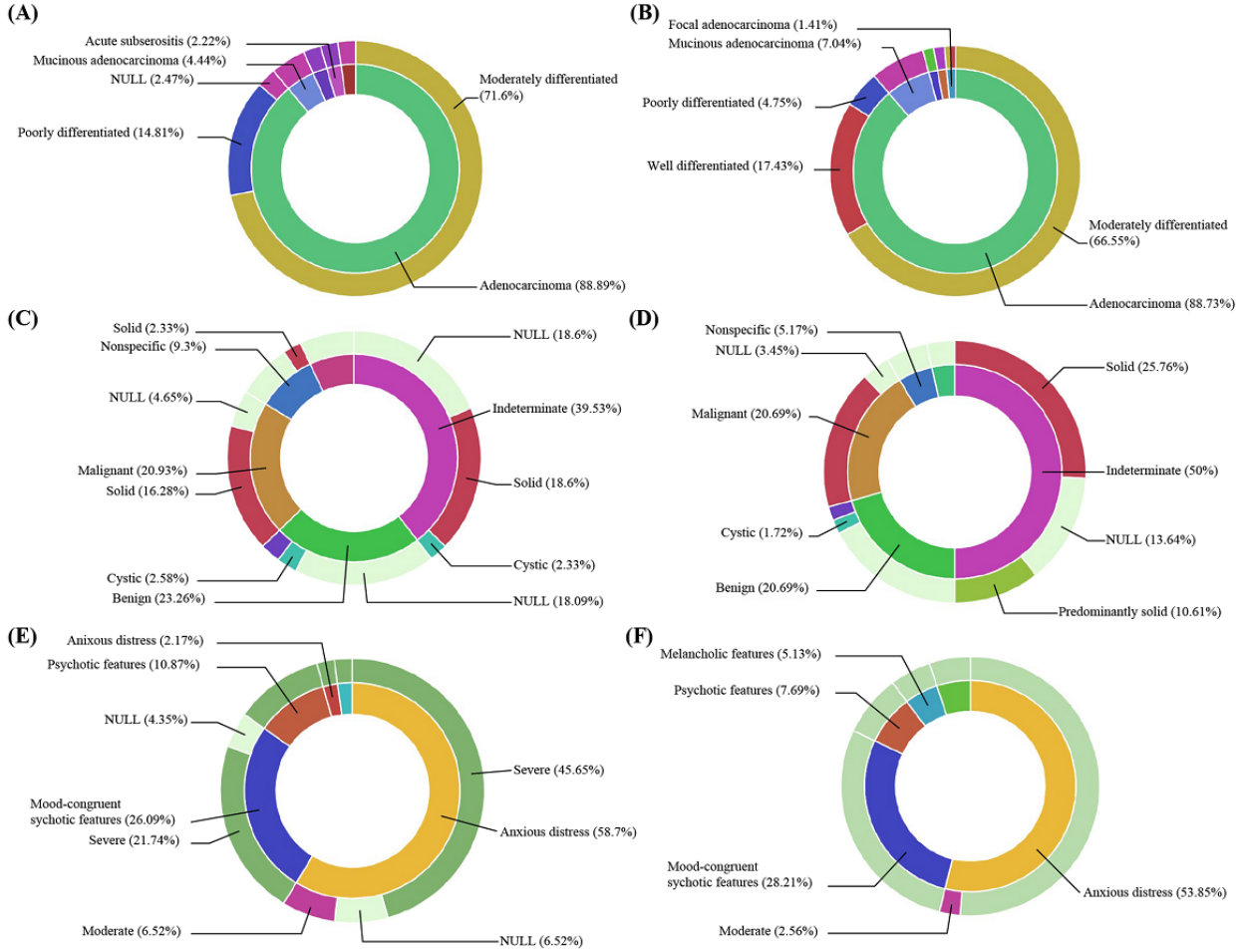
Figure 4. Histograms of annotation entities derived from pathology reports (A), radiology reports (B), and admission notes (C). (A) shows the number of observed histologies, differentiations, procedures, and biomarkers; (B) shows the number of locations, impressions, contents, and diameters of the observed thyroid nodules; and (C) shows the specifiers, episodes, severities, and used medications in major depressive disorder patients.



Hierarchical annotations can show further relationships between the entities. Figure 5 describes the hierarchical relations of the entities. Using Kibana queries, we classified each annotation body into two categories. First, the observed tumors and the differentiation from pathology report findings are described (Figure 5A and 5B). Both are distinguished by lymph node positivity. Among moderately differentiated adenocarcinomas, 29 (71.6%) lymph node–positive cases and 45 (66.6%) lymph node–negative cases were observed. Second, frequent differentiation of adenocarcinoma differed according to lymph node involvement as follows: 6 (14.8%) poorly differentiated in lymph node–positive cases and 11 (17.4%) well differentiated in lymph node–negative cases. Additionally, relations of thyroid nodule types and contents by anatomic locations are described

in Figure 5C and 5D. The results show that solid nodules with malignancy were observed in 7 (16.3%) cases in the left thyroid and 10 (14.1%) cases in the right thyroid. On the contrary, benign cystic thyroid nodules were observed in only 2 (2.6%) cases in the left thyroid and 2 (1.7%) cases in the right thyroid. Third, the specifiers and severities of major depressive disorder were identified (Figure 5E and 5F). The results were divided according to single or recurrent episodes of major depressive disorder. Among the patients with single episodes of the disease, 21 (45.7%) cases were identified as involving severe major depressive disorder with an anxious distress specifier. Additionally, 20 (51.3%) cases of multiple episodes were identified as involving an anxious distress specifier with severe symptoms.

Figure 5. Sunburst plots generated using the Kibana interface. (A) and (B) show the observed histologies and their differentiation from pathology reports. (A) shows the results of lymph node–positive cases, and (B) shows the results of lymph node–negative cases. (C) and (D) are observed from radiology reports. Each of the plots indicates the left and right thyroid in order. (E) and (F) show the disease specifier and its severity from the admission notes. (E) shows the results of single-episode patients, and (F) shows the results of multiple-episode patients.

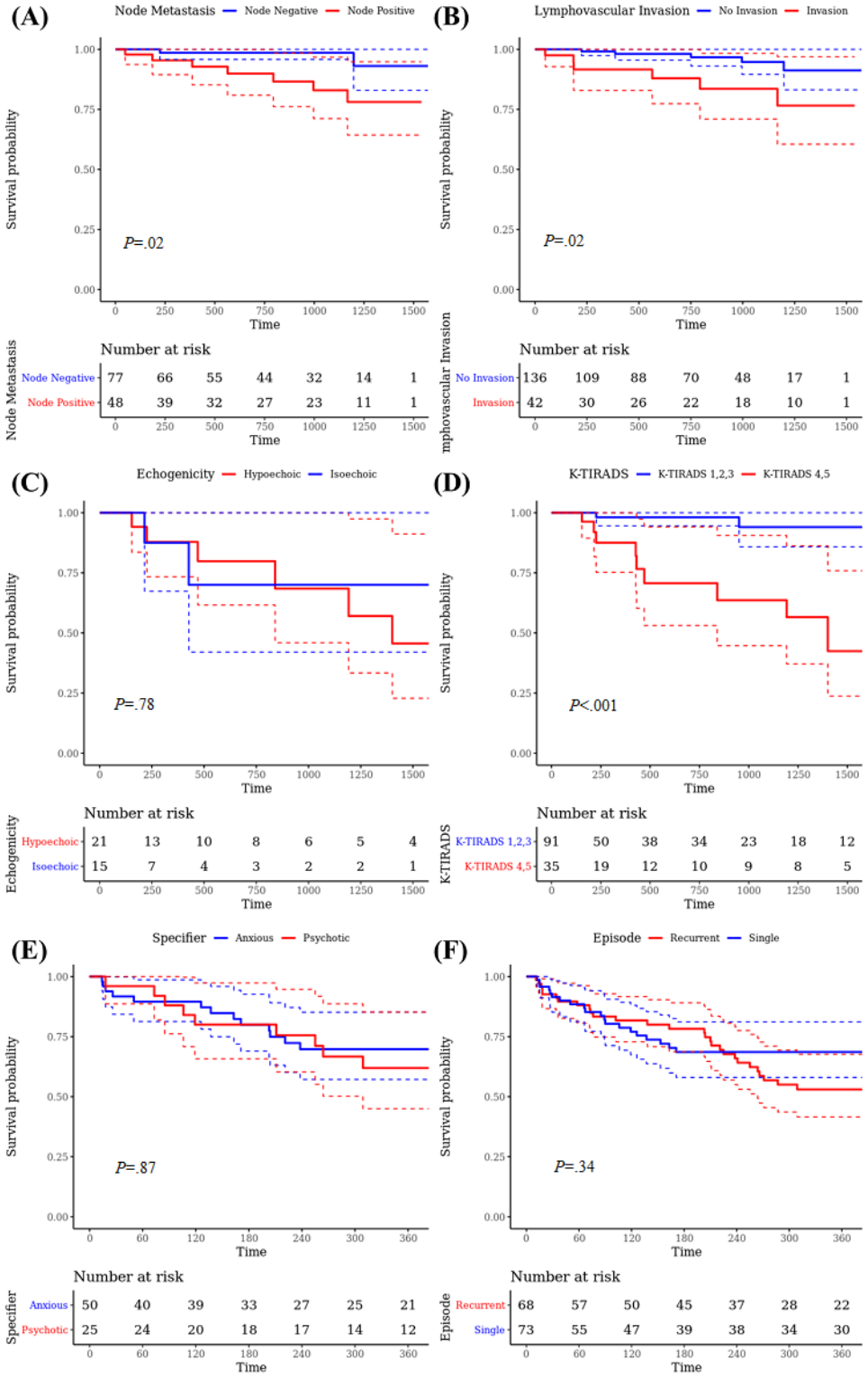


The annotation results of the other tertiary hospital database are described in Multimedia Appendix 5. Tubular adenoma observed at the sigmoid colon was the most frequent histology with 131 cases among 720 observed lesions (20.1%), and hyperplastic polyps represented the second most frequent histology in the sigmoid colon with 76 (11.7%) cases.

Association of Features From Clinical Notes and Structured Data

For patients diagnosed with malignant neoplasm of the colon and rectum, 5-year survival analyses were performed (Figure 6A and 6B). The analyses measured mortality rate according to node metastasis and lymphovascular tumor invasion. We found that patients with lymph node involvement had significantly worse survival rates than those without involvement (HR 5.22, 95% CI 1.08-25.22; *P*=.04). Lymphovascular invasion was also associated with significantly higher mortality in patients with colorectal cancer (HR 3.75, 95% CI 1.14-12.32; *P*=.03).

Figure 6. Kaplan-Meier curves with P values of the log-rank test. Survival analyses were performed. (A) and (B) measure 5-year mortality rates of patients with colorectal cancer by node metastasis and lymphovascular tumor invasion, respectively. (C) and (D) measure thyroid cancer recurrence by echogenicity of thyroid nodules and K-TIRADS scores, respectively. (E) and (F) measure 30-day readmission of patients with major depressive disorder by disease specifiers and episodes, respectively. K-TIRADS: Korean Thyroid Imaging Reporting and Data System.



Recurrence risk of thyroid cancer stratified by the echogenicity of thyroid nodules and the K-TIRADS score was measured (Figure 6C and 6D). In our analysis, recurrence of thyroid cancer was not significantly associated with the echogenicity of thyroid nodules (HR 0.80, 95% CI 0.16-3.98; *P*=.78). On the other hand, we found that high K-TIRADS scores (K-TIRADS 3 and 4) were associated with a higher risk of thyroid cancer recurrence compared with low K-TIRADS scores (K-TIRADS 1-3) (HR 12.43, 95% CI 2.73-56.60; *P*<.001).

Among patients with major depressive disorder, we measured 30-day readmission according to disease specifiers and episodes, which were measured based on the DSM-5 (Figure 6E and 6F). The specifiers were classified into anxious distress and psychotic

features. Disease episodes were classified into single or recurrent episodes. The results showed that 30-day readmission was not significantly associated with the specifiers of major depressive disorder (HR 1.07, 95% CI 0.50-2.26; *P*=.87). Single or recurrent episodes of major depressive disorder were not significantly associated with 30-day readmission (HR 0.78, 95% CI 0.47-1.29; *P*=.34).

Discussion

Principal Findings

The framework succeeded in hierarchically annotating unstructured clinical documents and integrating them into standardized structured data. Through proof-of-concept studies, three different types of clinical documents (ie, pathology reports, radiology reports, and admission notes) were extracted and processed with topic modeling to identify medical concepts. The hierarchical schemas were defined with efficient chart review by sampling documents according to semantic topics. Overall, 1055 documents were manually annotated using the schemas and indexed in the search engine. We attempted multidimensional validation by identifying the characteristics of the hierarchical annotations and by performing survival analyses with integrated data of structured and unstructured textual information. The following were identified through validation: (1) the association of node positivity with mortality in patients with colorectal cancer, (2) the association of the K-TIRADS score with thyroid cancer mortality, and (3) medication usage patterns according to depression episodes.

SOCRATex uses flexible annotation schemas for clinical text annotation that can include complex information in free-text documents (Multimedia Appendix 7). The narrow annotation schema can only extract the entities of disease, treatment, and test [15,16]. These simple entities are effective to annotate and train the model, but difficult to explain their relationships. On the contrary, the annotation schema on pathology reports successfully contained the relationships among tumor type, dimension, location, and invasion. Consequently, we identified that more than 42% (253/588) of colorectal cancer patients had moderately differentiated adenocarcinoma and underwent a microsatellite instability test. In the radiology reports, 23% (34/148) of thyroid nodules were identified as having solid content. The hierarchical schema of admission notes identified medication usage patterns by disease episodes, showing that alprazolam and escitalopram were the most frequently prescribed medications in both patient groups.

Through proof-of-concept studies, we demonstrated that the generated hierarchical annotations could be used in various settings of clinical research. The survival analyses of patients with colorectal cancer showed that node positivity and lymphovascular invasion were significantly associated with a higher mortality rate, which is consistent with the findings of previous studies [49,50]. The analyses of radiology reports found that higher K-TIRADS scores were significantly associated with the recurrence of thyroid cancer, which is consistent with previous reports [48].

Limitations

This study has several limitations that can direct future research. First, interesting clinical implications were not determined from our proof-of-concept studies. To discover novel medical evidence, a sophisticated study design is required. However, our aim here was to demonstrate that the generated textual data repository could be used for clinical research. Second, the feasibility of the framework in the distributed research network was not fully validated. Still, we distributed the annotation schema of pathology reports to the other institution and were able to annotate 1000 colonoscopy pathology reports. Third, the defined annotation schema was not systemically evaluated. Three annotation schemas were defined with domain experts according to their related clinical domains. However, systematic validation of the schemas is still required. Moreover, the applicability of FHIR standards in the system of this study will be investigated to test its extensibility.

Although the generated annotations can be reused for clinical analyses of various purposes, the initial manual annotation of documents is still a time-consuming and costly process. In future work, state-of-the-art algorithms, such as BERT, XLNet, and GPT-3, could be applied to automatic information extraction processes to reduce the annotation burden and cost [51-53].

Conclusions

We propose a clinical text processing framework to generate flexible hierarchical annotations and integrate them with the standardized structured data of the OMOP-CDM. The proof-of-concept studies demonstrated that the generated annotations were integrated with the structured data and were successfully used for various clinical research approaches with efficient chart review processes. The conformance with CDM allows the application of a standard annotation schema to generate homogeneous annotations from different institutions.

Acknowledgments

This work was supported by the Bio Industrial Strategic Technology Development Program (20001234, 20003883) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI16C0992, HI19C0872).

Authors' Contributions

SCY, JP, and RWP contributed to the study design. JR, DYL, JYC, JWC, MK, and RWP obtained the relevant data used for the study. JP and DP contributed to the development and evaluation of SOCRATex. JP, SCY, EJ, CW, and RWP contributed to

writing and revising the paper. All authors contributed to the writing and final approval of this manuscript. JP and SCY contributed equally to this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Clinical note data extraction, processing, and validation.
[DOCX File , 745 KB - medinform_v9i3e23983_app1.docx]

Multimedia Appendix 2
Evaluating the latent Dirichlet allocation model performance and defining annotation schemas.
[DOCX File , 1067 KB - medinform_v9i3e23983_app2.docx]

Multimedia Appendix 3
Staged Optimization of Curation, Regularization, and Annotation of clinical text (SOCRATex) annotation and information retrieval system.
[DOCX File , 519 KB - medinform_v9i3e23983_app3.docx]

Multimedia Appendix 4
Protecting and deidentifying patient information.
[DOCX File , 434 KB - medinform_v9i3e23983_app4.docx]

Multimedia Appendix 5
Study results from Samsung Medical Center.
[DOCX File , 239 KB - medinform_v9i3e23983_app5.docx]

Multimedia Appendix 6
Comparison between Staged Optimization of Curation, Regularization, and Annotation of clinical text (SOCRATex) annotation and traditional chart review.
[DOCX File , 14 KB - medinform_v9i3e23983_app6.docx]

Multimedia Appendix 7
Comparison between Staged Optimization of Curation, Regularization, and Annotation of clinical text (SOCRATex) and other natural language processing systems.
[DOCX File , 16 KB - medinform_v9i3e23983_app7.docx]

References

1. Blumenthal D. Launching HITECH. N Engl J Med 2010 Feb 04;362(5):382-385. [doi: 10.1056/NEJMp0912825] [Medline: 20042745]
2. Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med 1999 Aug;74(8):890-895. [doi: 10.1097/00001888-199908000-00012] [Medline: 10495728]
3. Gans D, Kralewski J, Hammons T, Dowd B. Medical groups' adoption of electronic health records and information systems. Health Aff (Millwood) 2005 Sep;24(5):1323-1333. [doi: 10.1377/hlthaff.24.5.1323] [Medline: 16162580]
4. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
5. Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. The Lancet 2019 Nov 16;394(10211):1816-1826. [doi: 10.1016/S0140-6736(19)32317-7] [Medline: 31668726]
6. Weng C, Shah NH, Hripcsak G. Deep phenotyping: Embracing complexity and temporality-Towards scalability, portability, and interoperability. J Biomed Inform 2020 May;105:103433 [FREE Full text] [doi: 10.1016/j.jbi.2020.103433] [Medline: 32335224]
7. Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. Improving reproducibility by using high-throughput observational studies with empirical calibration. Philos Trans A Math Phys Eng Sci 2018 Sep 13;376(2128):20170356 [FREE Full text] [doi: 10.1098/rsta.2017.0356] [Medline: 30082302]

8. Reps J, Schuemie M, Suchard M, Ryan P, Rijnbeek P. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc 2018 Aug 01;25(8):969-975 [FREE Full text] [doi: 10.1093/jamia/ocy032] [Medline: 29718407]
9. Ford E, Nicholson A, Koeling R, Tate AR, Carroll J, Axelrod L, et al. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? BMC Med Res Methodol 2013 Aug 21;13(1):105 [FREE Full text] [doi: 10.1186/1471-2288-13-105] [Medline: 23964710]
10. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011 Mar 01;18(2):181-186. [doi: 10.1136/jamia.2010.007237] [Medline: 21233086]
11. Uzuner Ö, Stubbs A, Lenert L. Advancing the state of the art in automatic extraction of adverse drug events from narratives. J Am Med Inform Assoc 2020 Jan 01;27(1):1-2 [FREE Full text] [doi: 10.1093/jamia/ocz206] [Medline: 31841150]
12. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 2011 Oct;18(5):540-543 [FREE Full text] [doi: 10.1136/amiajnl-2011-000465] [Medline: 21846785]
13. Chowdhury GG. Natural language processing. Ann. Rev. Info. Sci. Tech 2005 Jan 31;37(1):51-89. [doi: 10.1002/aris.1440370103]
14. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
15. Stubbs A, Kotfila C, Xu H, Uzuner Ö. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. J Biomed Inform 2015 Dec;58 Suppl:S67-S77 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.001] [Medline: 26210362]
16. Styler WF, Bethard S, Finan S, Palmer M, Pradhan S, de Groen PC, et al. Temporal Annotation in the Clinical Domain. Trans Assoc Comput Linguist 2014 Apr;2:143-154 [FREE Full text] [Medline: 29082229]
17. Hong N, Wen A, Mojarad MR, Sohn S, Liu H, Jiang G. Standardizing Heterogeneous Annotation Corpora Using HL7 FHIR for Facilitating their Reuse and Integration in Clinical NLP. AMIA Annu Symp Proc 2018;2018:574-583 [FREE Full text] [Medline: 30815098]
18. Stewart M, Liu W, Cardell-Oliver R. Redcoat: A Collaborative Annotation Tool for Hierarchical Entity Typing. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. 2019 Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations; November 3-7, 2019; Hong Kong, China p. 193-198. [doi: 10.18653/v1/d19-3033]
19. Nye B, Li J, Patel R, Yang Y, Marshall I, Nenkova A, et al. A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018 Presented at: 56th Annual Meeting of the Association for Computational Linguistics; July 15-20, 2018; Melbourne, Australia. [doi: 10.18653/v1/p18-1019]
20. Campillos L, Deléger L, Grouin C, Hamon T, Ligozat A, Névéol A. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). Lang Resources & Evaluation 2017 Feb 15;52(2):571-601. [doi: 10.1007/s10579-017-9382-y]
21. Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. Pharmacoepidemiol Drug Saf 2012 Jan;21 Suppl 1:23-31. [doi: 10.1002/pds.2336] [Medline: 22262590]
22. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016 Dec;64:333-341 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]
23. Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). J Biomed Inform 2015 Jun;55:290-300 [FREE Full text] [doi: 10.1016/j.jbi.2015.05.003] [Medline: 25979153]
24. Wu H, Toti G, Morley K, Ibrahim Z, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc 2018 May 01;25(5):530-537 [FREE Full text] [doi: 10.1093/jamia/ocx160] [Medline: 29361077]
25. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
26. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc 2018 Mar 01;25(3):331-336 [FREE Full text] [doi: 10.1093/jamia/ocx132] [Medline: 29186491]

27. Liu S, Wang Y, Wen A, Wang L, Hong N, Shen F, et al. Implementation of a Cohort Retrieval System for Clinical Data Repositories Using the Observational Medical Outcomes Partnership Common Data Model: Proof-of-Concept System Validation. JMIR Med Inform 2020 Oct 06;8(10):e17376 [FREE Full text] [doi: 10.2196/17376] [Medline: 33021486]
28. Sharma H, Mao C, Zhang Y, Vatani H, Yao L, Zhong Y, et al. Developing a portable natural language processing based phenotyping system. BMC Med Inform Decis Mak 2019 Apr 04;19(Suppl 3):78 [FREE Full text] [doi: 10.1186/s12911-019-0786-z] [Medline: 30943974]
29. Zheng K, Vydiswaran VGV, Liu Y, Wang Y, Stubbs A, Uzuner, et al. Ease of adoption of clinical natural language processing software: An evaluation of five systems. J Biomed Inform 2015 Dec;58 Suppl:S189-S196 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.008] [Medline: 26210361]
30. Park J. ABMI / SOCRATex. GitHub. URL: https://github.com/ABMI/SOCRATex [accessed 2021-03-23]
31. Hripcsak G, Shang N, Peissig PL, Rasmussen LV, Liu C, Benoit B, et al. Facilitating phenotype transfer using a common data model. J Biomed Inform 2019 Aug;96:103253 [FREE Full text] [doi: 10.1016/j.jbi.2019.103253] [Medline: 31325501]
32. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J Biomed Inform 2012 Aug;45(4):689-696 [FREE Full text] [doi: 10.1016/j.jbi.2012.05.002] [Medline: 22683994]
33. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research 2003;3:993-1022. [doi: 10.5555/944919.944937]
34. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 2018 Nov 28;78(11):15169-15211. [doi: 10.1007/s11042-018-6894-4]
35. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci U S A 2004 Apr 06;101 Suppl 1:5228-5235 [FREE Full text] [doi: 10.1073/pnas.0307752101] [Medline: 14872004]
36. Cao J, Xia T, Li J, Zhang Y, Tang S. A density-based method for adaptive LDA model selection. Neurocomputing 2009 Mar;72(7-9):1775-1781. [doi: 10.1016/j.neucom.2008.06.011]
37. Arun R, Suresh V, Madhavan C, Murthy M. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In: PAKDD 2010: Advances in Knowledge Discovery and Data Mining. 2010 Presented at: Pacific-Asia Conference on Knowledge Discovery and Data Mining; July 21-24, 2010; Hyderabad, India. [doi: 10.1007/978-3-642-13657-3_43]
38. Deveaud R, SanJuan E, Bellot P. Accurate and effective latent concept modeling for ad hoc information retrieval. Document numérique 2014 Apr 30;17(1):61-84. [doi: 10.3166/dn.17.1.61-84]
39. Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei D. Reading tea leaves: how humans interpret topic models. In: Proceedings of the 22nd International Conference on Neural Information Processing Systems. 2009 Presented at: 22nd International Conference on Neural Information Processing Systems; December 2009; Vancouver p. 288-296. [doi: 10.5555/2984093.2984126]
40. Pezoa F, Reutter J, Suarez F, Ugarte M, Vrgoč D. Foundations of JSON Schema. In: Proceedings of the 25th International Conference on World Wide Web. 2016 Presented at: 25th International Conference on World Wide Web; April 2016; Montréal p. 263-273. [doi: 10.1145/2872427.2883029]
41. Kononenko O, Baysal O, Holmes R, Godfrey M. Mining modern repositories with elasticsearch. In: Proceedings of the 11th Working Conference on Mining Software Repositories. 2014 Presented at: 11th Working Conference on Mining Software Repositories; May 2014; Hyderabad, India p. 328-331. [doi: 10.1145/2597073.2597091]
42. Yoon D, Ahn EK, Park MY, Cho SY, Ryan P, Schuemie MJ, et al. Conversion and Data Quality Assessment of Electronic Health Record Data at a Korean Tertiary Teaching Hospital to a Common Data Model for Distributed Network Research. Healthc Inform Res 2016 Jan;22(1):54-58 [FREE Full text] [doi: 10.4258/hir.2016.22.1.54] [Medline: 26893951]
43. Srigley JR, McGowan T, Maclean A, Raby M, Ross J, Kramer S, et al. Standardized synoptic cancer pathology reporting: a population-based approach. J Surg Oncol 2009 Jun 15;99(8):517-524. [doi: 10.1002/jso.21282] [Medline: 19466743]
44. Voigt P, von dem Bussche A. Practical Implementation of the Requirements Under the GDPR. In: The EU General Data Protection Regulation (GDPR). Cham: Springer; 2017:245-249.
45. Centers for Disease ControlPrevention (CDC). HIPAA privacy rule and public health. Guidance from CDC and the U.S. Department of Health and Human Services. MMWR Suppl 2003 May 02;52:1-17, 19 [FREE Full text] [Medline: 12741579]
46. Shin S, Park YR, Shin Y, Choi HJ, Park J, Lyu Y, et al. A De-identification method for bilingual clinical texts of various note types. J Korean Med Sci 2015 Jan;30(1):7-15 [FREE Full text] [doi: 10.3346/jkms.2015.30.1.7] [Medline: 25552878]
47. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5®). Washington, DC: American Psychiatric Association Publishing; 2013.
48. Shin JH, Baek JH, Chung J, Ha EJ, Kim J, Lee YH, Korean Society of Thyroid Radiology (KSThR)Korean Society of Radiology. Ultrasonography Diagnosis and Imaging-Based Management of Thyroid Nodules: Revised Korean Society of Thyroid Radiology Consensus Statement and Recommendations. Korean J Radiol 2016;17(3):370-395 [FREE Full text] [doi: 10.3348/kjr.2016.17.3.370] [Medline: 27134526]
49. Lim S, Yu CS, Jang SJ, Kim TW, Kim JH, Kim JC. Prognostic Significance of Lymphovascular Invasion in Sporadic Colorectal Cancer. Diseases of the Colon & Rectum 2010;53(4):377-384. [doi: 10.1007/dcr.0b013e3181cf8ae5]

50. Lykke J, Roikjaer O, Jess P, Danish Colorectal Cancer Group. The relation between lymph node status and survival in Stage I-III colon cancer: results from a prospective nationwide cohort study. Colorectal Dis 2013 May 25;15(5):559-565. [doi: 10.1111/codi.12059] [Medline: 23061638]
51. Devlin J, Chang M, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/N19-1423]
52. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019). 2019 Presented at: 2019 Conference on Neural Information Processing Systems; December 8-14, 2019; Vancouver, Canada.
53. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020). 2020 Presented at: 2020 Conference on Neural Information Processing Systems; December 6-12, 2020; Virtual.

Abbreviations

CDM: common data model
DSM-5: Diagnostics and Statistical Manual of Mental Disorder
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HIPAA: Health Insurance Portability and Accountability Act
HR: hazard ratio
JSON: JavaScript object notation
K-TIRADS: Korean Thyroid Imaging Reporting and Data System
LDA: latent Dirichlet allocation
NER: named entity recognition
NLP: natural language processing
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Observational Medical Outcomes Partnership
PHI: protected health information
SOCRATex: Staged Optimization of Curation, Regularization, and Annotation of clinical text

Edited by G Eysenbach; submitted 31.08.20; peer-reviewed by Y Chu, Y Yu; comments to author 22.09.20; revised version received 14.11.20; accepted 23.01.21; published 30.03.21.

Please cite as:
Park J, You SC, Jeong E, Weng C, Park D, Roh J, Lee DY, Cheong JY, Choi JW, Kang M, Park RW
A Framework (SOCRATex) for Hierarchical Annotation of Unstructured Electronic Health Records and Integration Into a Standardized Medical Database: Development and Usability Study
JMIR Med Inform 2021;9(3):e23983
URL: https://medinform.jmir.org/2021/3/e23983
doi:10.2196/23983
PMID:33783361

Original Paper

Human–Computer Agreement of Electrocardiogram Interpretation for Patients Referred to and Declined for Primary Percutaneous Coronary Intervention: Retrospective Data Analysis Study

Aleeha Iftikhar[1], MS; Raymond Bond[1], BSc, PhD; Victoria Mcgilligan[2], BSc, PhD; Stephen J Leslie[3], BSc, MBChB, PhD, FRCP; Charles Knoery[3], MBChB; James Shand[4], MBBSMD, FRCP; Adesh Ramsewak[4], MBBS, FRCP; Divyesh Sharma[4], MBBS, MSc, FRCP; Anne McShane[5], MSc; Khaled Rjoob[1], MSc; Aaron Peace[4], MBChB, MRCPE, PhD

[1]Computing Engineering and Build Environment, Ulster University, Belfast, United Kingdom

[2]Centre for Personalised Medicine, Ulster University, Londonderry, United Kingdom

[3]Cardiac Unit, Raigmore Hospital, Inverness, United Kingdom

[4]Department of Cardiology, Altnagelvin Hospital, Western Health and Social Care Trust, Londonderry, United Kingdom

[5]Letterkenny University Hospital, Letterkenny, Ireland

Corresponding Author:
Aleeha Iftikhar, MS
Computing Engineering and Build Environment
Ulster University
Jordanstown
Belfast, BT37 0QB
United Kingdom
Phone: 44 7496635353
Email: Iftikhar-a1@ulster.ac.uk

Abstract

Background: When a patient is suspected of having an acute myocardial infarction, they are accepted or declined for primary percutaneous coronary intervention partly based on clinical assessment of their 12-lead electrocardiogram (ECG) and ST-elevation myocardial infarction criteria.

Objective: We retrospectively determined the agreement rate between human (specialists called activator nurses) and computer interpretations of ECGs of patients who were declined for primary percutaneous coronary intervention.

Methods: Various features of patients who were referred for primary percutaneous coronary intervention were analyzed. Both the human and computer ECG interpretations were simplified to either "suggesting" or "not suggesting" acute myocardial infarction to avoid analysis of complex heterogeneous and synonymous diagnostic terms. Analyses, to measure agreement, and logistic regression, to determine if these ECG interpretations (and other variables such as patient age, chest pain) could predict patient mortality, were carried out.

Results: Of a total of 1464 patients referred to and declined for primary percutaneous coronary intervention, 722 (49.3%) computer diagnoses suggested acute myocardial infarction, whereas 634 (43.3%) of the human interpretations suggested acute myocardial infarction (P<.001). The human and computer agreed that there was a possible acute myocardial infarction for 342 out of 1464 (23.3%) patients. However, there was a higher rate of human–computer agreement for patients not having acute myocardial infarctions (450/1464, 30.7%). The overall agreement rate was 54.1% (792/1464). Cohen κ showed poor agreement (κ=0.08, P=.001). Only the age (odds ratio [OR] 1.07, 95% CI 1.05-1.09) and chest pain (OR 0.59, 95% CI 0.39-0.89) independent variables were statistically significant (P=.008) in predicting mortality after 30 days and 1 year. The odds for mortality within 1 year of referral were lower in patients with chest pain compared to those patients without chest pain. A referral being out of hours was a trending variable (OR 1.41, 95% CI 0.95-2.11, P=.09) for predicting the odds of 1-year mortality.

Conclusions: Mortality in patients who were declined for primary percutaneous coronary intervention was higher than the reported mortality for ST-elevation myocardial infarction patients at 1 year. Agreement between computerized and human ECG interpretation is poor, perhaps leading to a high rate of inappropriate referrals. Work is needed to improve computer and human decision making when reading ECGs to ensure that patients are referred to the correct treatment facility for time-critical therapy.

(JMIR Med Inform 2021;9(3):e24188) doi:10.2196/24188

KEYWORDS

ECG interpretation; agreement between human and computer; primary percutaneous coronary intervention service; acute myocardial infarction; scan; electrocardiogram; heart; intervention; infarction; human-computer; diagnostic

Introduction

Background

According to the British Heart Foundation, circulatory diseases cause more than one-quarter (27%) of all deaths in the United Kingdom [1]. In the United Kingdom, more than 100,000 hospital admissions each year are due to heart attacks (280 admissions per day) [1]. Acute coronary syndrome occurs due to a restriction in blood flow in the coronary arteries [2]. Acute coronary syndromes are subdivided into (1) ST-elevation myocardial infarctions, (2) non–ST-elevation myocardial infarctions, and (3) unstable angina [3]. ST-elevation myocardial infarction is generally more serious when there is total occlusion of a coronary blood vessel leading to extensive damage to a large area of the heart [4]. Once a blocked artery is suspected, a patient is typically referred for reperfusion therapy which can include a primary percutaneous coronary intervention [5]. The preferred treatment for an acute myocardial infarction with ST-segment elevation is angioplasty (primary percutaneous coronary intervention) given that this is an effective therapy for opening occluded arteries [6-8]. The admission criteria for primary percutaneous coronary intervention are often variable and partly based on electrocardiogram (ECG) interpretation and patient symptoms, hence not all referrals are accepted. Even if ST-elevation myocardial infarction is present, ECG interpretation can be difficult because of different factors, including misleading computerized interpretations, signal noise, poor confidence or competency in reading ECGs, human error, and indeed, borderline ECGs (not precisely normal, but not significantly abnormal either), that make it difficult for clinicians to make a binary decision. A strict criterion may result in patients with acutely occluded coronary arteries not getting the treatment in time. It has been reported that several patients not meeting ST-elevation myocardial infarction criteria who were nevertheless referred for primary percutaneous coronary intervention did indeed require angioplasty [9].

ECG interpretation is central to deciding whether patients should be declined or accepted for primary percutaneous coronary intervention. The ECG is the most widely used diagnostic tool for patients with suspected acute myocardial infarction [10,11]. Many prehospital protocols require the acquisition of a single 12-lead ECG when assessing a patient for a ST-elevation myocardial infarction or ischemia. However, if necessary, a second or third prehospital ECG is recorded to correctly identify a ST-elevation myocardial infarction due to the number of ECGs (15% in [5]) that are nonspecific, ambiguous, and perhaps borderline [5]. When arriving at an emergency, paramedics are often first to record and interpret the ECG. Different studies [12,13] have been conducted to compare ECG interpretation accuracy between paramedics and physicians. Mencl et al [12] found no correlation between training, experience, or confidence in the ability of paramedics to recognize ST-elevation myocardial infarctions. The paramedics in the study were only able to identify inferior ST-elevation myocardial infarctions and normal ECGs; paramedics' ECG interpretations cannot be solely relied on (low sensitivity and specificity) for activating the catheterization laboratory (CathLab), in which diagnostic imaging equipment used to visualize the arteries and the chambers of the heart and to treat any stenosis or abnormality, in a primary percutaneous coronary intervention service [12].

Identification of patients with acute myocardial infarction continues to be challenging, especially when automated ECG interpretation is inconclusive or misleading. However, a study [13] has shown that, when the ECG exhibits vagueness, clinician input (using the internet) can improve diagnostic performance and reduce time to treatment. It is well documented that misinterpretation of the ECG can lead to incorrect decision making regarding treatment, such as false activations (rates of up to 36% [14]) or patients being declined. According to Degheim et al [15], 12.5% of all CathLab activations were false activations for misinterpreted ST-elevation myocardial infarction. These false activations have both clinical and financial costs.

Prior Work

Given the challenges of reading ECGs, computer interpretation has been used for many years to assist human interpreters. In a retrospective cross-sectional study [16] of 200 prehospital ECGs, computer interpretation for detecting ST-elevation myocardial infarction achieved a specificity of 100% (100/100; 95% CI 0.96-1.00) and a sensitivity of 58% (58/100; 95% CI 0.48-0.67). This illustrates that this computer algorithm would have incorrectly declined 42% of patients but had zero inappropriate activations [16]; the most common incorrect computer statements for false negatives were “data quality prohibits interpretation” and “abnormal ECG unconfirmed.” Another study [17] concluded that computer-interpretation failed to identify a number of patients with ST-elevation myocardial infarction. This shows that prehospital computerized ECG interpretation is suboptimal for ST-elevation myocardial infarction detection and should not be used as a single method for prehospital activation of the CathLab. Cardiologists are the most accurate diagnosticians and are the least likely to falsely activate the CathLab [18]. Nevertheless, other physicians, paramedics, and specialized nurses (activator nurse) are expected to competently read ECGs.

Study Goals

Having summarized the research to date, we have identified that ECG interpretation is challenging for both humans and computers, and there is a need to better understand the characteristics of the patients who are declined for primary percutaneous coronary intervention, especially given that there are a number of likely false negatives (patients who are declined but needed an emergency intervention).

We aimed to analyze agreement between computer and human (activator nurses) ECG interpretations for patients who were referred to but declined for primary percutaneous coronary intervention.

Methods

Data Set

This study involved an analysis of an anonymized data set from Altnagelvin Hospital (Northern Ireland, United Kingdom) of consecutive patients who were declined for primary percutaneous coronary intervention from January 2015 to December 2017. The total study population consisted of 1464 patients who were referred but declined for a primary percutaneous coronary intervention.

Data Collection

When paramedics suspect acute myocardial infarction based on ECG findings, they contact the primary percutaneous coronary intervention department at the hospital and describe the symptoms and ECG findings to an activator nurse. The activator nurse routinely records this referral using a paper-based form, which is then digitized to a spreadsheet. Therefore, the data contained some inconsistencies and missing values.

Data Analysis

All statistical analyses were performed using R (version 3.5.2, RStudio). The time-series visualization of interpretations was generated using an R package for visual analytics (ggplot2; version 3.3.2). Data were interrogated for missing values and completeness. There were no missing values in the most important data columns (ie, computer ECG interpretation, activator nurse ECG interpretation); however, to overcome data inconsistencies, the required fields were manually cleaned. There were typographical issues such as the inconsistent use of mixed upper and lower case, spelling mistakes, use of shorthand, and abbreviations used in the computer and human ECG interpretation columns. Comparisons between the distinct groups were investigated for significance using chi-square tests for categorical dichotomous variables. One-tailed Student *t* or Mann-Whitney tests were used for continuous variables depending upon whether the variables were normally distributed. Logistic multivariate regression analysis was performed on independent variables such as gender, age, out of hours, chest pain, activator nurse interpretation, computer interpretation, and computer and activator nurse agreement where the response variables included 30-day and 1-year mortality (encoded as 1 or 0, where 1=mortality). We also investigated mutual agreement and disagreement over the 24-hour day. To analyze the agreement between the computer and activator nurse, all interpretations were simplified and re-encoded as either suggesting or not suggesting acute myocardial infarction. To achieve this binary encoding of ECG interpretations, 3 medical doctors (AP, SL, and CK—2 of whom were clinical lead and consultant cardiologists) reviewed the original interpretations. The 3 medical doctors independently reclassified these statements as either suggesting acute myocardial infarction or not suggesting acute myocardial infarction, then they met as a team to arrive at consensus when there were discrepancies.

Ethical Aspects

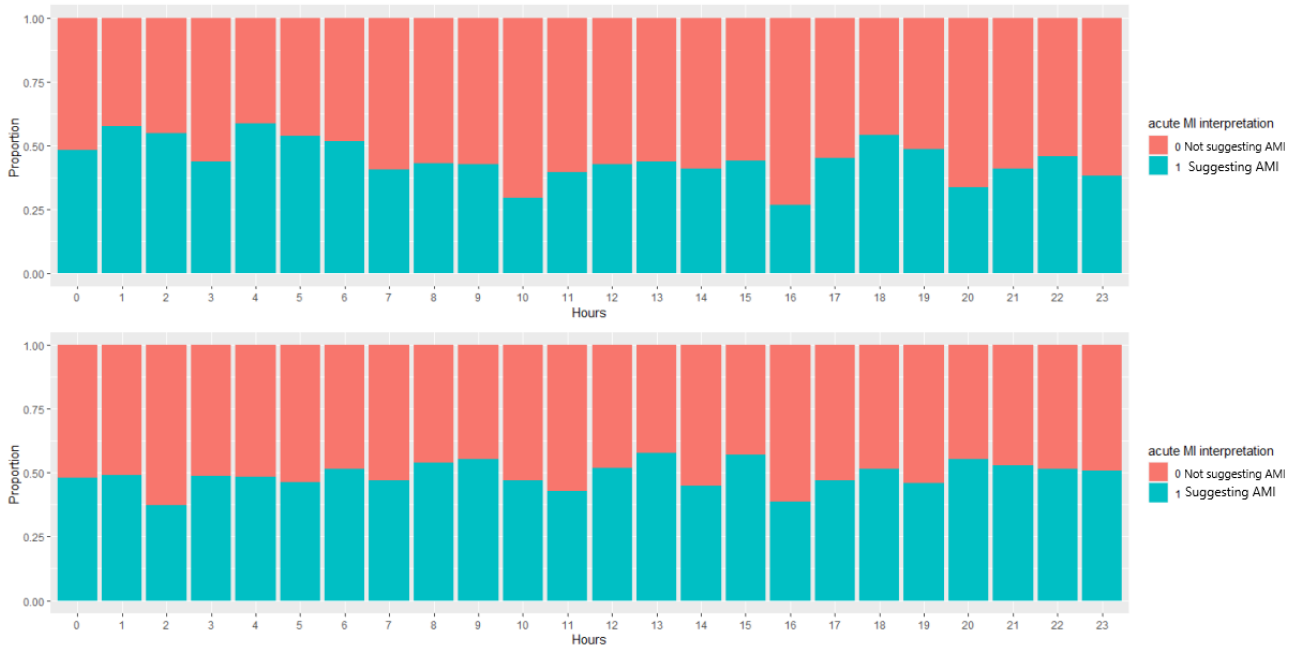
Permission for the study was obtained from the Regional Ethical Review Board (IRAS 251710) of the National Health Service Office for Research Ethics Committees Northern Ireland. The study complied with the Declaration of International Research Integrity Association. After the study received ethical approval for secondary data analysis, the staff nurse removed all personal identifiable information such as names, date of birth, and unique patient identifiers.

Results

Activator Nurse and Computer ECG Interpretations

The computer suggested acute myocardial infarction more often than the activator nurses (722/1464, 49.3% vs 634/1464, 43.3%; *P*=.001). Figure 1 depicts the acute myocardial infarction interpretation rate per hour for both the activator nurses and the computer. The highest relative rate of acute myocardial infarction interpretation by activator nurses occurred at 1 AM (26/45, 57.8%) and 4 AM (17/29, 58.6%). The activator nurses seemed to interpret more acute myocardial infarctions during the middle of the night (12 AM to 6 AM) with a mean of 53% (SD 5.3%) compared to during daytime hours (mean 41%, SD 6.6%; *P*=.001). In contrast, computer interpretation did not show much variation with respect to hours of the day; for the middle of the night (12 AM to 6 AM), the average acute myocardial infarction interpretation rate was a mean of 47% (SD 4.7%) compared with a mean 50% (SD 5.2%) for the daytime hours. There was slightly more variation in activator nurse interpretations than in those of the computer over the hours of the day.

Figure 1. (a) Activator nurse and (b) computer interpretations of acute myocardial infarction rate by the hour. AMI: acute myocardial infarction; MI: myocardial infarction.

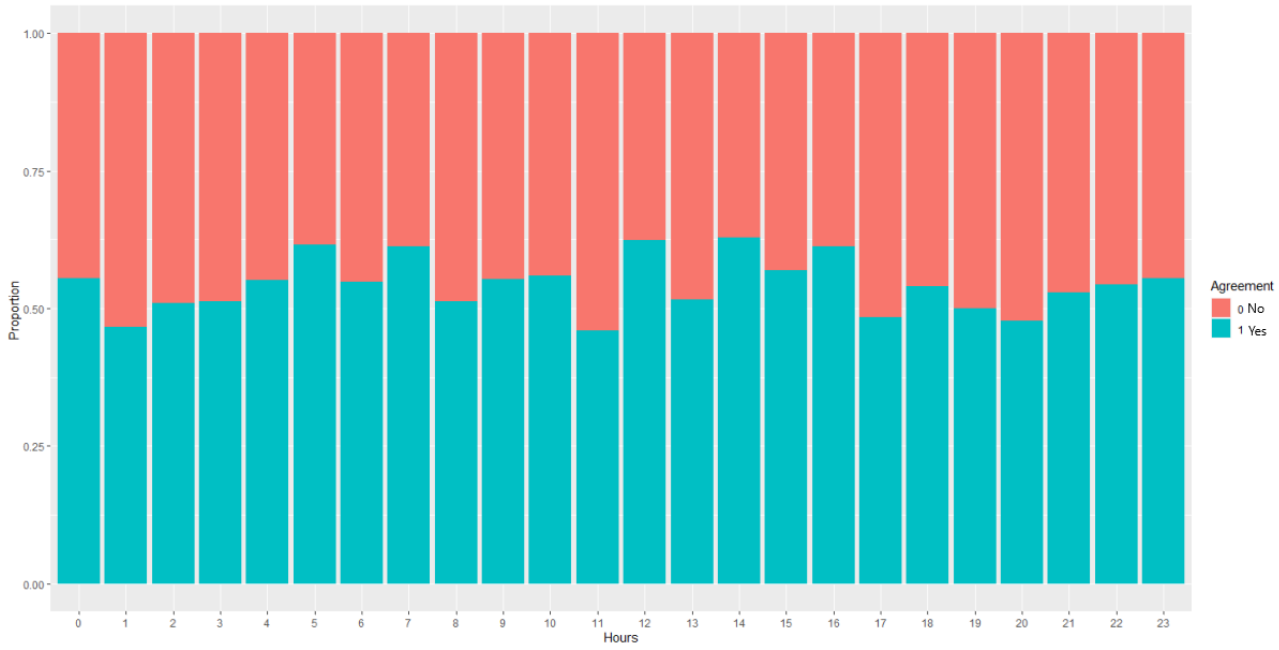


Activator Nurse and Computer Overall Agreement

The human and computer ECG interpretations agreed for 54.1% of patients (792/1464; *P*<.001). This statistic includes suggesting and not suggesting acute myocardial infarction (Figure 2). The human–computer agreement rates were analyzed per hour; Figure 3 shows that the maximum agreement occurred at 12 PM and 2 PM during the daytime. Whereas in the middle of the night, the peak agreement occurred at 5 AM and 7 AM. Figure 2b shows that there was more variation in activator nurse and computer agreement not suggesting acute myocardial infarction than in those suggesting acute myocardial infarction (mean 57%, SD 7.5% vs mean 43% SD 4.7%; *P*<.001). There was more uncertainty *out of hours* when compared to *in hours*. Activator nurses suggested more acute myocardial infarctions during the middle of the night than in the daytime.

Figure 2. Activator nurse and computer agreement of (a) acute myocardial infarction and (b) not acute myocardial infarction. AMI: acute myocardial infarction; MI: myocardial infarction.

Figure 3. Activator nurse and computer agreement by the hour.



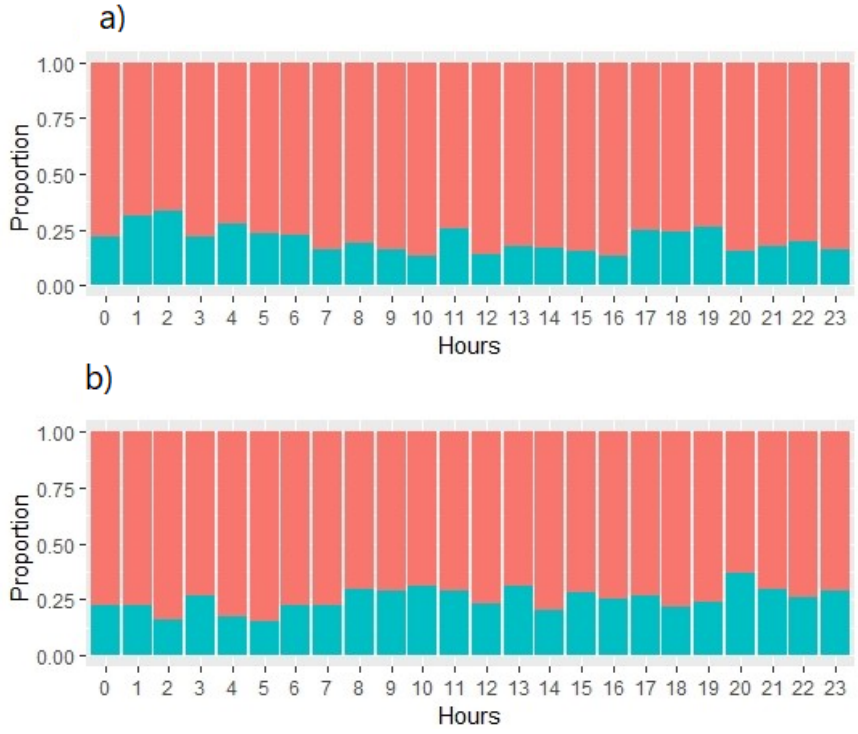
Activator Nurse and Computer Overall Disagreement

The analysis of disagreement between human and computer interpretations was performed by first analyzing instances where activator nurses suggested acute myocardial infarction and the computer did not, and then vice versa. Maximum disagreement occurred at 11 AM.

Activator Nurse Suggested Acute Myocardial Infarction

The number of patients for whom the activator nurse suggested acute myocardial infarction but the computer did not were selected and displayed per hour. The total number of such instances was 292/1464 (19.9%). Activator nurse interpretations suggested acute myocardial infarctions and the computer interpretation disagreed for more patients during the middle of the night (between 1 AM and 2 AM; Figure 4a).

Figure 4. (a) Activator nurse interpretation suggesting acute myocardial infarction and computer disagreed; (b) computer interpretation suggesting acute myocardial infarction and activator nurse disagreed.



Computer Suggested Acute Myocardial Infarction

Computer interpretation suggested acute myocardial infarction and the corresponding activator nurses' interpretation disagreed for 26.0% of patients (380/1464). The maximum disagreement occurred in the evening at 8 PM (P<.001; Figure 4b).

Analysis of Other Variables

Patients With Chest Pain

More males (1002/1464, 68.4%) were referred to primary percutaneous coronary intervention than females. More than half (769/1464, 52.5%) of the patients had either chest pain (n=556) or resolved chest pain (n=213). Most of these patients were male (385/556, 69.2%). More patients reported chest pain during the middle of the night (4 AM to 5 AM: 34/55, 61.8%; P=.02; Figure 5).

Logistic regression analysis was performed on independent variables including gender, age, out of hours, chest pain, activator nurse interpretation, computer interpretation, and computer–activator nurse agreement with the response variables being 30-day (Table 1) and 1-year mortality (Table 2). Age and chest pain were the only independent variables that were statistically significant (P<.001) for predicting mortality after 30 days or 1 year. Another trending variable was out of hours which increased the chance of mortality within 1 year (odds ratio [OR] 1.41, 95% CI 0.95-2.11). Being referred out of hours was more predictive for 1-year mortality than 30-day mortality. Being older (OR 1.07, 95% CI 1.05-1.09) increased the probability of 30-day and 1-year mortality. Activator nurse and computer agreement of acute myocardial infarction and having chest pain reduced the odds of mortality after 1 year. The odds of mortality within 30 days and 1 year of referral were lower in patients with chest pain compared to those patients without chest pain.

Figure 5. Proportion of patients with chest pain by the hour.

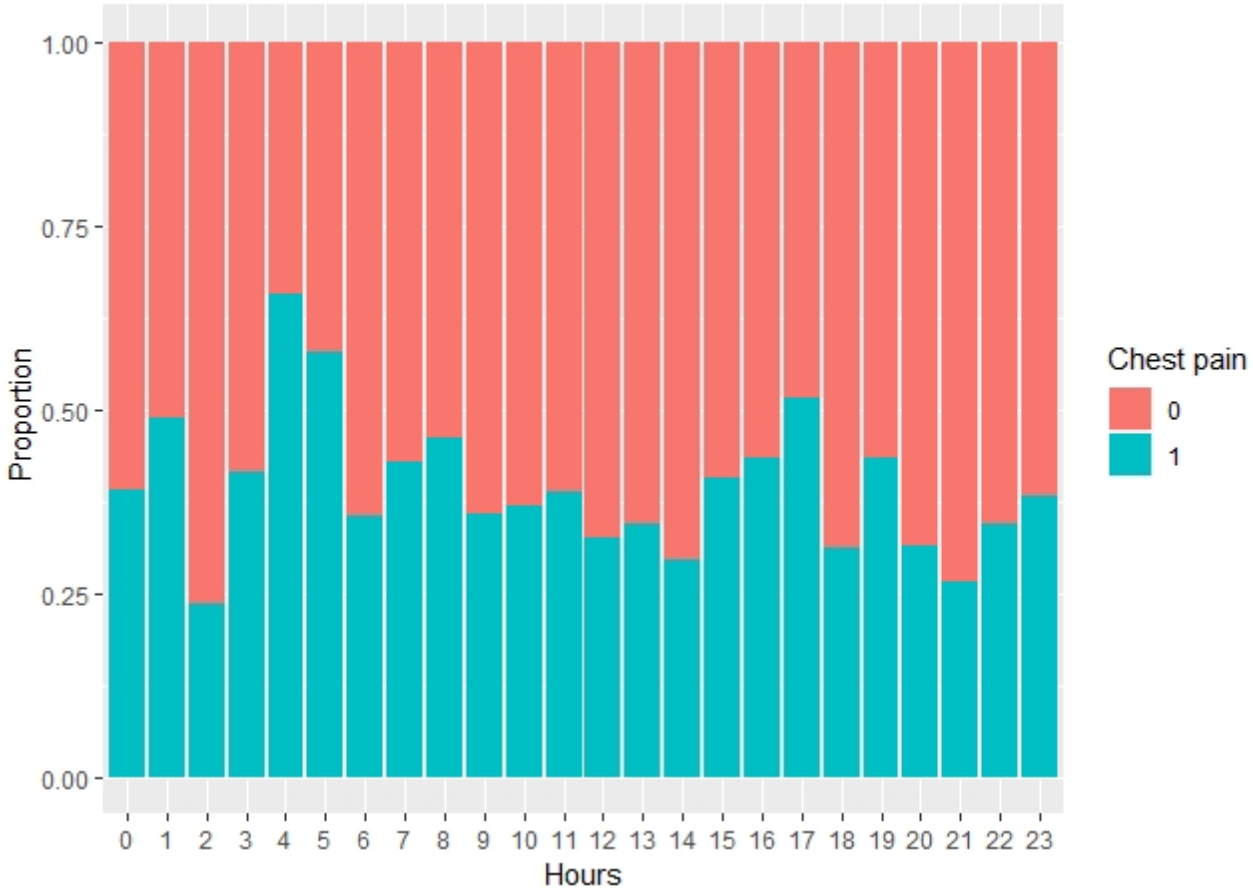


Table 1. Odds ratios of variables derived from multiple logistic regression where the response variable was mortality after 1 year.

Variable	Odds ratio (95% CI)	SE	*P* value
Out of hours (true/false)	1.41 (0.95-2.11)	0.012	.09
Age	1.07 (1.05-1.09)	0.434	<.001
Chest pain (true)[a]	0.59 (0.39-0.89)	0.012	.008
Activator nurse diagnosis suggesting acute myocardial infarction (true)	1.26 (0.73-2.16)	0.012	.39
Computer diagnosis suggesting acute myocardial infarction (true)	1.30 (0.78-2.17)	0.013	.31
Activator nurse–computer acute myocardial infarction agreement (true)	0.97 (0.47-2.03)	0.011	.95

[a]42 patients with chest pain died after 1 year, whereas 130 patients without chest pain died after 1 year.

Table 2. Odds ratios of variables derived from multiple logistic regression where the response variable was mortality after 30 days.

Variable	Odds ratio (95% CI)	SE	*P* value
Out of hours (true/false)	1.39 (0.90-2.20)	0.012	.17
Age	1.06 (1.04-1.08)	0.434	<.001
Chest pain (true)[a]	0.47 (0.29-0.74)	0.012	.001
Activator nurse diagnosis suggesting acute myocardial infarction (true)	1.06 (0.59-1.87)	0.012	.84
Computer diagnosis suggesting acute myocardial infarction (true/false)	0.86 (0.49-1.57)	0.013	.68
Activator nurse–computer acute myocardial infarction agreement (true)	1.46 (0.65-3.31)	0.011	.35

[a]25 patients with chest pain died after 30 days, whereas 92 patients without chest pain died after 30 days.

Acute Myocardial Infarction Terminology

Table 3 shows the most frequently used terms by the computer and activator nurses for ECG interpretation to suggest acute myocardial infarction or not suggest acute myocardial infarction. The computer used the term abnormal ECG most frequently, which we classified as not suggesting acute myocardial infarction, whereas activator nurses used the term high take-off for interpreting the ECG, which we classified as not suggesting acute myocardial infarction. Moreover, the computer used the term acute myocardial infarction most frequently for suggesting acute myocardial infarction, and activator nurses used the terms ST depression or ST-elevation for suggesting acute myocardial infarction. Overall, the activator nurses used 45 unique terms to interpret the ECG as not suggestive of acute myocardial infarction and used 19 different terms in suggesting acute myocardial infarction. In contrast, the computer used 59 different terms to interpret the ECG as not suggestive of acute myocardial infarction and 60 unique terms in suggesting acute myocardial infarction.

Table 3. Frequently used terms by computer and activator nurses for suggesting or not suggesting acute myocardial infarction.

Classification and rank[a]	Computer		Activator nurse	
	Interpretation term	Patients, n (%)	Interpretation term	Patients, n (%)
Suggests acute myocardial infarction[b]				
1	“acute myocardial infarction”	337 (47)	“Ste”	159 (25)
2	“inferior infarct”	23 (3)	“St depression”	125 (20)
3	“anterior injury”	34 (5)	“twi”	129 (20)
Does not suggest acute-myocardial infarction[c]				
1	“abnormal ECG”	382 (51)	“nil acute”	377 (45)
2	“LBBB”	108 (15)	“high take-off”	187 (23)
3	“borderline ECG”	41 (5.5)	“RBBB”	59 (7)

[a]Terms with low frequencies (1 or 2) are not included.

[b]n=722 for Computer; n=634 for Activator nurse.

[c]n=742 for Computer; n=830 for Activator nurse.

Interpretation Terminology

Activator nurses were more consistent in their nomenclature in suggesting acute myocardial infarction. In contrast to the activator nurse, the computer used a greater range of nomenclature in suggesting acute myocardial infarction (Table 3). The terms with low frequencies (1 or 2 instances) are not included.

Discussion

Principal Findings

The level of agreement between human and computer ECG interpretation for acute myocardial infarction regarding patients who were declined for primary percutaneous coronary intervention is an interesting research area for clinicians. It unveils useful insights. In this study, we analyzed an anonymized data set from Altnagelvin Hospital (Northern Ireland, United Kingdom) of patients who were declined for primary percutaneous coronary intervention from January 2015 to December 2017. The total study population consisted of 1464 patients who were declined for a primary percutaneous coronary intervention (996/1464, 68.0% men). The decision was appropriate for all patients; none of the patients who were declined for primary percutaneous coronary intervention experienced an acute ST-elevation myocardial infarction. More declined patients were referred out of hours 66.3% (971/1464). Out of all 1464 declined patients, 117 (8.0%) patients died within 30 days, and a total of 174 (11.8%) patients died within 1 year. Furthermore, the 1-year mortality rate was highest if the patient was referred at 4 AM (7/12, 58.3%). This is not surprising as patients who are less sick are less likely to present in the middle of the night.

Human and computer ECG interpretations did not have a high level of agreement, and the computer tended to suggest acute myocardial infarction more often than the specialist activator nurses, especially for the declined patients. A total of 722/1464 (49.3%) computerized diagnoses suggested acute myocardial infarction, whereas only 634/1464 (43.3%) activator nurse diagnoses suggested acute myocardial infarction (P=.001). However, the activator nurse interpreted that ECGs suggested acute myocardial infarction more often during the middle of the night (12 AM to 6 AM: mean 53%, SD 5.3%) than in daytime hours (mean 41%, SD 6.6%; P=.001). In contrast, the computer interpretation did not show much difference for hours of the day; for the middle of the night (12 AM to 6 AM), the average acute myocardial infarction ECG interpretation was 47% (SD 4.7%), and for the rest of the hours of the day, the average acute myocardial infarction ECG interpretation was 50% (SD 5.2%). We speculate that there may be human bias at night—the activator nurses tend to overidentify acute myocardial infarction during the night possibly because they are forced to make a decision when there are fewer consultants or clinicians available for a second opinion.

Prior research stated that major problems in computer interpretation were the interpretation of rhythm disturbances and the diagnosis of acute myocardial infarction, T-wave changes, and ventricular hypertrophy [19]. Researchers also found that there was a considerable difference in accuracy between 3 different computer systems [19].

There were only 342/1464 (23.3%) patients for whom there was human and computer agreement that there was an acute myocardial infarction. There was agreement more often for not being acute myocardial infarction (450/1464, 30.7%; P<.001). The overall agreement rate was only 54.1% (792/1464). The maximum agreement between activator nurses and the computer occurred from 2 PM to 4 PM (139/231, 60.2%). There were 292/1464 (19.9%) patients for whom the computer did not suggest an acute myocardial infarction but the activator nurse did, and 380/1464 (26.0%) patients for whom the activator nurse identified an acute myocardial infarction but the computer did not. The peak disagreement rate between activator nurse and computer occurred at 11 AM (53/98, 54.1%). The result shows that the computer interpreted ECGs as suggesting acute myocardial infarction more often than activator nurses. Activator nurse–computer agreement was poor (Cohen κ=0.08, P=.001). Activator nurses seemed to use fewer terms, whereas the computer used almost 60 different terms suggesting acute myocardial infarction. Previous studies [20] show that there is significant interobserver variability that results in false positives and false negatives. There is a higher rate of discordance among clinically significant ECGs [21].

Additionally, 556 out of 1464 (38.0%) patients who were declined had chest pain. More patients reported chest pain during the middle of the night, between 4 AM and 5 AM (34/55, 61.8%; P=.001). This could be because underlying medical conditions and obstructive sleep apnea can be a trigger for myocardial infarction [22]. For logistic regression analysis, both age and chest pain were the only independent variables that were statistically significant in predicting mortality after 30 days (P<.001 and P=.001, respectively) and 1 year (P<.001 and P=.008, respectively). Another trending variable was *out of hours*, which increased the odds of 1-year mortality. Being referred out of hours was more predictive for 1-year mortality than 30-day mortality. This could be because not all referral resources were available out of hours. The odds of mortality within 30 days and 1 year of referral were lower in patients with chest pain than in those patients without chest pain. This might be because people with chest pain call for help sooner and receive the appropriate treatment. People without chest pain are more likely to be misdiagnosed.

Limitations

This was a retrospective analysis. The results are based on a single data set from one hospital in Northern Ireland, which can limit the results; the results may not be generalizable for the overall population and primary percutaneous coronary intervention services.

Policy and Practical Implications

Algorithms to detect acute myocardial infarction need to be improved. More ECG data are needed for training ECG interpretation algorithms. Perhaps deep learning and neural networks can be used with the ECG interpretation algorithms for more accurate results. In addition, enhanced training and education can provide nurses and activator nurses with support

for enhanced ECG interpretation capabilities. ECG interpretation in a primary percutaneous coronary intervention service should be more sophisticated and rely upon more than ST-elevation myocardial infarction criteria. Algorithms could be trained to read ECGs using ECG data sets that have a better ground truth for a fully occluded artery. This label could be based on immediate angiographic findings from ST-elevation myocardial infarction and non–ST-elevation myocardial infarction patients.

Conclusion

The agreement between computerized and human ECG interpretation was poor for patients who were declined for primary percutaneous coronary intervention. This uncertainty makes it difficult to accept or decline referrals. The results show that the computer suggests acute myocardial infarction more often than activator nurses for declined patients. Work is needed to improve computer and human decision making to ensure that patients are referred to the correct treatment facility for time-critical therapy. In future, there might be comparison among the computer human agreement between male and female patients and various age groups. We believe that this might be an interesting research question.

Clinical Perspectives

The 12-lead ECG remains the mainstay in assessing patients with suspected coronary artery occlusion. However, despite improvements in the quality of data acquisition and computer-generated reports, the accuracy of using ECG to diagnose occluded coronary arteries remains suboptimal. There remains a need for improved computer-generated interpretation, which may need to consider patient factors such as sex, age, risk factors, and ongoing symptoms. Including these factors could improve diagnostic accuracy and help triage patients to the best possible treatment. What is unknown is whether this would lead to better clinical outcomes in terms of reduced infarction size and better survival in patients having a myocardial infraction. This study described the interaction and ECG interpretation agreement rate between humans and computers and how they might have an impact on outcomes.

Acknowledgments

The authors disclose receipt of the following financial support for the research, authorship, and/or publication of this article: This research is supported by the European Union's Interreg VA Programme, managed by the Special European Union Programmes Body. The views and opinions expressed in this paper do not necessarily reflect those of the European Commission or the Special European Union Programmes Body.

Conflicts of Interest

None declared.

References

1. Heart statistics. British Heart Foundation. URL: https://www.bhf.org.uk/what-we-do/our-research/heart-statistics [accessed 2020-05-01]
2. Amsterdam EA, Wenger NK, Brindis RG, Casey DE, Ganiats TG, Holmes DR, ACC/AHA Task Force Members. 2014 AHA/ACC guideline for the management of patients with non-ST-elevation acute coronary syndromes: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 2014 Dec 23;130(25):e344-e426. [doi: 10.1161/CIR.0000000000000134] [Medline: 25249585]
3. Torres M, Moayedi S. Evaluation of the acutely dyspneic elderly patient. Clin Geriatr Med 2007 May;23(2):307-25, vi. [doi: 10.1016/j.cger.2007.01.007] [Medline: 17462519]
4. Diagnosis-heart attack. UK National Health Service. URL: https://www.nhs.uk/conditions/heart-attack/diagnosis/ [accessed 2020-04-02]
5. Verbeek PR, Ryan D, Turner L, Craig AM. Serial prehospital 12-lead electrocardiograms increase identification of ST-segment elevation myocardial infarction. Prehosp Emerg Care 2012 Jan 05;16(1):109-114. [doi: 10.3109/10903127.2011.614045] [Medline: 21954895]
6. Widimský P. Long distance transport for primary angioplasty vs immediate thrombolysis in acute myocardial infarction Final results of the randomized national multicentre trial—PRAGUE-2. Eur Heart J 2003 Jan;24(1):94-104. [doi: 10.1016/s0195-668x(02)00468-2]
7. Silber S, Albertsson P, Avilés FF, Camici PG, Colombo A, Hamm C, Task Force for Percutaneous Coronary Interventions of the European Society of Cardiology. Guidelines for Percutaneous Coronary Interventions: The Task Force for Percutaneous Coronary Interventions of the European Society of Cardiology. Eur Heart J 2005 Apr;26(8):804-847. [doi: 10.1093/eurheartj/ehi138] [Medline: 15769784]
8. Keeley EC, Boura JA, Grines CL. Primary angioplasty versus intravenous thrombolytic therapy for acute myocardial infarction: a quantitative review of 23 randomised trials. The Lancet 2003 Jan;361(9351):13-20. [doi: 10.1016/s0140-6736(03)12113-7]
9. Apps A, Malhotra A, Tarkin J, Smith R, Kabir T, Lane R, et al. High incidence of acute coronary occlusion in patients without protocol positive ST segment elevation referred to an open access primary angioplasty programme. Postgrad Med J 2013 Jul 30;89(1053):376-381. [doi: 10.1136/postgradmedj-2012-130818] [Medline: 23542430]

10. Meek S, Morris F. ABC of clinical electrocardiography. BMJ 2002 Feb 16;324(7334):415-418 [FREE Full text] [doi: 10.1136/bmj.324.7334.415] [Medline: 11850377]
11. Miranda DF, Lobo AS, Walsh B, Sandoval Y, Smith SW. New insights into the use of the 12-lead electrocardiogram for diagnosing acute myocardial infarction in the emergency department. Can J Cardiol 2018 Feb;34(2):132-145. [doi: 10.1016/j.cjca.2017.11.011] [Medline: 29407007]
12. Mencl F, Wilber S, Frey J, Zalewski J, Maiers JF, Bhalla MC. Paramedic ability to recognize ST-segment elevation myocardial infarction on prehospital electrocardiograms. Prehosp Emerg Care 2013 Feb 12;17(2):203-210. [doi: 10.3109/10903127.2012.755585] [Medline: 23402376]
13. Anroedh SS, Kardys I, Akkerhuis KM, Biekart M, van der Hulst B, Deddens GJ, et al. e-Transmission of ECGs for expert consultation results in improved triage and treatment of patients with acute ischaemic chest pain by ambulance paramedics. Neth Heart J 2018 Nov;26(11):562-571 [FREE Full text] [doi: 10.1007/s12471-018-1187-0] [Medline: 30357611]
14. McCabe JM, Armstrong EJ, Kulkarni A, Hoffmayer KS, Bhave PD, Garg S, et al. Prevalence and factors associated with false-positive ST-segment elevation myocardial infarction diagnoses at primary percutaneous coronary intervention–capable centers: a report from the Activate-SF registry. Arch Intern Med 2012 Jun 11;172(11):864-871. [doi: 10.1001/archinternmed.2012.945] [Medline: 22566489]
15. Degheim G, Berry A, Zughaib M. False activation of the cardiac catheterization laboratory: the price to pay for shorter treatment delay. JRSM Cardiovasc Dis 2019 Apr 08;8:2048004019836365 [FREE Full text] [doi: 10.1177/2048004019836365] [Medline: 31007905]
16. Bhalla MC, Mencl F, Gist MA, Wilber S, Zalewski J. Prehospital electrocardiographic computer identification of ST-segment elevation myocardial infarction. Prehosp Emerg Care 2013;17(2):211-216. [doi: 10.3109/10903127.2012.722176] [Medline: 23066910]
17. Mawri S, Michaels A, Gibbs J, Shah S, Rao S, Kugelmass A, et al. The comparison of physician to computer interpreted electrocardiograms on ST-elevation myocardial infarction door-to-balloon times. Crit Pathw Cardiol 2016 Mar;15(1):22-25. [doi: 10.1097/HPC.0000000000000067] [Medline: 26881816]
18. Huitema AA, Zhu T, Alemayehu M, Lavi S. Diagnostic accuracy of ST-segment elevation myocardial infarction by various healthcare providers. Int J Cardiol 2014 Dec 20;177(3):825-829. [doi: 10.1016/j.ijcard.2014.11.032] [Medline: 25465827]
19. Saner H, Lindeland A, Scallen R, Paule W, Lange HW, Gobel FL. [Comparison of 3 ECG computer programs with interpretation by physicians]. Schweiz Med Wochenschr 1987 Jul 07;117(27-28):1035-1039. [Medline: 3303319]
20. Brosnan M, La Gerche A, Kumar S, Lo W, Kalman J, Prior D. Modest agreement in ECG interpretation limits the application of ECG screening in young athletes. Heart Rhythm 2015 Jan;12(1):130-136. [doi: 10.1016/j.hrthm.2014.09.060] [Medline: 25285648]
21. Wathen JE, Rewers AB, Yetman AT, Schaffer MS. Accuracy of ECG interpretation in the pediatric emergency department. Ann Emerg Med 2005 Dec;46(6):507-511. [doi: 10.1016/j.annemergmed.2005.03.013] [Medline: 16308065]
22. Kuniyoshi FHS, Garcia-Touchard A, Gami A, Romero-Corral A, van der Walt C, Pusalavidyasagar S, et al. Day-night variation of acute myocardial infarction in obstructive sleep apnea. J Am Coll Cardiol 2008 Jul 29;52(5):343-346 [FREE Full text] [doi: 10.1016/j.jacc.2008.04.027] [Medline: 18652941]

Abbreviations

CathLab: catheterization laboratory
ECG: electrocardiogram
OR: odds ratio

Edited by G Eysenbach; submitted 08.09.20; peer-reviewed by AUR Bacha; comments to author 08.10.20; revised version received 13.10.20; accepted 17.01.21; published 02.03.21.

Please cite as:
Iftikhar A, Bond R, Mcgilligan V, Leslie SJ, Knoery C, Shand J, Ramsewak A, Sharma D, McShane A, Rjoob K, Peace A
Human–Computer Agreement of Electrocardiogram Interpretation for Patients Referred to and Declined for Primary Percutaneous Coronary Intervention: Retrospective Data Analysis Study
JMIR Med Inform 2021;9(3):e24188
URL: https://medinform.jmir.org/2021/3/e24188
doi:10.2196/24188
PMID:33650984

Original Paper

A Personal Health System for Self-Management of Congestive Heart Failure (HeartMan): Development, Technical Evaluation, and Proof-of-Concept Randomized Controlled Trial

Mitja Luštrek[1], PhD; Marko Bohanec[2], PhD; Carlos Cavero Barca[3], BSc; Maria Costanza Ciancarelli[4], MSc; Els Clays[5], PhD; Amos Adeyemo Dawodu[4], MD; Jan Derboven[6], PhD; Delphine De Smedt[5], PhD; Erik Dovgan[1], PhD; Jure Lampe[7], BSc; Flavia Marino[8], PhD; Miha Mlakar[1], PhD; Giovanni Pioggia[8], PhD; Paolo Emilio Puddu[4], PhD, MD; Juan Mario Rodríguez[3], BSc; Michele Schiariti[4], PhD, MD; Gašper Slapničar[1], MSc; Karin Slegers[9], PhD; Gennaro Tartarisco[8], PhD; Jakob Valič[1], MTh; Aljoša Vodopija[1], MSc

[1]Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia

[2]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

[3]Health Unit, Atos Research and Innovation (ARI), Atos Spain S.A., Madrid, Spain

[4]Department of Cardiovascular, Respiratory, Nephrological, Anesthesiological and Geriatric Sciences, Sapienza University of Rome, Rome, Italy

[5]Department of Public Health and Primary Care, Ghent University, Ghent, Belgium

[6]Meaningful Interactions Lab, KU Leuven, Leuven, Belgium

[7]SenLab d.o.o., Ljubljana, Slovenia

[8]Institute for Biomedical Research and Innovation, National Research Council of Italy, Messina, Italy

[9]Department of Communication & Cognition, Tilburg School of Humanities and Digital Sciences, Tilburg University, Tilburg, Netherlands

Corresponding Author:
Mitja Luštrek, PhD
Department of Intelligent Systems
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, 1000
Slovenia
Phone: 386 1 477 3380
Email: mitja.lustrek@ijs.si

Abstract

Background: Congestive heart failure (CHF) is a disease that requires complex management involving multiple medications, exercise, and lifestyle changes. It mainly affects older patients with depression and anxiety, who commonly find management difficult. Existing mobile apps supporting the self-management of CHF have limited features and are inadequately validated.

Objective: The HeartMan project aims to develop a personal health system that would comprehensively address CHF self-management by using sensing devices and artificial intelligence methods. This paper presents the design of the system and reports on the accuracy of its patient-monitoring methods, overall effectiveness, and patient perceptions.

Methods: A mobile app was developed as the core of the HeartMan system, and the app was connected to a custom wristband and cloud services. The system features machine learning methods for patient monitoring: continuous blood pressure (BP) estimation, physical activity monitoring, and psychological profile recognition. These methods feed a decision support system that provides recommendations on physical health and psychological support. The system was designed using a human-centered methodology involving the patients throughout development. It was evaluated in a proof-of-concept trial with 56 patients.

Results: Fairly high accuracy of the patient-monitoring methods was observed. The mean absolute error of BP estimation was 9.0 mm Hg for systolic BP and 7.0 mm Hg for diastolic BP. The accuracy of psychological profile detection was 88.6%. The F-measure for physical activity recognition was 71%. The proof-of-concept clinical trial in 56 patients showed that the HeartMan system significantly improved self-care behavior (P=.02), whereas depression and anxiety rates were significantly reduced (P<.001), as were perceived sexual problems (P=.01). According to the Unified Theory of Acceptance and Use of Technology questionnaire, a positive attitude toward HeartMan was seen among end users, resulting in increased awareness, self-monitoring, and empowerment.

Conclusions: The HeartMan project combined a range of advanced technologies with human-centered design to develop a complex system that was shown to help patients with CHF. More psychological than physical benefits were observed.

Trial Registration: ClinicalTrials.gov NCT03497871; https://clinicaltrials.gov/ct2/history/NCT03497871.

International Registered Report Identifier (IRRID): RR2-10.1186/s12872-018-0921-2

(JMIR Med Inform 2021;9(3):e24501) doi:10.2196/24501

KEYWORDS

congestive heart failure; personal health system; mobile application; mobile phone; wearable electronic devices; decision support techniques; psychological support; human centered design

Introduction

Background and Motivation

Congestive heart failure (CHF) is a disease in which the heart cannot pump enough blood to supply oxygen and nutrients to the body. The main symptoms are shortness of breath (dyspnea), diminished ability to exercise, fatigue, and swelling in the feet and legs (edema). The lifetime risk of developing CHF ranges from 20% to 33%, and only approximately half of patients survive for more than 5 years after diagnosis [1]. As CHF is frequently the end stage of various conditions that affect left ventricular function and cannot be cured, the focus of the treatment is to prevent deterioration, manage symptoms, and maintain a good quality of life [2].

The management of CHF includes multiple medications, appropriate exercise, diet (paying particular attention to fluids and salt), management of body weight, and abstaining from alcohol and smoking. As the average age at CHF diagnosis is 74 (SD 14) years [3], 25% to 80% of the patients are affected by cognitive impairment [4], a third of them have depression or anxiety [5], and other comorbidities are also common, they often find it difficult to manage the disease on their own [6]. Cardiac rehabilitation programs are either not available or poorly attended—participation in Europe is approximately 20% [7]. Therefore, the relevant alternatives are technological solutions to support the management of CHF.

Approximately 64 million people live with CHF globally [1], and the economic burden of their disease amounts to more than 100 billion US $ annually [8]. This is a strong incentive to improve CHF management. In addition to medications, implantable devices (mainly pacemakers and defibrillators) are already established treatment options [9]. Another option is telemonitoring, but its benefits in CHF are uncertain [9]. Another option is mobile health (mHealth) solutions, whose benefits in CHF are poorly explored (see the Related Work section) but have a strong backing of the market: the mHealth market in 2019 was US $46 billion and grew by 22% annually [10] (compared with the telemonitoring market of US $2 billion with 13% growth [11] and the more mature implantable devices market of US $23 billion and 8% growth [12]).

In the HeartMan project, we developed a comprehensive personal health system for the self-management of physical and psychological aspects of CHF. The first step was to analyze evidence-based medical requirements and—following the human-centered design process—to elicit requirements related to everyday management of CHF from the patients themselves. We then developed a mobile app comprising a decision support system (DSS) and several intelligent data analysis modules. A web application for medical professionals has also been developed. Finally, the system was evaluated in a proof-of-concept trial that assessed both its effectiveness and patient perception.

Related Work

In 2018, a systematic review was devoted to mobile apps supporting the self-management of CHF [13]. The authors surveyed 10 leading paper repositories for papers on interventions that used a mobile platform, evaluated them with a randomized controlled trial or a similar design, and provided usability or efficacy results. Papers on telecare and structured telephone support were excluded. In total, 18 papers meeting the inclusion and exclusion criteria were included in the review. The authors also searched Google Play and Apple App Store for health care apps by including “heart failure” as a keyword. After excluding apps that track only blood pressure (BP) and/or heart rate, a total of 26 apps were downloaded and evaluated with respect to the quality of self-management components included in the apps and quality of the user experience provided by the apps.

According to the authors of the review [13], most apps are poorly designed and do not include all the necessary components for the self-management of CHF. Indeed, only 2 apps—Heart Failure Storylines [14] and HeartMapp [15]—include exercise interventions, which is one of the most important aspects of CHF management. The Heart Failure Storylines app is perhaps the most complete one that can be currently found in the market. It provides medication reminders, a symptom tracker, keeps a record of vital signs, and tracks physical activity and daily moods. Nevertheless, the interventions provided by the app are poorly personalized (except for medication reminders) because the app does not consider the patients’ psychophysical state, making the usefulness of such interventions questionable [9,16]. The HeartMapp app provides personalized interventions, but it is quite basic and is not adapted to the patients’ psychophysical state. The app was tested in a randomized controlled trial with only 18 participants (intervention group, n=9) [17].

We searched the Google Play and Apple App Store for apps that were not included in the review. We found 6 apps that were published after the review. Five of these apps include only educational materials [18-23], whereas 1 app provides only guidance on medication therapy [24]. In short, no new apps provide a comprehensive solution for CHF management.

Methods

Collection of Requirements and Human-Centered Design

Medical Requirements

The first step in designing the HeartMan system was to study the state-of-the-art medical knowledge on CHF self-management. A systematic review of the available literature was performed to identify parameters that predict the hard outcomes of mortality and hospitalization in patients with CHF as well as variables that affect the patient-reported outcome of quality of life in this patient group [25]. We further selected those parameters that are modifiable by self-care behaviors that the HeartMan system can recommend. These modifiable parameters are primarily clinical parameters (eg, body mass index, BP, heart rate), physical capacity, medication use, characteristics of CHF (eg, fluid retention), and mental health (eg, depression, anxiety). We then screened relevant medical guidelines for CHF, focusing on nonpharmacological recommendations and lifestyle advice, to identify the best approaches for modifying these parameters [26] and incorporated these into the HeartMan DSS. We designed an exercise training and nutrition program (including diet and fluid intake restrictions) to influence physical capacity, clinical parameters, and fluid retention. Medication adherence is expected to be enhanced through DSS, providing reminders, disease education, and self-monitoring. Finally, cognitive behavioral therapy and mindfulness exercises were included to improve mental health and self-management. Management guidelines for comorbidities were also taken into account, as many patients with CHF have conditions such as diabetes, atrial fibrillation, and chronic obstructive pulmonary disease.

An additional source for developing the medical requirements was data from the Chiron project [27], a previous telemonitoring study in patients with CHF focusing on short-term outcomes of subjective well-being on a daily basis. Data mining analysis suggested environmental parameters, that is, ambient conditions such as temperature and humidity, to play a role in predicting day-to-day changes in perceived health. This was incorporated into an additional module of DSS.

User Requirements

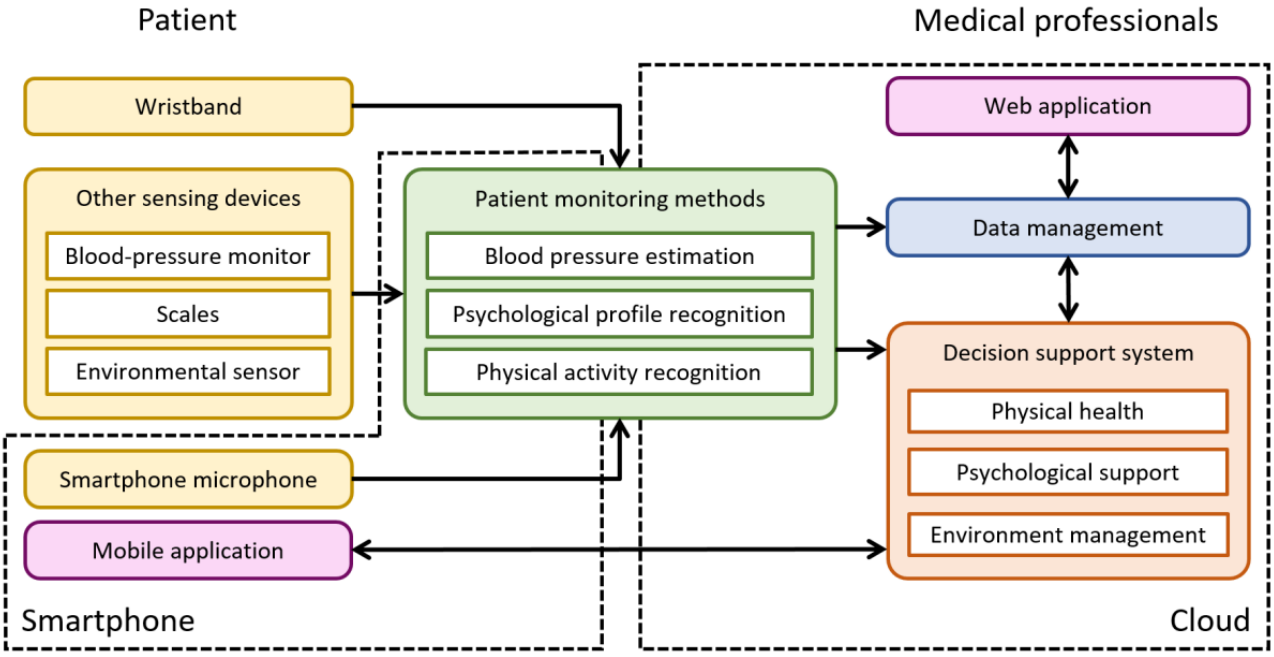
As our goal was not only to provide medically relevant advice but also to design the HeartMan system to be useful and well accepted by the patients, we adopted a human-centered design [28]. This approach involves users throughout the design process, focusing on their perspective and needs. In our case, it consisted of a thorough analysis of patients' context of use, which took place in three stages in Belgium and Italy. The first stage was a diary study, in which patients kept a diary for a period of 10-14 days (n=19 in Belgium; n=18 in Italy). The diary contained questions and assignments related to everyday activities and habits, such as patients' experience, disease management, and their social network. The second stage was a follow-up interview study conducted with most patients who participated in the diary study (n=14 in Belgium; n=15 in Italy). In this interview study, patients participated in semistructured interviews in which the output of the diary study was discussed in detail. This analysis resulted in a rich, qualitative description of patient characteristics as well as the patient experience regarding disease management, the challenges related to therapy adherence, lifestyle changes as a result of being a CHF patient, and relationships with caregivers. These insights were translated into concrete user requirements for the HeartMan system, which served, together with the medical requirements, as the starting point for the third stage: the design and evaluation of a series of prototypes with both patients and caregivers. In this process, several design trade-offs were made regarding patient autonomy, technology appropriation, and patient well-being [29]. The main patient characteristics that were found to impact the design of the HeartMan system were the patient's digital literacy, perception of empowerment, and existing therapy adherence habits.

For medical professionals, a web portal was developed, allowing them to follow up on the patients' data gathered by the HeartMan system. This prototype was developed and evaluated using a separate human-centered design process. In this process, various stakeholders (including cardiologists, nurses, dieticians, psychologists, and physiotherapists) offered insights into the needs and requirements related to the follow-up of patients with CHF based on the HeartMan monitoring data.

System Overview

In the HeartMan system designed as described in the previous section, sensing devices collect information about the patient, patient monitoring methods further interpret some of this information, and a DSS recommends actions based on the (interpreted) information. The recommendations are presented to the patient via a mobile app, and medical professionals have access to the system via a web application. A diagram presenting an overview of the system is presented in Figure 1.

Figure 1. The logical architecture of the HeartMan system.



The sensing devices (yellow in Figure 1) are custom sensing wristbands, off-the-shelf BP monitors, weight scales, and environmental sensors that measure temperature and humidity. According to the medical requirements, heart rate (obtained from the photoplethysmogram [PPG] signal), BP, weight, and ambient temperature and humidity are important determinants of the health and well-being of patients with CHF. As it would be relevant to monitor BP more frequently than once per day (which can be expected with a regular BP monitor), we developed a method to estimate BP continuously from the PPG signal (green). Owing to the importance of psychological support for patients with CHF, we also developed a method to recognize their psychological profile from the heart rate, heart-rate variability, and voice recorded with the smartphone. Finally, the accelerometer in the wristband is used to recognize the patient's physical activities, which allows the initiation of psychological interventions at the appropriate moment. As the accelerometer provides the greatest volume of data of all the sensors, this last method is implemented on the smartphone, whereas the previous 2 reside in the cloud.

All patient information is fed into the DSS and stored in the cloud (blue in Figure 1). The DSS has three components, the first of which is an expert system that helps patients manage their physical health (exercise, nutrition, medications, and self-monitoring). The second is another expert system that provides psychological support (elements of cognitive behavioral therapy and mindfulness). The third uses predictive models (based on the previously mentioned Chiron data) to recommend actions related to temperature and humidity that are expected to improve patients' well-being. The first 2 components rely on expert knowledge because it is well established how the aspects of the CHF management they address should be tackled. The last one relies on data and predictive modeling because we had relevant data available, but there is little expert knowledge on the effect of the environment on the well-being of patients with CHF.

The recommendations provided by the DSS are shown in the mobile app (purple in Figure 1), which also collects inputs from the patients. Medical professionals can use a web application to view information collected from sensing devices as well as the patients' adherence to recommendations. Although the content of recommendations is mostly based on the medical requirements, the way information is presented via the 2 applications was heavily influenced by the users' inputs obtained during the human-centered design process.

Patient-Monitoring Methods

The HeartMan Wristband

The wristband used by the system includes a PPG sensor, which provides information on the heart rate and beat-to-beat intervals in addition to the raw data, tri-axial accelerometer, and temperature sensor. It communicates with the HeartMan app via Bluetooth Low Energy 4.1. Its battery life is sufficient for a full day of operation, while continuously streaming sensor data to the phone. It features a liquid crystal display and vibration motor, which can be used to deliver urgent notifications to the user, such as about too high or low heart rate during exercise.

BP Estimation

Continuous BP estimation is well researched when 2 signals, typically ECG and PPG, are available, as the pulse transit time between 2 points on the body is highly correlated with the BP [30,31]. In HeartMan, we aimed to use a single wristband PPG sensor [32,33], as this is the most convenient for the patients. However, such a sensor typically has a modest sampling frequency, the sensor-to-skin contact is often compromised due to movement, and the wrist area exhibits less pulsatility compared with a fingertip, making this approach challenging.

To obtain high-quality parts of the PPG waveform, the signal was preprocessed. The first step was zero-mean unit-variance normalization. Outlier samples above 3 SDs from the local

median (10-sample window) were removed using a Hampel filter. Afterward, the signal was filtered using a fourth-order Butterworth band-pass (0.5-4.0 Hz) filter. Then, a transformation based on the first-order derivative was used to detect systolic peaks and diastolic valleys in between. Once the valleys were detected, the signal was traversed with a sliding window, and a template was created as the average of all cycles in a window. Following this, each individual cycle was compared with the template using several metrics. This allowed for the detection of segments where the signal was stable with only a few artifacts, while also allowing for individual bad cycles in an otherwise good segment to be discarded [34].

After preprocessing, per-cycle temporal features describing the cycle shape were computed based on related work [35] and further expanded with some features from the frequency domain. The latter were computed from a window centered on a cycle and extending 5 seconds before the cycle start point and 5 seconds after the end point. Most of the temporal morphologic features rely on high-quality waveform, exhibiting a clear systolic and diastolic peak, as they were designed for fingertip PPG devices in a controlled setting. The HeartMan wristband signal is generally of lower quality, so we focused on frequency domain features, which are more robust, as they are computed from longer windows and not on a per-cycle basis. In addition, as some morphological features were infeasible to compute from the HeartMan wristband data, we additionally leveraged information from the accelerometer, which tells us about the person's physical activity. We considered some commonly used features computed from the three-axis accelerometer, which are known to work well in separating a person's activities [36]. We decided on this because having information about a person's activity might prove useful for BP estimation, as the cardiovascular response of the body changes during intense physical activity compared with the state. This fact differentiates this work from previous work dealing with similar problems, as related work often focuses on PPG signals without considering the person's activity, which can be reflected in the accelerometer signal [37]. Finally, heart rate was also used as a feature to inform us about a person's cardiac activity. All these features were fed into regression models that estimated systolic BP (SBP) and diastolic BP (DBP). Several algorithms implemented in the Scikit-learn toolbox [38] were used to train the models, some of which are compared in the Results section.

Psychological Profile Recognition

The development of technological interventions for behavior changes as well as growing interest in affective computing have resulted in various attempts to recognize psychological states from sensor data. Some authors [39] used mobile phones to analyze user voices and classify their emotions (happy, sad, fear, anger, and neutral). Others have focused on stress, dementia, and cognitive dysfunctions, relying more on wearable devices that sense the heart rate, electrodermal activity, skin temperature, and acceleration [40,41].

The HeartMan system combines the patient's voice obtained during a structured weekly phone interview with an informal caregiver with heart rate features, which can be obtained from the HeartMan wristband. The speech data were preprocessed to normalize the different acoustic properties, such as higher volume and background noise, using standard techniques [42]. The features extracted from the speech are the fundamental frequency (pitch), mel-frequency cepstral coefficients, and the smoothed energy. The mean, SD, range, maximum, and minimum were computed for each base speech feature. In addition, the heart rate and heart rate variability represented by the root mean square of successive differences between heartbeats were extracted. The features are then fed into a machine learning model that recognizes motivated, anxious, and depressed psychological profiles. All the data were preprocessed and analyzed using MATLAB and R software.

Physical Activity Recognition

Physical activity recognition is a relatively mature field, although the requirements of HeartMan present some challenges. As the purpose was to initiate psychological interventions, it was most relevant to recognize eating and to distinguish resting from walking and more intense activities. Eating recognition is quite difficult and rarely addressed in the literature, whereas wrist—being able to move independently from the body—is not the best location for recognizing the intensity of activity.

Similar to the previous 2 patient-monitoring methods, this method also uses machine learning. The stream of acceleration data is first low-pass filtered to remove noise and then band-pass filtered to remove the gravitational component, retaining the component due to dynamic human motion. The stream was then segmented into 2-second windows. In each window, the low-pass filtered data are used to compute features related to the orientation of the sensor, whereas the band-pass filtered data are used to compute the features related to the motion of the sensor. A total of 90 features were extracted [37]. Some describe the intensity and shape of the acceleration signal, such as the mean, variance, skewness, and kurtosis. Others have a physics-based interpretation, such as changes in velocity and kinetic energy. The rest are based on expert knowledge, such as the number of peaks in the signal and the number of times the signal crosses its mean value. The features are fed into a machine learning model that returns one of the following activities: rest, standing, walking, Nordic walking, running, other exercise, eating, washing hands or face, household chores (whole-body movement), and light hand activities (hand movement). The model was built using the random forest algorithm implemented in the Weka toolkit [43].

DSS

Expert System for Physical Health Management

Exercise

The HeartMan DSS administers a comprehensive exercise program [44] according to the established medical guidelines [16]. Before starting the exercise program, the patients were expected to perform a cardiopulmonary exercise (cycloergometry) or a 6-min walking test to assess their physical capacity. On this basis, the physical capacity of each patient is assessed as *low* or *normal*, which affects the exercise planning.

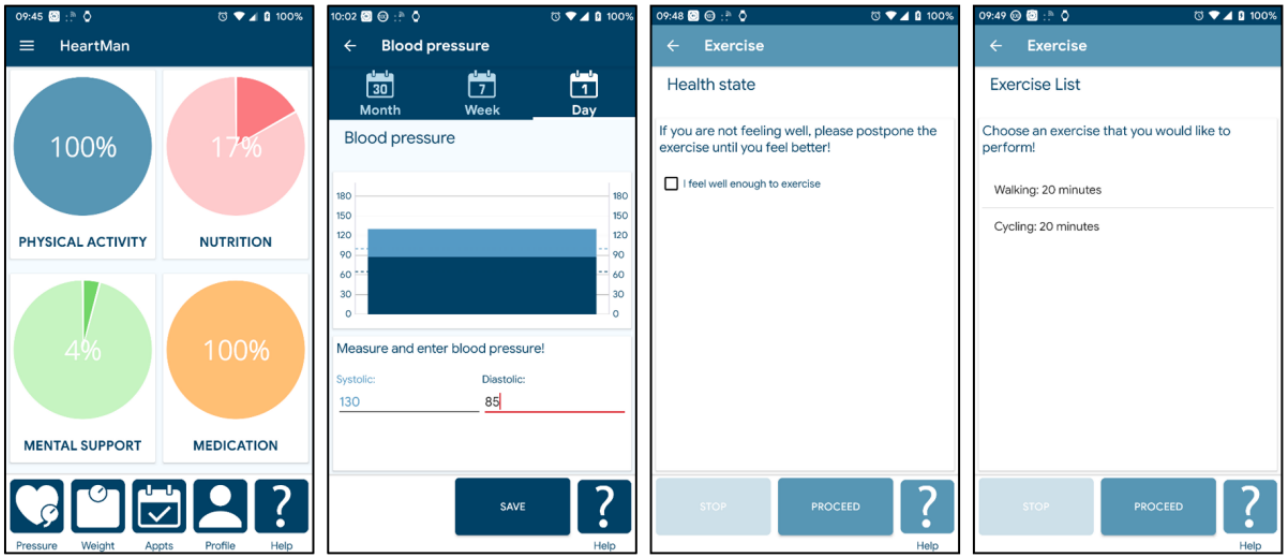
Weekly Exercise Planning

The DSS proposes a weekly exercise plan for each patient, consisting of endurance and resistance exercises. The DSS suggests the frequency (times per week), intensity, and duration of each exercise type. The suggestions are based on the patient's physical capacity, the number of active weeks in the program, and the current frequency and intensity. They are also based on the patient's psychological profile: the difficulty increases more gradually for depressed patients, which is in line with the *shaping* technique suitable for this profile. For instance, low-capacity patients start with very light 10- to 15-min endurance exercises twice per week. According to the patient's progress, these parameters may change with time, typically by increasing the frequency and intensity of exercises, if the patient agrees. The planning process is governed by an expert system that consists of 2 rule-based models, developed using a qualitative multicriteria method decision expert [45] and described in more detail in our earlier work [44].

Exercise Sessions

Before the start of each exercise session, the HeartMan DSS checks whether the patient's BP and heart rate are in a safe range and whether the patient feels well enough to exercise. If the exercise is allowed, a list of exercises is shown to the patient, who can then select the preferred exercise. This is illustrated in Figure 2. Typical endurance exercises involve walking and cycling, whereas resistance exercises aim to strengthen the patient's arms, legs, and body. After selecting the exercise, a detailed description (text or graphical) was provided. During the exercise, the heart rate and SBP were continuously measured using the wristband. Patients are advised to stop the exercise in cases of symptoms or measurements outside a safe range. During endurance exercises, the DSS uses the wristband display to suggest an increase or decrease in pace based on the heart rate. After completing the exercise, the patients can rate their feeling of intensity, which is used in the weekly planning to decide whether to increase the intensity.

Figure 2. Exercise-related screens of the HeartMan app: the main screen, blood pressure input before the exercise, health check before the exercise, and exercise list.



Nutrition

To provide appropriate nutrition advice, the DSS requires the following medical information: the patient's BMI, whether the patient has diabetes, and the prescribed amount of liquid intake. Next, the DSS creates a personalized questionnaire to be answered by the patient; it includes general questions about healthy nutrition and specific questions about the patient's eating and drinking behavior. On this basis, the DSS assesses the level to which topics (about breakfast, lunch, dinner, fat and cholesterol, fluid intake, salt, diabetes, and medication) are understood by the patient. Finally, the patient received feedback in terms of positive reinforcement messages (for well-understood topics), educational statements (for misunderstood general topics), and advice on how to modify the diet to make it healthier (for misunderstood eating behavior topics).

Self-Monitoring and Medication

Patients with CHF are required to measure their BP, heart rate, and daily weight. The HeartMan system reminds them of this and warns if the measurements are outside the safe ranges. It also reminds the patients to take their medications and helps them fill the weekly pillbox (if they use one). It periodically asks the patient about the number of pills remaining in the pillbox and assesses medication adherence based on the deviation from the expected number.

Expert System for Psychological Support

In most cases, CHF diagnosis requires substantial changes in daily life and habits, such as dietary modifications and increased physical activity. Combined with psychological distress, which also often follows the diagnosis, patients can face an intrusion of distorted beliefs and negative automated thoughts that cause them to feel unable to pursue a goal [46]. Sometimes a vicious circle called cognitive dissonance is triggered: a conflict between their desire to be healthy on one hand and practicing unhealthy behaviors for short-term comfort on the other hand. In the long run, this results in poor adherence to self-management guidelines as well as psychological discomfort [47].

The psychological DSS is designed to select the appropriate strategy to improve patients' psychological well-being and adherence to physical exercise and dietary guidelines. The strategy is adapted to the user's psychological profile, as discussed in the section on psychological profile recognition. The DSS provides cognitive behavioral interventions and mindfulness exercises that are modified according to a weekly plan. These exercises are suggested daily, at a time when the user engaged in a physical activity expected to make them receptive to the suggestion. The relevant activities are eating, walking, and sitting, as discussed in the Physical Activity Recognition section.

Cognitive Behavioral Therapy

This is a combination of behavioral and cognitive techniques developed to reduce anxiety and depressive symptoms, which tend to make patients less motivated, tired, and less energetic. The DSS provides specially designed messages intended to align the patients' actions with their desires, as shown in the examples in Table 1. These messages are formulated according to the principles by Festinger [48] of *cognitive consequences of forced compliance* for the motivated profile, *free choice* for the anxious profile, and *effort justification* for the depressed profile.

Table 1. Examples of cognitive behavioral therapy messages about physical exercises for three different psychological profiles.

Psychological profile	Festinger principle	Example message
Motivated profile	Cognitive consequences of forced compliance	I should perform physical exercise to obtain benefits similar to those from medications
Anxious profile	Free choice	Walking for 10 min and watching TV[a] are two ways to relax. Walking improves your heart health, whereas TV does not
Depressed profile	Effort justification	Walking for 10 min will bring benefits similar to those obtained from medication

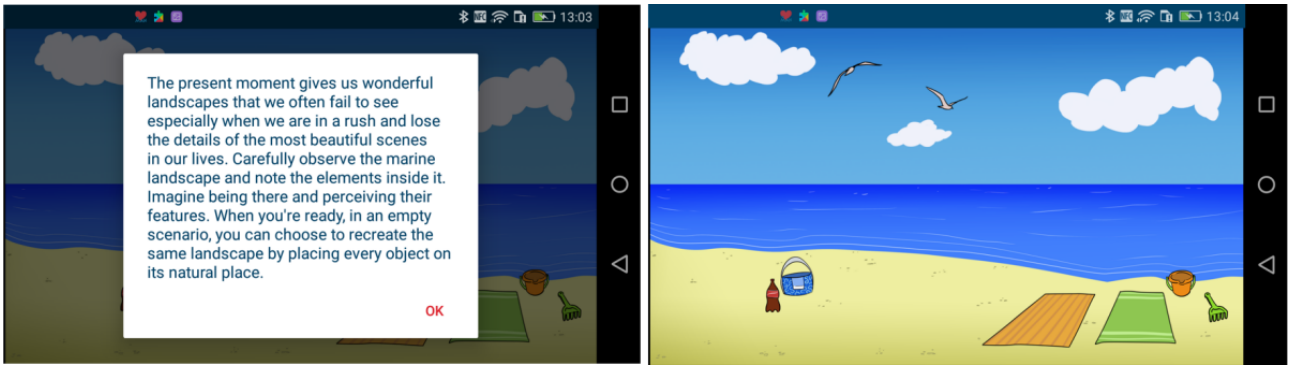
[a]TV: television.

Mindfulness

Mindfulness exercises enhance patients' awareness of their present condition and help them disassociate (unhealthy) emotional and behavioral responses from physical sensations and thoughts. Mindfulness exercises consisted of the following:

- Games to deal with intrusive thoughts (eg, loss of independence, feeling restricted in daily activities), as shown in Figure 3.
- Audio recordings dealing with the perception of the patient's body and breathing exercises.
- Mindful messages that help the patient focus on a mindful moment. These messages are contextualized as follows: mindful walking when the user is walking, mindful breathing when the user is sitting, mindful eating when the user is eating, and mindful listening and observing when the user is either walking or sitting.

Figure 3. Mindful game "World Sense".



Predictive Models for Environment Management

Unlike the DSS approaches used for physical health management and mental support, which mainly rely on expert knowledge, a data-based approach was developed for environment management. We used data from the Chiron project [27], which consists of features describing the patient's situation and their self-reported feeling of health. The features are physiological (eg, heart rate, BP) and environmental (eg, temperature, humidity) and very similar to those available to the HeartMan system.

In the first step, we built a machine learning model that could predict the feeling of health from the features. We used the random forest algorithm implemented in the Weka toolkit [43]. The accuracy of distinguishing between good and bad feelings of health was 83.2%. We also divided the features into modifiable, correlated (with modifiable), and uncorrelated. We build linear regression models that can predict each correlated feature from the modifiable ones.

In the second step, we set up a multi-objective optimization problem, where we searched for minimal modifications of modifiable features that change the feeling of health from bad to good. For each solution, the correlated features were predicted using linear models, and the admissibility of the solution was checked using the feeling-of-health model. The objectives were the sum of the volumes of modifications needed and the number

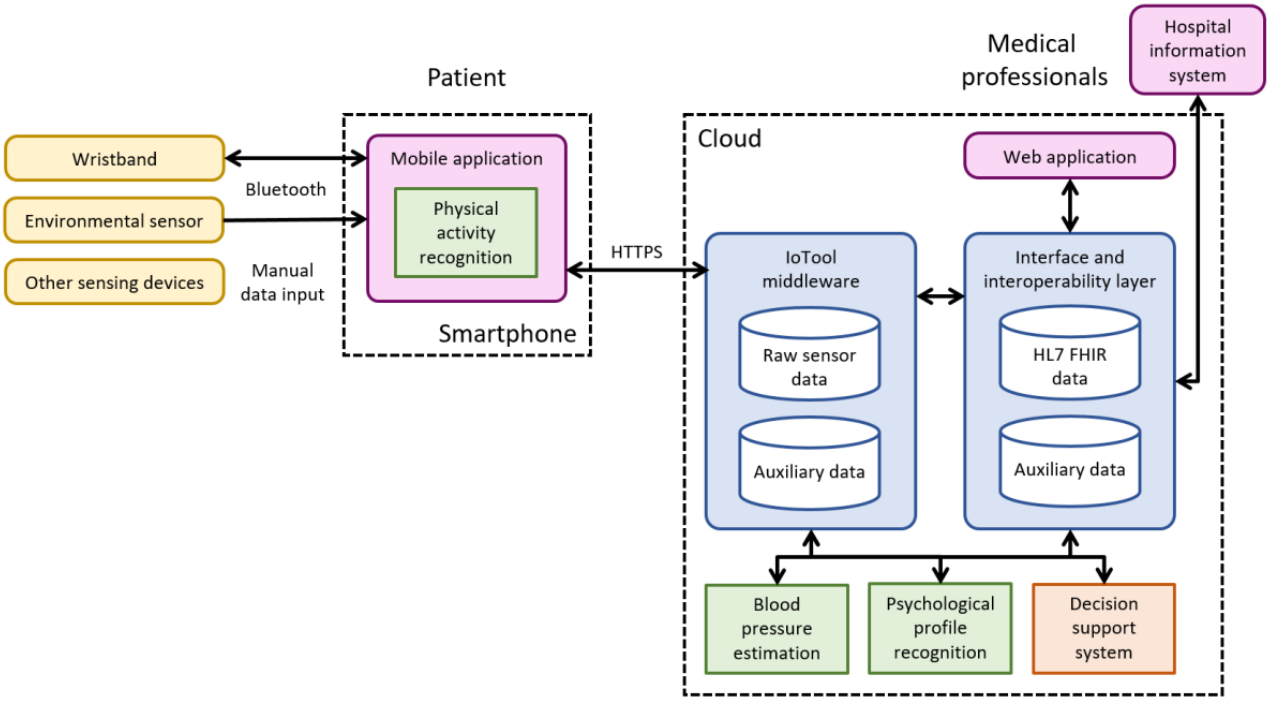
of modified features, as making smaller modifications to a smaller number of features is easier. To solve this problem, we used the multi-objective evolutionary algorithm Nondominated Sorting Genetic Algorithm-II [49].

For more than half of the cases, we were able to find a solution where changing only 1 or sometimes 2 modifiable features would improve the patient's feeling of health. For some cases, we needed to change more features, and for a minority of the cases, no suitable modification could be found. More detailed results can be found in our previous study [50].

Implementation

All the patient-monitoring and decision support modules were integrated into the HeartMan system together with apps for patients and medical professionals. The architecture of the integrated system is illustrated in Figure 4. The wristband and environmental sensor are connected to the mobile app via Bluetooth Low Energy. The mobile app, which includes the physical activity recognition, runs on the smartphone. Physical activity recognition was placed there because it was more efficient to do so than to transmit all the raw accelerometer data to the cloud. On the right side are cloud services, which include BP estimation, psychological profile recognition, and DSS. These were placed entirely in the cloud because they required less raw sensor data, and implementation was easier. Cloud services were installed inside the hospital to comply with the general data protection regulation.

Figure 4. The physical architecture of the HeartMan system.



The data from the mobile app were received by the IoTool middleware [51], whose main purpose was the retrieval of sensor data from smartphones and connected devices, and its storage in a database in the cloud. As it can send data in both directions, it was also used to synchronize application data (such as exercise plans, patient inputs, and push notifications) between the smartphone and the cloud. In this way, the app received the information needed to support each patient on a weekly basis and was then largely independent from the internet for a week. Finally, IoTool can apply arbitrary transformations to sensor data, creating so-called virtual sensors: this capability was used for physical activity recognition, which was implemented as an IoTool virtual sensor transforming acceleration data into physical activities.

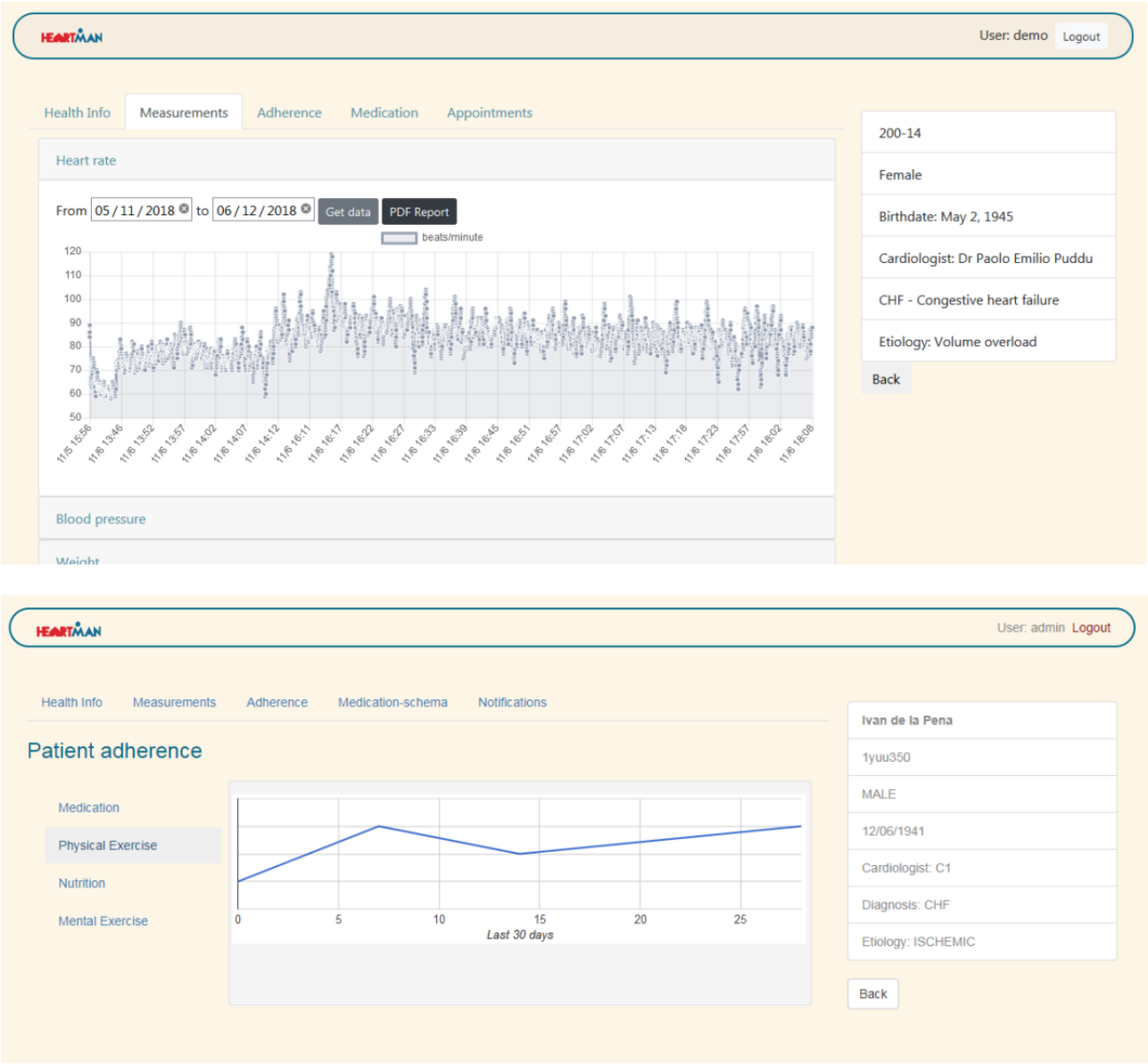
Most raw sensor data were retained in the IoTool database for offline analysis, whereas the data required for HeartMan operation were passed through the interface and interoperability layer, stored using the HL7 FHIR (fast health care interoperability resources) standard for health data exchange [52] if applicable and made available to other services: BP estimation, psychological profile detection, and the DSS. Each of these services reads inputs from and writes outputs to the central storage via the interface and interoperability layer. The data that needed to be sent back to the smartphone were stored in the IoTool database for synchronization. The interface and interoperability layer also provided data to the web application for medical professionals and enabled interoperability with hospital information systems. To do so, it complied with the FHIR REST (representational state transfer) API (application programming interface) specification [52].

The HeartMan mobile app is divided into four sections according to the main topics identified in the medical and user requirements. The respective dashboards are shown in Figure 5. They prominently show the percentage of monthly or weekly activities already performed, which corresponds to the adherence to the HeartMan-suggested self-management at the end of the month or week. The buttons at the bottom trigger various activities, and there is also an Insights section that provides general education on CHF.

The web application for medical professionals shows the patients' clinical information, measurements of heart rate, BP, and weight, and their adherence to the HeartMan-suggested self-management. It also enables the management of medications, with the updated medication plan displayed in the mobile app. Screenshots of the web application are shown in Figure 6.

Figure 5. Dashboards of the HeartMan mobile app for physical activity, nutrition, mental support, and medication management.

Figure 6. Screenshots of the HeartMan web application for medical professionals: heart rate measurements (upper) and adherence to the HeartMan-suggested self-management (lower).



Results

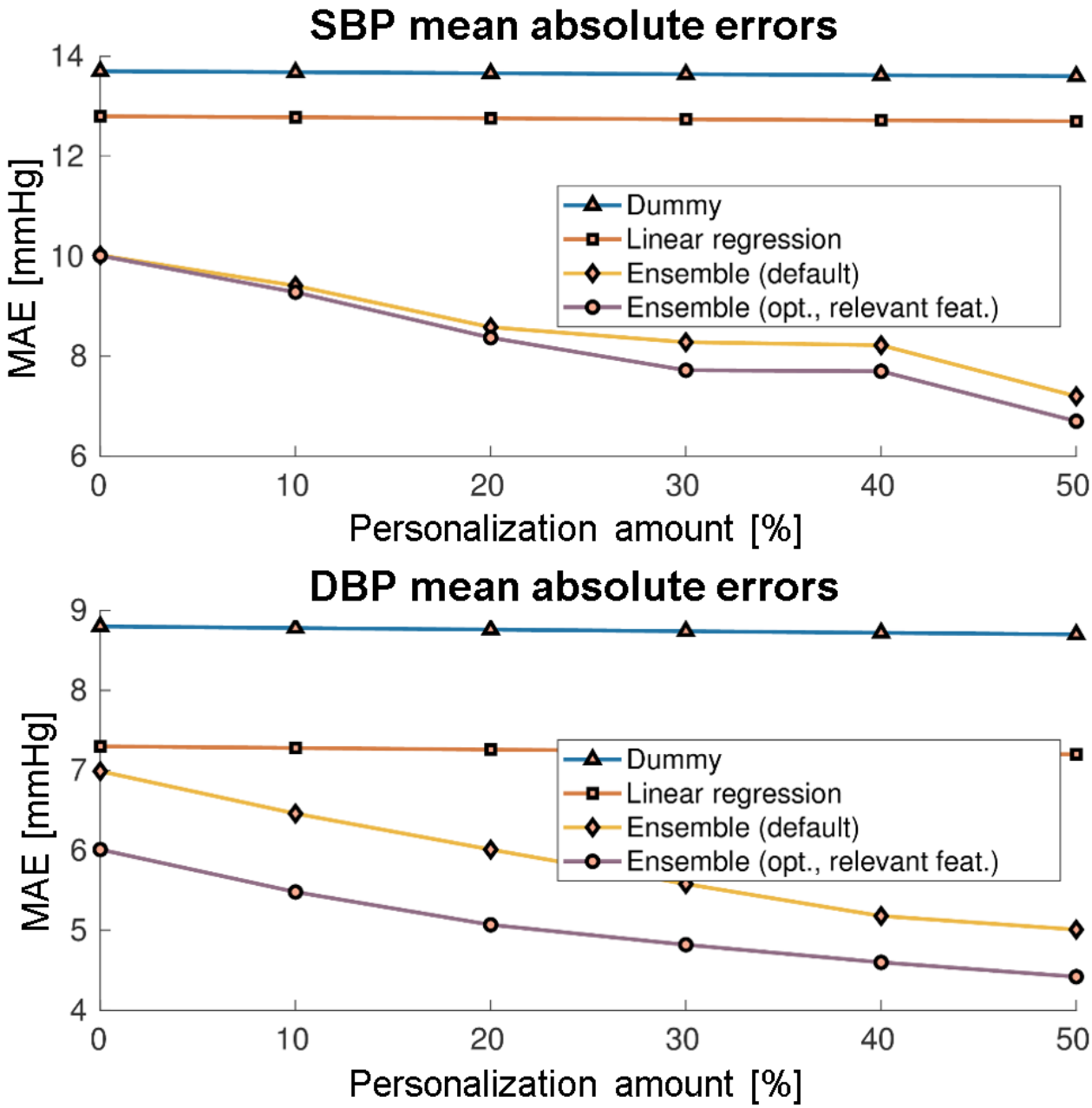
Accuracy of the Patient-Monitoring Methods

BP Estimation

For the first BP estimation test, we collected a data set from 22 healthy subjects (ages 22 to 39 years, 6 women and 16 men) using the Empatica E4 wristband [53]. They wore the wristbands continuously throughout the day and were told to measure their ground truth BP with a certified Omron device every 30 minutes. Each ground truth BP value was attributed to the PPG signal 30 seconds before and after each measurement was made. Leave-one-subject-out evaluation was conducted, and the mean absolute error (MAE) between the estimated and ground truth SBP and DBP was used as the evaluation metric. Several regression algorithms were compared against a baseline dummy regression model, which always outputs the average SBP and DBP of the training set.

Using the Empatica E4 data, the initial errors of ensembles of regression trees were approximately 10 mm Hg for SBP and 6 mm Hg for DBP, as shown in Figure 7. The results were further improved using personalization, achieving errors of 6.70 mm Hg for SBP and 4.42 mm Hg for DBP, suggesting that the connection between PPG and BP is person-specific and that a general model is difficult to derive.

Figure 7. Mean absolute error of systolic blood pressure and diastolic blood pressure estimation in the leave-one-subject-out experiment using the Empatica E4 wristband. DBP: diastolic blood pressure; MAE: mean absolute error; SBP: systolic blood pressure.



As the HeartMan wristband was a prototype intended for wide use by patients (as opposed to the Empatica E4, which is a high-cost research device), the quality of the PPG signal was lower. Therefore, we built person-specific models using the data collected from the HeartMan trials. The patients wore the wristband and were instructed to measure their BP daily with a certified device, so we matched the PPG and BP data as in the previous experiment. We used a train-test split of 70% to 30% to ensure no data leakage. We compared a number of regression algorithms with random forest performing the best, as shown in Table 2.

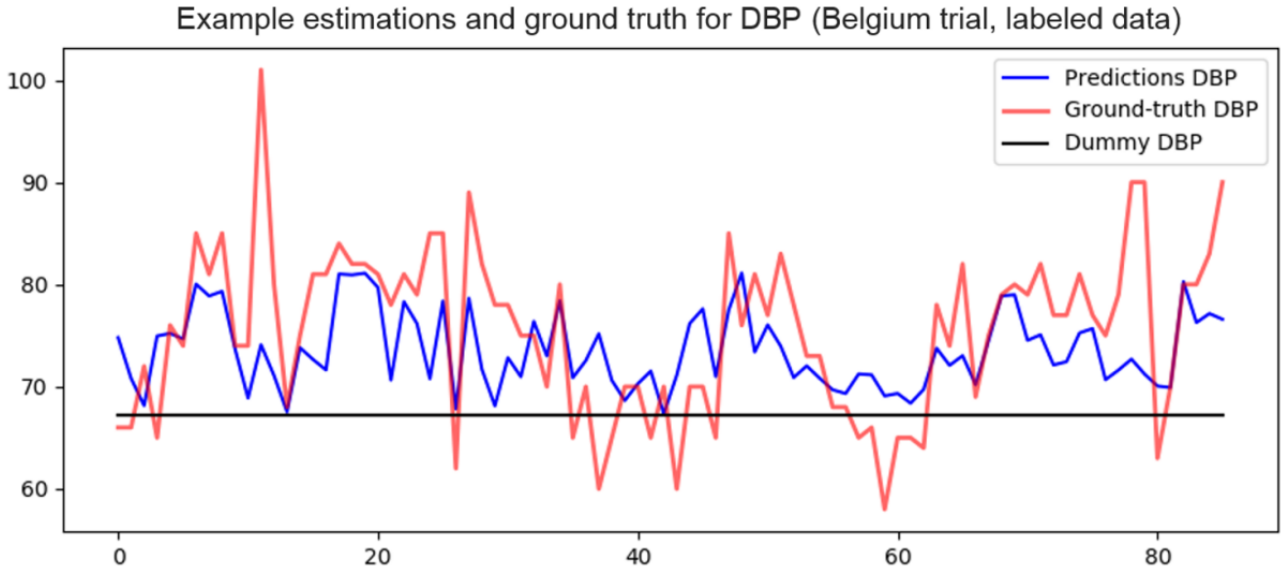
Table 2. MAEs of systolic blood pressure and diastolic blood pressure estimation of personalized models from the HeartMan trial.

Algorithm	MAE[a] of systolic blood pressure (mm Hg)	MAE of diastolic blood pressure (mm Hg)
Baseline dummy (mean)	11.4	8.9
Decision tree	13.1	10.1
k-nearest neighbors	10.6	7.5
Support vector regression	11.3	8.5
Random forest	9.0	7.0

[a]MAE: mean absolute error.

An example segment of the DBP estimates is shown in Figure 8. The results show that BP estimation is feasible; however, most state-of-the-art methods are highly dependent on high signal quality to obtain precise morphological features on a per-cycle basis, which is difficult to achieve with an affordable wristband.

Figure 8. Segment of example estimates and ground truth diastolic blood pressure from the HeartMan trial. DBP: diastolic blood pressure.



Psychological Profile Recognition

To test the psychological profile recognition, we collected a data set from 30 healthy subjects (mean age 68, SD 2 years, 6 women and 23 men). The subjects used the HeartMan mobile app for psychophysiological data collection. Leave-one-subject-out evaluation was conducted, and classification accuracy into depressed, anxious, and motivated profiles was used as the evaluation metric. Classification models trained with four machine learning algorithms were compared against a baseline dummy model, which always returned the majority class.

As shown in Table 3, the support vector machine (SVM) model performed best, achieving a fairly high accuracy, especially considering that this is a subject-independent result. In Table 4, we can see the results in terms of precision, recall, and F1-score for the SVM model. The percentages of the confusion matrix as a result of the cross-validation procedure showed that SVM can classify all 3 classes with precisions of 93%, 86%, and 84%, respectively. From the results, it can be observed that the motivated profile was recognized most accurately, whereas most of the misclassifications came from the anxious and depressed profiles, which are sometimes very similar.

Table 3. Classification accuracies of the psychological profile detection.

Algorithm	Classification accuracy (%)
Baseline dummy (majority)	37.9
Naïve Bayes	79.7
Multilayer perceptron	75.1
Random forest	62.6
Support vector machine	88.6

Table 4. Precision, recall, and F-measures of the psychological profile detection.

Psychological profile	Precision (%)	Recall (%)	F-measure (%)
Motivated profile	93	94	94
Anxious profile	86	83	85
Depressed profile	84	87	86

Physical Activity Recognition

The model for physical activity recognition was built and evaluated on recordings of 10 healthy subjects (mean age 59, SD 5 years, 6 women and 4 men). The subjects performed a scenario consisting of all the activities to be recognized with several variations: walking at different speeds, uphill and carrying a burden, eating various foods, and performing a wide range of chores (cooking, sweeping floor, gardening tasks, etc) and hand activities (writing, using a computer, knitting, etc). Similar to the previous cases, a leave-one-subject-out evaluation was conducted. Precision (the fraction of the instances recognized as a certain activity that in fact belong to that activity), recall (the fraction of the instances belonging to a certain activity that are recognized as such), and F-measure (harmonic mean of precision and recall) were used as the evaluation metrics.

Table 5 shows that most of the activities can be recognized reliably. Standing has the smallest F-measure, because it is often misclassified as rest. This is understandable because in both cases, the hand with the wristband does not move much and is not overly problematic because most people rarely stand still for a long time. The second largest problem is confusing eating with hand activities, which is also understandable but makes accurately triggering psychological interventions during eating difficult.

Table 5. Precision, recall, and F-measure of the physical activity recognition.

Activity	Precision (%)	Recall (%)	F-measure (%)
Rest	84	89	87
Standing	48	32	38
Walking	75	86	80
Nordic walking	67	78	72
Running	74	62	67
Exercise	72	77	74
Eating	62	61	61
Washing	73	77	75
Chores	84	81	82
Hand activities	67	65	66
Macro average	71	71	71

General Effectiveness of the System

A proof-of-concept trial was set up to evaluate the effects of the HeartMan intervention on health-related quality of life and disease management (self-care) as primary endpoints [54]. The secondary endpoints we targeted were clinical parameters, illness perception, and mental and sexual health. The clinical trial was registered on NCT03497871 on 2018-04-13. It was implemented in two countries: three hospitals were involved in Belgium, and one hospital and a local health authority participated in Italy. A randomized controlled design was used with a 1:2 ratio of the control and intervention groups. Eligible patients were recruited by the treating cardiologist or general practitioner at the time of regular consultation. After providing informed consent, participants underwent a baseline data collection, containing medical record data registration, questionnaire assessments, and some clinical assessments, including a 6-min walking test. Patients were then randomly assigned to either the control group receiving the usual care or the intervention condition additionally receiving the HeartMan personal health system that they used in their home setting for a period of 3-6 months. All outcome measurements were repeated in both the intervention and control groups at the end of the trial.

The intervention effects were evaluated in a final sample of 56 patients (ie, 34 in the intervention group and 22 in the control group). Trial results showed that the HeartMan system was successful in improving self-care behavior, resulting in a higher quality of disease management, as indicated by the significant (P=.02) improvement of 11% in the Self-Care of Heart Failure Index [55]. No such effect was observed on health-related quality of life, as assessed with the Minnesota Living with Heart Failure Questionnaire [56]. Regarding secondary endpoints, using HeartMan significantly (P<.001) improved psychological outcomes, that is, intervention patients decreased their level of depression (Beck Depression Inventory II [57]) and anxiety (State Trait Anxiety Inventory Form Y [58]) by 15%, and these reductions were even higher in the patients who had used the

mental support module in the app more intensely. The HeartMan intervention also significantly (P=.01) reduced the experience of sexual problems, that is, by 26% on the Sexual Adjustment Scale [59]. No effects were shown for illness perception or clinical outcome of exercise capacity. However, additional data available in a subgroup of the trial sample showed a significant (P=.04) improvement of 11% in the left ventricular ejection fraction. A more extensive publication of trial results is pending.

Patients' Perception of the System

The user experience of HeartMan was investigated both qualitatively and quantitatively in the intervention group. Quantitatively, the Unified Theory of Acceptance and Use of Technology (UTAUT) questionnaire was used [60], adapted to the objectives of the HeartMan system and to the population of older adult users [61]. This questionnaire assesses users' intentions to use the HeartMan system and their usage behavior. The UTAUT questionnaire pointed out that HeartMan users' attitude toward the system was generally positive, with low scores on technology anxiety related to this positive attitude and relatively high-performance expectancy ("the degree to which the user expects that using the system will help him or her to attain gains in job performance" [60]).

Qualitatively, semistructured interviews were performed with 10 patients (7 men and 3 women) and their informal caregivers after having participated in the trial for 3-4 months [62]. The results of an in-depth analysis of sociotechnical complexities in home-based health monitoring systems [63] showed some potential for the HeartMan system as a tool for self-management. Although stressful for some participants, collecting health data such as weight and BP in the HeartMan trial generally raised awareness among the patients of their lifestyle and health. Monitoring their health parameters enabled them to be more aware of their bodies, intervene, and ask for help in a timely manner. The evaluations also showed that the HeartMan system positively affected patients' dietary knowledge and that they felt stimulated to engage in physical activities. This suggests that self-monitoring and empowerment goals are generally achieved. Some weaknesses were also found, such as the need for increased flexibility regarding the interface and interactions with the system.

Discussion

Technology

The HeartMan system is complex, spanning sensing devices, a mobile app, and the cloud; combining diverse technologies; and featuring extensive content to comprehensively address CHF management. The challenge of integrating all this was tackled by an architecture with independent components connected through the IoTool middleware as well as the interface and interoperability layer. A lesson learned was that there is a tradeoff between too tight integration, which makes changes difficult, and too many layers between components, which makes integration testing difficult.

Individual components largely performed as expected. BP estimation from PPG proved the most difficult, as this is a difficult research problem even in ideal conditions, when high-quality PPG signals from a clinical or research device are available. Thus, this technology is not yet sufficiently mature for everyday use by patients. In the DSS, we mainly relied on expert knowledge, and only recommendations regarding temperature and humidity were provided by data-based methods. Although we believe data-based decisions will play a greater role in health management in the future, the amount of raw data currently available to support the range of decisions needed to manage a disease such as CHF cannot yet rival the expert knowledge available in the literature and medical practice. Although that knowledge is ultimately based on data, these data are simply not available in one place (and possibly not at all in some cases).

Medical Perspective

Although the use of telemonitoring systems in cardiac patients has increased tremendously, evidence regarding their effectiveness in managing patients with CHF remains to be mixed [64]. HeartMan, however, is different from most telemonitoring systems: it focuses on empowering patients to properly manage their disease, rather than remote monitoring by health care professionals. It mainly aims to improve the quality of life and self-management in patients by integrating several intervention modalities in the domains of physical health management and psychological support. The trial results showed that the obtained beneficial effects were mostly psychological, more than physical, which is in line with the predefined primary outcomes. A possible explanation is that the system did not achieve sufficient adherence to the advanced and gradually progressive exercise program, which would probably be the most effective way to improve physical health. Nonetheless, before drawing definite conclusions, we need to investigate the effectiveness of the HeartMan system in a wider context, that is, in a larger sample of patients with CHF over a longer intervention period.

User Perspective

As early as during the analysis of the patients' context of use, the HeartMan concept was presented to patients and their initial reactions were captured. Several insights gathered in this phase remained relevant during later evaluation phases and applied to patient-monitoring systems in general. One of the most important such insights was the fact that patients tend to have high and not necessarily correct expectations of automatic patient-monitoring systems such as HeartMan. Patients tend to expect their caregivers to be continuously aware of what the system detects. Although this can lead to a positive motivation to monitor health parameters, it can also lead to a false sense of safety. In addition, while many patients were motivated to monitor these health parameters, they were closely related to lifestyle choices, such as nutrition and physical exercise. We learned that several patients disliked the fact that HeartMan monitors these lifestyle choices and are concerned about a possible loss of control and autonomy in this respect.

These observations lead to a nuanced view of the patients' perspective on self-monitoring technology, with both perceived benefits (feeling of reassurance, increased awareness) and drawbacks (false perception of safety and loss of autonomy). This view suggests that patient empowerment truly is the correct

goal, in the sense that patients should not rely on the supervision of caregivers (as it may not be available) and should also not feel judged and controlled by the system (but should be making healthy lifestyle choices for themselves). We also observe that although HeartMan started on the way to this goal, further improvements can still be made.

On a more practical level, we learned that a distinction between patients regarding digital literacy can be useful [29]. The patients with high literacy received a full explanation of HeartMan functionality at the beginning of the trial. They were encouraged to be proactive and to navigate through the various functions of the application, which was empowering. Such use was feasible because the interface, particularly the information hierarchy of the application, was designed, tested, and refined in collaboration with the patients. Patients with lower digital literacy were asked to react primarily to notifications in the app. In this way, they were able to cope with the app that, even though it was designed to be simple, it was still relatively complex for some users.

Conclusions

We developed HeartMan, a personal health system for the comprehensive self-management of CHF. It uses a wristband and other sensing devices to obtain information on the patient's BP, physical activity, and psychological profile by means of machine learning as well as some other parameters by more mundane means. All this information is fed into a DSS, which provides recommendations on physical health and psychological support. These translate into a detailed physical exercise program, mindfulness exercises, games, and other forms of support for the patient. This is adapted to the patient's physical capacity, current activity, and psychological profile. A web application for medical professionals is also a part of the system. Patients with CHF were involved throughout the development of the system to ensure the system meets their needs. The final prototype was evaluated in a proof-of-concept trial in 56 patients, showing significantly improved disease management while reducing depression, anxiety, and sexual problems. Although illness perception and exercise capacity did not improve, a significant improvement in left ventricular ejection fraction was observed in a subgroup. Overall, the patients' perception of the system was positive.

The HeartMan system was designed with both patients and medical professionals. It works best when integrated with a hospital information system to have access to the users' up-to-date health records and to provide information on the users to their treating clinicians. As such, it bridges the gap between user-friendly mHealth solutions and medical devices, but it can only be offered to patients through a health provider. Therefore, we are also working on a simplified version of the system that will not be a medical device from a regulatory perspective and will not require connection to a hospital or any kind of backend. This will make it easily deployable via mobile app stores and widely accessible to patients with CHF.

In summary, the HeartMan project combined a range of advanced technologies with human-centered design to develop a complex system that was shown to help patients with CHF. Its benefits were psychological more than physical, which may be because the system did not manage to cause difficult behavioral changes such as increased exercise. The reason for this may be that the system was designed to be more supportive than persuasive. Thus, a key area for future development should be behavior change techniques. Nevertheless, the system is ready to be used, and we are pursuing multiple paths to the market.

Acknowledgments

The HeartMan project received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 689660. The project partners are Jožef Stefan Institute, Sapienza University, Ghent University, National Research Council, Atos Spain SA, SenLab, KU Leuven, Bittium Biosignals Ltd, and European Heart Network. The authors acknowledge financial support from the Slovenian Research Agency (research core funding no. P2-0209).

Conflicts of Interest

None declared.

References

1. Roger VL. Epidemiology of heart failure. Circ Res 2013 Aug 30;113(6):646-659. [doi: 10.1161/circresaha.113.300268]
2. Puddu PE, Menotti A. Natural history of coronary heart disease and heart disease of uncertain etiology: findings from a 50-year population study. Int J Cardiol 2015 Oct 15;197:260-264. [doi: 10.1016/j.ijcard.2015.06.046] [Medline: 26148769]
3. Christiansen MN, Køber L, Weeke P, Vasan RS, Jeppesen JL, Smith JG, et al. Age-specific trends in incidence, mortality, and comorbidities of heart failure in Denmark, 1995 to 2012. Circulation 2017 Mar 28;135(13):1214-1223. [doi: 10.1161/circulationaha.116.025941]
4. Leto L, Feola M. Cognitive impairment in heart failure patients. J Geriatr Cardiol 2014 Dec;11(4):316-328 [FREE Full text] [doi: 10.11909/j.issn.1671-5411.2014.04.007] [Medline: 25593581]
5. Celano CM, Villegas AC, Albanese AM, Gaggin HK, Huffman JC. Depression and anxiety in heart failure: a review. Harv Rev Psychiatry 2018;26(4):175-184. [doi: 10.1097/hrp.0000000000000162]
6. Corotto PS, McCarey MM, Adams S, Khazanie P, Whellan DJ. Heart failure patient adherence: epidemiology, cause, and treatment. Heart Fail Clin 2013 Jan;9(1):49-58. [doi: 10.1016/j.hfc.2012.09.004] [Medline: 23168317]

7. Bjarnason-Wehrens B, McGee H, Zwisler A, Piepoli MF, Benzer W, Schmid J, et al. Cardiac rehabilitation in Europe: results from the European Cardiac Rehabilitation Inventory Survey. Eur J Cardiovas Preve and Rehabili 2010 Aug;17(4):410-418. [doi: 10.1097/hjr.0b013e328334f42d]
8. Cook C, Cole G, Asaria P, Jabbour R, Francis DP. The annual global economic burden of heart failure. Int J Cardiol 2014 Feb 15;171(3):368-376. [doi: 10.1016/j.ijcard.2013.12.028] [Medline: 24398230]
9. Ponikowski P, Voors A, Anker S. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Eur J Heart Fail 2016;37(27):2200. [doi: 10.3410/f.718489795.793497182]
10. Allied Market Research. Mhealth market by type (mhealth devices and mhealth services), stakeholders (mobile operators, device vendors, healthcare providers, and content players), and application (cardiovascular diseases, diabetes, respiratory diseases, neurological disorders, and others): global opportunity analysis and industry forecast, 2020-2027. 2020. URL: https://www.alliedmarketresearch.com/mobile-health-market [accessed 2021-02-11]
11. Mordor Intelligence. Telemonitoring systems market - growth, trends, and forecasts (2020 - 2025). 2020. URL: https://www.mordorintelligence.com/industry-reports/global-patient-monitoring-market-industry [accessed 2021-02-11]
12. Grand View Research. Congestive heart failure treatment devices market size, share & trends analysis report by product (ventricular assist devices, counter pulsation devices, implantable cardioverter defibrillators, pacemakers) and segment forecasts to 2016 - 2024. 2016. URL: https://www.grandviewresearch.com/industry-analysis/congestive-heart-failure-treatment-devices-market [accessed 2021-02-11]
13. Athilingam P, Jenkins B. Mobile phone apps to support heart failure self-care management: integrative review. JMIR Cardio 2018 May 02;2(1):e10057 [FREE Full text] [doi: 10.2196/10057] [Medline: 31758762]
14. Heart Failure Storylines. URL: https://play.google.com/store/apps/details?id=com.selfcarecatalyst.healthstorylines.hf&hl=en [accessed 2021-02-04]
15. Athilingam P, Labrador MA, Remo EFJ, Mack L, San Juan AB, Elliott AF. Features and usability assessment of a patient-centered mobile application (HeartMapp) for self-management of heart failure. Appl Nurs Res 2016 Nov;32:156-163. [doi: 10.1016/j.apnr.2016.07.001] [Medline: 27969021]
16. Piepoli MF, Conraads V, Corrà U, Dickstein K, Francis DP, Jaarsma T, et al. Exercise training in heart failure: from theory to practice. A consensus document of the Heart Failure Association and the European Association for Cardiovascular Prevention and Rehabilitation. Eur J Heart Fail 2011 Apr 18;13(4):347-357 [FREE Full text] [doi: 10.1093/eurjhf/hfr017] [Medline: 21436360]
17. Athilingam P, Jenkins B, Johansson M, Labrador M. A mobile health intervention to improve self-care in patients with heart failure: pilot randomized control trial. JMIR Cardio 2017 Aug 11;1(2):e3 [FREE Full text] [doi: 10.2196/cardio.7848] [Medline: 31758759]
18. Aintree heart failure passport. URL: https://play.google.com/store/apps/details?id=com.s3kdevelopers.aintreeheartfailurepassport [accessed 2021-02-04]
19. FAQs in heart failure. URL: https://play.google.com/store/apps/details?id=com.focusmedica.ccfaqheartfailure [accessed 2021-02-04]
20. Heart failure A-Z discussion. URL: https://play.google.com/store/apps/details?id=com.andromo.dev728517.app800941 [accessed 2021-02-04]
21. Heart failure info. URL: https://m.apkpure.com/heart-failure-info/com.programmingisfun.heartfailure [accessed 2021-02-04]
22. Heart Foundation. URL: https://apps.apple.com/us/app/hf-smart-heart-guidelines/id1458070545 [accessed 2021-02-04]
23. HF Path. URL: https://m.apkpure.com/hf-path/com.wellness.aha [accessed 2021-02-04]
24. TreatHF. URL: https://apps.apple.com/us/app/treathf/id1321599852 [accessed 2021-02-04]
25. Baert A, De Smedt D, De Sutter J, De Bacquer D, Puddu PE, Clays E, et al. Factors associated with health-related quality of life in stable ambulatory congestive heart failure patients: systematic review. Eur J Prev Cardiol 2018 Mar 31;25(5):472-481. [doi: 10.1177/2047487318755795] [Medline: 29384392]
26. Dickstein K, Cohen-Solal A, Filippatos G. ESC guidelines for the diagnosis and treatment of acute and chronic heart failure 2008: the Task Force for the diagnosis and treatment of acute and chronic heart failure 2008 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association of the ESC (HFA) and endorsed by the European Society of Intensive Care Medicine (ESICM). Eur J Heart Fail 2008;10:933-989. [doi: 10.3410/f.718489795.793497182]
27. Mlakar M, Puddu PE, Somrak M, Bonfiglio S, Luštrek M, ChironHeartMan research projects. Mining telemonitored physiological data and patient-reported outcomes of congestive heart failure patients. PLoS One 2018 Mar 1;13(3):e0190323 [FREE Full text] [doi: 10.1371/journal.pone.0190323] [Medline: 29494601]
28. Bazzano AN, Martin J, Hicks E, Faughnan M, Murphy L. Human-centred design in global health: a scoping review of applications and contexts. PLoS One 2017;12(11):e0186744 [FREE Full text] [doi: 10.1371/journal.pone.0186744] [Medline: 29091935]
29. Derboven J, Voorend R, Slegers K. Design trade-offs in self-management technology: the HeartMan case. Behaviour & Information Technology 2019 Jul 03;39(1):72-87. [doi: 10.1080/0144929x.2019.1634152]
30. Geddes LA, Voelz MH, Babbs CF, Bourland JD, Tacker WA. Pulse transit time as an indicator of arterial blood pressure. Psychophysiology 1981 Jan;18(1):71-74. [doi: 10.1111/j.1469-8986.1981.tb01545.x] [Medline: 7465731]

31. Mukkamala R, Hahn J, Inan OT, Mestha LK, Kim C, Toreyin H, et al. Toward ubiquitous blood pressure monitoring via pulse transit time: theory and practice. IEEE Trans Biomed Eng 2015 Aug;62(8):1879-1901. [doi: 10.1109/tbme.2015.2441951]
32. Xing X, Sun M. Optical blood pressure estimation with photoplethysmography and FFT-based neural networks. Biomed Opt Express 2016 Aug 01;7(8):3007-3020 [FREE Full text] [doi: 10.1364/BOE.7.003007] [Medline: 27570693]
33. Slapničar G, Mlakar N, Luštrek M. Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network. Sensors (Basel) 2019 Aug 04;19(15):3420 [FREE Full text] [doi: 10.3390/s19153420] [Medline: 31382703]
34. Slapničar G, Luštrek M, Kukar M. Continuous blood pressure estimation from PPG signal. Faculty of Computer and Information Science, University of Ljubljana. Ljubljana: University of Ljubljana, Faculty of Computer and Information Science; 2018. URL: http://www.informatica.si/index.php/informatica/article/view/2229 [accessed 2021-02-11]
35. Kurylyak Y, Lamonaca F, Grimaldi D. A neural network-based method for continuous blood pressure estimation from a PPG signal. 2013 Presented at: The IEEE International instrumentation and measurement technology conference; 2013; Minneapolis, USA p. 6-9. [doi: 10.1109/i2mtc.2013.6555424]
36. Gjoreski M, Janko V, Slapničar G, Mlakar M, Reščič N, Bizjak J, et al. Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors. Information Fusion 2020 Oct;62:47-62. [doi: 10.1016/j.inffus.2020.04.004]
37. Cvetković B, Szeklicki R, Janko V, Lutomski P, Luštrek M. Real-time activity monitoring with a wristband and a smartphone. Information Fusion 2018 Sep;43:77-93. [doi: 10.1016/j.inffus.2017.05.004]
38. Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
39. Rachuri K, Musolesi M, Mascolo C. A mobile phones based adaptive platform for experimental social psychology research. In: Proceedings of the ACM Conference on Ubiquitous Computing. 2010 Presented at: ACM Conference on Ubiquitous Computing; 2010; Copenhagen, Denmark. New York p. 26-29. [doi: 10.1145/1864349.1864393]
40. Grossman JT, Frumkin MR, Rodebaugh TL, Lenze EJ. mHealth assessment and intervention of depression and anxiety in older adults. Harv Rev Psychiatry 2020 Apr 20;28(3):203-214. [doi: 10.1097/hrp.0000000000000255]
41. Zenonos A, Khan A, Kalogridis G. HealthyOffice: Mood recognition at work using smartphones and wearable sensors. In: Proceedings of the IEEE International Conference on Pervasive Computing and Communication Workshops. 2016 Presented at: IEEE International Conference on Pervasive Computing and Communication Workshops; March 14-18, 2016; Sydney, Australia p. 14-18. [doi: 10.1109/percomw.2016.7457166]
42. Gideon J, Provost E, McInnis M. Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. 2016 Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing; March 20-25, 2016; Shanghai, China p. 2359-2363. [doi: 10.1109/icassp.2016.7472099]
43. Frank E, Hall M, Witten I. Online appendix for data mining: practical machine learning tools and techniques. Fourth edition. In: The WEKA workbench. Burlington, MA: Morgan Kaufmann; 2016.
44. Bohanec M, Dovgan E, Maslov P. Designing a personal decision support system for congestive heart failure management. In: In Proceedings of the 20th International Conference Information Society. 2017 Presented at: 20th International Conference Information Society; Oct 9-13, 2017; Ljubljana, Slovenia p. 67-70.
45. Bohanec M, Znidarsic M, Rajkovic V, Bratko I, Zupan B. DEX methodology: three decades of qualitative multi-attribute modelling. - 2013;37(1):49-54 [FREE Full text]
46. Ekman I, Andersson G, Boman K, Charlesworth A, Cleland JG, Poole-Wilson P, et al. Adherence and perception of medication in patients with chronic heart failure during a five-year randomised trial. Patient Educ Couns 2006 Jun;61(3):348-353. [doi: 10.1016/j.pec.2005.04.005] [Medline: 16139468]
47. van der Wal MHL, Jaarsma T, Moser D, Veeger NJGM, van Gilst WH, van Veldhuisen DJ. Compliance in heart failure patients: the importance of knowledge and beliefs. Eur Heart J 2006 Feb;27(4):434-440. [doi: 10.1093/eurheartj/ehi603] [Medline: 16230302]
48. Festinger L. A theory of cognitive dissonance. California: Stanford University Press; 1957.
49. Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Computat 2002;6(2):182-197. [doi: 10.1109/4235.996017]
50. Vodopija A, Mlakar M, Luštrek M. Predictive models to improve the wellbeing of heart-failure patients. In: In Proceedings of 16th Conference on Artificial Inteligence in Medicine, Workshop on advanced predictive model in healthcare. 2017 Presented at: 16th Conference on Artificial Inteligence in Medicine, Workshop on advanced predictive model in healthcare; June 21-24, 2017; Vienna, Austria p. 21-24.
51. IoTool. URL: https://iotool.io/ [accessed 2021-02-04]
52. HL7 FHIR. URL: https://www.hl7.org/fhir/ [accessed 2021-02-04]
53. Empatica E4 wristband. URL: https://www.empatica.com/research/e4/ [accessed 2021-02-04]
54. Baert A, Clays E, Bolliger L, De Smedt D, Lustrek M, Vodopija A, HeartMan consortium. A Personal Decision Support System for Heart Failure Management (HeartMan): study protocol of the HeartMan randomized controlled trial. BMC Cardiovasc Disord 2018 Sep 27;18(1):186 [FREE Full text] [doi: 10.1186/s12872-018-0921-2] [Medline: 30261836]

55. Riegel B, Carlson B, Moser DK, Sebern M, Hicks FD, Roland V. Psychometric testing of the self-care of heart failure index. J Card Fail 2004 Aug;10(4):350-360. [doi: 10.1016/j.cardfail.2003.12.001] [Medline: 15309704]
56. Rector T, Francis G, Cohn J. Patients self-assessment of their congestive heart failure. Part 1: patient perceived dysfunction and its poor correlation with maximal exercise tests. Heart Fail 1987;3:192-196 [FREE Full text]
57. Beck AT, Steer RA, Brown GK. Manual for the Beck Depression Inventory-II. 1996. URL: https://www.brown.edu/academics/public-health/research/mens-health-initiative/bdiii [accessed 2021-02-11]
58. Spielberger CD. State-trait anxiety inventory for adults. Palo Alto: Mind Garden; 1983. URL: https://www.mindgarden.com/145-state-trait-anxiety-inventory-for-adults [accessed 2021-02-11]
59. Derogatis LR. The psychosocial adjustment to illness scale (PAIS). J Psychosomatic Res 1986 Jan;30(1):77-91. [doi: 10.1016/0022-3999(86)90069-3]
60. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425-478. [doi: 10.2307/30036540]
61. Lu HK, Lin PC. Toward a modified UTAUT model for IT acceptance by senior citizens: using technology life style as an individual difference factor. Advanced Materials Research 2014;905:757-763. [doi: 10.4028/www.scientific.net/amr.905.757]
62. Derboven J, Slegers K, Baert A, Clays E. Human agency in self-management tools. In: Proceedings of 13th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2019 Presented at: 13th EAI International Conference on Pervasive Computing Technologies for Healthcare; May 2019; Trento, Italy p. 20-23. [doi: 10.1145/3329189.3329242]
63. Grönvall E, Verdezoto N. Understanding non-functional aspects of home-based healthcare technology. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2013 Presented at: 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing; Sep 2013; Zürich, Switzerland p. 8-12. [doi: 10.1145/2493432.2493495]
64. Monzo L, Schiariti M, Puddu P. Health monitoring systems: an enabling technology for patient care. Boca Raton: CRC Press; 2019:A.

Abbreviations

BP: blood pressure
CHF: congestive heart failure
DBP: diastolic blood pressure
DSS: decision support system
FHIR: fast health care interoperability resource
MAE: mean absolute error
mHealth: mobile health
PPG: photoplethysmogram
SBP: systolic blood pressure
SVM: support vector machine
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by C Lovis; submitted 02.10.20; peer-reviewed by H Papadopoulos, P Athilingam; comments to author 15.11.20; revised version received 30.12.20; accepted 11.01.21; published 05.03.21.

Please cite as:
Luštrek M, Bohanec M, Cavero Barca C, Ciancarelli MC, Clays E, Dawodu AA, Derboven J, De Smedt D, Dovgan E, Lampe J, Marino F, Mlakar M, Pioggia G, Puddu PE, Rodríguez JM, Schiariti M, Slapničar G, Slegers K, Tartarisco G, Valič J, Vodopija A
A Personal Health System for Self-Management of Congestive Heart Failure (HeartMan): Development, Technical Evaluation, and Proof-of-Concept Randomized Controlled Trial
JMIR Med Inform 2021;9(3):e24501
URL: https://medinform.jmir.org/2021/3/e24501
doi:10.2196/24501
PMID:33666562

Original Paper

A Clinical Decision Support System (KNOWBED) to Integrate Scientific Knowledge at the Bedside: Development and Evaluation Study

Alicia Martinez-Garcia[1], PhD; Ana Belén Naranjo-Saucedo[1], MSc; Jose Antonio Rivas[1], BSc; Antonio Romero Tabares[2], PhD; Ana Marín Cassinello[3], BSc; Anselmo Andrés-Martín[3], PhD; Francisco José Sánchez Laguna[4], MSc; Roman Villegas[1], MSc; Francisco De Paula Pérez León[1], BSc; Jesús Moreno Conde[1], MSc; Carlos Luis Parra Calderón[1,5], MSc

[1]Group of Research and Innovation in Biomedical Informatics, Biomedical Engineering and Health Economy, Institute of Biomedicine of Seville, IBiS / Virgen del Rocío University Hospital / CSIC / University of Seville, Seville, Spain

[2]Publications Department, Andalusian Institute of Emergencies and Public Safety, Seville, Spain

[3]Paediatrics Unit, Virgen Macarena University Hospital, Seville, Spain

[4]Department of Information Systems Coordination, Andalusian Health Service, Seville, Spain

[5]Department of Innovation Technology, Virgen del Rocío University Hospital, Seville, Spain

Corresponding Author:
Alicia Martinez-Garcia, PhD
Group of Research and Innovation in Biomedical Informatics, Biomedical Engineering and Health Economy
Institute of Biomedicine of Seville
IBiS / Virgen del Rocío University Hospital / CSIC / University of Seville
Av Manuel Siurot
Seville, 41013
Spain
Phone: 34 955 01 36 16
Email: alicia.martinez.garcia@juntadeandalucia.es

Abstract

Background: The evidence-based medicine (EBM) paradigm requires the development of health care professionals' skills in the efficient search of evidence in the literature, and in the application of formal rules to evaluate this evidence. Incorporating this methodology into the decision-making routine of clinical practice will improve the patients' health care, increase patient safety, and optimize resources use.

Objective: The aim of this study is to develop and evaluate a new tool (KNOWBED system) as a clinical decision support system to support scientific knowledge, enabling health care professionals to quickly carry out decision-making processes based on EBM during their routine clinical practice.

Methods: Two components integrate the KNOWBED system: a web-based knowledge station and a mobile app. A use case (bronchiolitis pathology) was selected to validate the KNOWBED system in the context of the Paediatrics Unit of the Virgen Macarena University Hospital (Seville, Spain). The validation was covered in a 3-month pilot using 2 indicators: usability and efficacy.

Results: The KNOWBED system has been designed, developed, and validated to support clinical decision making in mobility based on standards that have been incorporated into the routine clinical practice of health care professionals. Using this tool, health care professionals can consult existing scientific knowledge at the bedside, and access recommendations of clinical protocols established based on EBM. During the pilot project, 15 health care professionals participated and accessed the system for a total of 59 times.

Conclusions: The KNOWBED system is a useful and innovative tool for health care professionals. The usability surveys filled in by the system users highlight that it is easy to access the knowledge base. This paper also sets out some improvements to be made in the future.

(JMIR Med Inform 2021;9(3):e13182) doi:10.2196/13182

KEYWORDS

evidence-based medicine; clinical decision support system; scientific knowledge integration

Introduction

Currently, in developed countries, the concept of evidence-based medicine (EBM) is part of medicine itself. In the beginning, the EBM meant a paradigm change in the way that clinical practice was accomplished, leaving a process regarding learning and practice based on static knowledge and authority. However, the EBM concept assumes that the scientific-medical knowledge must emerge from clinical experimentation, and must be used, criticized, and qualitatively interpreted with the best available methodology. Consequently, this knowledge must be essentially dynamic. In the EBM general approach, this knowledge, in conjunction with the clinical experience and the patient's preferences and data, should directly influence the clinical decision-making process at all the levels of care, considering that the goal of EBM is to improve the patient's health care quality through enhanced clinical practice [1,2].

Clinical practice is carried out at many complexity levels, so the necessary knowledge to perform it according to the EBM concept must adapt to the real conditions to use the highest quality information possible. The EBM knowledge sources are categorized according to the usability that allows them to be incorporated into the clinical decision-making process at any level in which this process takes place. The usability of the knowledge source shows a direct relationship with the complexity of its methodology, and is therefore better assimilated by the decision process. As a result, the products with detailed information are also more difficult to be incorporated in the health system environment.

Although the EBM supposed a change of attitude in clinical systems, ensuring efficient support to the clinical decisions that must be taken in the patient–doctor relationship context (where it is not easy to consult nor perform an in-depth reading of the original research) was always difficult. In parallel with the EBM conceptual consolidation, some clinical researchers proposed systematic methodologies to achieve products based on the knowledge, to reduce the distance between the research and the practice, thereby saving time for health care professionals in the critical interpretation of the evidence during decision making.

These products were hierarchized in models in 3 proposals. The first one, in 2001, developed a 4-level classification [3]. The second, in 2007, proposed a classification of 5 levels [4]. And the more recent one, in 2009, developed a 6-level classification [5], and has been recently used in relevant research [6,7].

More concretely, the first proposal [3] defined a classification of the following 4 levels:

- Studies: original papers published in journals.
- Syntheses: recompilation of the existing evidence about a specific issue (eg, systematic revisions).
- Synopses: the most relevant elements of a set of evaluated primary studies, including evaluating the methodological quality (eg, the ACP Journal Club).
- Systems: integrate information about the rest of the levels with electronic health records (EHRs).

Comparing this model with the more recent proposal, the 6-level model [5], the main differences between these are (1) the synopses repositories on systematic studies published in scientific reviews that some institutions maintain, and (2) the editorial products that integrate the best practice, in terms of efficacy, according to the explicit and rigorous methods, such as clinical practice guidelines (CPGs) or evidence-based manuals. Probably, CPGs have been the most relevant attempt to inform about the quotidian clinical decisions, and their institutional adoption and individual use are currently accepted criteria for good practice. However, CPGs are complex so their adoption and use are difficult, even in the best of circumstances.

At the top of the pyramid in the 3 proposals, the clinical decision support systems (CDSSs) appear. The CDSSs are the clinical information systems in charge of integrating and summarizing the relevant information about the clinical problems, actualizing, and connecting this information with the patient's situation. The generalization of the EHR makes possible knowledge integration and records management, allowing the habitual use of the evidence at the patient's bedside. Incorporating the CDSS in the EHR is a tough challenge that is not solved yet [8,9].

Adopting an EHR by a health care organization involves making organizational decisions to register and maintain patients' health data, including changes. However, this adoption also makes possible the approach of other types of choices, such as integrating the evidence-based decision support [10].

Consequently, EBM-based interventions improve patient safety. Any clinical intervention must comply with the beneficence principle to the patient, and it is an obligation not to add damage that exceeds the initial clinical condition. Effectiveness and safety are the 2 dimensions that determine the degree of quality of the interventions because no intervention should be assumed to be ineffective even if its cost is zero. The context in which patient care is practiced—the health system—requires improving the effectiveness of interventions and optimizing the efficiency of resources, because health care, whatever its nature, offers a balance between benefits, risks, inconveniences, and costs. Areas such as public health, nursing, and even health policies (called evidence-based health care) have been incorporated into the EBM to ensure the optimal functioning of health systems. It is necessary to extend the dissemination of systematic reviews and clinical guidelines to include electronic access to EHR for all devices, including smartphones [11,12].

EBM contributes to the knowledge in all these dimensions to increase the quality of the intervention. This knowledge is dispersed in many CPGs applied to generalize those actions of proven effectiveness within a specialty or a clinical condition. Despite this, a substantial variation in the provision of services and patient management is documented, between institutions and between professionals of the same institution. The result is known as variability in clinical practice, which can compromise

the quality of the services themselves beyond the health care professionals' actions and the resource allocation equity [13-15]. EBM tends to reduce this variability, promoting the adoption of the most effective, safe, and efficient practices. CDSSs to support translational medicine have been proposed by some researchers [16,17].

The scientific knowledge integration at the bedside with a mobile platform enables health care professionals to make faster and more effective decisions based on validated clinical practice experience. In this sense, a CDSS called the KNOWBED system [18] has been designed, which provides to the health care professional clinically relevant questions concerning the pathology of interest. These questions are associated with recommendations at the bedside, based on the scientific evidence in different existing knowledge bases (eg, massive reference bases, CPGs, systematic reviews). The global architecture of the KNOWBED system is designed as a secure, scalable, standards-based, and EBM service–oriented architecture. Regarding scalability, the infrastructure in which the KNOWBED system has been developed supports more than a dozen similar projects, so it is prepared to receive an even more significant number of users, providing service to all physicians who want to use it within a health system. The fact that the KNOWBED system generates and indexes a set of recommendations from existing scientific evidence, offering intelligent assistance for health professionals, makes it a system based on EBM.

This paper aims to disseminate the KNOWBED project results, highlighting the benefits identified using a CDSS to integrate scientific knowledge at the bedside, encouraging the scientific community to use this kind of system.

The paper is structured as follows: after this introduction, where a brief review of relevant EBM work is presented, we expose the methods carried out. Then, the study results obtained are discussed. Finally, the discussion and conclusions are presented.

Methods

Overview of System Components

Functionally, 2 components integrate the KNOWBED system: a web-based knowledge station and a mobile app.

The knowledge station's actors are the knowledge managers who use the system to manage all the information shown in the mobile app. In other words, the knowledge managers collect the information coming from the existing scientific knowledge in the bibliography and, at the same time, index the clinical recommendations and questions that usually arise throughout the clinical practice, which will be accessible by context based on the HL7 Infobutton standard [19,20]. OpenInfobutton service, from the University of Utah, was used for this task. This service uses contextual information (based on the HL7 Infobutton standard) about the patient, user, clinical setting, and EHR task to anticipate clinicians' and patients' information needs. Furthermore, this service retrieves information from online provider reference and patient education resources that may help meet their information needs. This web service exposes an endpoint that receives all the previously detailed information, and returns a JSON format response. This response is processed to offer access through links to the different sources of information provided by it. The effort to deal with the conflict between recommendations from different sources is made by the knowledge manager technologically supported by the knowledge station.

The mobile app allows health care professionals to have access to scientific knowledge from their mobiles device—both smartphones and tablets—as it provides access to questions and clinical recommendations to follow regarding patients' diagnosis, admission, treatment, etc., indexed by knowledge managers.

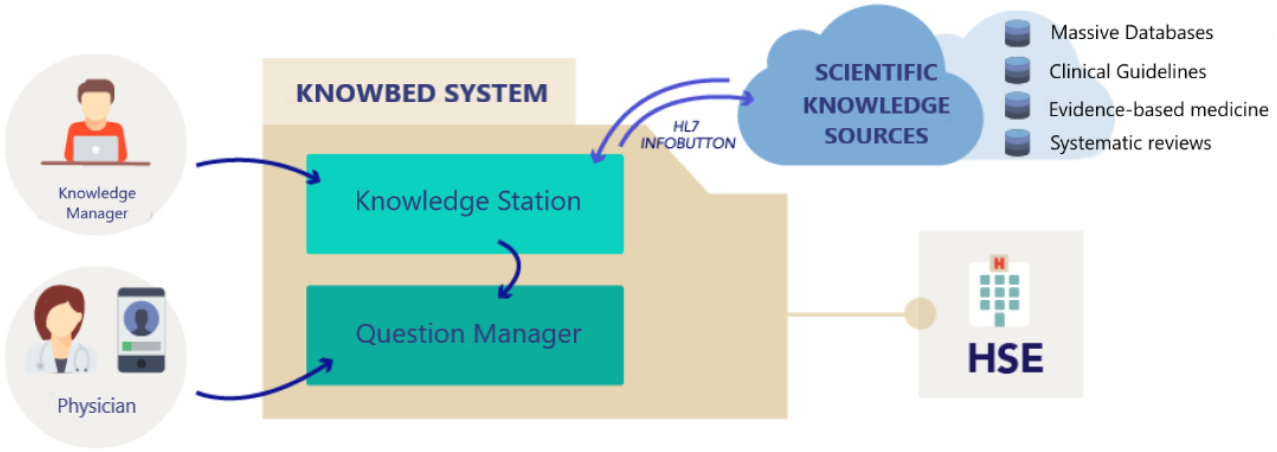
System Technological Architecture

From a technological point of view, the KNOWBED system is based on the development and deployment of 2 different modules (Figure 1): the knowledge station and the question manager.

The knowledge station is a web application that allows access to health professionals from the health care centers through their workstations. This web application will be responsible for visualizing, managing, and maintaining the information associated with the knowledge bases defined in the KNOWBED system. For the development of this web application, the Angular 2 framework has been used which, through HTML5, Sass, and TypeScript, allows the development of web applications based on the SPA paradigm (Single-Page-Application). The PostgreSQL relational database engine supported storage and knowledge management, which stores the information associated with questions, recommendations, and suggestions defined for each of the knowledge areas established in the KNOWBED system. The communication between the application and the server uses the HTTP protocol utilizing the Angular 2 built-in HTTP library. The system has communications security based on tokens generated on the server, and managed in the application through the JWT library; these tokens are renewed in each new connection or after a while.

The question manager offers a multiplatform hybrid mobile app. This app has been developed following the IONIC development framework's premises, capable of developing apps through Angular and Apache Cordova, which provide access to the native mobile phone capabilities. This tool offers professionals the ability to, using their Android mobile devices, access and visualize the set of recommendations defined within the KNOWBED project through a comfortable and intuitive user interface.

Figure 1. KNOWBED architecture.



For the integration of these different modules, a service-oriented architecture has been implemented. This system focuses on the use of an integration gateway based on the Mirth Connect enterprise service bus. Through this integration gateway, services offering the ability to interoperate remotely with the knowledge have been developed. Several mechanisms have been implemented to access the system from outside of the hospital network and invoke the services published in this integration gateway, based on the corporative LDAP system and the generation of random access tokens. Besides, a set of specific routing rules associated with a reverse proxy working as a gateway to the hospital's corporate network has been implemented.

Regarding security aspects, the queries to the knowledge bases are based on general parameters such as gender, age, other conditions, the disease, or inpatient/outpatient. The recommendations are generic for this condition, so no personal data of the patient critical to the possibility of identifying the patient are provided. The "Patient data" section was developed as a link to the EHR application, and this can be used only when connected to the secure corporate network.

Selected Use Case

A specific use case was selected to validate the KNOWBED system: the bronchiolitis pathology from the Paediatrics Unit of the Virgen Macarena University Hospital (Seville).

Bronchiolitis is a common viral infection of the lower respiratory tract that affects children under 2 years of age. This pathology is characterized by acute infection and inflammation of the small airways in the lungs [21,22]. It is the most frequent cause of non-elective admission to the intensive care unit [23,24]. Other researchers have performed studies to improve bronchiolitis management using the technology [25,26].

Based on these considerations, and considering this pathology has a greater incidence during the winter months [27], the pilot was carried out between December and February. In this way, the system was more frequently used and more useful for health care professionals' clinical decision making.

System Evaluation

The system was evaluated using 2 indicators: usability and efficacy.

The KNOWBED system usability was assessed to evaluate the health care professionals' acceptance, using an ad hoc survey asking users regarding the functionalities (Multimedia Appendix 1). The survey recorded sociodemographic information (sex, date of birth, and job title) as well as 13 items that were answered with a 10-point Likert scale (1=strongly disagree; 10=strongly agree) [28]. The survey was administered in 2 phases: phase 1, before using the technological system to know their expectations of the system before interacting with it; and phase 2, to learn about their experience after using the mobile app. Likewise, when new health care professionals joined the Paediatrics Unit, they were informed about the mobile app and were invited to use it.

By contrast, the system's efficacy was studied by analyzing the percentage of acceptance of the recommendations generated by the system. This acceptance was studied by launching the following question when leaving the system: "You are going to leave the App, was this App useful to you?"

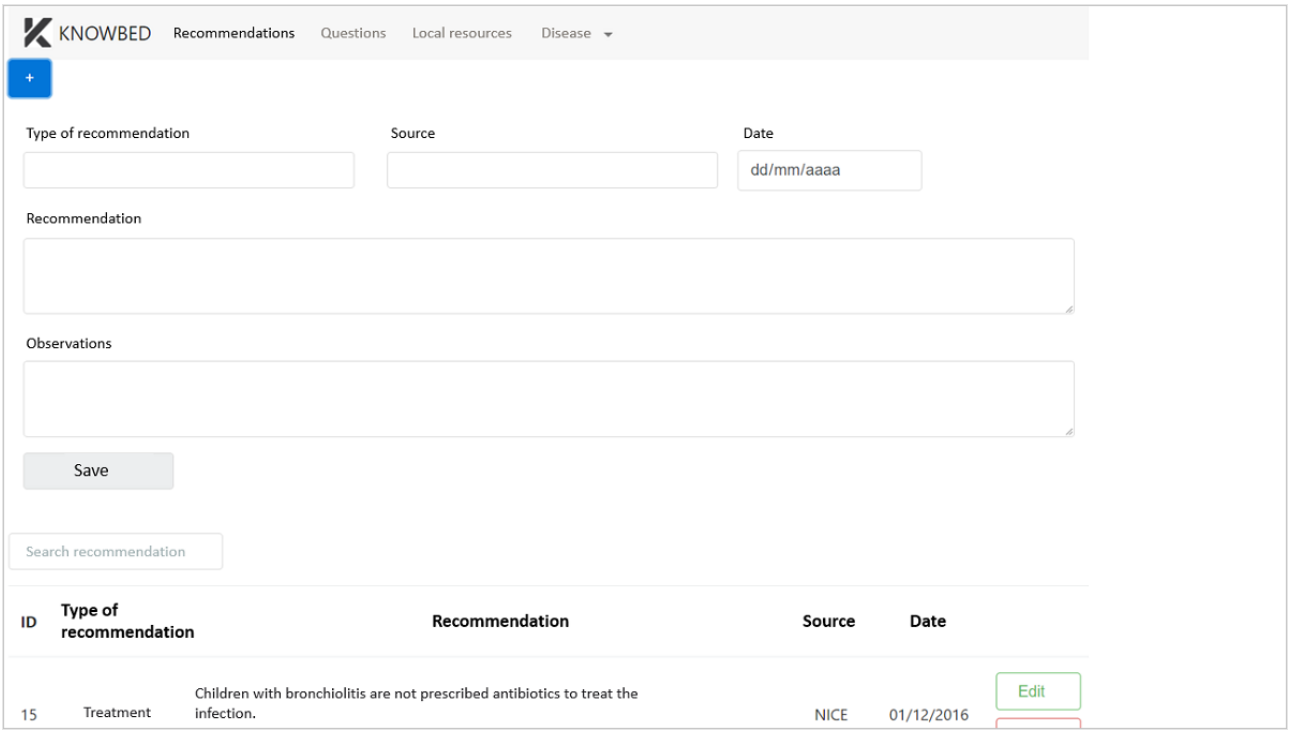
Results

Development and Evaluation of the KNOWBED System

The KNOWBED system has been developed to incorporate scientific evidence into daily clinical practice, improving patient care and providing health care professionals with recommendations based on up-to-date and relevant scientific knowledge.

A screenshot of the KNOWBED knowledge station is shown in Figure 2, which presents the section to add new recommendations, specifying the type, the source, the date, and possible observations.

Figure 2. KNOWBED knowledge station.

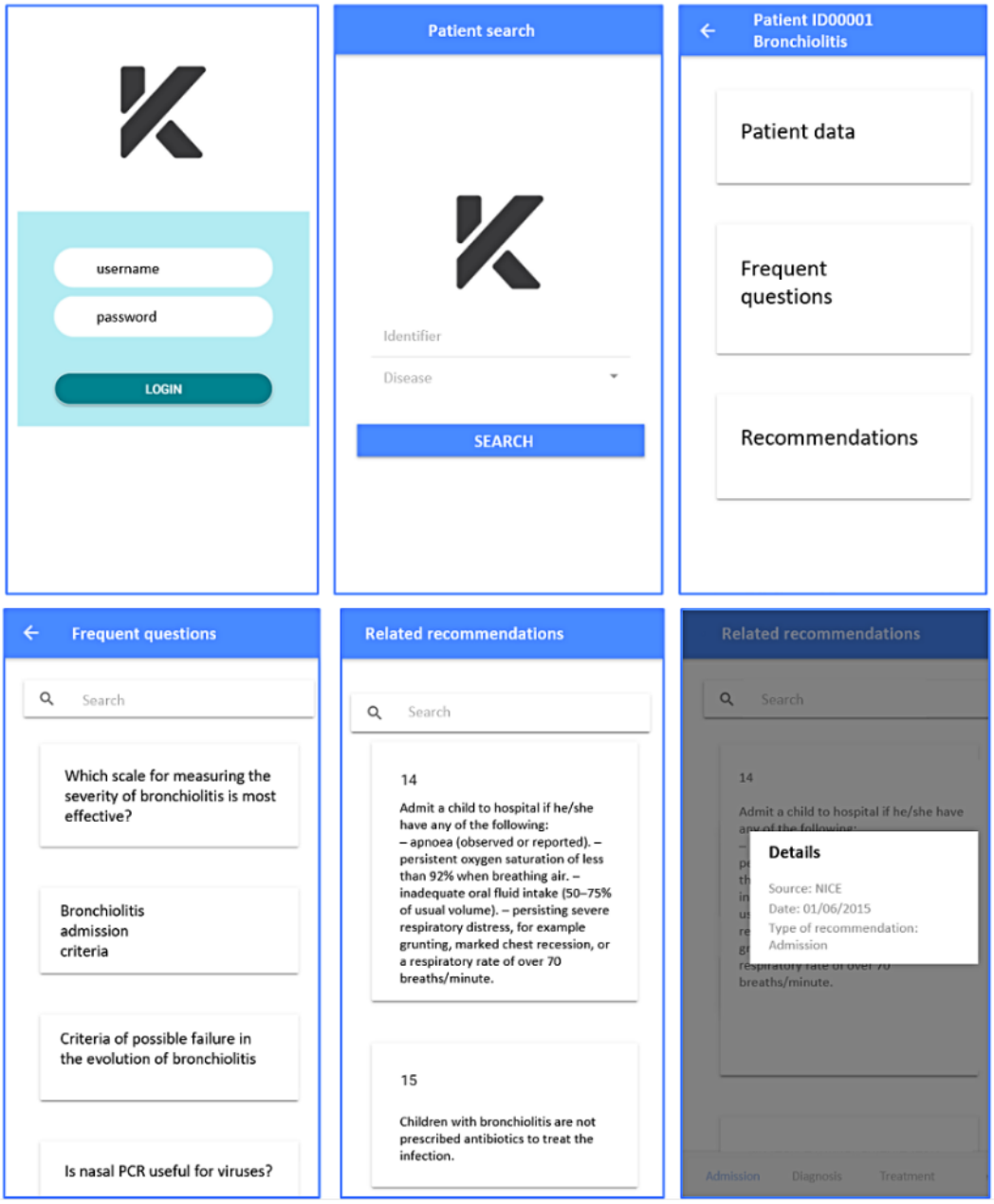


Some screenshots of the KNOWBED mobile app are shown in Figure 3: On the upper left side, the login section is shown. The patient search is displayed in the upper middle section. In the upper right section, the main menu regarding a specific patient is shown. The list of frequent questions regarding this pathology created by the knowledge manager is shown in the bottom left. In the bottom middle, the list of recommendations related to a specific question is displayed. The details of a particular recommendation, including the source, the date, and the type, are shown in the bottom right.

It is also relevant to mention that the KNOWBED system can be integrated for its exploitation in other health care centers.

Furthermore, a methodology to incorporate a new pathology into the knowledge station has been defined. In this sense, a knowledge manager can include further information, and new questions and recommendations could support a health care professional regarding other pathologies.

Figure 3. KNOWBED mobile app.



System Evaluation

To assess the system usage, the number of times users have used the mobile app was analyzed. During the 3-month pilot, the results show that 15 health care professionals made use of it, having registered up to 59 accesses, 23 of which took place after the pilot period.

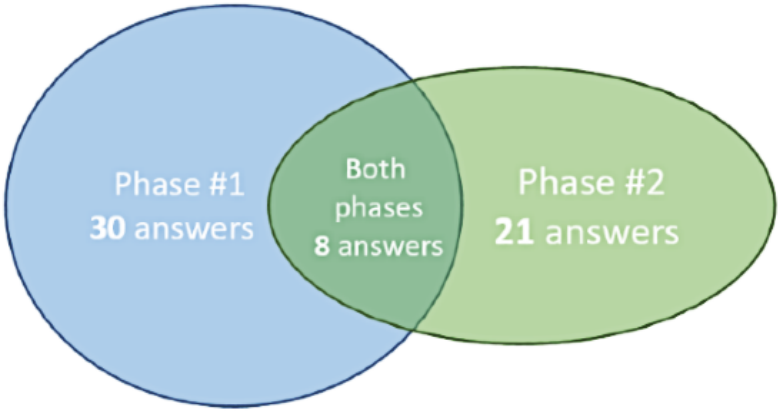
Regarding the usability survey (Multimedia Appendix 1), in phase 1, 30 health care professionals answered the survey, but of them, only 8 completed it in phase 2. However, as mentioned in the "System Evaluation" section, new health care professionals joined the Paediatrics Unit during the pilot, and they used the system. More specifically, 13 health care professionals were incorporated, and they answered the survey after using the system (ie, only in phase 2). In this way, 8 surveys were filled-in for both phase 1 and phase 2, while 13 surveys were filled-in only for phase 2. Figure 4 shows the groups and numbers of health care professionals who responded to phases 1 and 2.

Additionally, 5 health care professionals interested in using the system could not use it because they had iOS phones.

In those users where a comparative analysis can be done (ie, in cases where they have answered in both phases), the results show the following:

- In 7 of 13 questions, the expectations were somewhat higher.
- In 2 of the 13 questions, the expectations were lower.
- In 3 of the 13 questions, the expectations coincided with what was experienced after using the system.

Figure 4. Usability survey: Groups and quantities of responses.



On the one hand, for the health care professionals who have answered in the second phase exclusively (n=13), a comparison of the expectation before and after using the system was not possible. Upon reviewing their opinion after using the system, it is relevant to highlight that the best-scored questions related to the organization support (item number 13; score 8.63/10), and to the improvement in the time spent for decision making (item number 10; score 8.46/10):

- (item number 13) "Overall, I think the organization where I work would support the use of the KNOWBED App."
- (item number 10) "The KNOWBED App can help me resolve some clinical decisions quickly."

On the other hand, the worst-scored question (item number 9; score 7.46/10) was: "I think I will have the technical assistance available to solve problems associated with the KNOWBED App."

The system's efficacy has not been revealed because none of the users have answered the question when leaving the mobile app, so no data on efficacy are available.

Discussion

The study's main findings are the design, development, deployment, and validation of a CDSS called the KNOWBED system to integrate scientific knowledge at the bedside. This system can be presented as an innovative and useful tool due to clinical decision making being offered, allowing health care professionals to access recommendations based on scientific evidence at the bedside by using a mobile device.

A limitation of this study is that the number of answered usability surveys has been small. However, 23 of the accesses that health care professionals made (out of 59 total accesses) have taken place after the pilot period. Consequently, the affirmation of the "KNOWBED system is useful even in months of a lower incidence of this pathology" has been concluded.

This experience with the KNOWBED system concludes that if pathologies with more incidence than bronchiolitis are included, the technological system will be useful for clinical decision making. Furthermore, bronchiolitis is a pathology whose clinical protocols are very well defined, so consulting the literature based on evidence is perhaps less relevant than other pathologies for which clinical protocols are less defined. This fact explains why the 15 users have only registered 59 accesses to the mobile app.

As future work, to continue analyzing the system's usability, encouraging health care professionals' consciousness-raising about the importance of answering the usability survey is relevant, both in the preuse phase of the technology and in the postuse phase, to obtain important data on the usability of technologies.

Furthermore, as future work, it should be stressed that it is required to answer the final question about the usefulness of the mobile app. This indicator was not utilized in this first pilot because the health care professionals have not answered the final question.

The next stage will be extending the experience to more health care centers and including other pathologies, making it possible to increase the number of health care professionals for whom the KNOWBED system's use may be useful and relevant.

The pilot has highlighted a technological-level limitation: the KNOWBED system should have been developed for the iOS operating system as well. During the pilot execution, 5 of the potential users interested in using the mobile app could not make use of it as it was not available for Apple devices.

As an improvement to the knowledge station, the acceptance of all knowledge managers of a specific pathology will be required to validate any information inclusion/modification in the system, and this validation must be done before that new information is reflected in the mobile app. Moreover, nonfree bibliographic bases will be included to improve the knowledge base by feeding their information as well.

Currently, HL7 International is working on an HL7 project called The Fast Healthcare Interoperability Resources (FHIR) for EBM Knowledge Assets project (EBMonFHIR), sponsored by the HL7 Clinical Decision Support Work Group and co-sponsored by the HL7 Clinical Quality Information Work Group and Biomedical Research and Regulation Work Group. The goal of EBMonFHIR is to provide interoperability for those producing, analyzing, synthesizing, disseminating, and implementing evidence of clinical research and recommendations for clinical care included in the CPGs.

EBMonFHIR could be a new relevant standard to take into account in the KNOWBED system.

Acknowledgments

This project has received funding from the Andalusian Ministry of Health from Spain (reference PIN-0213-2016), and FEDER funds.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Usability survey.
[PDF File (Adobe PDF File), 127 KB - medinform_v9i3e13182_app1.pdf]

References

1. Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA 1992 Nov 04;268(17):2420-2425. [doi: 10.1001/jama.1992.03490170092032] [Medline: 1404801]
2. Montori VM, Guyatt GH. Progress in evidence-based medicine. JAMA 2008 Oct 15;300(15):1814-1816. [doi: 10.1001/jama.300.15.1814] [Medline: 18854545]
3. Haynes RB. Of studies, summaries, synopses, and systems: the "4S" evolution of services for finding current best evidence. Evid Based Nurs 2005 Jan;8(1):4-6 [FREE Full text] [doi: 10.1136/ebn.8.1.4] [Medline: 15688480]
4. Haynes B. Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based healthcare decisions. Evid Based Nurs 2007 Jan;10(1):6-7. [doi: 10.1136/ebn.10.1.6] [Medline: 17218282]
5. Dicenso A, Bayley L, Haynes RB. Accessing pre-appraised evidence: fine-tuning the 5S model into a 6S model. Evid Based Nurs 2009 Oct;12(4):99-101. [doi: 10.1136/ebn.12.4.99-b] [Medline: 19779069]
6. Spilsbury K, Devi R, Griffiths A, Akrill C, Astle A, Goodman C, et al. Seeking Answers for Care Homes during the COVID-19 pandemic (COVID SEARCH). Age Ageing 2020 Sep 15:afaa201 [FREE Full text] [doi: 10.1093/ageing/afaa201] [Medline: 32931544]
7. Delvaux N, Goossens M, Van Royen P, Van de Velde S, Vanderstichele R, Cloetens H, et al. Involving general practice trainees in clinical practice guideline adaptation. BMC Med Educ 2018 Jun 22;18(1):148 [FREE Full text] [doi: 10.1186/s12909-018-1252-9] [Medline: 29929504]
8. McGinn T. Putting Meaning into Meaningful Use: A Roadmap to Successful Integration of Evidence at the Point of Care. JMIR Med Inform 2016;4(2):e16 [FREE Full text] [doi: 10.2196/medinform.4553] [Medline: 27199223]
9. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. J Med Internet Res 2018 May 29;20(5):e185 [FREE Full text] [doi: 10.2196/jmir.9134] [Medline: 29844010]
10. Powell J, Buchan I. Electronic health records should support clinical research. J Med Internet Res 2005;7(1):e4 [FREE Full text] [doi: 10.2196/jmir.7.1.e4] [Medline: 15829476]
11. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet 2017 Dec 22;390(10092):415-423. [doi: 10.1016/S0140-6736(16)31592-6] [Medline: 28215660]
12. Celi LA, Marshall JD, Lai Y, Stone DJ. Disrupting Electronic Health Records Systems: The Next Generation. JMIR Med Inform 2015;3(4):e34 [FREE Full text] [doi: 10.2196/medinform.4192] [Medline: 26500106]
13. Love TE. Variation in Medical Practice: Literature Review and Discussion. Wellington, New Zealand: Health Quality & Safety Commission New Zealand; 2013.
14. Shashar S, Codish S, Ellen M, Davidson E, Novack V. Determinants of Medical Practice Variation Among Primary Care Physicians: Protocol for a Three Phase Study. JMIR Res Protoc 2020 Oct 20;9(10):e18673 [FREE Full text] [doi: 10.2196/18673] [Medline: 33079069]
15. Hisham R, Ng CJ, Liew SM, Hamzah N, Ho GJ. Why is there variation in the practice of evidence-based medicine in primary care? A qualitative study. BMJ Open 2016 Mar 09;6(3):e010565 [FREE Full text] [doi: 10.1136/bmjopen-2015-010565] [Medline: 26962037]
16. Velickovski F, Ceccaroni L, Roca J, Burgos F, Galdiz JB, Marina N, et al. Clinical Decision Support Systems (CDSS) for preventive management of COPD patients. J Transl Med 2014 Nov 28;12 Suppl 2:S9 [FREE Full text] [doi: 10.1186/1479-5876-12-S2-S9] [Medline: 25471545]
17. Maier D, Kalus W, Wolff M, Kalko SG, Roca J, Marin de Mas I, et al. Knowledge management for systems biology a general and visually driven framework applied to translational medicine. BMC Syst Biol 2011 Mar 05;5:38 [FREE Full text] [doi: 10.1186/1752-0509-5-38] [Medline: 21375767]

18. Martínez-García A, Rivas-González J, Romero-Tabares A, Marín-Cassinello A, Andrés-Martín A, Sánchez-Laguna F, et al. A mobile Clinical Decision Support System based on the integration of Scientific Knowledge at bedside. 2018 Presented at: Medical Informatics Europe; April 2018; Gothenburg, Sweden p. 24-26.
19. HL7 International. HL7 Version 3 Standard: Context Aware Knowledge Retrieval Application. URL: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=208 [accessed 2021-02-26]
20. Teixeira M, Cook DA, Heale BSE, Del Fiol G. Optimization of infobutton design and Implementation: A systematic review. J Biomed Inform 2017 Dec;74:10-19 [FREE Full text] [doi: 10.1016/j.jbi.2017.08.010] [Medline: 28838801]
21. Meissner HC. Viral Bronchiolitis in Children. N Engl J Med 2016 Jan 07;374(1):62-72. [doi: 10.1056/NEJMra1413456] [Medline: 26735994]
22. Schaller A, Galloway CS. Bronchiolitis in Infants and Children. S D Med 2017 Jun;70(6):274-277. [Medline: 28813765]
23. Schlapbach LJ, Straney L, Gelbart B, Alexander J, Franklin D, Beca J, et al. Burden of disease and change in practice in critically ill infants with bronchiolitis. Eur Respir J 2017 Jun;49(6):1601648 [FREE Full text] [doi: 10.1183/13993003.01648-2016] [Medline: 28572120]
24. Mecklin M, Heikkilä P, Korppi M. The change in management of bronchiolitis in the intensive care unit between 2000 and 2015. Eur J Pediatr 2018 Jul;177(7):1131-1137. [doi: 10.1007/s00431-018-3156-4] [Medline: 29766326]
25. Luo G, Nkoy FL, Gesteland PH, Glasgow TS, Stone BL. A systematic review of predictive modeling for bronchiolitis. Int J Med Inform 2014 Oct;83(10):691-714. [doi: 10.1016/j.ijmedinf.2014.07.005] [Medline: 25106933]
26. Luo G, Stone BL, Nkoy FL, He S, Johnson MD. Predicting Appropriate Hospital Admission of Emergency Department Patients with Bronchiolitis: Secondary Analysis. JMIR Med Inform 2019 Jan 22;7(1):e12591 [FREE Full text] [doi: 10.2196/12591] [Medline: 30668518]
27. Wu P, Dupont WD, Griffin MR, Carroll KN, Mitchel EF, Gebretsadik T, et al. Evidence of a causal role of winter virus infection during infancy in early childhood asthma. Am J Respir Crit Care Med 2008 Dec 01;178(11):1123-1129 [FREE Full text] [doi: 10.1164/rccm.200804-579OC] [Medline: 18776151]
28. Likert R. A technique for the measurement of attitudes. Archives of Psychology 1932:5-55 [FREE Full text]

Abbreviations

CDSS: clinical decision support system
CPG: clinical practice guideline
EBM: evidence-based medicine
EBMonFHIR: FHIR for EBM Knowledge Assets project
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources

Edited by C Lovis; submitted 18.12.18; peer-reviewed by R Guan, N Miyoshi, S Zheng; comments to author 22.03.19; revised version received 18.12.20; accepted 23.01.21; published 10.03.21.

Please cite as:
Martinez-Garcia A, Naranjo-Saucedo AB, Rivas JA, Romero Tabares A, Marín Cassinello A, Andrés-Martín A, Sánchez Laguna FJ, Villegas R, Pérez León FDP, Moreno Conde J, Parra Calderón CL
A Clinical Decision Support System (KNOWBED) to Integrate Scientific Knowledge at the Bedside: Development and Evaluation Study
JMIR Med Inform 2021;9(3):e13182
URL: https://medinform.jmir.org/2021/3/e13182
doi:10.2196/13182
PMID:33709932

Original Paper

Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study

Alberto Tena[1], MSc; Francec Claria[2], PhD; Francesc Solsona[2], PhD; Einar Meister[3], PhD; Monica Povedano[4], PhD

[1]Information and Communication Technologies Group, International Centre for Numerical Methods in Engineering, Barcelona, Spain

[2]Department of Computer Science, Universitat de Lleida, Lleida, Spain

[3]Institute of Cybernetics, Tallinn University of Technology, Tallinn, Estonia

[4]Motoneuron Functional Unit, Hospital Universitari de Bellvitge, Barcelona, Spain

Corresponding Author:
Francesc Solsona, PhD
Department of Computer Science
Universitat de Lleida
Jaume II, 69
Lleida
Spain
Phone: 34 973702735
Email: francesc.solsona@udl.cat

Abstract

Background: Bulbar involvement is a term used in amyotrophic lateral sclerosis (ALS) that refers to motor neuron impairment in the corticobulbar area of the brainstem, which produces a dysfunction of speech and swallowing. One of the earliest symptoms of bulbar involvement is voice deterioration characterized by grossly defective articulation; extremely slow, laborious speech; marked hypernasality; and severe harshness. Bulbar involvement requires well-timed and carefully coordinated interventions. Therefore, early detection is crucial to improving the quality of life and lengthening the life expectancy of patients with ALS who present with this dysfunction. Recent research efforts have focused on voice analysis to capture bulbar involvement.

Objective: The main objective of this paper was (1) to design a methodology for diagnosing bulbar involvement efficiently through the acoustic parameters of uttered vowels in Spanish, and (2) to demonstrate that the performance of the automated diagnosis of bulbar involvement is superior to human diagnosis.

Methods: The study focused on the extraction of features from the phonatory subsystem—jitter, shimmer, harmonics-to-noise ratio, and pitch—from the utterance of the five Spanish vowels. Then, we used various supervised classification algorithms, preceded by principal component analysis of the features obtained.

Results: To date, support vector machines have performed better (accuracy 95.8%) than the models analyzed in the related work. We also show how the model can improve human diagnosis, which can often misdiagnose bulbar involvement.

Conclusions: The results obtained are very encouraging and demonstrate the efficiency and applicability of the automated model presented in this paper. It may be an appropriate tool to help in the diagnosis of ALS by multidisciplinary clinical teams, in particular to improve the diagnosis of bulbar involvement.

(JMIR Med Inform 2021;9(3):e21331) doi:10.2196/21331

KEYWORDS

amyotrophic lateral sclerosis; bulbar involvement; voice; diagnosis; machine learning

Introduction

Background

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease with an irregular and asymmetric progression, characterized by a progressive loss of both upper and lower motor neurons that leads to muscular atrophy, paralysis, and death, mainly from respiratory failure. The life expectancy of patients with ALS is between 3 and 5 years from the onset of symptoms. ALS produces muscular weakness and difficulties

of mobility, communication, feeding, and breathing, making the patient heavily dependent on caregivers and relatives and generating significant social costs. Currently, there is no cure for ALS, but early detection can slow the disease progression [1].

The disease is referred to as spinal ALS when the first symptoms appear in the arms and legs (limb or spinal onset; 80% of cases) and bulbar ALS when it begins in cranial nerve nuclei (bulbar onset; 20% of cases). Patients with the latter form tend to have a shorter life span because of the critical nature of the bulbar muscle function that is responsible for speech and swallowing. However, 80% of all patients with ALS experience dysarthria, or unclear, difficult articulation of speech [2]. On average, speech remains adequate for approximately 18 months after the first bulbar symptoms appear [3]. These symptoms usually become noticeable at the beginning of the disease in bulbar ALS or in later stages of spinal ALS. Early identification of bulbar involvement in people with ALS is critical for improving diagnosis and prognosis and may be the key to effectively slowing progression of the disease. However, there are no standardized diagnostic procedures for assessing bulbar dysfunction in ALS.

Speech impairment may begin up to 3 years prior to diagnosis of ALS [3], and as ALS progresses over time there is significant deterioration in speech [4]. Individuals with ALS with severe dysarthria present specific speech production characteristics [5-7]. However, it is possible to detect early, often imperceptible, changes in speech and voice through objective measurements, as suggested in previous works [8-11]. The authors concluded that phonatory features may be well suited to early ALS detection.

Related Work

Previous speech production studies have revealed significant differences in specific acoustic parameters in patients with ALS. Carpenter et al [7] studied the articulatory subsystem of individuals with ALS and found different involvement of articulators—that is, the tongue function was more involved than the jaw function. In a recent study, Shellikeri et al [5] found that the maximum speed of tongue movements and their duration were only significantly different at an advanced stage of bulbar ALS compared with the healthy control group. Connaghan et al [12] used a smartphone app to identify and track speech decline. Lee et al [6] obtained acoustic patterns for vowels in relation to the severity of the dysarthria in patients with ALS.

Other works have demonstrated the efficiency of features obtained from the phonatory subsystem for detecting early deterioration in ALS [8-11,13-15]. Studies have shown significant differences between jitter, shimmer, and the harmonics-to-noise ratio (HNR) in patients with ALS [8,10,11]. More specifically, Silbergleit et al [8] obtained these features from a steady portion of sustained vowels that provided information regarding changes in the vocal signal that are believed to reflect physiologic changes of the vocal folds. Alternative approaches used formant trajectories to classify the ALS condition [13], correlating formants with articulatory patterns [14], fractal jitter [15], Mel Frequency Cepstral Coefficients (MFCCs) [16], or combined acoustic and motion-related features [9] at the expense of introducing more invasive measurements to obtain data. Besides, the findings revealed significant differences in motion-related features only at an advanced stage of bulbar ALS.

Other related studies, such as one by Frid et al [17], used speech formants and their ratios to diagnose neurological disorders. Teixeira et al [18] and Mekyska et al [19] suggested jitter, shimmer, and HNR as good parameters to be used in intelligent diagnosis systems for dysphonia pathologies.

Garcia-Gancedo et al [20] demonstrated the feasibility of a novel digital platform for remote data collection of digital speech characteristics, among other parameters, from patients with ALS.

In the literature, classification models are widely used to test the performance of acoustic parameters in the analysis of pathological voices. Norel et al [21] identified acoustic speech features in naturalistic contexts and machine learning models developed for recognizing the presence and severity of ALS using a variety of frequency, spectral, and voice quality features. Wang et al [9] explored the classification of the ALS condition using the same features with support vector machine (SVM) and neuronal network (NN) classifiers. Rong et al [22] used SVMs with two feature selection techniques (decision tree and gradient boosting) to predict the intelligible speaking rate from speech acoustic and articulatory samples.

Suhas et al [16] implemented SVMs and deep neuronal networks (DNNs) for automatic classification by using MFCCs. An et al [23] used convolutional neuronal networks (CNNs) to compare the intelligible speech produced by patients with ALS to that of healthy individuals. Gutz et al [24] merged SVM and feature filtering techniques (SelectKBest). In addition, Vashkevich et al [25] used linear discriminant analysis (LDA) to verify the suitability of the sustain vowel phonation test for automatic detection of patients with ALS.

Among feature extraction techniques, principal component analysis (PCA) [26] shows good performance in a wide range of domains [27,28]. Although PCA is an unsupervised technique, it can efficiently complement a supervised classifier in order to achieve the objective of the system. In fact, any classifier can be used in conjunction with PCA because it does not make any kind of assumption about the subsequent classification model.

Hypothesis

Based on previous works, our paper suggests that the acoustic parameters obtained through automated signal analysis from a steady portion of sustained vowels may be used efficiently as predictors for the early detection of bulbar involvement in patients with ALS. For that purpose, the main objectives (and contributions) of this research were (1) to design a methodology for diagnosing bulbar involvement efficiently through the acoustic parameters of uttered vowels in Spanish; and (2) to demonstrate that the performance of the automated diagnosis of bulbar involvement is superior to human diagnosis.

To fulfill these objectives, 45 Spanish patients with ALS and 18 control subjects took part in the study. They were recruited by a neurologist, and the five Spanish vowel segments were

elicited from each participant. The study focused on the extraction of features from the phonatory subsystem—jitter, shimmer, HNR, and pitch—from the utterance of each Spanish vowel.

Once the features were obtained, we used various classification algorithms to perform predictions based on supervised classification. In addition to traditional SVMs [9,16,21,22,24], NNs [9,16,23], and LDA [25], we used logistic regression (LR), which is one of the most frequently used models for classification purposes [29,30]; random forest (RF) [31], which is an ensemble method in machine learning that involves the construction of multiple tree predictors that are classic predictive analytic algorithms [22]; and naïve Bayes (NaB), which is still a relevant topic [32] and is based on applying Bayes' theorem.

Prior to feeding the models, PCA was applied to the features obtained due to the good performance observed of this technique in a wide range of domains.

Methods

Participants

The study was approved by the Research Ethics Committee for Biomedical Research Projects (CEIm) at the Bellvitge University Hospital in Barcelona, Spain. A total of 45 participants with ALS (26 males and 19 females) aged from 37 to 84 (mean 57.8, SD 11.8) years and 18 control subjects (9 males and 9 females) aged from 21 to 68 (mean 45.2, SD 12.2) years took part in this transversal study. All participants with ALS were diagnosed by a neurologist.

Bulbar involvement was diagnosed by following subjective clinical approaches [33], and the neurologist made the diagnosis of whether a patient with ALS had bulbar involvement. Of the 45 participants with ALS, 5 reported bulbar onset and 40 reported spinal onset, but at the time of the study 14 of them presented bulbar symptoms.

To summarize, of the 63 participants in the study, 14 were diagnosed with ALS with bulbar involvement (3 males and 11 females; aged from 38 to 84 years, mean 56.8 years, SD 12.3 years); 31 were diagnosed with ALS but did not display this dysfunction (23 males and 8 females, aged from 37 to 81 years, mean 58.3 years, SD 11.7 years); and 18 were control subjects (9 males and 9 females; aged from 21 to 68 years, mean 45.2 years, SD 12.2 years).

The severity of ALS and its bulbar presentation also varied among participants, as assessed by the ALS Functional Rating Scale-Revised (ALSFRS-R). The ALSFRS-R score (0-48) was obtained from 12 survey questions that assess the degree of functional impairment, with the score of each question ranging from 4 (least impaired) to 0 (most impaired). The scores of the 45 participants in this study ranged from 6 to 46 (mean 31.3, SD 8.6; 3 patients' scores were reported as not available). Within the subgroups, the scores of patients diagnosed with bulbar involvement ranged from 6 to 46 (mean 23.1, SD 9.8), and the scores of participants with ALS who did not present this dysfunction ranged from 17 to 46 (mean 30.2, SD 8.0; 3 patients' scores reported as not available).

The main clinical records of the participants with ALS are summarized in Multimedia Appendix 1.

Vowel Recording

The Spanish phonological system includes five vowel segments—a, e, i, o, and u. These were obtained and analyzed from each patient with ALS and each control participant, all of whom were Spanish speakers.

Sustained samples of the Spanish vowels a, e, i, o, and u were elicited under medium vocal loudness conditions for 3-4 s. The recordings were made in a regular hospital room using a USB GXT 252 Emita Streaming Microphone (Trust International BV) connected to a laptop. The speech signals were recorded at a sampling rate of 44.100 Hz and 32-bit quantization using Audicity, an open-source application [34].

Feature Extraction

Each individual phonation was cut out and anonymously labeled. The boundaries of the speech segments were determined with an oscillogram and a spectrogram using the Praat manual [35] and were audibly checked. The starting point of the boundaries was established as the onset of the periodic energy in the waveform observed in the oscillogram and checked by the apparition of the formants in the spectrogram. The end point was established as the end of the periodic oscillation when a marked decrease in amplitude in the periodic energy was observed. It was also identified by the disappearance of the waveform in the oscillogram and the formants in the spectrogram.

Acoustic analysis was done by taking into account the following features: jitter, shimmer, HNR, and pitch. Once the phonations of each participant had been segmented, the parameters were obtained from each vowel through the standard methods used in Praat [35]; they are explained in detail in this section and consist of a short-term spectral analysis and an autocorrelation method for periodicity detection.

Jitter and shimmer are acoustic characteristics of voice signals. Jitter is defined as the periodic variation from cycle to cycle of the fundamental period, and shimmer is defined as the fluctuation of the waveform amplitudes of consecutive cycles. Patients with lack of control of the vibration of the vocal folds tend to have higher values of jitter. A reduction of glottal resistance causes a variation in the magnitude of the glottal period correlated with breathiness and noise emission, causing an increase in shimmer [18].

To compute jitter parameters, some optional parameters in Praat were established. Period floor and period ceiling, defined as the minimum and maximum durations of the cycles of the waveform that were considered for the analysis, were set at 0.002 s and 0.025 s, respectively. The maximum period factor—the largest possible difference between two consecutive cycles—was set at 1.3. This means that if the period factor—the ratio of the duration of two consecutive cycles—was greater than 1.3, this pair of cycles was not considered in the computation of jitter.

The methods used to determine shimmer were almost identical to those used to determine jitter, the main difference being that

jitter considers periods and shimmer takes into account the maximum peak amplitude of the signal.

Once the previous parameters had been established, jitter and shimmer were obtained by the formulas shown below [35].

Jitter(absolute) is the cycle-to-cycle variation of the fundamental period (ie, the average absolute difference between consecutive periods):

where T_i is the duration of the ith cycle and N is the total number of cycles. If T_i or T_{i-1} is outside the floor and ceiling periods, or if  or  is greater than the maximum period factor, the term  is not counted in the sum, and N is lowered by 1 (if N ends up being less than 2, the result of the computation becomes "undefined").

Jitter(relative) is the average absolute difference between consecutive periods divided by the average period. It is expressed as a percentage:

Jitter(rap) is defined as the relative average perturbation—the average absolute difference between a period and the average of this and its two neighbors, divided by the average period:

Jitter(ppq5) is the five-point period perturbation quotient, computed as the average absolute difference between a period and the average of this and its four closest neighbors, divided by the average period:

Shimmer(dB) is expressed as the variability of the peak-to-peak amplitude, defined as the difference between the maximum positive and the maximum negative amplitude of each period in decibels (ie, the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20:

Where A_i is the extracted peak-to-peak amplitude data and N is the number of extracted fundamental periods.

Shimmer(relative) is defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude, expressed as a percentage:

Shimmer(apq3) is the three-point amplitude perturbation quotient. This is the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbors, divided by the average amplitude:

Shimmer(apq5) is defined as the five-point amplitude perturbation quotient, or the average absolute difference between the amplitude of a period and the average of the amplitudes of this and its four closest neighbors, divided by the average amplitude:

Shimmer(apq11) is expressed as the 11-point amplitude perturbation quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of this and its ten closest neighbors, divided by the average amplitude:

The HNR provides an indication of the overall periodicity of the voice signal by quantifying the ratio between the periodic (harmonics) and aperiodic (noise) components. The HNR was computed using Praat [35], based on the second maximum of normalized autocorrelation function detection, which is used in the following equation:

where $r(t)$ is the normalized autocorrelation function, $r(t = \tau)$ is the second local maximum of the normalized autocorrelation and τ is the period of the signal.

The time step, defined as the measurement interval, was set at 0.01 s, the pitch floor at 60 Hz, the silence threshold at 0.1 (time steps that did not contain amplitudes above this threshold, relative to the global maximum amplitude, were considered silent), and the number of periods per window at 4.5, as suggested by Boersma and Weenink [35].

For the purpose of this study, the mean and standard deviation of the HNR were used.

To obtain the pitch, the autocorrelation method implemented in Praat [35] was used. The pitch floor for males and females was set at 60 Hz and 100 Hz, respectively, and the pitch ceiling for males and females was set at 300 Hz and 500 Hz, respectively. The time step was set, according to Praat [35], at 0.0075 s and 0.0125 s for females and males, respectively. Pitch above pitch ceiling and below pitch floor were not estimated. The mean and standard deviation of the pitch, as well as the minimum and maximum pitch, were features obtained from the pitch metric.

Textbox 1 shows the procedure, inspired by Praat [35], that was used to obtain the features explained above. The full code is freely available online [36].

Textbox 1. Algorithm for obtaining the features (jitter, shimmer, harmonics-to-noise ratio [HNR], and pitch) for acoustic analysis.

1. Each individual phonation of each vowel was cut out and anonymously labeled to define the boundaries of the speech segments.
2. The values for the optional paramaters for analysis were set:
   - Optional parameters to obtain jitter and shimmer parameters
     - pitch floor: females 100 Hz and males 60 Hz
     - pitch ceiling: females 500 Hz and males 300 Hz
     - period floor: 0.002 s
     - period ceiling: 0.025 s
     - maximum period factor: 1.3
   - Optional parameters to obtain HNR
     - time step: 0.01 s
     - pitch floor: 60 Hz
     - silence threshold: 0.1
     - number of periods per windows: 4.5
   - Optional parameters to obtain pitch
     - pitch floor: females 100 Hz and males 60 Hz
     - pitch ceiling: females 500 Hz and males 300 Hz
     - time step: females 0.0075 s and males 0.0125 s
3. Compute jitter and shimmer features—jitter(absolute), jitter(relative), jitter(rap), jitter(ppq5), shimmer(dB), shimmer(relative), shimmer(apq3), shimmer(apq5), shimmer(apq11)—using the configuration parameters established and then obtain the mean of each of these parameters for each vowel.
4. Compute HNR using the configuration parameters established and then obtain the mean (HNR[mean]) and standard deviation (HNR[SD]) values.
5. Compute pitch using the configuration parameters established and then obtain the mean (pitch[mean]), standard deviation (pitch[SD]), minimum (pitch[min]), and maximum (pitch[max]) values.
6. Obtain a data set with the 15 features computed.

PCA

The PCA technique [37], a ranking feature extraction approach, was implemented in R [38] using the Stats package [38]. PCA was used to decompose the original data set into principal components (PCs) to obtain another data set whose data were linearly independent and therefore uncorrelated. It was performed by means of singular value decomposition (SVD) [39].

Prior to applying PCA, given that the mean age of control subjects was approximately 12 years younger than patients with ALS, we removed the age effects by using the data from the control subjects and applying the correction to all the participants as in the study by Norel et al [21]. We fitted the features extracted for healthy people and their age linearly. Then, the "normal aging" of each single feature of each participant was obtained by multiplying the age of the participants by the slope parameter obtained from the linear fit. Finally, the computed "normal aging" was removed from the features. Afterward, a standardized data set was obtained by subtracting the mean and centering the age-adjusted features at 0.

Then, by applying SVD to the standardized data set, a decomposition was obtained:  , where X is the matrix of the standardized data set, U is a unitary matrix and S is the diagonal matrix of singular values s_i. PCs are given by US, and V contains the directions in this space that capture the maximal variance of the features of the matrix X. The number of PCs obtained was the same as the original number of features, and the total variance of all of the PCs was equal to the total variance among all of the features. Therefore, all of the information contained in the original data was preserved.

From the PCA, a biplot chart was obtained for a visual appraisal of the data [40]. The biplot chart allowed us to visualize the data set structure, identify the data variability and clustering participants, and display the variances and correlations of the analyzed features. Then, the first eight PCs that explained almost 100% of the variance were selected to fit the classification models.

Supervised Models

The participants in this study belonged to three different groups: the control group (n=18), patients with ALS with bulbar involvement (n=14), and patients with ALS without bulbar involvement (n=31). Each participant was properly labeled as

control (C) if the subject was a control participant, ALS with bulbar (B) if the subject was a participant with ALS diagnosed with bulbar involvement, or ALS without bulbar (NB) if the subject was a participant with ALS without bulbar involvement. In addition, the ALS (A) label was added to every participant with ALS, with or without bulbar involvement.

Supervised models were built to obtain predictions by comparing the four labeled groups between them. Textbox 2 summarizes the procedure used to create proper classification models.

Textbox 2. Algorithm used to create the classification models.

1. Building the data set: each participant was classified as C (control), B (amyotrophic lateral sclerosis [ALS] with bulbar involvement), or NB (ALS without bulbar involvement) according to the features extracted from the utterance of the five Spanish vowels and the categorical attributes of the bulbar involvement.
2. "Undefined" values were found in few participants when computing the shimmer(apq11) for a specific vowel. They were handled by computing the mean of this parameter for the other vowels uttered by the same participant.
3. The age effects were removed from the data set.
4. The values of the features obtained from the acoustic analysis were zero centered and scaled by using the following equation: $(x_i - \square) / \sigma$, where x_i is the feature vector, $\square$ is the mean, and σ is the standard deviation. Scaling was performed to handle highly variable magnitudes of the features prior to computing primary component analysis (PCA).
5. The PCA was computed and a new data set was created with the first eight primary components (PCs).
6. A random seed was set to generate the same sequence of random numbers. They were used to divide the data set into chunks and randomly permute the data set. The random seed made the experiments reproducible and the classifier models comparable.
7. A 10-fold cross-validation technique was implemented and repeated for 10 trials. The data set was divided into ten contiguous chunks of approximately the same size. Then, 10 training-testing experiments were performed as follows: each chunk was held to test the classifier, and we performed training on the remaining chunks, applying upsampling with replacement by making the group distributions equal; the experiments were repeated for 10 trials, each trial starting with a random permutation of the data set.
8. Two different classification thresholds were established; 50% and 95% (more restrictive). The classification threshold is a value that dichotomizes the result of a quantitative test to a simple binary decision by treating the values above or equal to the threshold as positive and those below as negative.

Several supervised classification models were implemented in R [38] to measure the classification performance. The classification models were fitted with the first eight PCs that explained almost 100% of the data variability. Finally, 10-fold cross-validation was implemented in R using the caret package [41] to draw suitable conclusions. The upsampling technique with replacement was applied to the training data by making the group distributions equal to deal with the unbalanced data set, which could bias the classification models [42].

The first classifier employed was SVM, which is a powerful, kernel-based classification paradigm. SVM was implemented using the e1071 [43]. We used a C-support vector classification [44] and a linear kernel that was optimized through the tune function, assigning values of 0.0001, 0.0005, 0.001, 0.01, 0.1, and 1 to the C parameter, which controls the trade-off between a low training error and a low testing error. A C parameter value of 1 gave the best performance, and thus this was the SVM model chosen.

Next, a classical NN trained with the back propagation technique with an adaptive learning rate was implemented using the RSNNS package [45]. After running several trials to decide the NN architecture, a single hidden layer with three neurons was implemented because it showed the best performance. The activation function (transfer function) used was the hyperbolic tangent sigmoid function.

LDA was implemented using the MASS package [46]. It estimated the mean and variance in the training set and computed the covariance matrix to capture the covariance between the groups to make predictions by estimating the probability that the test set belonged to each of the groups.

LR was implemented by using the Gaussian generalized linear model applying the Stats package [38] for binomial distributions. A logit link function was used to model the probability of "success." The purpose of the logit link was to take a linear combination of the covariate values and convert those values into a probability scale.

Standard NaB based on applying Bayes' theorem was implemented using the e1071 package [43].

Finally, the RF classifier was implemented using the randomForest package [47] with a forest of 500 decision tree predictors. The optimal mtry—a parameter that indicated the number of PCs that were randomly distributed at each decision tree—was optimized for each classification problem by using the train function included in the caret package [41]. Each decision tree performed the classification independently and RF computed each tree predictor classification as one "vote." The majority of the votes computed by all of the tree predictors decided the overall RF prediction.

The code of these implementations is freely available online [48].

Performance Metrics

There are several metrics to evaluate classification algorithms [49]. The analysis of such metrics and their significance must be interpreted correctly to evaluate these algorithms.

There are four possible results in the classification task. If the sample is positive and it is classified as positive, it is counted as a true positive (TP), and when it is classified as negative, it is considered a false negative (FN). If the sample is negative and it is classified as negative or positive, it is considered a true negative (TN) or false positive (FP), respectively. Based on that, three performance metrics, presented below, were used to evaluate the performance of the classification models.

- Accuracy: ratio between the correctly classified samples.

- Sensitivity: proportion of correctly classified positive samples compared with the total number of positive samples.

- Specificity: proportion of correctly classified negative samples compared with the total number of negative samples.

Finally, paired Bonferroni-corrected Student *t* tests [50] were implemented to evaluate the statistical significance of the metrics results. To reject the null hypothesis, which entails considering that there is no difference in the performance of the classifiers, a significance level of α=.05 was established for all tests. The *P* values obtained by performing the tests with values below α=.05 rejected the null hypothesis.

Results

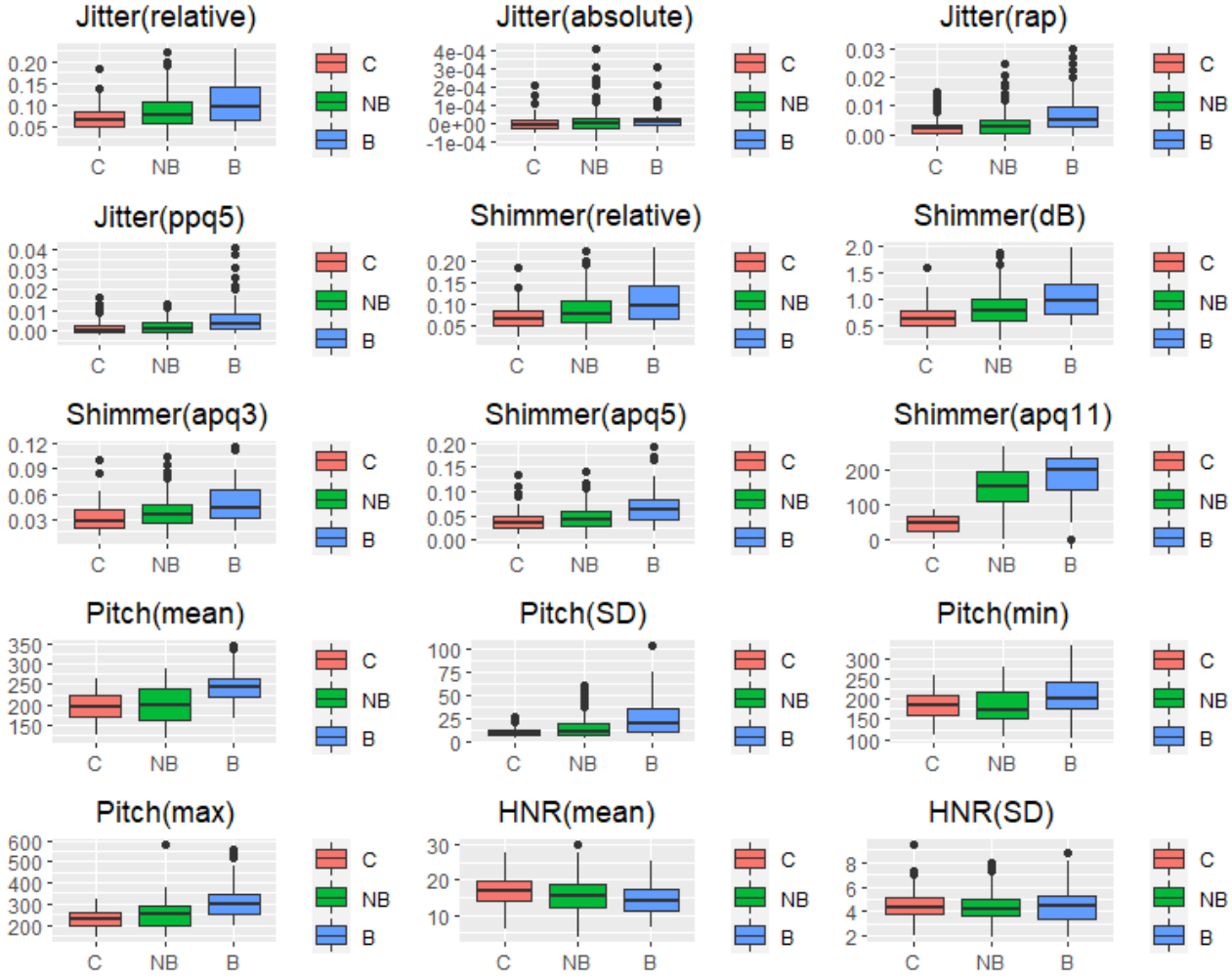
First, the distributions of the features obtained were examined. Then, the PCA was performed and the supervised models studied were evaluated.

Data Exploration

A total of 15 features were obtained in this study. These features were jitter(absolute), jitter(relative), jitter(rap), jitter(ppq5), shimmer(relative), shimmer(dB), shimmer(apq3), shimmer(apq5), shimmer(apq11), pitch(mean), pitch(SD), pitch(min), pitch(max), HNR(mean), and HNR(SD).

Figure 1 shows the box plot of the features obtained from the control (C) group, patients with ALS with bulbar involvement (B), and patients with ALS without bulbar involvement (NB). The means in the B group were higher than those in the C and NB groups. The means in the NB group were located in the middle of the means of the C and B groups. On the contrary, the B group obtained the lowest values for the mean HNR(mean) and HNR(SD). Differences in the standard deviation between the three groups were also observed. In general, features obtained from the B group presented the highest standard deviations.

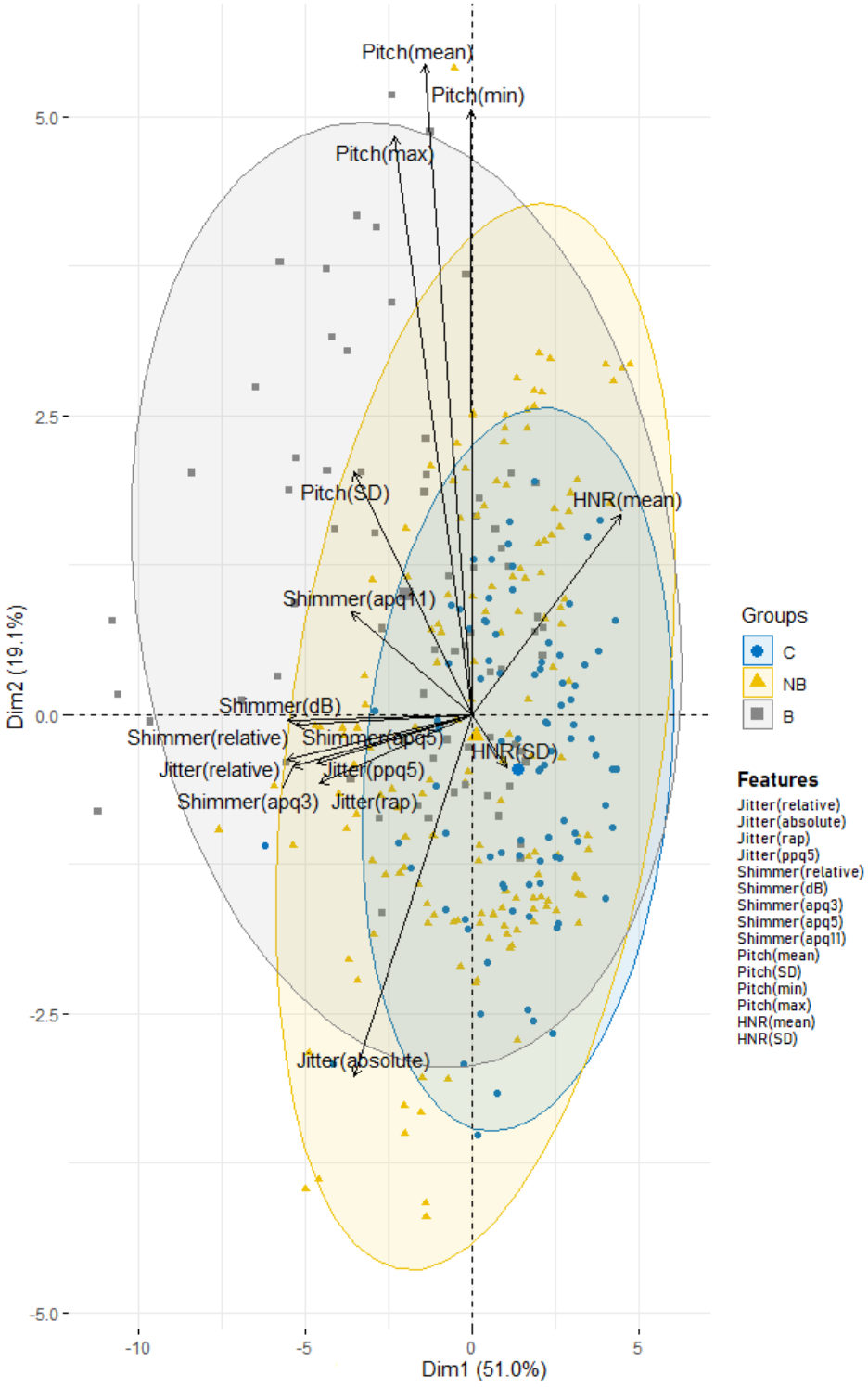
Figure 1. Box plots of features by group. B: patients with amyotrophic lateral sclerosis (ALS) with bulbar involvement; C: control group; HNR: harmonics-to-noise ratio; NB: patients with ALS without bulbar involvement.



PCA

PCA was performed using the data set that contained the 15 features extracted from all of the participants. Figure 2 shows the associated PCA biplot chart. The two axes represent the first (Dim1) and second (Dim2) PCs. The biplot uses the diagonalization method to give a graphical display of its dimensional approximation [51,52]. The interpretation of the biplot involves observing the lengths and directions of the vectors of the features, the data variability, and the clusterization of the participants.

Figure 2. Principal component analysis biplot chart representing the variance of the first (Dim1) and second (Dim2) principal components in the control group (C), patients with amyotrophic lateral sclerosis (ALS) without bulbar involvement (NB), and patients with ALS with bulbar involvement (B). HNR: harmonics-to-noise ratio.



It can be observed that a considerable proportion of variance (70.1%) of the shimmer, jitter, pitch, and HNR was explained. The relative angle between any two vector features represents their pairwise correlation. The closer the vectors are to each other (<90°), the higher their correlation. When vectors are perpendicular (angles of 90° or 270°), the variables have a small or null correlation. Angles approaching 0° or 180° (collinear vectors) indicate a correlation of 1 or –1, respectively. Thus, in this case, shimmer and jitter show a strong positive correlation. Another important observation reflected in Figure 2 is the spatial proximity of the groups in relation both to each other and to the set of features. The projection of the B group onto the vector for shimmer and jitter falls to the left of the vector features. This means that subjects labelled as the B group had higher

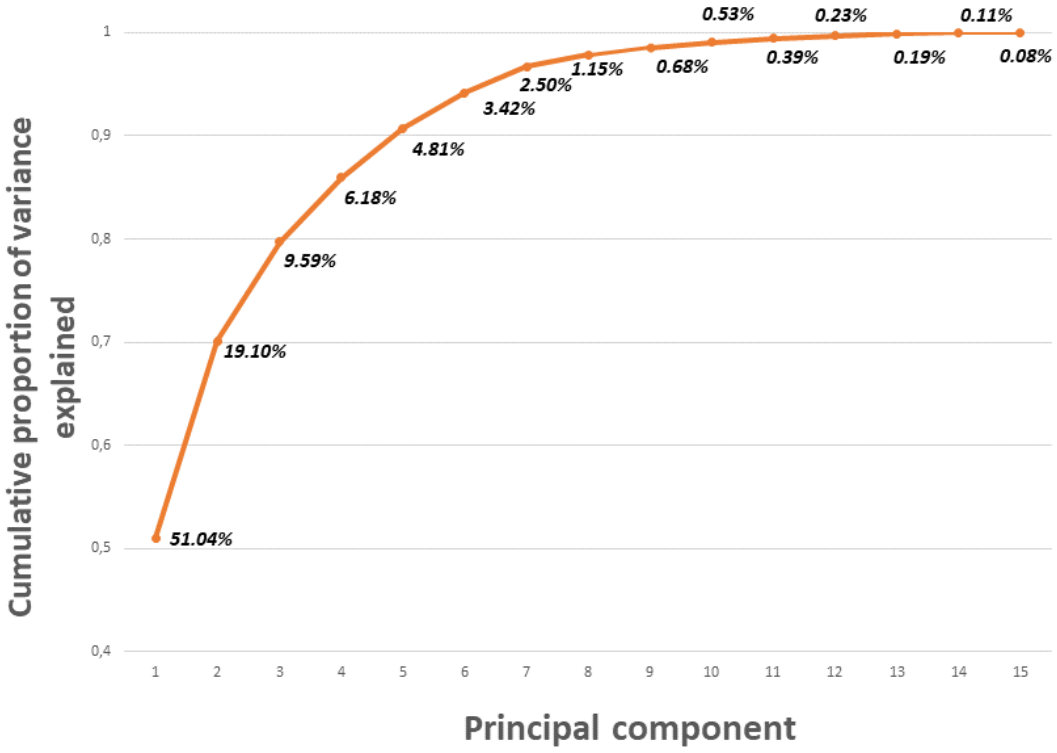
average values for those features than the average values of the other groups. Conversely, the projection of the C group onto those variables falls on the opposite side. In addition, the C and B groups are more distant from each other when projected onto shimmer and jitter. This indicates that shimmer and jitter features are the most important features for the classification of participants in the B and C groups.

The projection of subjects in the NB group requires special attention. Although the projection of these subjects has a spatial proximity with respect to the C group, their variability is higher, overflowing the gray circle corresponding to the B group.

This indicates that some features, especially shimmer and jitter, of some subjects in the NB group have similar projections to the features of the B group.

To fit the models, as explained in detail in the next section, the first eight PCs were selected in order to reduce the dimensionality but preserve almost 100% of the variability as shown in Figure 3.

Figure 3. Cumulative percentage of the explained variance using principal component analysis.



Supervised Model Evaluation

The first eight PCs were selected. Then, each classification model was applied to these PCs. Consequently, better results were obtained than when applying the classification models alone. The results of the classification methods alone are not shown because of their limited contribution to the analysis.

Tables 1 and 2 show the classification performance (accuracy, sensitivity, and specificity metrics) of the supervised models tested for the four cases with the classification threshold set at 50% and 95%, respectively.

Table 1. Classification performance of the supervised models with the classification threshold set at 50%.

Model and metrics	Classification performance (%)			
	C[a] vs B[b]	C vs NB[c]	B vs NB	C vs ALS[d]
Random forest				
Accuracy	93.6	91.1	75.5	90.3
Sensitivity	91.1	92.1	55.7	92.1
Specificity	95.5	89.6	88.4	85.7
Naïve Bayes				
Accuracy	91.0	87.9	75.4	90.3
Sensitivity	89.2	86.7	62.7	92.1
Specificity	93.2	90.0	81.2	85.7
Logistic regression				
Accuracy	93.8	91.4	70.1	91.1
Sensitivity	92.5	89.1	62.2	89.6
Specificity	94.8	95.6	73.5	93.3
Linear discriminant analysis				
Accuracy	94.3	91.6	71.2	91.6
Sensitivity	95.6	87.4	61.8	88.3
Specificity	90.0	98.8	75.4	87.8
Neuronal network				
Accuracy	94.8	92.5	70.4	92.2
Sensitivity	91.7	90.3	60.0	90.8
Specificity	97.2	96.4	75.2	95.6
Support vector machine				
Accuracy	95.8	91.5	69.9	91.6
Sensitivity	91.4	88.4	59.4	88.9
Specificity	99.3	97.0	74.6	98.2

[a]C: control group.

[b]B: patients with amyotrophic lateral sclerosis (ALS) with bulbar involvement.

[c]NB: patients with ALS without bulbar involvement.

[d]ALS: all patients with ALS.

Table 2. Classification performance of the supervised models with the classification threshold set at 95%.

Model and metrics	Classification performance (%)			
	C[a] vs B[b]	C vs NB[c]	B vs NB	C vs ALS[d]
Random forest				
Accuracy	58.3	56.1	68.8	75.1
Sensitivity	4.8	30.4	0.0	65.6
Specificity	100.0	100.0	100.0	98.8
Naïve Bayes				
Accuracy	82.3	68.8	72.8	75.1
Sensitivity	64.7	54.6	15.8	65.6
Specificity	96.1	93.3	98.6	98.8
Logistic regression				
Accuracy	92.8	77.7	74.1	76.0
Sensitivity	84.8	65.1	16.7	66.4
Specificity	99.0	99.6	100.0	100.0
Linear discriminant analysis				
Accuracy	88.1	70.6	71.7	71.1
Sensitivity	72.7	53.5	0.9	59.5
Specificity	100.0	100.0	100.0	100.0
Neuronal network				
Accuracy	92.6	84.8	73.1	86.8
Sensitivity	83.2	76.1	20.5	81.6
Specificity	100.0	100.0	96.8	99.8
Support vector machine				
Accuracy	86.3	71.1	70.7	71.1
Sensitivity	68.8	54.3	6.1	59.4
Specificity	100.0	100.0	100.0	100.0

[a]C: control group.

[b]B: patients with amyotrophic lateral sclerosis (ALS) with bulbar involvement.

[c]NB: patients with ALS without bulbar involvement.

[d]ALS: all patients with ALS.

In the case of the C group versus the B group, with the classification threshold set at 50%, the results indicated that all classifiers had a good classification performance. SVM obtained the best accuracy (95.8%). The tests of significance, which are reported in Multimedia Appendix 2, revealed statistically significant differences between SVM and the other models, with the exception of LDA, which obtained an accuracy (94.3%) that closely approximated that of the SVM model. NN also showed really good results (accuracy 94.8%).

Similar behavior was obtained in the C group versus the NB group and the C group versus all patients with ALS. In these cases, NN was the best model (92.5% for C vs NB and 92.2% for C versus ALS). Meanwhile, generally poor performance was obtained in the B group versus the NB group compared with the other cases. Although RF showed the best accuracy (75.5%), the performance of specificity and especially sensitivity dropped dramatically in comparison with the previous cases.

In general, the model performance dropped with a 95% threshold. In the C group versus the B group, the accuracy of the classification models (Table 2) was worse than when the classification threshold was set at 50%. LR shows the best accuracy (92.8%). LDA, SVM, and NaB obtained accuracies of 88.1%, 86.3%, and 82.3%, respectively. RF did not seem to be a good model for this threshold, with an accuracy of 58.3%.

Lower results were obtained in the C group versus the NB group and the C group versus the group with ALS. NN showed the best performance, with accuracies of 84.8% and 86.8%, respectively.

With the 95% threshold, the performance of sensitivity dropped in all cases, especially for the B group versus the NB group, where LR obtained the best performance with an accuracy of 74.1% but a sensitivity of 16.7%.

Discussion

Principal Findings

This study was guided by 2 objectives: (1) to design a methodology for diagnosing bulbar involvement efficiently through the acoustic parameters of uttered vowels in Spanish, and (2) to demonstrate the superior performance of automated diagnosis of bulbar involvement compared with human diagnosis. This was based on the accurate acoustic analysis of the five Spanish vowel segments, which were elicited from all participants. A total of 15 acoustic features were extracted: jitter(absolute), jitter(relative), jitter(rap), jitter(ppq5), shimmer(relative), shimmer(dB), shimmer(apq3), shimmer(apq5), shimmer(apq11), pitch(mean), pitch(SD), pitch(min), pitch(max), HNR(mean), and HNR(SD). Then, the PCs of these features were obtained to fit the most common supervised classification models in clinical diagnosis: SVM, NN, LDA, LR, NaB, and RF. Finally, the performance of the models was compared.

The study demonstrated the feasibility of automatic detection of bulbar involvement in patients with ALS through acoustic features obtained from vowel utterance. It also confirms that speech impairment is one of the most important aspects for diagnosing bulbar involvement, as was suggested by Pattee et al [33]. Furthermore, bulbar involvement can be detected using automatic tools before it becomes perceptible to human hearing.

Voice features extracted from the B group compared with those features extracted from the C group showed the best performance of the classification model for determining bulbar involvement in patients with ALS.

Accuracy for the C group versus the B group revealed values of 95.8% for SVM with the classification threshold established at 50%. However, on increasing the threshold to 95%, the accuracy values for SVM dropped (86.3%) and LR showed the best performance (accuracy 92.8%). NN also showed a good accuracy at 92.6%. This implies that NN and LR are more robust for finding accuracy.

For that case, the results obtained reinforce the idea that it is possible to diagnose bulbar involvement in patients with ALS using supervised models and objective measures. The SVM and LR models provided the best performance for the 50% and 95% thresholds, respectively.

Great uncertainty was found in the analysis regarding bulbar involvement in the NB group. The accuracy values of the C group versus the NB group and the C group versus the group with ALS with the classification threshold at 50% were 92.5% and 92.2%, respectively, for NN. That reveals that the features extracted from the NB group differed significantly from those of the C group. Lower performance should be expected because participants labeled as the C group and NB group should have similar voice performance. This may indicate that some of the participants in the NB group probably had bulbar involvement but were not correctly diagnosed because the perturbance in their voices could not be appreciated by the human ear. Alternatively, it could be simply that a classification threshold of 50% was too optimistic. With a 95% classification threshold, lower results were obtained in the C group versus the NB group and in the C group versus patients with ALS. NN showed the best performance with accuracies of 84.8% and 86.8%, respectively, for the two cases.

The performance between the B group and C group showed better results than between the NB group and C group. Despite this, the unexpectedly high performance of the models for the C group versus the NB group still suggests that some participants in the NB group could have had bulbar involvement. Changing the classification threshold to 95% worsened the results, especially for sensitivity, although this still remained significant.

The case of the B group versus the NB group revealed that the classification models did not distinguish B group and NB group participants as well as they did with the other groups. The accuracy with the 50% threshold showed the highest performance for RF (75.5%), but the models showed difficulties in identifying positive cases. That may be due to the small difference in the variation of the data among participants in the B and NB groups. The same occurred for the 95% threshold: LR obtained the highest accuracy (74.1%) but a sensitivity of only 16.7%. These values remain far from those in the case of the C group versus the B group. These results also reinforce the idea that participants in the NB group were misdiagnosed.

The good model performance obtained in comparing the C and NB groups supports these findings and underscores the importance of using objective measures for assessing bulbar involvement. This corroborated the results obtained in the data exploration and PCA, which were presented in the Results section.

The projection of the NB group in the PCA biplot chart requires special attention. Although the projection of these subjects has a spatial proximity with regard to the C group, their variability is higher, overflowing the circle corresponding to the B group. This indicates that some features, especially shimmer and jitter, of some patients in the NB group have similar projections to those in the B group. This may reveal that these patients in the NB group could have bulbar involvement but were not yet correctly diagnosed because the perturbance in their voices could still not be appreciated by human hearing.

Figure 1 also indicates that the means of the features of the patients in the NB group were between the means of the features of the C and B groups, thus corroborating these assumptions.

Limitations

This study has some limitations. First, using machine learning on small sample sizes makes it difficult to fully evaluate the significance of the findings. The sample size of this study was heavily influenced by the fact that ALS is a rare disease. At the time of the study, 14 of the patients with ALS presented bulbar symptoms. The relatively small size of this group was because ALS is a very heterogeneous disease and not all patients with ALS present the same symptomatology. Additionally, the control subjects were approximately 12 years younger than the patients with ALS. Vocal quality changes with age, and comparing younger control subjects' vocalic sounds with those of older participants with ALS might introduce additional variations.

Although upsampling techniques were used in this study to correct the bias and age adjustments have been applied to correct the vocal quality changes due to the age difference, it would be necessary in future studies to increase the number of participants, especially of patients with ALS with bulbar involvement and control participants of older ages, to draw definitive conclusions.

Second, the variability inherent in establishing the boundaries of the speech segments on spectrograms manually makes replicability challenging. Speakers will differ in their production, and even the same speaker in the same context will not produce two completely identical utterances. In this study, the recorded speech was processed manually in the uniform approach detailed in the Methods section. Automatic instruments have been developed, but unfortunately these methods are not yet accurate enough and require manual correction.

Comparison with Prior Work

The PCA biplot charts indicated that shimmer and jitter were the most important features for group separation in the 2-PC model for ALS classification; however, they also revealed pitch and HNR parameters as good variables for this purpose. These results are consistent with those of Vashkevich et al [25], who demonstrated significant differences in jitter and shimmer in patients with ALS. They are also consistent with Mekyska et al [19] and Teixeira et al [18], who mentioned pitch, jitter, shimmer, and HNR values as the most popular features describing pathological voices. Finally, Silbergleit et al [8] suggested that the shimmer, jitter, and HNR parameters are sensitive indicators of early laryngeal deterioration in ALS.

Concerning the classification models, Norel et al [21] recently implemented SVM classifiers to recognize the presence of speech impairment in patients with ALS. They identified acoustic speech features in naturalistic contexts, achieving 79% accuracy (sensitivity 78%, specificity 76%) for classification of males and 83% accuracy (sensitivity 86%, specificity 78%) for classification of females. The data used did not originate from a clinical trial or contrived study nor was it collected under laboratory conditions. Wang et al [9] implemented SVM and NN using acoustic features and adding articulatory motion information (from tongue and lips). When only acoustic data were used to fit the SVM, the overall accuracy was slightly higher than the level of chance (50%). Adding articulatory motion information further increased the accuracy to 80.9%. The results using NN were more promising, with accuracies of 91.7% being obtained using only acoustic features and increasing to 96.5% with the addition of both lip and tongue data. Adding motion measures increased the classifier accuracy significantly at the expense of including more invasive measurements to obtain the data. We investigated the means of optimizing accuracy in detecting ALS bulbar involvement by only analyzing the voices of patients. An et al [23] implemented CNNs to classify the intelligible speech produced by patients with ALS and healthy individuals. The experimental results indicated a sensitivity of 76.9% and a specificity of 92.3%. Vashkevich et al [25] performed LDA with an accuracy of 90.7% and Suhas et al [16] used DNNs based on MFCCs with an accuracy of 92.2% for automatic detection of patients with ALS.

Starting with the most widely used features suggested in the literature, the classification models used in this paper to detect bulbar involvement automatically (C group versus B group) performed better than the ones used by other authors, specifically the ones obtained using NN (Wang et al [9]) and DNNs based on MCCFs (Suhas et al [16]). We obtained the best-ever performance metrics. This suggests that decomposing the original data set of features into PCs to obtain another data set whose data (ie, PCs) were linearly independent and therefore uncorrelated improves the performance of the models.

Conclusions

This paper suggests that machine learning may be an appropriate tool to help in the diagnosis of ALS by multidisciplinary clinical teams. In particular, it could help in the diagnosis of bulbar involvement. This work demonstrates that an accurate analysis of the features extracted from an acoustic analysis of the vowels elicited from patients with ALS may be used for early detection of bulbar involvement. This could be done automatically using supervised classification models. Better performance was achieved by applying PCA previously to the obtained features. It is important to note that when classifying participants with ALS with bulbar involvement and control subjects, the SVM with a 50% classification threshold exceeded the performance obtained by other authors, specifically Wang et al [9] and Suhas et al [16].

Furthermore, bulbar involvement can be detected using automatic tools before it becomes perceptible to human hearing. The results point to the importance of obtaining objective measures to allow an early and more accurate diagnosis, given that humans may often misdiagnose this deficiency. This directly addresses a recent statement released by the Northeast ALS Consortium's bulbar subcommittee regarding the need for objective-based approaches [53].

Future Work

Future work is directed toward the identification of incorrectly undiagnosed bulbar-involvement in patients with ALS. A time-frequency representation will be used to detect possible deviations in the voice performance of patients in the time-frequency domain. The voice distributions of patients with ALS diagnosed with bulbar involvement and patients with ALS without that diagnosis will be compared in order to detect pattern differences between these two groups. That could provide indications to distinguish undiagnosed participants with ALS who could be misdiagnosed. Also, an improvement in the voice database by increasing the sample size is envisaged.

Acknowledgments

This work was supported by the Ministerio de Economía y Competitividad under contract TIN2017-84553-C2-2-R. Einar Meister's research has been supported by the European Regional Development Fund through the Centre of Excellence in Estonian Studies. The Neurology Department of the Bellvitge University Hospital in Barcelona allowed the recording of the voices of the participants in its facilities. The clinical records were illustrated by Carlos Augusto Salazar Talavera. Dr Marta Fulla and Maria Carmen Majos Bellmunt advised about the process of eliciting the sounds.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of the clinical records of participants with amyotrophic lateral sclerosis.
[PDF File (Adobe PDF File), 37 KB - medinform_v9i3e21331_app1.pdf]

Multimedia Appendix 2
Paired t test with Bonferroni correction.
[PDF File (Adobe PDF File), 55 KB - medinform_v9i3e21331_app2.pdf]

References

1. Carmona C, Gómez P, Ferrer MA, Plamondon R, Londral A. Study of several parameters for the detection of amyotrophic lateral sclerosis from articulatory movement. loquens 2017 Dec 18;4(1):038 [FREE Full text] [doi: 10.3989/loquens.2017.038]
2. Tomik B, Guiloff R. Dysarthria in amyotrophic lateral sclerosis: A review. Amyotroph Lateral Scler 2010;11(1-2):4-15. [doi: 10.3109/17482960802379004] [Medline: 20184513]
3. Makkonen T, Ruottinen H, Puhto R, Helminen M, Palmio J. Speech deterioration in amyotrophic lateral sclerosis (ALS) after manifestation of bulbar symptoms. Int J Lang Commun Disord 2018 Mar;53(2):385-392. [doi: 10.1111/1460-6984.12357] [Medline: 29159848]
4. Tomik B, Krupinski J, Glodzik-Sobanska L, Bala-Slodowska M, Wszolek W, Kusiak M, et al. Acoustic analysis of dysarthria profile in ALS patients. J Neurol Sci 1999 Oct 31;169(1-2):35-42. [doi: 10.1016/s0022-510x(99)00213-0] [Medline: 10540005]
5. Shellikeri S, Green JR, Kulkarni M, Rong P, Martino R, Zinman L, et al. Speech Movement Measures as Markers of Bulbar Disease in Amyotrophic Lateral Sclerosis. J Speech Lang Hear Res 2016 Oct 01;59(5):887-899 [FREE Full text] [doi: 10.1044/2016_JSLHR-S-15-0238] [Medline: 27679842]
6. Lee J, Dickey E, Simmons Z. Vowel-Specific Intelligibility and Acoustic Patterns in Individuals With Dysarthria Secondary to Amyotrophic Lateral Sclerosis. J Speech Lang Hear Res 2019 Jan 30;62(1):34-59. [doi: 10.1044/2018_JSLHR-S-17-0357] [Medline: 30950759]
7. Carpenter RJ, McDonald TJ, Howard FM. The otolaryngologic presentation of amyotrophic lateral sclerosis. Otolaryngology 1978;86(3 Pt 1):ORL479-ORL484. [doi: 10.1177/019459987808600319] [Medline: 112540]
8. Silbergleit AK, Johnson AF, Jacobson BH. Acoustic analysis of voice in individuals with amyotrophic lateral sclerosis and perceptually normal vocal quality. J Voice 1997 Jun;11(2):222-231. [doi: 10.1016/s0892-1997(97)80081-1] [Medline: 9181546]
9. Wang J, Kothalkar PV, Kim M, Bandini A, Cao B, Yunusova Y, et al. Automatic prediction of intelligible speaking rate for individuals with ALS from speech acoustic and articulatory samples. Int J Speech Lang Pathol 2018 Nov;20(6):669-679 [FREE Full text] [doi: 10.1080/17549507.2018.1508499] [Medline: 30409057]
10. Chiaramonte R, Di Luciano C, Chiaramonte I, Serra A, Bonfiglio M. Multi-disciplinary clinical protocol for the diagnosis of bulbar amyotrophic lateral sclerosis. Acta Otorrinolaringol Esp 2019;70(1):25-31 [FREE Full text] [doi: 10.1016/j.otorri.2017.12.002] [Medline: 29699694]
11. Tomik J, Tomik B, Wiatr M, Składzień J, Stręk P, Szczudlik A. The Evaluation of Abnormal Voice Qualities in Patients with Amyotrophic Lateral Sclerosis. Neurodegener Dis 2015;15(4):225-232. [doi: 10.1159/000381956] [Medline: 25967115]
12. Connaghan K, Green J, Paganoni S. Use of beiwe smartphone app to identify and track speech decline in amyotrophic lateral sclerosis (ALS). Graz, Austria; 2019 Presented at: Interspeech 2019; September 15-19; Graz, Austria. [doi: 10.21437/interspeech.2019-3126]
13. Horwitz-Martin R, Quatieri T, Lammert A. Relation of automatically extracted formant trajectories with intelligibility loss and speaking rate decline in amyotrophic lateral sclerosis. 2016 Presented at: Interspeech 2016; September 16; San Francisco. [doi: 10.21437/interspeech.2016-403]
14. Rong P. Parameterization of articulatory pattern in speakers with ALS. 2014 Presented at: Interspeech 2014; September 18; Singapore.

15. Spangler T, Vinodchandran N, Samal A, Green J. Fractal features for automatic detection of dysarthria. 2017 Presented at: IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); 2017; Orlando, FL. [doi: 10.1109/bhi.2017.7897299]
16. Suhas, Patel D, Rao N. Comparison of speech tasks and recording devices for voice based automatic classification of healthy subjects and patients with amyotrophic lateral sclerosis. 2019 Presented at: Interspeech 2019; September 15-19; Graz, Austria. [doi: 10.21437/interspeech.2019-1285]
17. Frid A, Kantor A, Svechin D, Manevitz L. Diagnosis of Parkinson?s disease from continuous speech using deep convolutional networks without manual selection of features. 2016 Presented at: IEEE International Conference on the Science of Electrical Engineering (ICSEE); 2016; Eilat, Israel. [doi: 10.1109/icsee.2016.7806118]
18. Teixeira JP, Fernandes PO, Alves N. Vocal Acoustic Analysis – Classification of Dysphonic Voices with Artificial Neural Networks. Procedia Comput Sci 2017;121:19-26. [doi: 10.1016/j.procs.2017.11.004]
19. Mekyska J, Janousova E, Gomez-Vilda P, Smekal Z, Rektorova I, Eliasova I, et al. Robust and complex approach of pathological speech signal analysis. Neurocomputing 2015 Nov;167:94-111. [doi: 10.1016/j.neucom.2015.02.085]
20. Garcia-Gancedo L, Kelly ML, Lavrov A, Parr J, Hart R, Marsden R, et al. Objectively Monitoring Amyotrophic Lateral Sclerosis Patient Symptoms During Clinical Trials With Sensors: Observational Study. JMIR Mhealth Uhealth 2019 Dec 20;7(12):e13433 [FREE Full text] [doi: 10.2196/13433] [Medline: 31859676]
21. Norel R, Pietrowicz M, Agurto C, Rishoni S, Cecchi G. Detection of amyotrophic lateral sclerosis (ALS) via acoustic analysis. In: Interspeech 2018. 2018 Presented at: Interspeech 2018; September 2-6; Hyderabad, India. [doi: 10.21437/interspeech.2018-2389]
22. Rong P, Yunusova Y, Wang J, Zinman L, Pattee GL, Berry JD, et al. Predicting Speech Intelligibility Decline in Amyotrophic Lateral Sclerosis Based on the Deterioration of Individual Speech Subsystems. PLoS One 2016;11(5):e0154971 [FREE Full text] [doi: 10.1371/journal.pone.0154971] [Medline: 27148967]
23. An K, Kim M, Teplansky K. Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks. 2018 Presented at: Interspeech 2018; September 2-6; Hyderabad, India. [doi: 10.21437/interspeech.2018-2496]
24. Gutz S, Wang J, Yunusova Y, Green J. Early identification of speech changes due to amyotrophic lateral sclerosis using machine classification. 2019 Presented at: Interspeech 2019; September 15-19; Graz, Austria. [doi: 10.21437/interspeech.2019-2967]
25. Vashkevich M, Petrovsky A, Rushkevich Y. Bulbar ALS detection based on analysis of voice perturbation and vibrato. 2019 Presented at: IEEE 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA); September 18-20; Poznan, Poland. [doi: 10.23919/spa.2019.8936657]
26. Jolliffe I. Principal Component Analysis. In: Lovric M, editor. International Encyclopedia of Statistical Science. Berlin Heidelberg: Springer; 2011:1094-1096.
27. Rodriguez-Lujan I, Bailador G, Sanchez-Avila C, Herrero A, Vidal-de-Miguel G. Analysis of pattern recognition and dimensionality reduction techniques for odor biometrics. Knowl-Based Syst 2013 Nov;52:279-289. [doi: 10.1016/j.knosys.2013.08.002]
28. Zhao W, Chellappa R, Krishnaswamy A. Discriminant analysis of principal components for face recognition. 1998 Presented at: Third IEEE International Conference on Automatic Face and Gesture Recognition; April 14-16; Nara, Japan. [doi: 10.1109/afgr.1998.670971]
29. Hosmer D, Lemeshow S. Applied Logistic Regression, Second Edition. Hoboken, NJ: John Wiley & Sons, Inc; 2000.
30. Dingen D, van't Veer M, Houthuizen P, Mestrom EHJ, Korsten EH, Bouwman AR, et al. RegressionExplorer: Interactive Exploration of Logistic Regression Models with Subgroup Analysis. IEEE T Vis Comput Gr 2019 Jan;25(1):246-255. [doi: 10.1109/tvcg.2018.2865043]
31. Flaxman A, Vahdatpour A, Green S, James S, Murray C, Population Health Metrics Research Consortium (PHMRC). Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. Popul Health Metr 2011 Aug 04;9:29 [FREE Full text] [doi: 10.1186/1478-7954-9-29] [Medline: 21816105]
32. Bermejo P, Gámez JA, Puerta JM. Speeding up incremental wrapper feature subset selection with Naive Bayes classifier. Knowl-Based Syst 2014 Jan;55:140-147. [doi: 10.1016/j.knosys.2013.10.016]
33. Pattee GL, Plowman EK, Focht Garand KL, Costello J, Brooks BR, Berry JD, Contributing Members of the NEALS Bulbar Subcommittee. Provisional best practices guidelines for the evaluation of bulbar dysfunction in amyotrophic lateral sclerosis. Muscle Nerve 2019 May;59(5):531-536. [doi: 10.1002/mus.26408] [Medline: 30620104]
34. Audacity Manual Contents. Audacity. 2019. URL: https://manual.audacityteam.org/ [accessed 2021-02-01]
35. Moltu C, Stefansen J, Svisdahl M, Veseth M. Negotiating the coresearcher mandate - service users' experiences of doing collaborative research on mental health. Disabil Rehabil 2012;34(19):1608-1616. [doi: 10.3109/09638288.2012.656792] [Medline: 22489612]
36. Voice features extraction. Alberto Tena. URL: https://github.com/atenad/greco [accessed 2021-02-01]
37. Phaladiganon P, Kim SB, Chen VC, Jiang W. Principal component analysis-based control charts for multivariate nonnormal distributions. Expert Syst Appl 2013 Jun;40(8):3044-3054. [doi: 10.1016/j.eswa.2012.12.020]
38. The R Project for Statistical Computing. URL: https://www.R-project.org/ [accessed 2021-02-01]

39. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Berlin: Springer; 2009.
40. Gabriel KR, Odoroff CL. Biplots in biomedical research. Stat Med 1990 May;9(5):469-485. [doi: 10.1002/sim.4780090502] [Medline: 2349401]
41. Kuhn M. Building Predictive Models in Using the caret Package. J Stat Softw 2008;28(5):1-26. [doi: 10.18637/jss.v028.i05]
42. Kuhn M, Johnson K. Applied Predictive Modeling. New York: Springer; 2013.
43. Meyer D. Misc Functions of the Department of Statistics, Probability Theory Group. R package version 1. 2019. URL: https://CRAN.R-project.org/package=e1071 [accessed 2021-02-01]
44. Boser B, Guyon I, Vapnik V. A Training Algorithm for Optimal Margin Classifiers. : Association for Computing Machinery; 1992 Presented at: COLT'92; July; Pittsburgh, Pennsylvania p. 144-152. [doi: 10.1145/130385.130401]
45. Bergmeir C, Benítez JM. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. J Stat Softw 2012;46(7):1-26. [doi: 10.18637/jss.v046.i07]
46. Venables W, Ripley B. Modern Applied Statistics with S, Fourth Edition. USA: Springer; 2002.
47. Liaw A, Wiener M. Classification and regression by randomforest. R News 2002;2(3):18-22.
48. Supervised classification models for automated detection of bulbar involvement in als patients. Alberto Tena. URL: https://github.com/atenad/greco [accessed 2021-02-01]
49. Tharwat A. Classification assessment methods. ACI 2020 Aug 03;ahead-of-print(ahead-of-print) [FREE Full text] [doi: 10.1016/j.aci.2018.08.003]
50. Hothorn T, Leisch F, Zeileis A, Hornik K. The Design and Analysis of Benchmark Experiments. J Comput Graph Stat 2005;14(3):675-699. [doi: 10.1198/106186005x59630]
51. Gabriel KR. The biplot graphic display of matrices with application to principal component analysis. Biometrika 1971;58(3):453-467. [doi: 10.1093/biomet/58.3.453]
52. Underhill LG. Two Graphical Display Methods for Ecological Data Matrices. In: McLachlan A, Erasmus T, editors. Sandy Beaches as Ecosystems. Dordrecht: Springer; 1983:433-439.
53. Plowman EK, Tabor LC, Wymer J, Pattee G. The evaluation of bulbar dysfunction in amyotrophic lateral sclerosis: survey of clinical practice patterns in the United States. Amyotroph Lateral Scler Frontotemporal Degener 2017 Aug;18(5-6):351-357 [FREE Full text] [doi: 10.1080/21678421.2017.1313868] [Medline: 28425762]

Abbreviations

ALS: amyotrophic lateral sclerosis
ALSFRS-R: ALS Functional Rating Scale-Revised
CNN: convolutional neuronal network
DNN: deep neuronal network
FN: false negative
FP: false positive
HNR: harmonics-to-noise ratio
LDA: linear discriminant analysis
LR: logistic regression
MFCC: Mel Frequency Cepstral Coefficient
NaB: naïve Bayes
NN: neuronal network
PC: principal component
PCA: principal component analysis
RF: random forest
SVD: singular value decomposition
SVM: support vector machine
TN: true negative
TP: true positive

Edited by G Eysenbach; submitted 11.09.20; peer-reviewed by E Toki, E Beneteau; comments to author 12.10.20; revised version received 26.10.20; accepted 17.01.21; published 10.03.21.

Please cite as:
Tena A, Claria F, Solsona F, Meister E, Povedano M
Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study
JMIR Med Inform 2021;9(3):e21331
URL: https://medinform.jmir.org/2021/3/e21331
doi:10.2196/21331
PMID:33688838

Original Paper

Clinical Decision Support for Traumatic Brain Injury: Identifying a Framework for Practical Model-Based Intracranial Pressure Estimation at Multihour Timescales

J N Stroh[1], PhD; Tellen D Bennett[2], MD; Vitaly Kheyfets[1], PhD; David Albers[1,2], PhD

[1]Department of Bioengineering, University of Colorado Denver | Anschutz Medical Campus, Aurora, CO, United States

[2]Department of Pediatrics, University of Colorado School of Medicine, Aurora, CO, United States

Corresponding Author:
J N Stroh, PhD
Department of Bioengineering
University of Colorado Denver | Anschutz Medical Campus
12705 E Montview Blvd
Aurora, CO, 80045
United States
Phone: 1 303 724 4619
Email: jn.stroh@cuanschutz.edu

Abstract

Background: The clinical mitigation of intracranial hypertension due to traumatic brain injury requires timely knowledge of intracranial pressure to avoid secondary injury or death. Noninvasive intracranial pressure (nICP) estimation that operates sufficiently fast at multihour timescales and requires only common patient measurements is a desirable tool for clinical decision support and improving traumatic brain injury patient outcomes. However, existing model-based nICP estimation methods may be too slow or require data that are not easily obtained.

Objective: This work considers short- and real-time nICP estimation at multihour timescales based on arterial blood pressure (ABP) to better inform the ongoing development of practical models with commonly available data.

Methods: We assess and analyze the effects of two distinct pathways of model development, either by increasing physiological integration using a simple pressure estimation model, or by increasing physiological fidelity using a more complex model. Comparison of the model approaches is performed using a set of quantitative model validation criteria over hour-scale times applied to model nICP estimates in relation to observed ICP.

Results: The simple fully coupled estimation scheme based on windowed regression outperforms a more complex nICP model with prescribed intracranial inflow when pulsatile ABP inflow conditions are provided. We also show that the simple estimation data requirements can be reduced to 1-minute averaged ABP summary data under generic waveform representation.

Conclusions: Stronger performance of the simple bidirectional model indicates that feedback between the systemic vascular network and nICP estimation scheme is crucial for modeling over long intervals. However, simple model reduction to ABP-only dependence limits its utility in cases involving other brain injuries such as ischemic stroke and subarachnoid hemorrhage. Additional methodologies and considerations needed to overcome these limitations are illustrated and discussed.

(JMIR Med Inform 2021;9(3):e23215) doi:10.2196/23215

KEYWORDS

intracranial pressure; traumatic brain injury; intracranial hypertension; patient-specific modeling; theoretical models

Introduction

Background

Traumatic brain injury (TBI) is a major public health problem. Intracranial hypertension (ICH) is common after TBI and can cause secondary injury by decreasing local or global cerebral perfusion [1,2]. Cerebral autoregulation governs cerebral blood flow (CBF) by changing local artery diameter [3-5] and usually provides autonomic control of intracranial pressure (ICP). The capacity of this mechanism to adapt to pressure changes may be exhausted by sufficiently acute or prolonged hypertension, which can lead to insufficient perfusion following TBI. Impaired autoregulation also affects a patient's response to drug therapies

to reduce ICP [6]. Therefore, clinical management of ICH after brain injury is crucial for improving patient outcomes.

TBI is often accompanied by elevated systemic arterial blood pressure (ABP) and loss of cranial volume due to cerebral edema. The Monro-Kellie doctrine [7] postulates a constant volume of intracranial (IC) parenchyma (functional brain tissue) and fluids (blood and cerebrospinal fluid [CSF]), so changes in net fluid yield changes in ICP. Consequently, ABP is the primary external ICP driver under this hypothesis, together with changes in volume and fluid [8]. Therefore, clinical protocols seek to control ICP while maintaining cerebral perfusion pressure (CPP, the difference between ABP and ICP) [9] or risk cerebral hypoxia, which may result in death or permanent brain injury.

Important changes in patient ICP occur at minute-to-hour timescales, and clinicians need to know about them quickly. Decisions regarding the escalation of care and intervention for TBI patients are often driven by elevated ICP, typically defined as exceeding 20 mm Hg (1 mm Hg=133.3 Pa approximately) [10]. This underscores the need to monitor the ICP and identify critical changes. An ideal form of clinical decision support would predict ICP many minutes to a few hours in advance, as seconds or minutes might not provide adequate warning for timely intervention.

The Need for ICP Estimation

ICP is measured in situ via an external ventricular drain (gold standard) or a fiberoptic intraparenchymal catheter. Both modalities are invasive and may adversely affect patient outcomes through the risks of infection and hemorrhage [11]. In some patients, the risks associated with monitoring are outweighed by the benefits of ICP- and CPP-guided therapy, but patient selection is critical. Alternatively, noninvasive intracranial pressure (nICP) estimation is less risky and could both inform patient selection and timing for monitor placement (eg, early for those who are predicted to benefit). It may also be paired with invasive ICP monitoring as a powerful clinical decision support tool. Methods of nICP estimation generally involve identifying relationships between ICP and proxies that may be more easily observable in real time. These relationships may be explored empirically or on the basis of explicit models representing underlying physiology; a recent comprehensive survey of nICP estimation modalities is available [12].

Data and Clinical Availability

Estimation of ICP using models and/or proxy data is highly dependent on the availability of specific data, which limits its use. For example, nICP may be statistically estimated from ABP and concurrent measurements of CBF velocity or cerebral oxygenation via empirical relationships [13,14] or physiological models [15,16]. The collection of such data requires advanced techniques such as transcranial Doppler sonography or near-infrared spectroscopy, which are limited by the availability of instruments and trained technicians. These data must also typically undergo quality control, delaying their availability for nICP estimation. Although nICP may be estimated using various modalities, practical considerations such as clinical logistics and data timeliness render their applications difficult.

TBI Modeling for Decision Support

An ideal model for clinical decision support of TBI management is one that quickly provides nICP forecasts at multihour timescales from commonly available data and includes IC (as an adjective) process resolution. Such a model does not currently exist. Fast methods based on machine learning and signal processing [17-21] provide empirical nICP forecasts but rely on an abundance of training and/or patient history data that may not be widely available. Real-time models that empirically approximate physiological relationships [15,16] are also fast, but they still require uncommon data and do not provide IC mechanism resolution useful for diagnosis or patient-specific tuning. Mechanistic modeling approaches [22-24] emphasize either broad systemic dynamics or short-time resolution of IC processes and may be too coarse or slow for the purpose of clinical nICP estimation.

Two recent models [15,16,23,24] have been cited extensively in this document. The more anatomically representative model of Ryu et al [23] estimates nICP from ABP without additional data but emphasizes pulse-scale pressure signals rather than hour-scale dynamics. The fast nICP estimation schemes of Kashif et al [15] track ICP at suitable multihour timescales but have stringent requirements for uncommon data, which limits their applicability. Although contrapuntal to one another, both models are foundational to this study, which focuses on the limits and extension of these methodologies for long-time nICP estimation from ABP.

Objectives of This Paper

The different methodologies of Kashif et al [15] and Ryu et al [23] present two feasible options for nICP estimation: full systemic integration of the former's simple model with a systemic hemodynamics model or unintegrated use of the more complex model of the latter. This investigation considers which model development strategy is a better initial step toward an ideal nICP estimation tool in a clinical setting. We present the advantages and disadvantages of each approach: to better inform the development of a tool representative of the ideal model, and to identify the input requirements for each model in relation to clinically available data.

This study has three primary objectives. The first is to extend the simplified nICP estimation framework [15,16] by using a coupled arterial vasculature model to eliminate its dependence on jointly measured CBF. The second is to evaluate the ABP-only simulation of this model and the one developed by Ryu et al [23] for nICP estimation over a duration of hours. The third goal is to clarify the additional model machinery, such as case-specific parameter estimation and inference, needed to implement nICP estimation for complex, clinically important situations. These goals aim to inform the development of a practical tool capable of providing timely support in the clinical decision-making process for TBI patients on a broader timescale than those considered in the literature.

The remainder of this paper is organized as follows: the *Methods* section presents the models and methods of investigation, describes the model experiments, and establishes the model assessment criterion; the *Results* section presents the results of

the experiments and model comparison and discusses the simulations of cases involving other brain injuries such as ischemic stroke, which are poorly simulated without optimization; and the *Discussion* section summarizes the analysis and motivates ongoing work toward modeling nICP estimation in a particular direction on the basis of the results and implications.

Methods

Overview

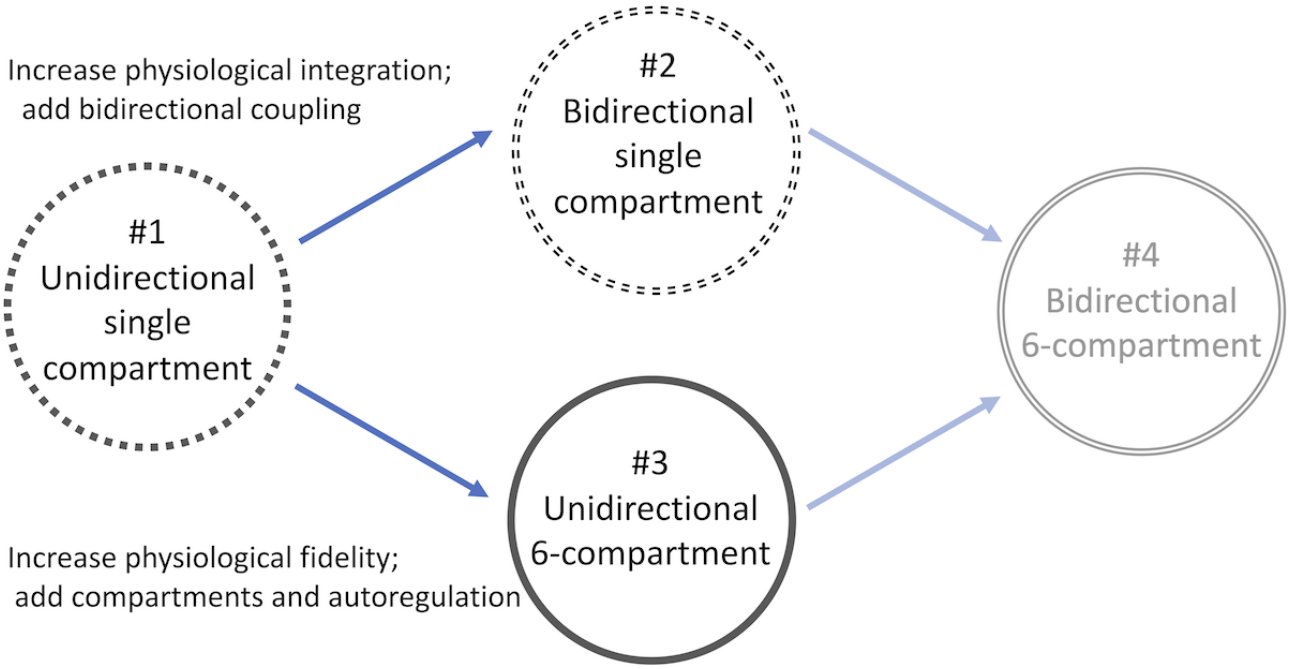
The comparison of nICP estimation schemes involves three essential parts: model configurations, aortic inflow data that drive the system, and metrics used to compare models on the basis of various aspects of performance, which are presented in the following subsections.

Numerical nICP Estimation Frameworks

Model Components

The models considered here are algorithms that transform aortic ABP data into nICP estimates using two components that may be coupled or independent. The first component is a vascular hemodynamics model that distributes ABP forcing through the systemic arterial network (AN) to the anatomical Circle of Willis (CoW), and is referred to as AN-CoW. The second component, referred to as the intracranial model (ICM), estimates nICP estimates using the outflow of the AN-CoW at the cranial arteries. We evaluated ICMs that either considered the cerebral perfusion system as a single compartment or as 6 interacting compartments defined by flow distributions of the anterior, middle, and posterior cerebral arteries. These compartments correspond to unresolved cerebrovascular territories perfused by the cerebral arteries [23], and these ICMs therefore differ in anatomical fidelity. The considered model formulations are differentiated by whether they interact unidirectionally or bidirectionally with the AN and by the complexity of the ICM component. The possible configurations are shown in Figure 1. In the unidirectional configurations, the AN-CoW boundary outflow at the middle cerebral artery (MCA) was prescribed to the ICM as an inflow boundary condition. The AN-CoW calculates this pressure and flow for the entire simulation, which is then applied to the ICM. Bidirectional coupling of the AN-CoW and ICM enforces interactive agreement of flow volumes and pressures at the interface of the components (enforced as conservation of current and voltage).

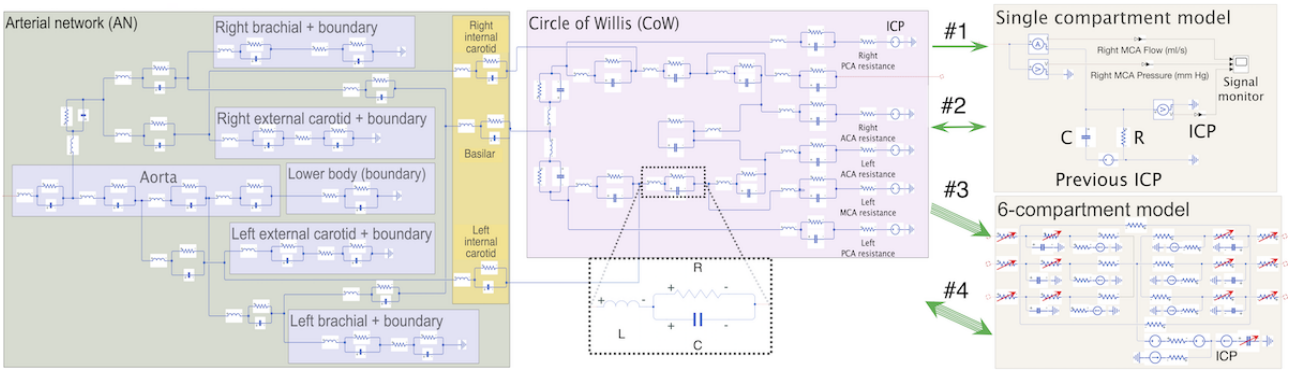
Figure 1. Conceptual overview of the relation among 4 models. The single-compartment model forced by prescribed hemodynamic time series (model #1) is the baseline model for comparison. Model #2 bidirectionally integrates the lower arterial network with the single-component intracranial model. In contrast, model #3 uses a more complex 6-compartment intracranial model with prescribed hemodynamic forcing. Model #4 represents of a multicompartment intracranial model fully integrated with the systemic arteries.



Two directions for refining the base model are proposed as possible steps toward achieving a preferred but demanding model. Figure 1 shows the relationships of the models using model #1 as the most basic form and models #2 and #3 as parallel steps toward ideal model #4. Models #2 and #3 extend model #1 either by a bidirectionally coupled interface between the AN-CoW and ICM or by increasing the physiological complexity of the ICM component, respectively. This perspective also tests which choice yields the highest gain in improvement over model #1 and the cost of implementing it. Model #4 reflects the ultimate goal of a fully integrated bidirectional model featuring an anatomically accurate ICM. However, such a model is not presented here because of its difficult implementation and impractical computational cost for the simulation timescales considered. Bidirectional coupling is difficult for multicompartment models because of the codependency of the ICM state and the common pressure at each CoW terminal interface. Solving the ICM state equations at each time step requires several iterations, and each iterate requires recalculation of the entire upstream AN-CoW system

constrained by pressure equality among the interfaces. The modeling framework used in this study is shown in Figure 2.

Figure 2. Diagram of model configurations 1–4. Schematic view of the various model configurations where green and pink boxes identify the AN and Circle of Willis vascular components, respectively, and intracranial models at right. Purple and orange boxes in the AN identify represented anatomy for reference. The vascular component is structured as in the source studies but uses 3-element electrical representations of each vessel, shown in the dashed white box. The single-compartment intracranial model is shown in the upper tan box; below it is a conceptual illustration of the 6-compartment model where red arrows indicate variable state components related to autoregulation and adaptive capacity. Unidirectional and bidirectional green arrows indicate the type of coupling between vascular and intracranial model components to distinguish configurations #1-4. ACA: anterior cerebral artery; ICP: intracranial pressure; MCA: middle cerebral artery; PCA: posterior cerebral artery.



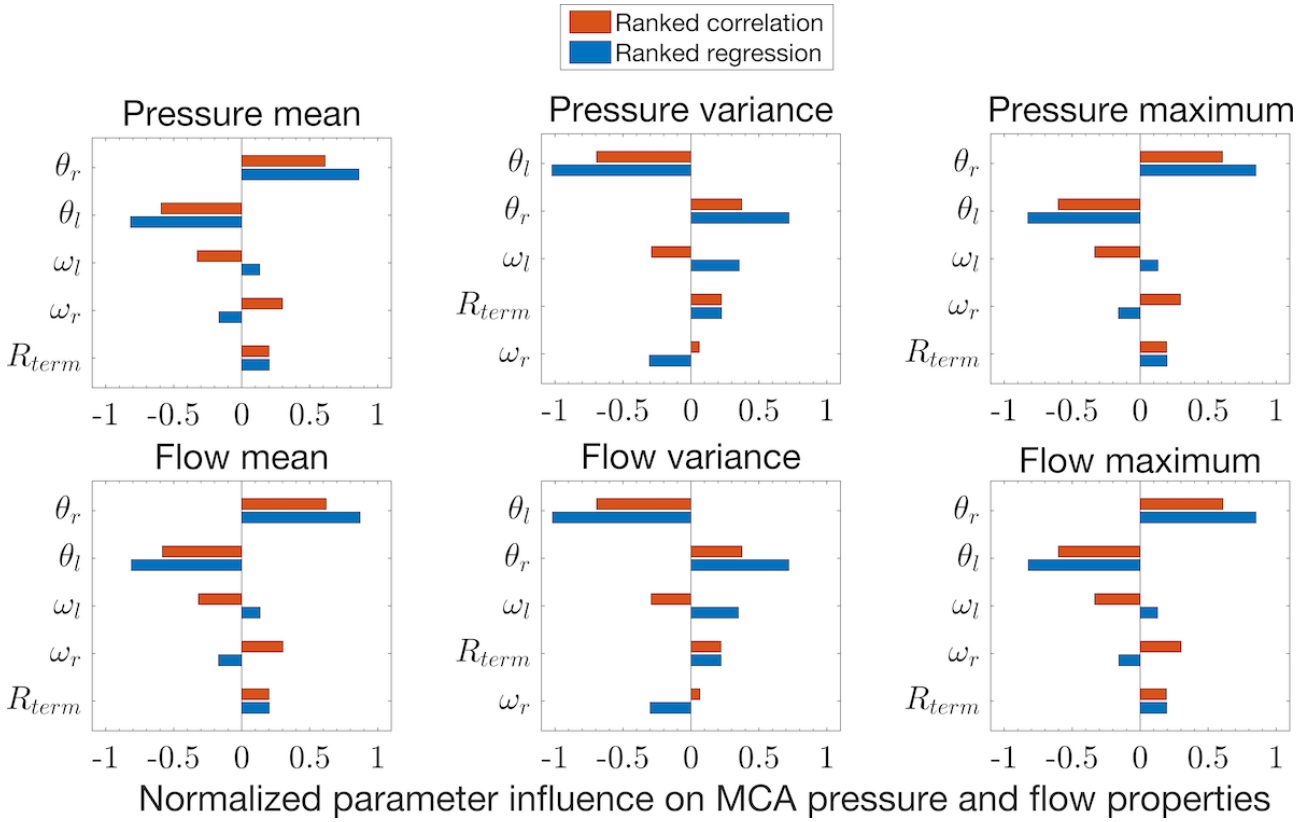
Each model comprises two separate model components, which are described below. The AN-CoW for resolving hemodynamics outside the cerebral territories is presented in the *Hemodynamical Modeling of Subcranial Arteries* subsection, whereas the ICMs for estimating ICP are presented in the *ICP Model Components* subsection.

Hemodynamical Modeling of Subcranial Arteries

Figure 2 depicts the AN-CoW model component, which comprises a subcranial AN (green box) and CoW vessels (pink box), as part of the modeling framework. As the spatial resolution of vessels is unnecessary, AN-CoW is modeled by a zero-dimensional framework of electrical analogs [25,26]. Each of the constituent 33 vessels was represented using a 3-element electrical analog (white inset box). This so-called lumped parameter approach has several advantages, including a relatively small number of patient-specific parameters. Furthermore, conservation laws at vessel interfaces reduce at each time step to algebraic systems rather than high-dimensional nonlinear functional representations [27] when spatially resolved.

Vascular network parameters total more than 100 but may be approximated by physically consistent functions of vessel length l and radius r [28]. A simple assumption of uniform dimensional scaling among the AN vessels is also applied to 3-element Windkessel boundaries and to the terminal resistances at CoW outflows. As CoW and adjacent vessel radii are approximately adult-sized by approximately 5 years of age [29], we did not scale vessels within the CoW model component. This reduces the large number of model parameters to only five effective parameters describing the scaling factors (proportions) of the base model values, which were adopted from a previous study [23] and references therein. This nonlinear reparametrization simplifies the AN-CoW component identification and is effective within realistic ranges of parameter values, as shown in Figure 3. Further details of the component definition, parametrization, boundaries, and sensitivity analysis are provided in Multimedia Appendix 1.

Figure 3. Ranked sensitivities of arterial network scaling parameters. Normalized empirical estimates of sensitivity ranking, shown here for key signal features (mean, variance, and maximum) of pressure (top row) and flow (bottom row) in the middle cerebral artery, summarize Monte Carlo experiments using global structured random uniform variations of scaling parameters (vertical axis of each panel). Parameter variations in vessel length (θ_l) and radius (θ_r) are most influential, whereas resistance scale (R_{term}) and Windkessel scales (ω_l, ω_r) had relatively little impact on the solutions. The vessel dimension parameters have considerable influence on intracranial model inflow signals and provide global control while reducing the number of parameters needed to specify the hemodynamic model. MCA: middle cerebral artery.



ICP Model Components

The ICM component is responsible for estimating nICP from the AN-CoW outflow to the cerebral arteries. The two ICM configurations considered are a 6-compartment model [23,30] and a single-compartment model [15,16], where each compartment represents a vascular perfusion territory. In addition to the number of represented cerebral perfusion territories, the models differ in their estimation approaches. The multicompartment model is more anatomically accurate and explicitly resolves IC hemodynamics with communicating arteries and autonomic pressure regulatory processes. In contrast, the single-compartment approach computes ICP using window-based statistical estimates of IC compliance and pressure determined through regression of the ICM inflow waveform properties. An overview of the multi- and single-compartment ICMs is presented in the following subsections.

Overview of the 6-Compartment Model

The complex model of Hu et al [30] and Ryu et al [23] presents an anatomical layout of the main cerebral pathways and their dependent mechanisms. Using six interacting territories, the model includes IC pressure and perfusion dynamics coupled by communicating arteries, dynamic autoregulation, and CSF balance. The autoregulatory processes are modeled as internal feedback mechanisms that regulate compartmental flow toward target values by controlling vessel radii [31]. This autonomic control influences the local pressure and flow balances between compartments, leading to intercompartmental blood flow via the communicating arteries. IC pressure and compliance are nonlinearly codetermined by volume changes resulting from autoregulation and net fluid change. The high degree of physiological fidelity resolves the IC dynamics at timescales inherited from ABP forcing. Furthermore, the 6-compartment nonlinear nICP component calculated numerous potentially clinically relevant diagnostic variables during the simulation. Unlike the source model, our implementation (model #3) is informed by the arterial inflow pressure and flow rate but does not provide feedback on systemic hemodynamics. A mathematical description, including a table of physiological and model parameters, is provided in Multimedia Appendix 2.

Overview of the Single-Compartment Model

The single-compartment ICM of Kashif et al [15] is a simple model that estimates ICP physiologically rather than anatomically modeling it. Here, nICP is constructed from linear regression estimates of bulk IC compliance (C) and resistance (R) over a temporal window containing several cardiac cycles. The algorithm estimates compliance C and resistance R by identifying the statistical relationships within a lumped parameter model representing IC physiology (details in Multimedia Appendix 3). These estimates and local ICP are related to MCA inflow and its applied pressure signal, from which nICP is deduced algebraically under the assumption of stationary parameter values.

The estimation process of this ICM requires no physiological parameters but requires algorithmic parameters that influence model behavior. Two required model hyper-parameters control the length of the temporal window over which each estimation occurs and the time step of the parameter updates. The first is limited by the stationarity assumption and determines the sample size for the regressions, whereas the second controls the output temporal resolution and coupling strength. Under bidirectional coupling, our implementation defines nICP as the simulated forecast based on the previous values of nICP and resistance *R*. These latter quantities were fixed in unidirectional coupling setups. Therefore, the length of the update time step affects the temporal coarseness of the nICP estimate in each model and defines the timescale of feedback between the ICM and upstream vascular model in the bidirectional model. Single-compartment model simulations use 1-minute windows and 1-minute updates, unless otherwise specified.

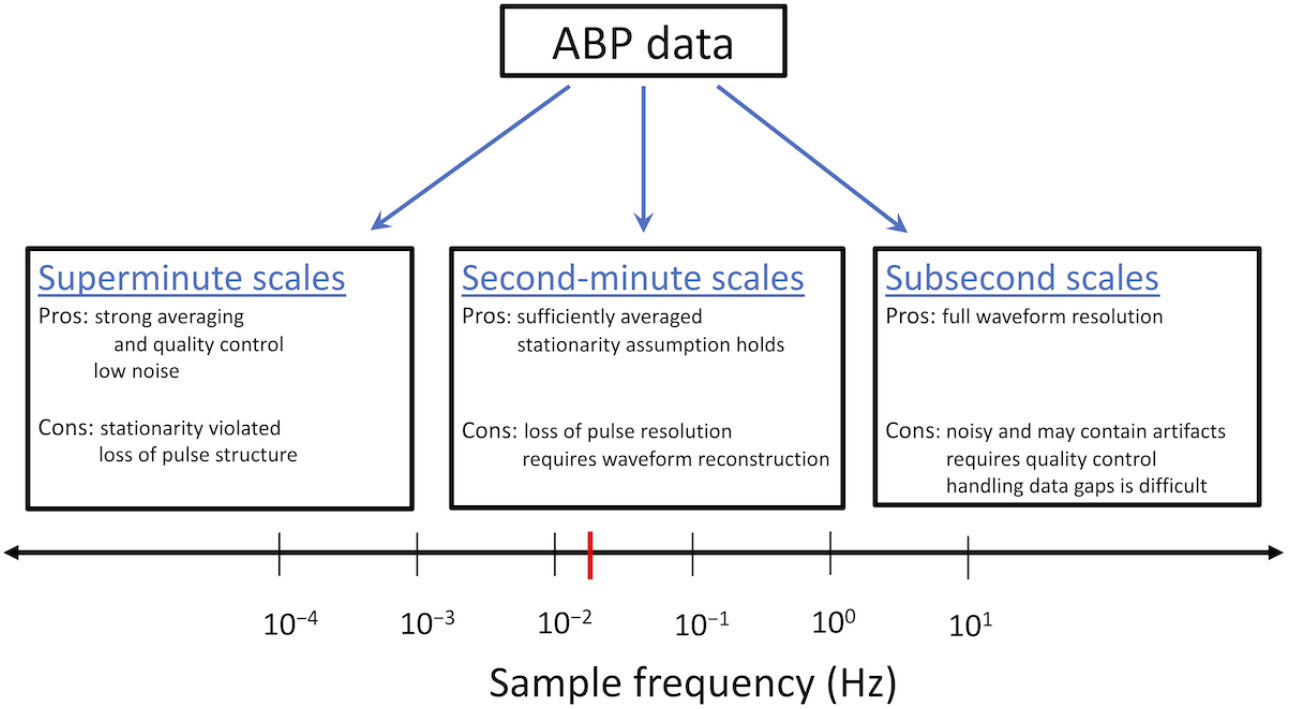
Observational Data and Patient Selection

The CHARIS v1.0.0 collection (Charis hereafter [32]), publicly available from PhysioNet [33], comprises 50 Hz joint radial ABP and ICP time series of 13 patients. These data satisfy the model requirements, including documentation of diagnosed IC injuries, and suffice for model input and evaluation. Using radial ABP data as aortic introduces biases against systolic pressure more than diastolic [34,35], and these errors are consistent among our experiments. Sophisticated transformations exist [36] for reconstructing aortic pressure from radial ABP, but the simple approach taken here avoids uncertainties associated with additional algorithmic processing.

For model comparison, this study focuses on Charis patient #6, a 20-year-old male with TBI, based on the simplicity of his injury, cleanliness of joint ABP-ICP signal, and representativeness of base parameters (eg, optimal scaling parameters for the AN-CoW were approximately 1). In addition, large-scale noise or corrupt signals are common in the records of the patients (Multimedia Appendix 4); most of their ABP and/or ICP data could not be used contiguously for 4- to 6-hour periods without extensive and uncertain preprocessing of the available data.

Figure 4 identifies the possible sampling frequencies for the aortic model inflow. Models #1 and #2 have stricter aortic inflow requirements than models #3 and #4, as their simpler ICMs require ABP sample frequency in the rightmost portion of the scale (<10 Hz) for waveform feature identification. For example, previous studies [15,16] validated the regressive method of the simple ICM using data sampled at 20-70 Hz and 125 Hz. Such data are obtainable but are not commonly available and typically require quality control.

Figure 4. Timescales of ABP inflow data. The complex models can run on data from any part of the sampling spectrum. Simple models require pulsatile inflow from the rightmost portion of the scale (above about 10 Hz), which may not be typically available. The central scale is desirable for hour-scale applications, as this resolution both qualitatively minimizes computational overhead and supports parameter stationarity assumed in the regressive single-compartment models. The quaque 1-min data sampling frequency is indicated in red. The left-most scale offers strong smoothing and low noise but fails to resolve pulsatile waveform and violates assumptions of the simple models. ABP: arterial blood pressure.



Lower-frequency ABP time series, which are more accessible and cleaner, are assumed to comprise nonoverlapping 1-minute (quaque 1-minute [q1m]) averages of systolic and diastolic pressures and heart rate. A waveform model (defined by a superposition of beta distribution probability density functions; Multimedia Appendix 5) projects these discrete q1m data into continuous time using patient-specific waveform parameters , which are then sampled for convenience at 60 Hz. In relation to Figure 4, this process maps q1m data (identified by the red mark) into the scale usable by the simpler models to test the robustness of their data requirements.

Measures of Quality and Efficiency for Models

Each experiment was evaluated using three scores: rating error, classification accuracy, and speed for simulations over time interval $[0,T]$ in N 1-minute intervals, which quantify the desirable properties of the nICP estimates [37,38] for the purposes of relative comparison. The symbol nICP* herein indicates nICP debiased against the observed ICP during the first hour of the simulation. The justification for this correction is that skill scores evaluate the model's ability to track variability in recorded ICP data rather than estimate the absolute pressure. It also accounts for some of the bias introduced through the misuse of radial blood pressure as aortic inflow pressure. Each evaluation is applied to an nICP estimate, the score of which is then associated with the model that produces it.

The first score is the time-averaged standard error between the ICP and the debiased model estimate:

This rates the ability of the model nICP to track the observed ICP changes and quantifies the general inaccuracy of the model nICP estimate in observed units. Scaling by the simulation length allows comparison over different simulation lengths.

The second evaluation is the mean percentage of time that nICP correctly agrees with the observed criticality (ICP>20 mm Hg) during the total N-minute simulation. The measure of model accuracy is defined as:

Although more qualitative than r_1, classification accuracy may be more relevant for clinical decision support as it quantifies the coherence between the model and observed critical ICP [39].

Finally, the third quantity is simply the ratio of the simulated time interval to the elapsed clock time:

with $r_3>1$ indicating a faster-than-real-time forward model integration. The values of t_{wall} correspond to serial run times using MATLAB R2020a with a 3.7 GHz Intel i5 central processing unit. This final evaluation measures the practicality of a model for providing timely clinical support as well as its utility in other applications, such as nonlinear parameter estimation or data assimilation methods that require extensive, repeated model simulation.

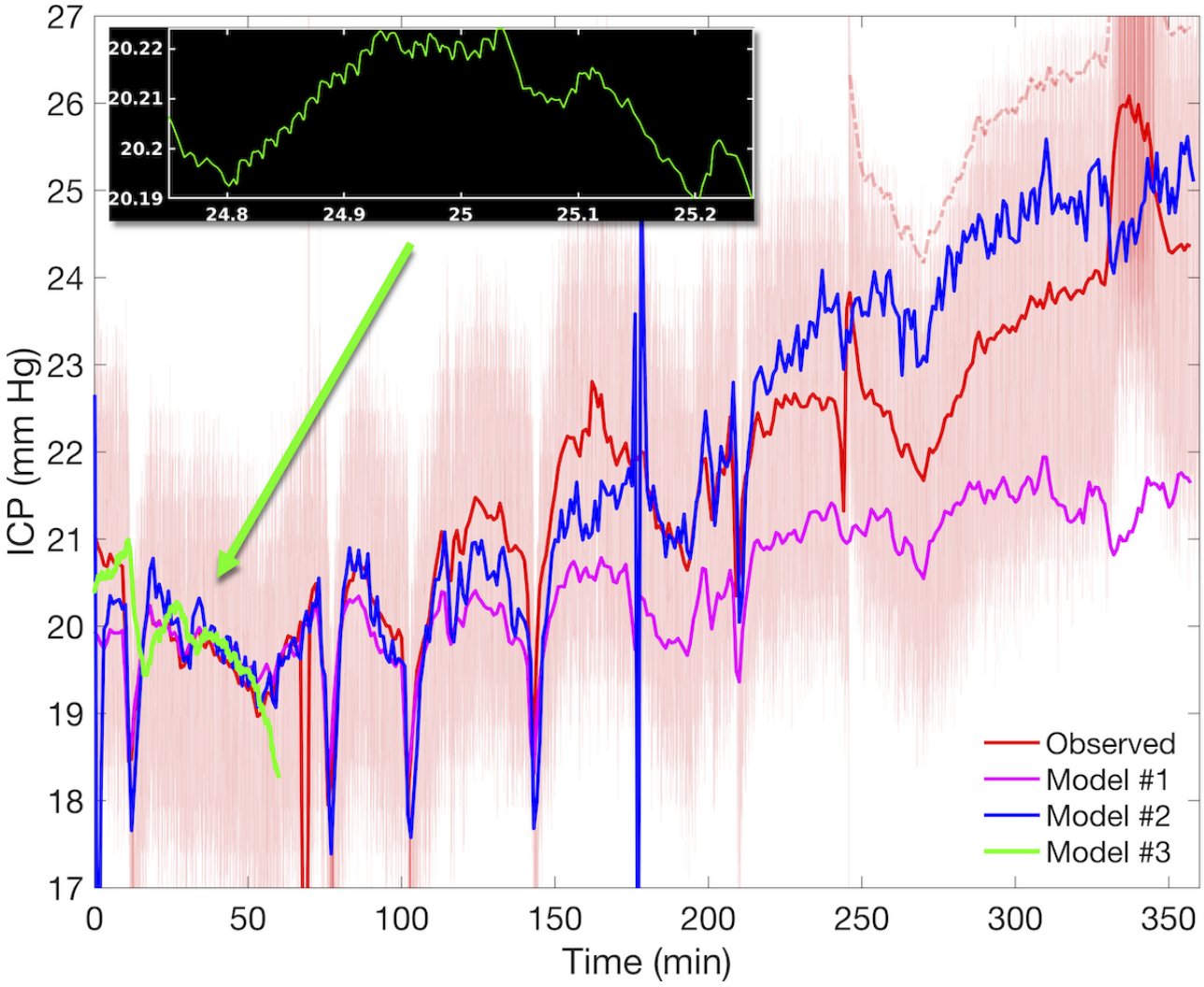
The number of necessary parameters required for realistic initialization and the input data fidelity were assessed in the context of model utility, but they were not evaluated quantitatively. Finally, all model simulations are initialized with zero flow within the AN-CoW system common to the various model configurations. A spin-up adjustment occurs in the first 2 to 3 minutes of simulation, and these errors are included in the skill calculation with negligible impact on comparative assessment.

Results

Comparative Assessment of Model Simulations

Assessment of nICP and model efficiency for the first hours of patient #6 indicates that model #2 has a lower error than model #1 and is more practical than model #3. Figure 5 shows the observed ICP signal along with the estimates from models #1 to #3. Patient ICP was initially stable near 20 mm Hg for approximately an hour and gradually increased to approximately 24 mm Hg during the final 2.5 hours. Temporary pressure drops near 10, 75, 105, and 142 minutes likely reflect interventions (eg, mannitol or hyperventilation treatments) [32]. The observed ICP signal used in the model evaluation is shown in solid red. This reference ICP is decreased by approximately 2.5 mm Hg after 243.5 minutes to compensate for the sharp 5+ mm Hg record discontinuity, which may be due to transducer recalibration. The original unaltered 1-minute average ICP observations (dashed light red) are shown for reference over the interval 244 to 360 minutes.

Figure 5. Observed and estimated noninvasive ICP for patient #6. Depicted are the observed (red) and estimated noninvasive ICP for Charis patient #6 using models #1-3, with model #2 showing the best accuracy. The noninvasive ICP estimated by model #1 (magenta) requires less than 5 minutes to run but has larger long-term errors. Model #2 (blue) takes approximately 45 minutes but produces a more accurate noninvasive ICP trend. Model #3 (green) estimates 1 hour of noninvasive ICP in approximately 6 hours of clock time; it requires variance inflation to obtain the curve shown. Model biases over the first hour are approximately 6.5 mm Hg, excluding spin-up errors. The black inset illustrates model #3 pulse resolution during a 30-second interval. Bidirectionality in model #2 has better low-frequency resolution and trend tracking than model #1, but makes it susceptible to feedback-driven instability under noisy inflow data (models #1 and #2 near 180 minutes). ICP: intracranial pressure.



Model comparison is organized into three subtopics: qualitative differences, quantitative differences, and observations about resolvable timescales and fidelity.

Qualitative Differences Between nICP Series

Models #1 and #2 produce qualitatively different pressure estimates, with the key difference being that model #2 follows the multihour trend of increasing ICP. Model #1 tracks the observations well for approximately 2 hours but fails to track the subsequent ICP elevation, as its bias falls from −1.8 mm Hg to nearly −3.2 mm Hg during 220 to 360 minutes. Model #2 tracks this observed pressure rise, although there is a roughly uniform bias of 1.02 mm Hg during this same period. One concludes that feedback from bidirectional coupling improves the estimation of low-frequency ICP signal components that are crucial in applications spanning several hours. Note that the observed 2 mm Hg pressure event (330-350 minutes) was resolved by neither model. This feature may be the result of a temporary change in patient posture, but no corresponding change occurs in the aortic ABP inflow signal (Multimedia Appendix 4, center left panel). This provides evidence that changes in ICP that do not arise from aortic ABP dynamics may not be resolved by simple ICMs.

The poorly identified parameters and long computation time hindered the simulation of model #3 for longer than 1 hour. The default ICM parameters [23] did not generate realistic ICP and required alteration of venous capillary conductance (G_{pv}) and reference pressure (P_{icn}) to obtain the reported nICP estimate. Small exploratory changes in parameter values often led to nICP divergence, indicating a strong dynamical dependence on parameters that must be inferred before useful simulation. The reported solution also includes a mean variance inflation of 26.3, which compensates for uncalibrated parameters, although the localized pulse amplitude (Figure 5, inset) is still too weak. This modified nICP estimates the observed trend well, although it lags behind the observations by approximately 4 minutes. This apparent delay, such as the reduced variance at several timescales, likely reflects poor representation by generic ICM parameters in the absence of additional inference. Attempts to

determine more accurate parameter values were limited by model speed, which is approximately 6 times slower than real time.

Quantitative Differences Between nICP Series

The qualitative advantages of the bidirectional simple model over the unidirectional complex model are borne out by model skills r_1-r_3, as shown in Table 1. Further classification metrics for this case are given in Multimedia Appendix 6. Scores for the commonly resolved first hour appear in parentheses to account for differences in the simulation period. Bidirectional coupling reduces the simple model error from approximately 5 mm Hg to approximately 3.5 mm Hg (an improvement of nearly 30%), whereas critical ICP is estimated more accurately by 9.3 percentage points (a 10.6% relative improvement). Over the first hour, there was a slight increase in model #2 error due to longer spin-up adjustment and a modest 6 percentage point improvement in critical ICP detection. The complex unidirectionally coupled model #3 shows less than 1% improvement in critical ICP identification over the base model, with mean error increasing to 3.83 mm Hg (a 57% increase relative to model #1) mostly due to the approximately 4-minute lag. Accounting for this delay reduces model #3 error to 2.75 mm Hg but also reduces accuracy to r_2=0.84; this affects neither skill ranking of model #2 over model #3. These results support that the feedback mechanism improves low-frequency tracking, which has little advantage over short timescales, and also suggests that model #2 has a practical advantage over model #3 in terms of error and accuracy.

Table 1. Model scores for principal comparison.[a]

Model	Error: r_1 6 hour (first hour)	Accuracy: r_2 6 hour (first hour)	Speed: r_3 6 hour (first hour)
Model #1	5.01 (2.42)	0.877 (0.92)	*116.129* [b]
Model #2	*3.53* (2.47)	*0.97* (0.98)	7.356
Model #3	(3.83)	(0.883)	(0.145)

[a]Scores for simulations of Charis patient #6 during initial hours of data. Scores r_1 and r_2 rate the nICP errors and accuracy in identifying critical ICP, respectively, whereas score r_3 rates the speed of the nICP estimation process. Parenthesized entries are calculated using only the first simulated hour.

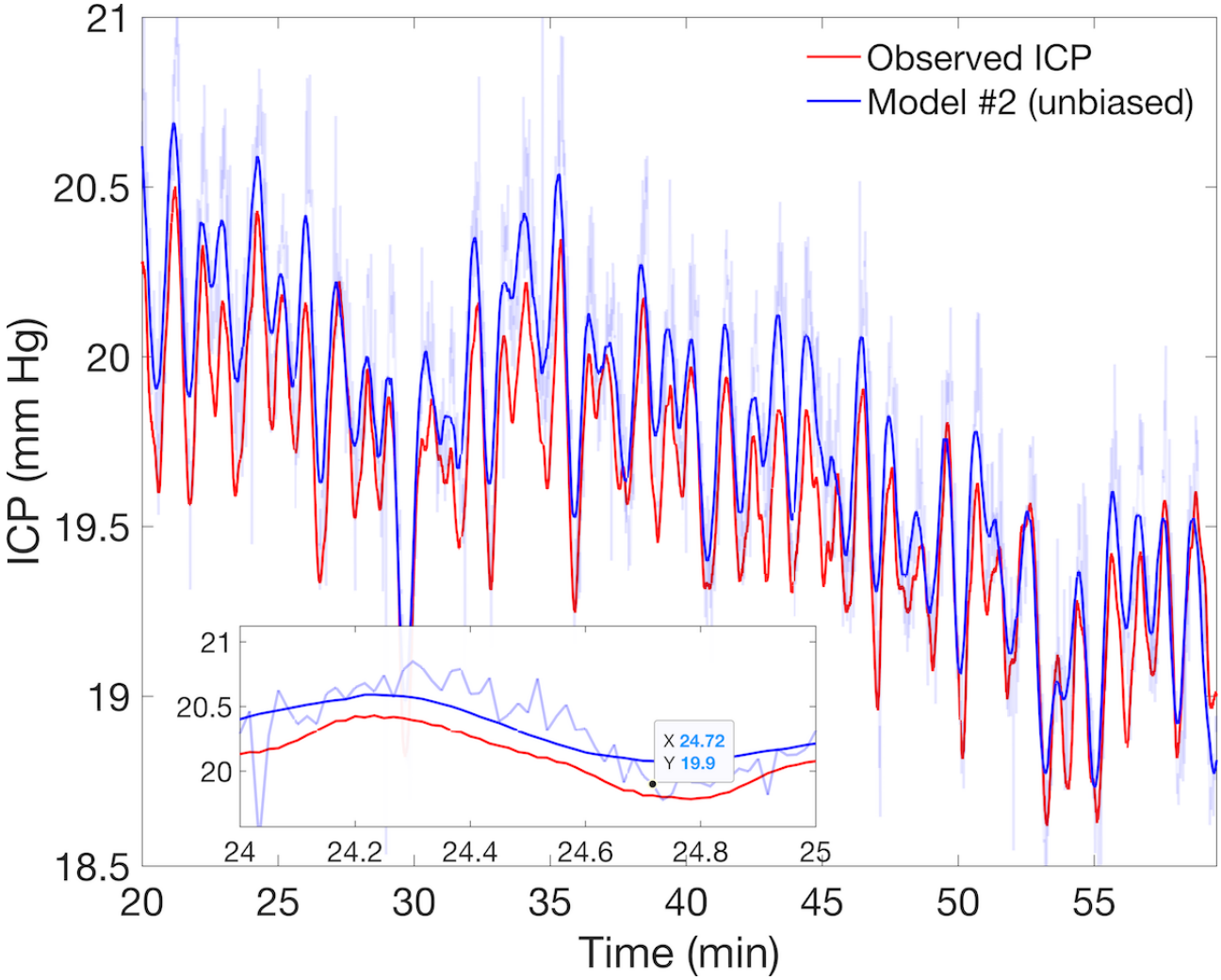
[b]Italic text indicates the best results for each score.

The most significant difference between models #2 and #3 for practical nICP estimation is in the simulation speed (measured by r_3). Both models #1 and #2 operate considerably faster than real time and are therefore suitable for an operational clinical support system. Model #3, however, is an order of magnitude slower than wall time under the same forcing and ill-suited for multihour simulation under pulsatile forcing. The speed of model #3 is limited by the calculation of many ($O(10^3)$) iterative solutions to its nonlinear ICM per cardiac cycle, primarily during systolic upswing. A previous study using this ICM [24] reported that each cardiac cycle required 40 seconds within their highly optimized numerical framework. As their implementation used a one-dimensional AN-CoW, the speed of model #4 had a lower bound of r_3=0.0225.

Resolution Versus Speed Considerations

Models also differ in their ABP data requirements, and one must consider the trade-off between the desired nICP temporal resolution and model efficiency. The complex ICM is defined by differential equations, so fine timescales inherited from pulsatile inflow boundary conditions require extensive, inflexible computation time to resolve the nICP pulse (black inset, Figure 5). The use of q1m mean ABP inflow increases model #3 speed considerably to r_3=1.15 (slightly faster than real time), and analysis suggests this should not impair the resolution of autoregulation effects, which manifest at timescales beyond 15 seconds. On the other hand, simple model hyper-parameters (window length and parameter update interval) can be adjusted to resolve higher-frequency nICP components with additional computational time. Figure 6 illustrates a model #2 simulation under raw ABP using a 30-second window and 1-second update period (ie, 29 second overlap). Additional computational overhead reduces speed (r_3=0.25 approximately), but there is considerable gain in nICP fidelity at high frequencies as well as strongly reduced error (r_1) and increased accuracy (r_2). This demonstrates the latent ability of model #2 to estimate higher-frequency components of ICP from ABP without additional ICM parameter inference, as in model #3.

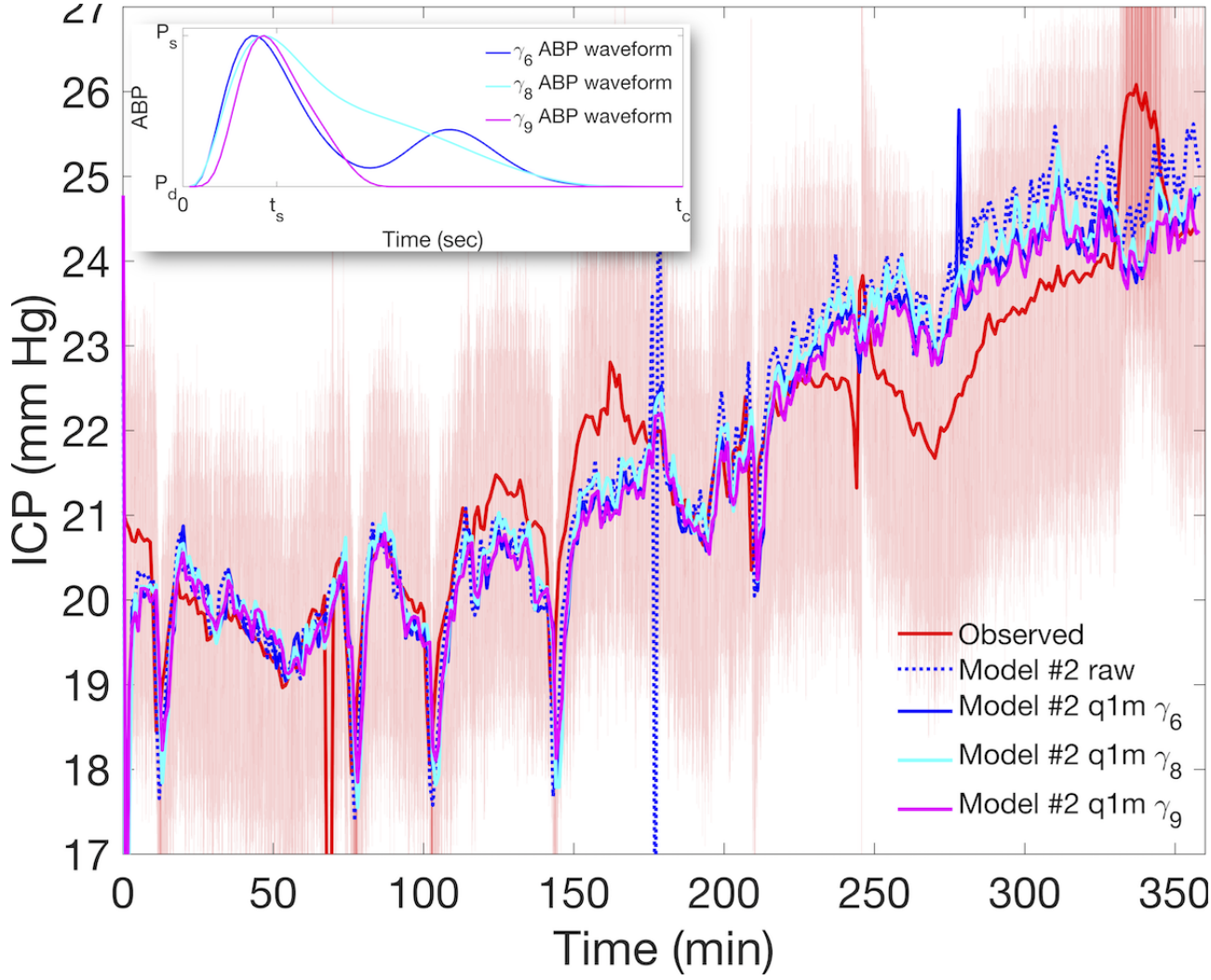
Figure 6. Strong local tracking of the ICP signal in model #2 at the expense of computational time. The mean noninvasive intracranial pressure estimates over 30-second intervals (blue curve) using the output of model #2 (light blue) with raw arterial blood pressure strongly track the observed ICP (red curve). The model simulation accurately reproduces local trends and $O(10^{-2})$ Hz waves of the averaged observed ICP. This simulation calculated resistance and compliance parameters at 1-second intervals using a 30-second moving window (ie, with a 29-second overlap). The corresponding mean ICP estimates are plotted as solid curves for comparison with the observed ICP, with an inset showing the lack of subminute resolution. Although 4 times slower than real time, this simulation is roughly twice as fast as model #3 under pulsatile aortic inflow and requires no additional data or external inference. ICP: intracranial pressure.



Simple Model Experiments With Low-Frequency Inflow Data

Models can use commonly available ABP summary records under appropriate representation without additional waveform data. The use of models #1 and #2 is limited by pulse-resolving ABP inflow, but this requirement may be weakened using waveform transformation of q1m ABP summary time series of phase pressures and heart rate (Multimedia Appendix 5). Figure 7 shows that the original nICP estimate of model #2 (blue dashed) and one using q1m inflow (γ_6, solid blue) are largely indistinguishable, although smoother inflow data of the latter avoids the instability around 175 minutes. The estimate from q1m ABP has a 3% larger error (r_1=3.7 mm Hg), although there is no qualitative difference in clinical accuracy or speed compared with the original estimate. Furthermore, the lack of patient-specific waveform parameters has little effect on simulated nICP in this model: all model scores are roughly preserved in 2 additional runs (cyan and magenta) using waveform parameters of Charis patients #8 and #9 (γ_8 and γ_9, respectively), which differ in postsystole shape (inset). However, model #2 nICP estimates based on q1m ABP without heart rate data (not shown) were highly inaccurate due to errors in the numerical calculation of the ICM inflow pressure derivative.

Figure 7. Model #2 performance using quaque 1-min (q1m) summary arterial blood pressure (ABP) data for Charis patient #6. Various simulations using q1m inflow data are compared with observed intracranial pressure (ICP; red curve) and noninvasive intracranial pressure (nICP) estimate based on raw 50 Hz data (dashed blue). Also shown are estimates using minute-wise constant continuous representatives of q1m ABP data generated by correct (blue) and incorrect (magenta and cyan) waveform parameters. The figure inset shows ABP waveform shapes for patients #6 (solid blue), #8 (cyan), and #9 (magenta), respectively, which yield qualitatively indistinguishable nICP estimates in the main plot. This shows that q1m ABP is sufficient for the aortic inflow and that patient-specific parametrization of ABP waveforms has little advantage in the simple model. ICP: intracranial pressure; P_s: systolic pressure; P_d: diastolic pressure; t_s: systolic upswing duration; t_c: cardiac cycle time.



Summary of Assessments and Experiments

Strengths and Weaknesses of the Bidirectionally Coupled Simple Model Approach

The model comparison suggests that bidirectional coupling strengthens the resolution of low-frequency nICP trends, which are crucial in multihour simulations, and improves the critical nICP classification accuracy by approximately 10%. Temporal estimation of $O(10^{-2})$ Hz ICP features is possible with no additional ICM parameter inference but requires additional computation time. Bidirectional coupling makes the model more prone to potential instabilities during spin-up and in the presence of noisy ABP inflow data. Using waveform projections of q1m summary ABP data as inflow data neither decreases nICP estimate quality nor requires patient-specific waveform parameterization, which both broadens applicability and decreases inflow noise. The simple model framework is still limited by its lack of internal process resolution and primarily responds to temporal variations in applied aortic inflow, but the fully coupled simple model approach is an order of magnitude faster than the clock time. Therefore, it has sufficient computational headroom to incorporate a more physiologically complex ICM (eg, [40,41]) and is still faster than real time.

Strengths and Weaknesses of the Unidirectionally Coupled Complex Model Approach

Increasing model complexity by resolving multiple interconnected IC compartments and autoregulatory feedback mechanisms offers physiological fidelity at the expense of strong parameter dependence and lengthy calculation time. Poor identifiability of dynamically balanced and representative ICM parameters required ad hoc nICP adjustment to obtain realistic results, but these were insufficient for multihour simulations. Additional inferences and/or data are required for practical applications. The nICP estimates, subject to additional posterior modification, qualitatively matched the observed ICP and still lacked realistic ICP pulse amplitude, as previously noted by Wang et al [24]. Although this type of resolution has a high

clinical diagnostic value [42-44], it is too computationally expensive for simulations of multiple hours. The use of nonpulsatile (eg, mean) ABP inflow increases computational overhead significantly, but model utility is still precluded by the need for ICM parameter estimation. Further exploration of model #3 is needed to evaluate the clinical diagnostic value of simulations driven by mean ABP over multihour periods.

Discussion

Summary

This study compared multicomponent modeling approaches to nICP estimation using commonly available data over multihour timescales to produce actionable clinical information. The purpose was to better inform the direction of estimation development by identifying the advantages, limits, and additional requirements of the 2 options. The choices were to integrate a simple ICM into a systemic hemodynamic model or to unidirectionally couple a more complex ICM to the hemodynamic model component. We assessed these methods based on error (r_1), clinical accuracy (r_2), and speed (r_3) of their estimates as well as on their dependence on data and parameter identification. The first key result is that the bidirectional coupling of the simple model is sufficiently fast and potentially accurate and can be implemented using commonly available q1m ABP data without patient-specific waveforms. Specifically, analysis of model performance during a slow ICH event revealed that inclusion of bidirectional coupling improved the low-frequency model resolution of ICP, improving estimation quality while remaining an order of magnitude faster than real time. The second main result is that the complex model approach is too slow for use in the targeted applications. In particular, model #3 required nearly 6 hours to perform a 1-hour simulation along with ad hoc changes to both input parameters and output solution, which can only be eliminated by parameter estimation from additional input data at additional computation time. Limited by publicly available data, the three model approaches considered here represent practical implementations of existing methods; therefore, this study is a comparison of existing models implemented in a typical, sparse data environment.

The stronger-performing simple model approach may use ABP summary data without patient specificity of the inflow waveform and is able to resolve minute-scale nICP variations at additional costs. Its ICM, originally designed to run on high-frequency joint ABP-CBF samples, was coupled to a hemodynamic model of upstream vasculature derived from the complex model to establish ABP-only data dependence. Our experiments show that simple model data dependence can be further reduced to coarse clinical summary data of phase pressures and heart rate, which is independent of the patient-specific postsystole waveform shape. The use of q1m summary data also serves to filter the aortic forcing, which is an important consideration given that the bidirectional setup is more prone to feedback instabilities originating from inflow noise. Furthermore, summary ABPs are less noisy and therefore reduce spurious feedback instabilities in fully coupled simple models (Figure 7 near 175 minutes).

Slow model speed and the need for ICM parameter identification limit the utility of the complex model. The estimation of nICP under model #3 is an order of magnitude slower than the clock time under pulsatile aortic inflow and is only slightly faster than real time under mean ABP. In both cases, strong parameter dependence renders model initialization difficult, and nICP estimates are inaccurate without posterior modification. Some ICM parameters may not be stationary over multihour timescales and may explain the difficulty in maintaining nondivergent behavior beyond the first hour of simulation. The inference necessary to identify these parameters results in additional computational overhead, making near real-time estimation an unrealistic expectation. However, these parameters provide extremely useful diagnostic information and make complex model estimations more suitable for retrospective analysis rather than operational support.

The main results of this work are summarized below:

1. The inclusion of feedback between ICM and AN-CoW components improves the tracking of higher-order trends over multihour timescales. The bidirectionally coupled single-compartment model #2 features a more accurate resolution of low-frequency ICP components than the unidirectionally coupled model at a lower computational cost than model #3.
2. The nICP estimates using q1m ABP data projected onto pulsatile waveforms are qualitatively similar to those obtained using high-frequency APB data. However, q1m summary data must include the heart rate in addition to diastolic and systolic pressures. This result broadens the applicability of simple models, as summary ABP data are more commonly and promptly available in a clinical setting.
3. Patient-specific waveforms are *not* required to use q1m ABP as simple model inflow data; the quality of nICP depends neither numerically nor empirically on resolving postsystole components of patient waveforms. Therefore, simple models do not require supplemental waveform-resolving data to use the q1m summary ABP.
4. Model #2 has a stronger potential for multihour applications because it does not require any parameters, can be run using commonly available data, and runs approximately seven times faster than real time. This makes it a suitable base for ongoing development, even if additional inference or control is required for practical use.
5. The large number of parameters within the complex, nonlinear ICM of model #3 experience difficult identifiability, and poorly specified parameters lead to divergent or unrealistic behavior. It cannot be adequately configured from available data for stable, multihour simulations and performs significantly slower than real time. This model requires sophisticated inference because its parameters, some of which may be nonstationary, need to be accurately specified.
6. The temporal resolution of model #3 was inherited from aortic inflow. With pulsatile inflow, nICP waveforms are resolved and data-optimized results can be used to characterize autoregulatory and adaptive capacity in retrospective studies. Quasi-operational nICP estimation is possible with a significant a priori investment of time for

parameter estimation but only under nonpulsatile forcing where the nICP pulse is not resolved.

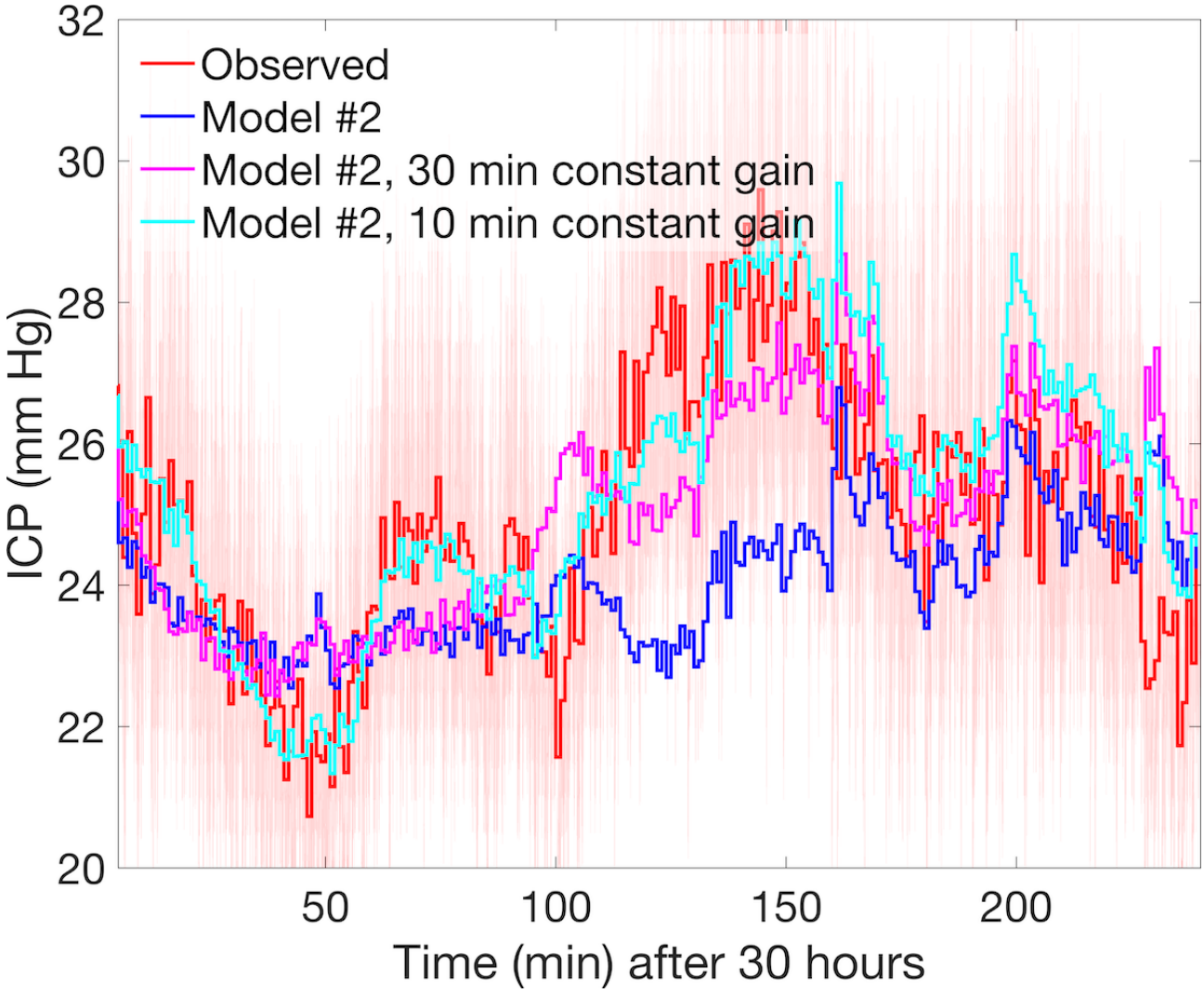
Overcoming Model Limitations

Refinement and Assimilation

The presented models have inherent limitations that are not fully realized, and a combination of parameter inference and/or data assimilation together with model improvements are necessary to meaningfully simulate clinically important scenarios. The need for accurate parameters in model #3 is evident, and the slow model speed retards this process. Although the simple bidirectional model (#2) is a strong candidate to build upon, it fails to accurately track the ICP trend and variability of patients with IC hemorrhage or stroke. The presence of raw ABP noise and large waveform variance may also play a confounding role in this limitation. However, failure likely results from omitting CBF data, which are independent of ABP, as well as the lack of parameters and simplified mechanism in the ICM that may not account for underlying IC physiological changes. For example, models #1 and #2 do not parameterize IC volume or impose upper bounds on IC compliance to reflect the thresholds of cerebral autoregulatory processes or other exhausted adaptability. Further limitations of all models include inability to account for many important aspects crucial to the clinical decision-making process, including patient age and other diagnoses; injury mechanism; imaging findings; or treatments such as sedation, neuromuscular blockade, osmolar therapy, and ventilation strategy.

Overcoming model #2 limitations to estimate nICP for some patients may require a more complex ICM or inclusion of additional dynamically controlled parameters. Many patients of clinical concern, like other Charis patients, have more complicated injuries, and their observed ICP occupies different dynamical regimes than those of patient #6 discussed above. For example, a critical hypertensive period is evident for patient #5, a 21-year-old female with TBI and identified subdural hematoma, whose ICP increased from 21 mm Hg to 29 mm Hg over a 47-minute period (Figure 8, red line) before gradually subsiding. For this patient, the local variability of q1m ICP relative to its 11-minute moving average is about 4 times larger than that of patient #6 (Multimedia Appendix 4). This increased variability is also present in the observed ABP serving as model inflow and may confound both the accuracy and stability of the model. The estimation here benefits from optimized scaling parameters, but additional machinery is necessary to drive model dynamics beyond its inherent ability to predict nICP from ABP. For the example above, the model #2 solution (blue line), using an optimized set of vascular parameters (θ_l, θ_r, ω_l, ω_r, R_{term})=(0.8, 1.0, 0.84, 0.93, 1.0) fails to follow the observed dynamics during the central ICH event and sequence of waves leading up to it. Two possible directions for ongoing research—increased fidelity and external parametric control—to improve the performance within the modeling framework are presented.

Figure 8. The ICP record for Charis patient #5 during hours 30–34 is shown in light red with its minute-to-minute mean traced in dark red. The observed signal includes stronger signal noise and high-frequency variability than that of patient #6. Slow wave pressure dynamics are observed, but they are absent from the model #2 solution (blue curve), which fails to track the rise and peak of the 7 mm Hg intracranial hypertensive event observed over 100–180 min. The solution using external inflow control specified at 10-minute intervals (cyan curve, using 24 independent parameters) features greatly improved trend tracking during these more dynamic regimes than the solution using parameters specified at 30-minute intervals (magenta curve, using 8 independent parameters). ICP: intracranial pressure.



Increased Sophistication

A simple model of increased complexity may account for changes in ICP arising from IC mechanisms, widening the applicability of the framework of model #2. To broaden the scope of potentially modelable cases, other lumped parameter ICMs that offer both increased physiological fidelity and low computational overhead may be considered.

In particular, Ursino and Lodi [41] and Czosnyka and Pickard [43] presented two simple models that offer increased IC process resolution and relevant internal parametrization. Both are directly representable within the electrical analog framework electrical circuit forms [45] and account for elements of autoregulation, varying volumes, and other pressure sources. Either may easily fit bidirectionally within the existing framework as alternate ICM components with sufficiently fast algorithms for the predictive desire discussed above. These models, specifically variations thereof, using the statistical simplification of Kashif et al [15], are part of continuing development within the general purview of this research.

Additional Parametrization

Another method of applying the existing simple model #2 to complex cases involves augmented boundary control as a proxy for unresolved processes within a statistical parameter estimation scheme. Although patient-specific optimization is beyond the scope of this study, additional experiments applying the model to ABP-ICP time series of interest show that model #2 is sufficiently robust to track ICP throughout these complex regimes. This requires the addition of modulation of the relationship between ABP inflow at the aorta and the ICM inflow from the MCA using a low-frequency nonstationary gain parameter G to vary ABP inflow: $\mathrm{ABP}(t) \leftarrow \mathrm{ABP}(t)\cdot(1+G(t))$. Figure 8 illustrates the potential of this approach by including 2 additional model #2 simulations using 8 and 24 equally spaced control parameters that define G piecewise to linearly vary the ABP inflow signal.

The simulation using eight additional parameters (cyan curve) is more dynamic than the base model; it resolves a portion of the central ICH event and decreases the mean error (r_1) by more

than 20% (from 8.75 mm Hg to 6.844 mm Hg) but misses its onset and underestimates peak pressure by approximately 1.5 mm Hg. Using 24 additional parameters (magenta curve) further improved this result, improving r_1 by 36% (to 5.633 mm Hg) and identifying the rising trend during the ICH onset as well as its maximum pressure. Determining the values of gain G involves placing the current ABP-to-nICP model into a data assimilation system, which provides a meaningful way of automatically constraining uncertainties due to inaccurate parameters and unresolved physiology. Such systems require extensive computational overhead, although some methods such as empirical (ie, ensemble) Kalman-type methods maintain operational estimation of faster-than-real-time models via parallelization. Practical applications require estimation of the parameters defining G, although they were specified a priori in this illustration, but underscore the need for simple, fast models to meet the goal of providing timely, relevant nICP estimation over multihour timescales.

Forecast Potential for Clinical Support

Bidirectional model #2 provides a basis for analyzing latent empirical relationships among patient signals and model parameters, including trends and covariances, which may be used to predict patient ICP changes. Such a system would greatly benefit both clinical decision support and care-level logistics by indicating possible changes in patient status with sufficient lead time to adjust room, equipment, and staff. This may also give practitioners advance warning with a timeframe for planning treatments, permitting earlier and lower-risk interventions to combat IC hypertension. Recent works [46-49] include machine learning approaches to ABP prediction and could be used in conjunction with the presented methods for short-term prediction of nICP. The application of these algorithms to low-sample-rate q1m ABP records has not been reported in the literature.

The speed of model #2 indicates that it is a plausible candidate for use within a statistical estimation and forecast scheme that requires many forward model integrations. The accurately identified parameters, together with acceptable simulation speed, add the possibility of practical forecast capabilities based on trends in diagnostically computed model parameters. For the applications discussed in this work, distributional trends and higher-order moments in ICM resistance and compliance may be inferred from robustly optimized model #2 simulations of a patient's relevant history. This statistical information may then be used to predict possible future ICP outcomes under current ABP measurements or ABP forecasts, potentially providing valuable and timely clinical decision support for caretakers and facility management.

Conclusions and Ongoing Work

This study identified the distinct advantages and disadvantages of the 2 paths within a modeling framework and clarified the applicability of each. Although model #2 was more successfully validated at multihour timescales, it required uninterpretable control parameters (G) in more complex cases. In contrast, the ICM of model #3 is highly parameter dependent and difficult to identify from accessible data, even for simple cases. These results ultimately motivate the development of a hybrid approach that strategically combines simplifications of the mechanistically resolved processes of model #3 with the speed advantages of locally stationary parameters in model #2. The desire to have an appropriate number of physiologically interpretable parameters for data-optimized modeling contextualizes the problem as one of mechanistic machine learning [50-52].

Our preliminary hypothesis of this work was that the high degree of anatomical fidelity offered by the complex multicompartment model would provide the most diagnostic information from available data. It also had numerous model parameters that could be inferred from patient data in the longer view of the research program, which is to aid in patient-specific clinical support. We pursued an implementation of model #4 using the spatially-resolved vascular system and complex ICM [23], which had recently been used within a data assimilation system [24]. Concern for speed motivated the elimination of the spatial resolution of vessels within the hemodynamic model by adopting the 0D electrical framework, but this approach could not be easily bidirectionally coupled to the analytical ICM. It remained unidirectionally coupled and became model #3. In contrast, the simple model (#1, [15]) was easily integrated bidirectionally into the AN-CoW system, becoming model #2, and this eliminated its dependence on localized CBF data. This fully incorporated model had better tracking of lower-frequency trends in ICP and could resolve higher-frequency ICP waves with additional computational cost, and importantly, it did not require the additional parameter identification of the offline complex ICM for simple cases. However, the lack of sophistication and parametrization in models #1 and #2 is the reason why external parameters for additional control are required for more complex patient cases.

Although the need for additional inference is clear for the application of models #2 and #3, there are substantive differences in methodology and potential benefits. Namely, simple model #2 is easily identified but is limited to applications where a strong correlation between systemic ABP and ICP response is present. This lack of internal parameters necessitates the use of nonstationary external controls for application in complex cases. Although these parameters may plausibly be estimated via ensemble filtering, they are not interpretable, and the necessary mapping between clinical data and the control parameters is unknown and requires further development. In contrast, model #3 has numerous highly interpretable and diagnostically informative parameters that must be properly inferred for meaningful simulations. These are likely estimable from historic patient data using traditional methods (eg, MCMC estimation or optimization), but the value of this investment may be limited if parameters are dynamic and/or only nICP estimation is sought. Given that the estimation of nICP, rather than clinical interpretability, is the primary objective of this project, the continued development of inference machinery for model #2 is the best choice.

The long-term vision of this project remains the development of a bidirectionally coupled model with anatomical fidelity (ie, model #4) fast enough for pre-emptive diagnostic uses such as nICP forecast and the identification of pathophysiology. One path toward this goal is the hybridization of methodologies that integrate an ICM of intermediate complexity under piecewise

stationarity assumptions akin to those of simple models. Possible ICMs include those mentioned previously and a simplified (eg, linearized) counterpart of model #3. This should reduce the computational burden of the complex model and allow it to be more easily coupled interactively with the upstream vascular component. Such a model would further benefit from highly interpretable inference based on data available when administering care, with the additional advantage of supporting summary ABP inflow. A remaining question is whether a model formulated in this way can be made fast enough to provide timely and clinically actionable information.

Acknowledgments

This project was supported by US National Institutes of Health (NIH)/National Library of Medicine 5R01LM012734; by NIH/Eunice Kennedy Shriver National Institute of Child Health and Human Development 5R03HD094912; and by NIH/National Heart, Lung, and Blood Institute 5K25H133481. JNS would like to thank CU Medicine building colleagues Deepak K. Agrawal, Seth Russell, Peter DeWitt, Meg Rebull, and Chan Voong.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed vessel-level parameterization of the hemodynamic model.
[PDF File (Adobe PDF File), 111 KB - medinform_v9i3e23215_app1.pdf]

Multimedia Appendix 2
Description of the 6-compartment intracranial model, including tables of parameters and variables and a description of the iterative numerical scheme used.
[PDF File (Adobe PDF File), 137 KB - medinform_v9i3e23215_app2.pdf]

Multimedia Appendix 3
Description of the single compartment intracranial model.
[PDF File (Adobe PDF File), 99 KB - medinform_v9i3e23215_app3.pdf]

Multimedia Appendix 4
Supplemental figure.
[PNG File , 707 KB - medinform_v9i3e23215_app4.png]

Multimedia Appendix 5
Explicit description of the parametric waveform model used to transform the arterial blood pressure summary data to a continuous waveform.
[PDF File (Adobe PDF File), 79 KB - medinform_v9i3e23215_app5.pdf]

Multimedia Appendix 6
Supplemental table.
[DOCX File , 13 KB - medinform_v9i3e23215_app6.docx]

References

1. Balestreri M, Czosnyka M, Hutchinson P, Steiner LA, Hiler M, Smielewski P, et al. Impact of intracranial pressure and cerebral perfusion pressure on severe disability and mortality after head injury. Neurocritical Care 2006;4(1):8-13. [doi: 10.1385/ncc:4:1:008]
2. Adams CA, Stein DM, Morrison JJ, Scalea TM. Does intracranial pressure management hurt more than it helps in traumatic brain injury? Trauma Surg Acute Care Open 2018 Jan 12;3(1):e000142 [FREE Full text] [doi: 10.1136/tsaco-2017-000142] [Medline: 29766131]
3. Lassen NA. Cerebral blood flow and oxygen consumption in man. Physiol Rev 1959 Apr 01;39(2):183-238 [FREE Full text] [doi: 10.1152/physrev.1959.39.2.183] [Medline: 13645234]
4. Baumbach GL, Heistad DD. Regional, segmental, and temporal heterogeneity of cerebral vascular autoregulation. Ann Biomed Eng 1985 May;13(3-4):303-310. [doi: 10.1007/bf02584248]
5. Armstead WM. Cerebral blood flow autoregulation and dysautoregulation. Anesthesiol Clin 2016 Sep;34(3):465-477 [FREE Full text] [doi: 10.1016/j.anclin.2016.04.002] [Medline: 27521192]

6. Rangel-Castilla L, Gasco J, Nauta HJW, Okonkwo DO, Robertson CS. Cerebral pressure autoregulation in traumatic brain injury. Neurosurg Focus 2008 Oct;25(4):E7. [doi: 10.3171/foc.2008.25.10.e7]
7. Wilson MH. Monro-Kellie 2.0: the dynamic vascular and venous pathophysiological components of intracranial pressure. J Cereb Blood Flow Metab 2016 May 12;36(8):1338-1350. [doi: 10.1177/0271678x16648711]
8. Cardim D, Robba C, Donnelly J, Bohdanowicz M, Schmidt B, Damian M, et al. Prospective study on noninvasive assessment of intracranial pressure in traumatic brain-injured patients: comparison of four methods. J Neurotrauma 2016 Apr 15;33(8):792-802 [FREE Full text] [doi: 10.1089/neu.2015.4134] [Medline: 26414916]
9. Rosner MJ, Rosner SD, Johnson AH. Cerebral perfusion pressure: management protocol and clinical results. J Neurosurg 1995 Dec;83(6):949-962. [doi: 10.3171/jns.1995.83.6.0949] [Medline: 7490638]
10. Stocchetti N, Maas AI. Traumatic intracranial hypertension. N Engl J Med 2014 May 29;370(22):2121-2130. [doi: 10.1056/nejmra1208708]
11. Tavakoli S, Peitz W, Ares W, Hafeez S, Grandhi R. Complications of invasive intracranial pressure monitoring devices in neurocritical care. Neurosurg Focus 2017:E6. [doi: 10.3171/2017.8.focus17450]
12. Khan M, Shallwani H, Khan M, Shamim M. Noninvasive monitoring intracranial pressure - a review of available modalities. Surg Neurol Int 2017;8(1):51 [FREE Full text] [doi: 10.4103/sni.sni_403_16] [Medline: 28480113]
13. Schmidt B, Klingelhöfer J, Schwarze JJ, Sander D, Wittich I. Noninvasive prediction of intracranial pressure curves using transcranial Doppler ultrasonography and blood pressure curves. Stroke 1997 Dec;28(12):2465-2472. [doi: 10.1161/01.str.28.12.2465] [Medline: 9412634]
14. Kim MN, Durduran T, Frangos S, Edlow BL, Buckley EM, Moss HE, et al. Noninvasive measurement of cerebral blood flow and blood oxygenation using near-infrared and diffuse correlation spectroscopies in critically brain-injured adults. Neurocrit Care 2010 Apr 12;12(2):173-180 [FREE Full text] [doi: 10.1007/s12028-009-9305-x] [Medline: 19908166]
15. Kashif FM, Verghese GC, Novak V, Czosnyka M, Heldt T. Model-based noninvasive estimation of intracranial pressure from cerebral blood flow velocity and arterial pressure. Sci Transl Med 2012 Apr 11;4(129):129 [FREE Full text] [doi: 10.1126/scitranslmed.3003249] [Medline: 22496546]
16. Fanelli A, Vonberg FW, LaRovere KL, Walsh BK, Smith ER, Robinson S, et al. Fully automated, real-time, calibration-free, continuous noninvasive estimation of intracranial pressure in children. J Neurosurg Pediatr 2019 Aug 23:1-11. [doi: 10.3171/2019.5.PEDS19178] [Medline: 31443086]
17. Swiercz M, Mariak Z, Krejza J, Lewko J, Szydlik P. Intracranial pressure processing with artificial neural networks: prediction of ICP trends. Acta Neurochir (Wien) 2000 Apr 12;142(4):401-406. [doi: 10.1007/s007010050449] [Medline: 10883336]
18. Zhang F, Feng M, Pan SJ, Loy LY, Guo W, Zhang Z, et al. Artificial neural network based intracranial pressure mean forecast algorithm for medical decision support. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2011 Presented at: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society; Aug 30 - Sept 3, 2011; Boston, MA, USA. [doi: 10.1109/iembs.2011.6091797]
19. Myers RB, Lazaridis C, Jermaine CM, Robertson CS, Rusin CG. Predicting intracranial pressure and brain tissue oxygen crises in patients with severe traumatic brain injury. Crit Care Med 2016;44(9):1754-1761. [doi: 10.1097/ccm.0000000000001838]
20. Hüser M, Kündig A, Karlen W, De Luca V, Jaggi M. Forecasting intracranial hypertension using multi-scale waveform metrics. Physiol Meas 2020 Feb 05;41(1):014001. [doi: 10.1088/1361-6579/ab6360] [Medline: 31851948]
21. Farhadi A, Chern J, Hirsh D, Davis T, Jo M, Maier F, et al. Intracranial pressure forecasting in children using dynamic averaging of time series data. Forecasting 2018 Aug 06;1(1):47-58. [doi: 10.3390/forecast1010004]
22. Lakin WD, Stevens SA, Tranmer BI, Penar PL. A whole-body mathematical model for intracranial pressure dynamics. J Math Biol 2003 Apr 1;46(4):347-383. [doi: 10.1007/s00285-002-0177-3] [Medline: 12673511]
23. Ryu J, Hu X, Shadden SC. A coupled lumped-parameter and distributed network model for cerebral pulse-wave hemodynamics. J Biomech Eng 2015 Oct;137(10):101009 [FREE Full text] [doi: 10.1115/1.4031331] [Medline: 26287937]
24. Wang JX, Hu X, Shadden SC. Data-augmented modeling of intracranial pressure. Ann Biomed Eng 2019 Mar;47(3):714-730 [FREE Full text] [doi: 10.1007/s10439-018-02191-z] [Medline: 30607645]
25. Jager GN, Westerhof N, Noordergraaf A. Oscillatory flow impedance in electrical analog of arterial system: representation of sleeve effect and non-newtonian properties of blood. Circ Res 1965 Feb;16(2):121-133. [doi: 10.1161/01.res.16.2.121]
26. Westerhof N, Bosman F, De Vries CJ, Noordergraaf A. Analog studies of the human systemic arterial tree. J Biomech 1969 May;2(2):121-143. [doi: 10.1016/0021-9290(69)90024-4]
27. Olufsen MS, Peskin CS, Kim WY, Pedersen EM, Nadim A, Larsen J. Numerical simulation and experimental validation of blood flow in arteries with structured-tree outflow conditions. Ann Biomed Eng 2000 Nov;28(11):1281-1299. [doi: 10.1114/1.1326031] [Medline: 11212947]
28. Milišić V, Quarteroni A. Analysis of lumped parameter models for blood flow simulations and their relation with 1D models. ESAIM: M2AN 2004 Aug 15;38(4):613-632. [doi: 10.1051/m2an:2004036]
29. He L, Ladner TR, Pruthi S, Day MA, Desai AA, Jordan LC, et al. Rule of 5: angiographic diameters of cervicocerebral arteries in children and compatibility with adult neurointerventional devices. J Neurointerv Surg 2016 Oct 06;8(10):1067-1071. [doi: 10.1136/neurintsurg-2015-012034] [Medline: 26546602]

30. Hu X, Nenov V, Bergsneider M, Glenn TC, Vespa P, Martin N. Estimation of hidden state variables of the intracranial system using constrained nonlinear kalman filters. IEEE Trans Biomed Eng 2007 Apr;54(4):597-610. [doi: 10.1109/tbme.2006.890130]
31. Ursino M, Lodi CA. Interaction among autoregulation, CO reactivity, and intracranial pressure: a mathematical model. Am J Physiol Heart Circ 1998 May 01;274(5):1715-1728. [doi: 10.1152/ajpheart.1998.274.5.h1715]
32. Kim N, Krasner A, Kosinski C, Wininger M, Qadri M, Kappus Z, et al. Trending autoregulatory indices during treatment for traumatic brain injury. J Clin Monit Comput 2016 Dec 7;30(6):821-831. [doi: 10.1007/s10877-015-9779-3] [Medline: 26446002]
33. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):215-220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
34. Pauca AL, Wallenhaupt SL, Kon ND, Tucker WY. Does radial artery pressure accurately reflect aortic pressure? Chest 1992 Oct;102(4):1193-1198. [doi: 10.1378/chest.102.4.1193] [Medline: 1395767]
35. Pauca AL, O'Rourke MF, Kon ND. Prospective evaluation of a method for estimating ascending aortic pressure from the radial artery pressure waveform. Hypertension 2001 Oct;38(4):932-937. [doi: 10.1161/hy1001.096106] [Medline: 11641312]
36. Chen C, Nevo E, Fetics B, Pak PH, Yin FC, Maughan WL, et al. Estimation of central aortic pressure waveform by mathematical transformation of radial tonometry pressure. Validation of generalized transfer function. Circulation 1997 Apr 01;95(7):1827-1836. [doi: 10.1161/01.cir.95.7.1827] [Medline: 9107170]
37. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Technometrics 2005 May;47(2):245. [doi: 10.1198/tech.2005.s278]
38. Jolliffe IT, Stephenson DB. Forecast verification: a practitioner's guide in atmospheric science, second edition. In: Wiley Online Library. New Jersey: John Wiley & Sons, Ltd; 2012.
39. Hogan RJ, Mason IB. Deterministic forecasts of binary events. In: Wiley Online Library. New Jersey: John Wiley & Sons, In; 2012:59.
40. Czosnyka M, Piechnik S, Richards HK, Kirkpatrick P, Smielewski P, Pickard JD. Contribution of mathematical modelling to the interpretation of bedside tests of cerebrovascular autoregulation. J Neurol Neurosurg Psychiatry 1997 Dec 01;63(6):721-731 [FREE Full text] [doi: 10.1136/jnnp.63.6.721] [Medline: 9416805]
41. Ursino M, Lodi CA. A simple mathematical model of the interaction between intracranial pressure and cerebral hemodynamics. J Appl Physiol (1985) 1997 Apr 01;82(4):1256-1269 [FREE Full text] [doi: 10.1152/jappl.1997.82.4.1256] [Medline: 9104864]
42. Lemaire JJ, Khalil T, Cervenansky F, Gindre G, Boire JY, Bazin JE, et al. Slow pressure waves in the cranial enclosure. Acta Neurochir (Wien) 2002 Mar;144(3):243-254. [doi: 10.1007/s007010200032] [Medline: 11956937]
43. Czosnyka M, Pickard JD. Monitoring and interpretation of intracranial pressure. J Neurol Neurosurg Psychiatry 2004 Jun 01;75(6):813-821 [FREE Full text] [doi: 10.1136/jnnp.2003.033126] [Medline: 15145991]
44. Jun-Yu Fan RN, Catherine Kirkness RN, Paolo V, Robert Burr MSEE, Pamela Mitchell CNRN. Intracranial pressure waveform morphology and intracranial adaptive capacity. Am J Crit Care 2008:545-554. [doi: 10.4037/ajcc2008.17.6.545]
45. Hawthorne C, Piper I. Monitoring of intracranial pressure in patients with traumatic brain injury. Front Neurol 2014 Jul 16;5:121 [FREE Full text] [doi: 10.3389/fneur.2014.00121] [Medline: 25076934]
46. Abbasi M. Long-term prediction of blood pressure time series using multiple fuzzy functions. In: Proceedings of the 21th Iranian Conference on Biomedical Engineering (ICBME). 2014 Presented at: 21th Iranian Conference on Biomedical Engineering (ICBME); Nov 26-28, 2014; Tehran, Iran p. 124-127. [doi: 10.1109/icbme.2014.7043906]
47. Sideris H. Building continuous arterial blood pressure prediction models using recurrent networks. In: Proceedings of the 2016 IEEE International Conference on Smart Computing (SMARTCOMP). 2016 Presented at: 2016 IEEE International Conference on Smart Computing (SMARTCOMP); May 18-20, 2016; St. Louis, MO, USA. [doi: 10.1109/smartcomp.2016.7501681]
48. Su P. Long-term blood pressure prediction with deep recurrent neural networks. In: Proceedings of 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). 2018 Presented at: 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); March 4-7, 2018; Las Vegas, NV, USA. [doi: 10.1109/bhi.2018.8333434]
49. Zhang B, Wei Z, Ren J, Cheng Y, Zheng Z. An empirical study on predicting blood pressure using classification and regression trees. IEEE Access 2018;6:21758-21768. [doi: 10.1109/access.2017.2787980]
50. Albers DJ, Levine ME, Stuart A, Mamykina L, Gluckman B, Hripcsak G. Mechanistic machine learning: how data assimilation leverages physiologic knowledge using Bayesian inference to forecast the future, infer the present, and phenotype. J Am Med Inform Assoc 2018 Oct 01;25(10):1392-1401 [FREE Full text] [doi: 10.1093/jamia/ocy106] [Medline: 30312445]
51. Zenker S, Rubin J, Clermont G. From inverse problems in mathematical physiology to quantitative differential diagnoses. PLoS Comput Biol 2007 Nov 9;3(11):e204. [doi: 10.1371/journal.pcbi.0030204]
52. Westwick DT, Kearney RE. Identification of nonlinear physiological systems. In: Wiley Online Library. New Jersey: John Wiley & Sons Inc; 2003.

Abbreviations

ABP: arterial blood pressure
AN: arterial network
CBF: cerebral blood flow
CoW: Circle of Willis
CPP: cerebral perfusion pressure
CSF: cerebrospinal fluid
IC: intracranial
ICH: intracranial hypertension
ICM: intracranial model
ICP: intracranial pressure
MCA: middle cerebral artery
nICP: noninvasive intracranial pressure
q1m: quaque 1-minute
TBI: traumatic brain injury

Edited by G Eysenbach; submitted 04.08.20; peer-reviewed by A James, K Malik, F Alam; comments to author 26.11.20; revised version received 21.01.21; accepted 22.01.21; published 22.03.21.

Please cite as:
Stroh JN, Bennett TD, Kheyfets V, Albers D
Clinical Decision Support for Traumatic Brain Injury: Identifying a Framework for Practical Model-Based Intracranial Pressure Estimation at Multihour Timescales
JMIR Med Inform 2021;9(3):e23215
URL: https://medinform.jmir.org/2021/3/e23215
doi:10.2196/23215
PMID:33749613

Original Paper

Applying Clinical Decision Support Design Best Practices With the Practical Robust Implementation and Sustainability Model Versus Reliance on Commercially Available Clinical Decision Support Tools: Randomized Controlled Trial

Katy E Trinkley[1,2,3,4], PharmD, PhD; Miranda E Kroehl[5], PhD; Michael G Kahn[6], MD, PhD; Larry A Allen[2,4], MHS, MD; Tellen D Bennett[2,6], MD, MS; Gary Hale[3], RPh; Heather Haugen[7], PhD; Simeon Heckman[3], RN, MS; David P Kao[3,4], MD; Janet Kim[1], PharmD; Daniel M Matlock[2,4,8], MPH, MD; Daniel C Malone[9], RPh, PhD; Robert L Page 2nd[1,4], PharmD, MSPH; Jessica Stine[1], PharmD; Krithika Suresh[2], PhD; Lauren Wells[1], PharmD; Chen-Tan Lin[3,4], MD

[1]Department of Clinical Pharmacy, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado Anschutz Medical Campus, Aurora, CO, United States

[2]Adult and Child Consortium for Outcomes Research and Delivery Science, University of Colorado, Aurora, CO, United States

[3]Department of Clinical Informatics, University of Colorado Health, Aurora, CO, United States

[4]Department of Medicine, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States

[5]Charter Communications Corporation, Greenwood Village, CO, United States

[6]Section of Informatics and Data Science, Department of Pediatrics, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States

[7]University of Colorado Clinical and Translational Sciences Institute, Aurora, CO, United States

[8]VA Eastern Colorado Geriastric Research Education and Clinical Center, Aurora, CO, United States

[9]Department of Pharmacotherapy, Skaggs College of Pharmacy, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Katy E Trinkley, PharmD, PhD
Department of Clinical Pharmacy
Skaggs School of Pharmacy and Pharmaceutical Sciences
University of Colorado Anschutz Medical Campus
12850 E Montview Blvd
Campus Box C238, Room V20-4125
Aurora, CO, 80045
United States
Phone: 1 3037246563
Fax: 1 3037240979
Email: katy.trinkley@cuanschutz.edu

Abstract

Background: Limited consideration of clinical decision support (CDS) design best practices, such as a user-centered design, is often cited as a key barrier to CDS adoption and effectiveness. The application of CDS best practices is resource intensive; thus, institutions often rely on commercially available CDS tools that are created to meet the generalized needs of many institutions and are not user centered. Beyond resource availability, insufficient guidance on how to address key aspects of implementation, such as contextual factors, may also limit the application of CDS best practices. An implementation science (IS) framework could provide needed guidance and increase the reproducibility of CDS implementations.

Objective: This study aims to compare the effectiveness of an enhanced CDS tool informed by CDS best practices and an IS framework with a generic, commercially available CDS tool.

Methods: We conducted an explanatory sequential mixed methods study. An IS-enhanced and commercial CDS alert were compared in a cluster randomized trial across 28 primary care clinics. Both alerts aimed to improve beta-blocker prescribing for heart failure. The enhanced alert was informed by CDS best practices and the Practical, Robust, Implementation, and Sustainability

Model (PRISM) IS framework, whereas the commercial alert followed vendor-supplied specifications. Following PRISM, the enhanced alert was informed by iterative, multilevel stakeholder input and the dynamic interactions of the internal and external environment. Outcomes aligned with PRISM's evaluation measures, including patient reach, clinician adoption, and changes in prescribing behavior. Clinicians exposed to each alert were interviewed to identify design features that might influence adoption. The interviews were analyzed using a thematic approach.

Results: Between March 15 and August 23, 2019, the enhanced alert fired for 61 patients (106 alerts, 87 clinicians) and the commercial alert fired for 26 patients (59 alerts, 31 clinicians). The adoption and effectiveness of the enhanced alert were significantly higher than those of the commercial alert (62% vs 29% alerts adopted, P<.001; 14% vs 0% changed prescribing, P=.006). Of the 21 clinicians interviewed, most stated that they preferred the enhanced alert.

Conclusions: The results of this study suggest that applying CDS best practices with an IS framework to create CDS tools improves implementation success compared with a commercially available tool.

Trial Registration: ClinicalTrials.gov NCT04028557; http://clinicaltrials.gov/ct2/show/NCT04028557

(JMIR Med Inform 2021;9(3):e24359) doi:10.2196/24359

KEYWORDS

PRISM; implementation science; clinical decision support systems; RE-AIM; congestive heart failure

Introduction

Background and Significance

Clinical decision support (CDS) tools within electronic health records (EHRs) hold the promise of improved patient care, but they are not always effective. To optimize effectiveness, developers are encouraged to apply CDS design best practices (eg, user-centered design) [1-4]. However, the comprehensive application of CDS best practices is resource intensive, and health care institutions are faced with an ever-growing list of CDS development projects. With limited resources, institutions often rely on commercially available CDS tools, which generally require fewer resources for deployment. Commercial CDS tools are created to meet the generalized needs of many institutions and thus may not integrate well into institution-specific workflows. Designing for the generalized needs of many institutions is not user centered. Thus, it violates a key CDS design best practice principle. Some have also asserted that commercial CDS tools may be based on content knowledge systems that are uninformative and not clinically relevant [1,5]; thus, they are less likely to be adopted [5,6]. However, these assertions have not been tested.

Although retrospective studies suggest that CDS best practices may improve CDS effectiveness [2-4,7,8], they are often minimally applied. Beyond resource availability, reasons for their minimal application may include skepticism about the evidence and insufficient guidance on how to apply them. Although CDS best practices acknowledge the importance of thoughtful implementation, they do not provide clear guidance regarding implementation considerations. Therefore, integration with evidence-based implementation science (IS) frameworks such as the Practical, Robust, Implementation, and Sustainability Model (PRISM) [9] can provide the direction needed to comprehensively apply CDS design best practices [10]. Such an integrated approach accounts for the many contextual factors that influence implementation success and makes CDS implementation more replicable. To maximize the quality of patient care, institutions need to understand the return on investment from allocating resources to apply CDS design best practices compared with relying on commercially available CDS tools.

Objective

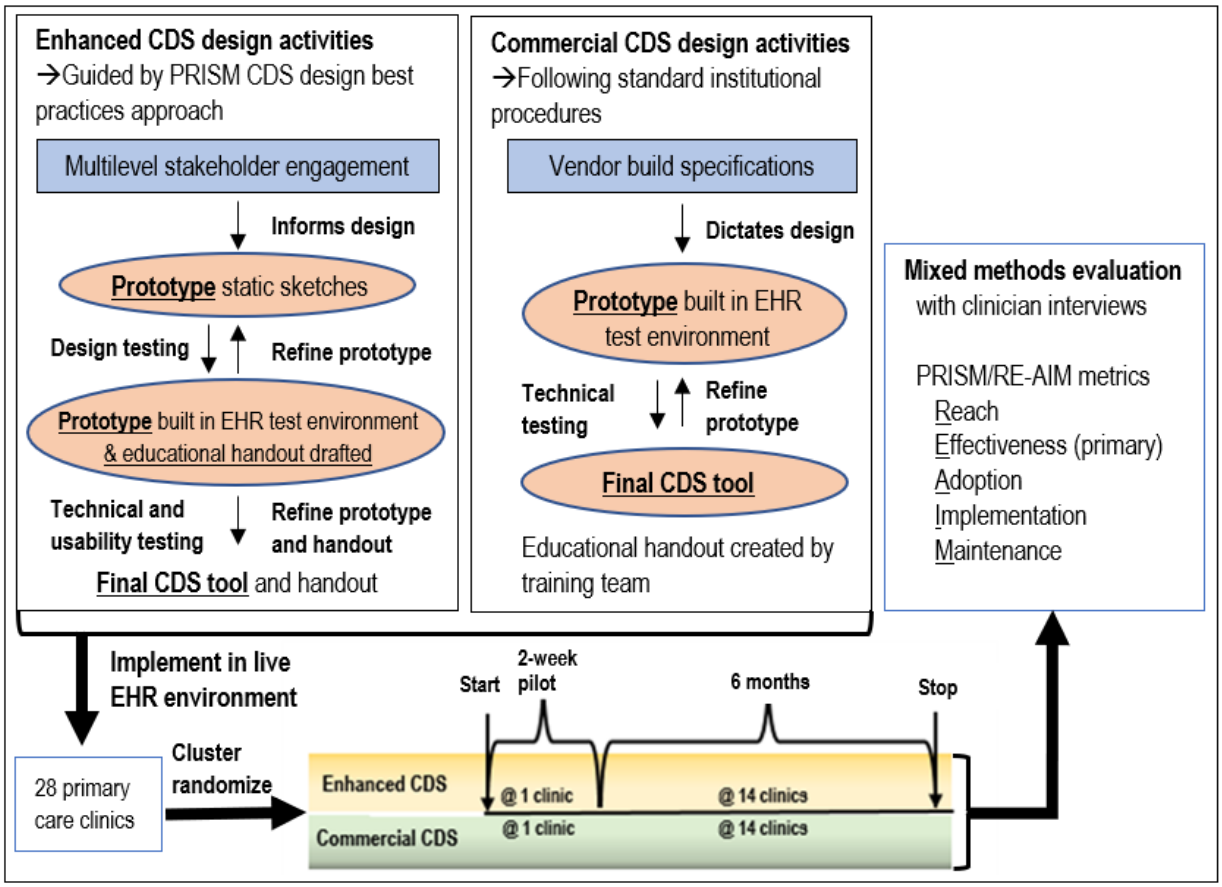
The objective of this study is to compare the effectiveness of an enhanced CDS tool informed by CDS design best practices and the PRISM IS framework with a prepackaged, commercially available CDS tool. The use case for this evaluation was an evidence-based beta-blocker (BB; bisoprolol, carvedilol, and metoprolol succinate) prescribed for patients with heart failure with reduced ejection fraction (HFrEF) in primary care. This use case was selected because it represents a national guideline recommendation with suboptimal adherence and both clear and compelling patient care implications [11-15]. Our hypothesis was that the enhanced CDS tool would result in greater clinician adoption and be more effective in changing prescribing than the commercial CDS tool.

Methods

Study Design

We conducted an explanatory sequential mixed methods study [16] at UCHealth, a large regional health system representing more than 5 million unique patients across diverse clinical settings. Since 2011, UCHealth has used the Epic EHR software program (Epic Systems). The first study phase was a cluster randomized controlled trial (RCT; NCT04028557), and the second phase consisted of a series of qualitative interviews with clinicians. Both phases were guided by the PRISM framework. Figure 1 provides an overview of the study design. The study design and reporting were guided by the CONSORT (Consolidated Standards of Reporting Trials) and best practices in complex trial interventions [17-19]. The study was approved by the Colorado Multiple Institutional Review Board.

Figure 1. Study design overview.



Description of the CDS Interventions

We evaluated 2 CDS tools within the EHR: a commercial alert and an enhanced alert. The automated alerts interrupted primary care providers (PCPs) when they opened a patient's chart during an office visit if the patient had a diagnosis of HFrEF and had not been prescribed evidence-based BB therapy. The CDS referred to the most recent ejection fraction (EF) value from an echocardiogram and/or a diagnosis of interest. Table 1 describes the build specifications and compares the way in which each CDS tool identifies an HFrEF diagnosis. Figures 2 and 3 depict the user interface for the enhanced and commercial CDS tools, respectively. Both alerts used the EHR-native CDS software *BestPractice Advisory* and underwent technical testing in EHR test environments. A 1-page educational handout on each alert was shared with clinician end users at the discretion of their respective clinic leaders or managers. There were some distinct differences in the design and implementation activities of each alert.

Table 1. Summary of build specifications for the enhanced and commercial alerts[a].

Enhanced CDS[b]	Commercial CDS
Inclusion criteria	
• ≥18 years old • A diagnosis that explicitly states an EF[c] ≤40% or an echocardiogram result indicating EF≤40%	• ≥18 years old • Any HF[d] diagnosis and an echocardiogram result indicating EF≤40%
Exclusion criteria	
• Prescribed or pended order for metoprolol succinate, carvedilol, or bisoprolol. Relied on knowledge management customized to the institution • BB[e] allergy using knowledge management customized to the institution	• Prescribed some versions of metoprolol tartrate, metoprolol succinate, carvedilol, or bisoprolol. Relied on vendor-supplied knowledge management, which did not comprehensively represent these BBs • BB or beta-agonist allergy using vendor-supplied knowledge management
Recommended action	
• Can pend evidence-based medication orders at starting doses without leaving the UI[f]	• Can open order set from UI, which opens a new screen and provides option to order any dose of BB, other drugs, labs, echo, and schedule follow-up visits
Response options (acknowledge reasons)	
• Options: Never appropriate, remind me later (1 month), provide comment • When a user selects a response option other than "never appropriate," it will not alert again for that user and patient for 28 days. If a user selects "never appropriate," it will not alert for that user and patient for >20 years • No dismiss button	• Options: Contraindicated, cost concern, patient declines • When a user selects a response option, it will not alert again for any user for that patient visit for 90 days • Dismiss button option
How to close	
• Easiest way to dismiss is to hit accept, which pends order for metoprolol succinate • Must select 1 of 3 acknowledge reasons or pend order for 1 of the BB options in the UI	• Must select "dismiss," open order set, or select 1 of 3 acknowledge reasons in the UI
Pertinent information displayed	
• Patient has HF and reduced EF • BB indicated • Values: most recent EF, last 3 BP[g] and HR[h] measurements • Benefit of starting BB—longevity • Parameters for caution: HR<50 and BP<90/60 • Asthma and chronic obstructive pulmonary disease are not contraindicated • Metoprolol tartrate is not evidence-based • Reminder to discontinue other BBs • Link to supporting reference	• Patient has HF and reduced EF • BB indicated • Values: most recent EF
Trigger	
• Open patient visit or encounter	• Open patient visit or encounter
Other features	
• Abnormal values of BP, HR, and EF are emphasized in red font	• None to note

[a]Key differences are italicized.

[b]CDS: clinical decision support.

[c]EF: ejection fraction.

[d]HF: heart failure.

[e]BB: beta-blocker.

[f]UI: user interface.

[g]BP: blood pressure.

[h]HR: heart rate.

Figure 2. Representative user interfaces of the enhanced clinical decision support alerts.

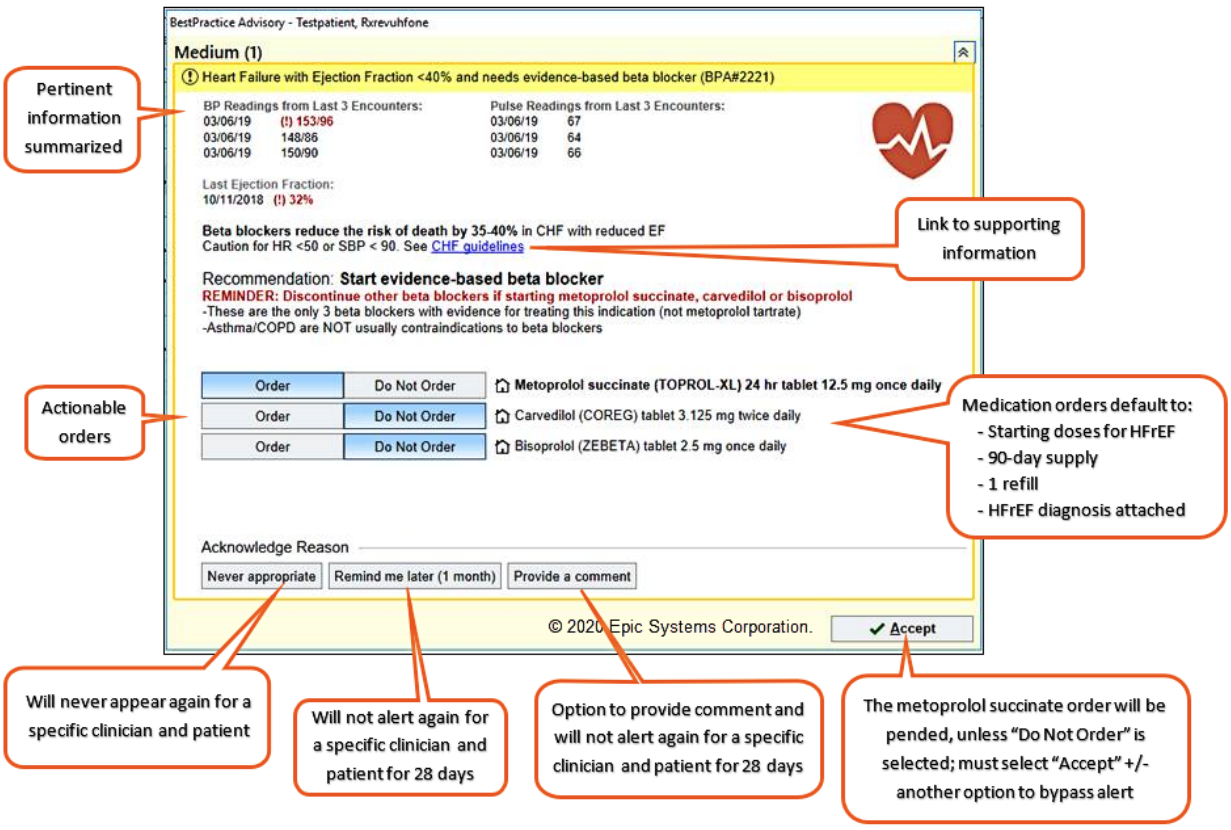
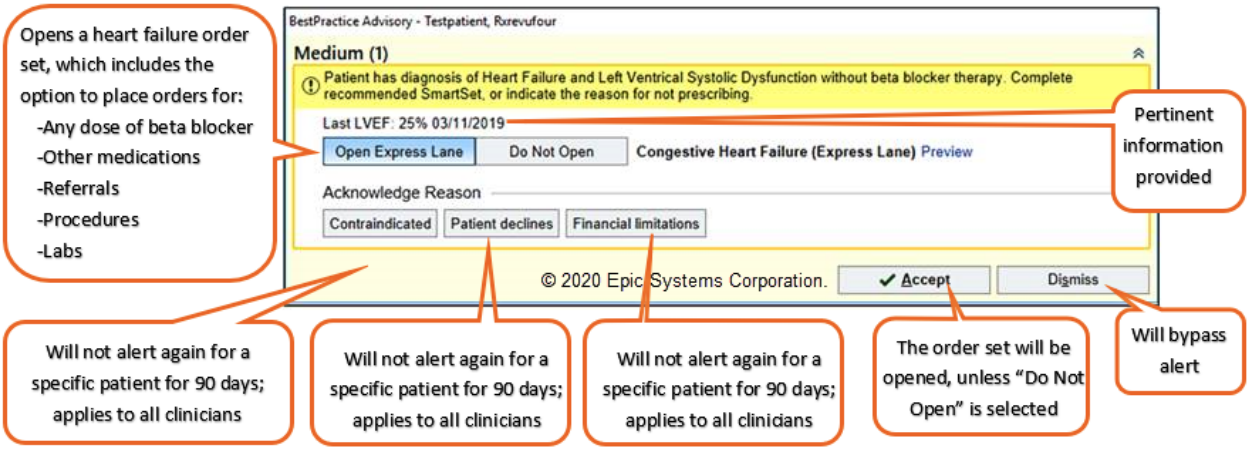


Figure 3. Representative user interfaces of the commercial clinical decision support alerts.



Enhanced CDS Alert

We designed and implemented the enhanced alert by applying the PRISM/CDS best practices approach, as previously described [10]. Briefly, PRISM is an expanded version of the widely used Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) measures [20,21] that includes additional contextual factors that influence implementation success [9]. The integrated PRISM/CDS best practices approach incorporates an iterative, user-centered design process with 5 phases: (1) multilevel stakeholder engagement, (2) designing the CDS tool, (3) design and usability testing, (4) thoughtful deployment, and (5) performance evaluation and maintenance [10]. Following PRISM, this approach considers the dynamic interactions between the internal and external environments [9]. We solicited extensive stakeholder input from clinicians [22] and patients to understand their needs, preferences, and values for the CDS design and treatment recommendations. Stakeholder input informed the enhanced alert design, which then underwent design and usability testing with clinicians. During usability testing, clinician end users refined the educational handout for the enhanced CDS alert.

Commercial CDS Alert

The commercial alert served as the active control. The EHR vendor provided the build specifications that are available for

use by all vendors' institutions. To be consistent with the way in which commercial alerts are commonly used, the build specifications were not informed by the PRISM/CDS best practices approach and did not change based on stakeholder input. However, we modified the build specifications to align with evidence-based and institution-specific practices. No modifications were made that would bias the results against the commercial alert. Multimedia Appendix 1 outlines the changes made to the commercial build specifications and the rationale for the changes. An educational handout on the commercial CDS was created following standard institutional procedures, including a review by the institution's dedicated training team.

Phase 1: Cluster RCT

Study Design and Randomization

Both alerts were deployed across 28 UCHealth primary care clinics (2 geriatrics, 17 family medicine, and 9 internal medicine) using a modified randomized parallel group design. Each alert was piloted in 1 clinic for 2 weeks before widespread implementation to facilitate clinic buy-in (Figure 1). Our initial plan was to implement the alerts in parallel at mutually exclusive clinics for 6 months. However, an a priori planned interim analysis revealed no instances of the commercial alert changing prescribing. Therefore, we determined that there was no added benefit of the commercial alert and stopped the trial early.

We performed cluster randomization, in which the cluster was defined as the clinic. Block randomization was used to allocate 1 of the 2 alerts (commercial or enhanced) to each clinic, with 6 blocks or strata, defined by geographical location (ie, North, South, or Metro) and volume of HFrEF patients (ie, small or large) [23-25]. Small-volume clinics had fewer than 25 patients with HFrEF. We used a random sampling scheme (function *sample_n* in R statistical software) to randomly assign half of the clinics in each block to the commercial alert. Only the study investigators knew to which group each clinic had been assigned.

Study Population

All 28 clinics that agreed to participate were included in the study. Clinicians in the participating clinics were evaluated if they were ever exposed to 1 of the alerts. A patient's EHR record was evaluated if 1 of the alerts was triggered.

Data Collection

Data were collected from the EHR via a CDS reporting analytic utility, chart review, and a secondary EHR virtual data warehouse. The EHR analytic utility allows postimplementation surveillance of CDS activity that includes when an alert is triggered, the identity of the patient and clinician, and the clinician's response to the alert (buttons clicked). Chart reviews were used to verify clinician-stated responses to the alert. Data collected from the data warehouse included patient and clinician characteristics. Comorbidities of interest were collected starting on the first day the alerts were deployed and are described in Multimedia Appendix 2. Concurrent medications included those prescribed within 12 months before the first day of deployment.

Primary RE-AIM Outcome: Effectiveness

Effectiveness was measured as the proportion of alerts that resulted in an evidence-based BB prescription when indicated.

Secondary RE-AIM Outcomes: Safety, Reach, Adoption, Implementation, and Maintenance

Effectiveness was also balanced by a safety evaluation that identified instances of bradycardia (heart rate<50 bpm), hypotension (blood pressure [BP]<90/60 mmHg), acute heart failure exacerbation requiring hospitalization or an emergency department visit, and unintended consequences, such as duplicate therapy. We evaluated the safety outcomes during the 1-month period after each alert was triggered. Reach was measured as the number of alerts for unique patient visits, unique patients, and unique clinicians. We also measured reach as the proportion of unique alerts relative to the number of patients with HFrEF. We selected this denominator based on data availability to allow for comparisons of representativeness between the 2 alerts. Adoption was measured as the proportion of times a clinician responded to an alert and was stratified by unique patients and clinicians. Overall, an alert was classified as *adopted* when the clinician paid attention to the information presented and did something other than dismiss it outright. Multimedia Appendix 3 provides details of how we defined adoption. Implementation was assessed by documenting the types and number of changes to the alert design or workflow integration [21]. Maintenance was assessed based on whether the intervention continued after the trial ended.

Data Analysis

Differences in overall effectiveness and adoption rates based on the number of alerts were tested using the chi-square test of independence. We compared the baseline characteristics of patients using a 2-sample *t* test for continuous variables and a chi-square test for independence for categorical characteristics. For chi-square tests in which any cell count was less than 5, a simulated P value with 2000 simulation replicates was calculated. For all analyses, R statistical software was used, and P<.05 was considered statistically significant.

Phase 2: Qualitative Interviews

Measures and Procedures

We invited clinicians exposed to 1 of the 2 alerts to participate in brief semistructured interviews. We used purposeful sampling to maximize the representativeness of exposure, practice setting, and clinician type. To capture the breadth of end user perspectives, interviews were conducted until saturation of ideas was reached. The participants provided informed consent. An investigator (KT) with domain expertise in primary care and CDS led each 30-min interview. The interviews followed a semistructured moderator guide, which was adapted as important concepts arose. Interviews were conducted in person or via a video conference. Audio recordings were transcribed and validated by independent investigators (JS and JK). Multiple strategies have been employed to maximize the rigor and quality of the methods and analysis [26], including triangulation, audit trails, and bracketing.

The interviews consisted of 3 consecutive components: (1) recall of the alert to which participants were exposed, (2) completion of the modified System Usability Scale (SUS) for that alert, and (3) introduction to the alert they had not been exposed, followed by a discussion comparing positive and negative features that influence the adoption of each alert. The validated SUS [27] terminology was modified to fit the HFrEF alert situation, and 2 questions were added to address ease of workflow integration and perceived impact on patient care.

Data Analysis

We used a thematic approach [28] and ATLAS.ti software (version 7, Scientific Software Development GmbH) to analyze the transcripts. A codebook was created a priori, which involved discussion among the 3 investigators (KT, JS, and DM). The changes to the codebook were documented. One investigator (JS) iteratively categorized the transcriptions into major themes that differentiated the 2 alerts using topic coding and analytical coding [29]. A second independent investigator (KT) reviewed the coding for the validation. Implementation measures of usability from the modified SUS survey were summarized according to the validated weighting system [27,30], and responses to the 2 additional questions were summarized descriptively.

Results

Patient Reach, Clinician Adoption, and Prescribing Effectiveness

The 2 alerts were deployed across 28 primary care clinics between March 15 and August 23, 2019. The mean age of patients exposed to an alert was 75.3 (SD 13.2) years, 70% (58/83) were male, and most were non-Hispanic and had Medicare insurance. Table 2 summarizes the characteristics of the patients who triggered the alerts.

The enhanced alert was triggered 106 times for 61 unique patients and 87 unique clinicians. The commercial alert was triggered 59 times for 26 unique patients and 31 unique clinicians. Patient visits were not always performed by the same clinician. Clinics allocated to the enhanced alert had 397 patients with HFrEF, compared with 307 patients in clinics allocated to the commercial alert; thus, reach was 26.7% (106/397) and 19.2% (59/307) for the 2 groups, respectively.

The overall adoption rate was significantly higher with the enhanced alert than with the commercial alert (62.3% vs 28.8%; P<.001). A total of 4 patients and 1 clinician were exposed to both commercial and enhanced alerts. None of these alerts led to a BB prescription, but adoption was higher with the enhanced alert. The 4 patients had a total of 7 visits with the enhanced alert, and 86% (6/7) resulted in adoption. The same 4 patients had a total of 12 visits with a commercial alert, and 2% (5/12) resulted in adoption. The single clinician who was exposed to both alerts adopted (did not outright dismiss) the enhanced alert each of the 2 times it was triggered (100%) and adopted the commercial alert 3 of the 6 times it was triggered (50%).

The enhanced alert was effective in changing prescribing for 15 of 61 unique patients (25%), whereas the commercial alert did not change prescribing at all. The overall rate of BB prescription was significantly higher when clinicians received the enhanced alert compared with the commercial alert (14.2% vs 0%; P=.006). Table 3 summarizes the results of the number of alerts, adoption, and effectiveness.

No adverse drug events were observed among patients who were prescribed a BB. When considering possible unintended consequences, a chart review revealed that 2 clinicians unintentionally ordered a BB from the customized alert, and the pharmacy processed the prescriptions, but neither patient picked up their prescription. These unintentional prescriptions were excluded from the effectiveness outcome. The enhanced alert was also triggered for 2 patients with a documented allergy to *beta-adrenergic blocking agts*, and neither led to a BB prescription. On the basis of clinician feedback, we identified and corrected an error in the build specification for the commercial alert during the first month of deployment; however, this error did not impact measures of reach, adoption, or effectiveness. Furthermore, upon completion of the study, the health system decided to continue the enhanced alert across all 28 clinics for operational and nonstudy purposes.

Table 2. Baseline characteristics of patients exposed to the alerts (N=83).

Characteristic	Enhanced alert (n=61)	Commercial alert (n=26)	Total[a] (N=83)	P value
Age (years), mean (SD)	74.8 (12.8)	76.6 (15)	75.3 (13.2)	.16
Male, n (%)	40 (66)	19 (73)	58 (70)	.66
White, n (%)	57 (93)	22 (85)	75 (90)	.23
Hispanic, n (%)	5 (8)	2 (7)	7 (8)	.99
Medicare, n (%)	50 (82)	22 (85)	69 (83)	.99
Primary care provider type: attending physician, n (%)	39 (64)	25 (96)	63 (76)	.01
Left ventricular ejection fraction, mean (SD)	31.7 (11)	34.7 (6)	32.7 (9)	.11
Heart rate, mean (SD)	78.7 (17)	73.4 (15)	76.9 (17)	.16
Heart rate<50, n (%)	1 (2)	1 (4)	2 (2)	.99
Systolic blood pressure, mean (SD)	123.7 (18)	121.0 (18)	123.3 (18)	.51
Diastolic blood pressure, mean (SD)	70.0 (12)	71.3 (9)	70.5 (12)	.59
Blood pressure <90/60, n (%)	1 (2)	1 (4)	2 (2)	.99
≥1 visit with cards[b] in past 1 year, n (%)	32 (53)	18 (69)	47 (57)	.23
≥1 visit with cards in past 2 years, n (%)	39 (64)	20 (77)	56 (68)	.35
Past BB[c], ever, n (%)	49 (80)	18 (69)	64 (77)	.40
BB allergy per chart review[d], n (%)	2 (3)	0 (0)	2 (2)	.59
BB intolerance or contraindication per chart review, n (%)	10 (16)	4 (15)	14 (17)	.99
Prescribed nonevidence-based BB[e], n (%)	29 (48)	12 (46)	38 (46)	.99
Prescribed metoprolol tartrate, n (%)	22 (36)	8 (31)	28 (34)	.82
Prescribed angiotensin converting enzyme inhibitor or angiotensin receptor blocker or ARNI[f], n (%)	37 (61)	18 (69)	55 (67)	.61
Prescribed ARNI, n (%)	1 (2)	0 (0)	1 (1)	.99
Prescribed mineralocorticoid receptor antagonist, n (%)	11 (18)	6 (23)	17 (20)	.80
Prescribed nondihydropyridine calcium channel blocker, n (%)	1 (2)	1 (4)	1 (1)	.99
Chronic obstructive pulmonary disease, n (%)	9 (15)	6 (23)	15 (18)	.53
Asthma, n (%)	7 (12)	3 (12)	10 (12)	.99
CAD[g] (myocardial infarction, percutaneous coronary intervention, bypass, CAD, angioplasty), n (%)	34 (56)	14 (54)	48 (58)	.99
Nonischemic cardiomyopathy, n (%)	19 (31.1)	10 (38.5)	29 (34.9)	.68
Atrial fibrillation, n (%)	25 (41.0)	15 (57.7)	40 (48.2)	.23

[a]Four patients were exposed to both the enhanced and commercial CDS.

[b]cards: outpatient cardiology provider.

[c]BB: beta-blocker.

[d]These patients were inadvertently not excluded from the alert.

[e]Other nonevidence-based beta blockers included atenolol, nebivolol, and sotalol.

[f]ARNI: angiotensin receptor-neprilysin inhibitor.

[g]CAD: coronary artery disease.

Table 3. Description of clinical decision support alerts, adoption, and effectiveness.

Characteristics	Enhanced	Commercial
Alerts for patients who had a visit with primary care during the evaluation period, n[a]		
Total number of alerts	106	59
Unique visits or encounters	104	59
Unique patients with alert	61	26
Unique clinicians alerted	87	31
Adoption (did not outright dismiss clinical decision support alert), n (%)		
Alerts adopted	66 (62.3)	17 (28)
Unique patients	44 (72)	13 (1)
Unique clinicians exposed to the alert	60 (69)	13 (41)
Clinicians who adopted with the first alert	55 (63)	11 (35)
Effectiveness, n (%)		
Alerts where BB[b] was prescribed	15 (14.2)	0 (0)
Unique patients where BB was prescribed	15 (25)	0 (0)
Unique patients prescribed with first alert	13 (87)	0 (0)
Unique patients prescribed BB by assigned primary care provider	7 (47)	0 (0)
Unique clinicians who ever prescribed BB	14 (16)	0 (0)
Clinicians who were attending physicians	9 (60)	0 (0)
Clinicians who were advanced practice clinicians	3 (21)	0 (0)
Clinicians who were a medical resident or fellow	2 (14)	0 (0)

[a]Four patients were exposed to both alerts, and 1 clinician was exposed to both alerts. One clinician prescribed a BB to 2 different patients.

[b]BB: beta-blocker.

Clinician Interviews: Usability, Satisfaction, and Design Features Influencing Adoption

The saturation of ideas was achieved after 21 interviews that included 15 clinicians exposed to the enhanced alert and 6 exposed to the commercial alert. One clinician was exposed to both alerts and did not recall either of the exposures. A total of 40% (6/15) of clinicians exposed to the enhanced alert and none exposed to the commercial alert stated that they recalled it, either before or after being prompted with a visual reminder. In total, 24% (5/21) of clinicians preferred the commercial alert, 2 because of brevity, 2 because of the dismiss option, and 1 because of the many options available within the order set. Most clinicians (19/21; 90%) stated that they felt an alert for BBs and HFrEF should be continued. Mean SUS scores were 65.7 (SD 14.2) and 53.4 (SD 14) for the enhanced and commercial alerts, respectively. The enhanced alert had higher median Likert scale scores for the survey questions related to workflow integration (3 vs 2.5) and perceived impact on patient care (4 vs 3.5). Multimedia Appendix 4 summarizes the SUS scores and survey questions with indices commonly used to interpret SUS scores.

During the open-ended discussions, the participants identified salient design features that influenced their alert preference. In general, clinicians preferred the enhanced alert because it was easier to digest the information presented and quickly determine its purpose. Clinicians liked the use of emphasis with different font sizes, bolding, and colors to draw their attention to key aspects of the enhanced alert. Furthermore, clinicians were unfamiliar with the *express lane* terminology of the commercial alert and stated that uncertainty about the consequences of selecting this option would deter adoption. Most clinicians felt that the commercial alert needed more information, so they could evaluate the appropriateness of the recommendation for a given patient. Although it was denser, most clinicians felt that the clinical information (eg, vital signs) in the enhanced alert was necessary and preferred. With one exception, the ability to *pend* a medication order within the enhanced alert was preferred over the order set in the commercial alert. The medication order option was preferred because it required fewer *clicks*, was specific to the recommendation, and provided information regarding which medications and doses were appropriate. Table 4 summarizes the representative quotes from clinicians that distinguish between the alerts.

Table 4. Representative clinician quotes distinguishing between the enhanced and commercial alerts.

Description of design features referred to	Quotes referring to the enhanced alert	Quotes referring to the commercial alert
Catching attention and use of emphasis	• "The color...different colors, catch our attention." • "The little heart icon gets your attention."	—[a]
Inclusion of a dismiss option	—	• "It encourages dismissal. It seems like the acknowledge reason is also a form of dismissal."
Clarity and uncertainty	• "It's much clearer in terms of what you're asking me is to order a bleeping [sic] beta-blocker, right? And you make it easy because you're clicking the most common starting doses."	• "Clicking on something where it goes to a black hole, or I don't know where it's going, especially if there is no training. I'm less likely to click on an unknown. Like this could end up 20 different ways that ends up with 10 different screens." • "I don't like this one as much, and I think it's because when I'm reading it, immediately, I have questions popping up, and while I think, I'm kind of in a hurry. And I don't know if I want to be clicking all these things to see what this is about. So express lane that makes me think of going to a gas station for an oil change."
Brevity and completeness of supporting information	• "It gives me the pieces of information that I would want to know to make a clinical decision and then it allows me to actually make that decision. You know, to pend up an order quickly."	• "I think this is more concise so I'm more prone to read it because this one [enhanced] vomited on me." • "This is nice and simple, but perhaps it's a little too simple."
Make it easy to do the right thing; ease of use	• "Yeah, I love that you picked the 3 medicines that I should be thinking about and kind of a typical starting dose, that's great." • "Easier to use. I don't have to leave the screen."	• "A little overwhelming for like labs now, labs in 3 months, labs in 6 months, echo now, 3, 6 months. And then medications, like every medication known to mankind."

[a]No relevant quote available

Discussion

Principal Findings

This study suggests that an enhanced CDS alert informed by CDS design best practices and an IS framework results in improved CDS adoption and effectiveness compared with a generic commercial alert. This conclusion is further supported by other findings related to the enhanced alert, including greater patient reach, higher usability scores, clinician-stated preference during the interviews, and the perceived impact on patient care and workflow integration. The commercial alert did not change prescribing, whereas the enhanced alert was associated with a 24% increase in BB prescriptions. Although 24% (5/21) of the interviewed clinicians preferred the commercial alert, their preference was driven by design features that were not prioritized by the majority of interview participants. Taken together, the results of this study suggest that applying the PRISM/CDS best practices approach [10] may improve the quality of care and, potentially, patient outcomes.

We achieved higher rates of adoption and effectiveness with our enhanced CDS tool than have previously been reported with other CDS tools designed to improve the prescription of similar medications for HFrEF. An RCT comparing a CDS tool with no CDS tool found no difference in effectiveness in changing HFrEF prescribing (23% vs 22%) [31], whereas another study found that a CDS tool improved prescribing by only 3.6% compared with 0.9% without a CDS (P=.01) [32]. These studies evaluating the effectiveness of changing HFrEF prescribing demonstrated minimal or no difference, whereas we found a 24% improvement in prescribing compared with an active control.

We identified aspects of the enhanced alert that need improvement, which developers should consider. For example, standardized drug vocabularies such as RxNorm should be used when possible and may have prevented the enhanced CDS from firing for patients with a *beta-adrenergic blocking agts* allergy. To mitigate the future risk of unintended prescriptions with the enhanced alert, we can reconsider the fundamental hard stop design. However, setting the default action of the enhanced alert to the desired change (ordering a BB) is aligned with CDS design best practices [3] and was done intentionally. Changing this design would likely minimize future instances of erroneous prescribing but might also deter effectiveness.

The reach (number of alerts, number of patients, and clinicians alerted) of the commercial alert was lower than that of the enhanced alert, which is likely because the build specifications were more constricting. Notably, the commercial alert had the following properties: (1) it required patients to have both a diagnosis of heart failure and a reduced EF, (2) it did not alert for some patients prescribed a nonevidence-based BB (ie, metoprolol tartrate), and (3) it excluded patients with an allergy to a beta-agonist, including albuterol inhalers. These 3 build specifications do not align with evidence-based clinical recommendations and limit the ability of the alert to reach the

intended patients. In our instance, inaccuracies in vendor-supplied knowledge content led to poor sensitivity and false negatives. However, such inaccuracies in knowledge management could also lead to poor specificity and worsen alert fatigue, as hypothesized by others [1,5,6].

The adoption of the commercial alert was also lower. On the basis of the interview findings, the adoption of the commercial alert could be improved by applying generalizable CDS design best practices that do not require input from the local setting. For example, any clinician considering BB initiation for HFrEF needs to know the patient's BP and heart rate. We estimate that 80% of the design decisions do not require input from the local context. Multimedia Appendix 5 describes examples of design features that do and do not require user-centered input from the local context.

Not all commercial CDS tools have the same limitations. The results of this study may have been different if the active control was a different commercial CDS tool. However, when relying on commercial CDS tools, this research highlights the need for institutions to carefully review the knowledge content and design features to ensure that they are accurate and appropriate for the local context. At a minimum, when resources are limited, institutions should review commercial CDS tools to evaluate unanticipated harm. There are advantages to relying on commercially available CDS, notably the need for fewer resources, but this may come at the cost of reduced implementation success. Similarly, there are advantages to customizing CDS for the institutional context, notably greater local ownership and implementation success, but this may come at the cost of greater resource burden. Our study demonstrates the difference accounting for the local context via an IS framework can have on implementation outcomes. By using an IS framework, CDS can be pragmatically customized to institution-specific contexts in a manner that is reproducible by other institutions and thereby generalizable.

Although adaptations to the local context may be inevitable to maximize implementation outcomes, additional efforts to share successful CDS tools across institutions are needed. Greater collaboration across institutions and repositories, such as Agency for Health care Research and Quality's CDS Authoring Tool and Connect Repository [33], can facilitate wider dissemination of well-designed CDS tools. Furthermore, given their influence over many health systems, EHR vendors could commit to increased surveillance and updates to the knowledge and design of CDS tools. Although external vendors may be unable to customize CDS tools to the local context, they should use CDS design best practices that are generalizable (Multimedia Appendix 5) and ideally consult with content experts to optimize the accuracy of knowledge content. External vendors should be transparent in the construction of their knowledge content and technologies and, where possible, apply CDS design best practices and IS frameworks such as PRISM.

Limitations

This study has several limitations. First, our measure of adoption aimed to identify clinicians who considered the information presented and relied on clinician responses to the alert, which can be imprecise. Similarly, our measure of effectiveness sought to capture clinicians who prescribed an evidence-based BB in response to the alert. We cannot say with certainty that the alert led to a prescribing change. It is possible that the clinician intended to prescribe the BB, and their actions were independent of the alert. However, a strength of this study is that we validated instances of BB prescriptions with chart review. Reliance on clinician-stated responses to alerts would have significantly overestimated the effectiveness of the enhanced alert. Although there are inherent limitations in our measures of adoption and effectiveness, our qualitative findings substantiate the validity of the quantitative methods.

In the initial design of this study, we planned to target 784 subjects and use generalized estimating equations to account for the within-clinic correlation in the analyses. However, due to a smaller-than-anticipated sample size and zero changes in prescribing behavior associated with the commercial alert, we needed to alter our plans. Although we were able to detect statistically meaningful differences, our small sample size warrants further research in larger populations and for different patient care scenarios. Similarly, the 21 clinicians we interviewed were not representative of all clinicians, but we did take measures to maximize credibility, transferability dependability, and confirmability of the qualitative methods [26,34]. Although the investigator (KT) who led the interviews also led the design of the enhanced CDS tool, biases were minimized by using a semistructured interview guide and documenting a priori preconceived ideas and biases. We also used a multidisciplinary approach for the thematic analysis in which an independent investigator (JS) led the coding with iterative input from 3 other investigators (JS, DM, and JK).

Finally, because much of our data were collected from the EHR, limitations inherent to secondary data sources and EHR data apply. One notable limitation is the inaccuracy and incompleteness of assigning PCPs to specific clinics within the EHR. Difficulty in accurately identifying PCP—and patient—clinic assignments prevented us from controlling for all potential cross-contamination of alert exposure. As we found, some clinicians practice at and some patients are seen at more than 1 clinic. Inaccuracy in the patient-clinic assignment also precluded us from defining the ideal denominator for reach. Furthermore, data limitations prevented us from characterizing clinician- and clinic-level characteristics that may have influenced implementation success and reporting a complete CONSORT or expanded CONSORT figure [18,35].

Conclusions

This study suggests that applying CDS design best practices with an IS framework to CDS tools leads to meaningful improvements in patient reach, clinician adoption, and effectiveness of behavior change, as compared with some commercially available CDS tools. Future research should assess the generalization of these results and consider how this IS-based approach to CDS implementation can be adapted to rapid prototyping of CDS to expedite the creation of widely adopted, effective, and sustainable CDS.

Acknowledgments

The authors thank Esther Langmack, MD, Langmack Medical Communications, LLC, for editorial assistance. This study is supported in part by National Institutes for Health/National Center for Advancing Translational Sciences Colorado Clinical and Translational Sciences Award Grant Number UL1 TR002535 and by National Heart, Lung, and Blood Institute K12 Training Grant Number K12HL137862.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Differences between vendor specifications and actual specifications for commercial.
[DOCX File , 13 KB - medinform_v9i3e24359_app1.docx]

Multimedia Appendix 2
Definitions for comorbidities of interest.
[DOCX File , 13 KB - medinform_v9i3e24359_app2.docx]

Multimedia Appendix 3
Categorization of clinician responses as an instance of adoption or not.
[DOCX File , 25 KB - medinform_v9i3e24359_app3.docx]

Multimedia Appendix 4
Usability, perceived impact, and workflow integration scores.
[DOCX File , 16 KB - medinform_v9i3e24359_app4.docx]

Multimedia Appendix 5
Examples of CDS design features that do and do not necessitate input from the local context.
[DOCX File , 15 KB - medinform_v9i3e24359_app5.docx]

Multimedia Appendix 6
General CONSORT checklist (2010).
[PDF File (Adobe PDF File), 156 KB - medinform_v9i3e24359_app6.pdf]

References

1. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc 2003 Nov 1;10(6):523-530. [doi: 10.1197/jamia.m1370]
2. Osheroff J, Teich J, Levick D, Saldana L, Velasco F, Sittig D. Improving Outcomes With Clinical Decision Support: an Implementers Guide. Chicago, IL: Healthcare Information Management Systems Society (HIMSS); 2012.
3. Horsky J, Schiff GD, Johnston D, Mercincavage L, Bell D, Middleton B. Interface design principles for usable decision support: a targeted review of best practices for clinical prescribing interventions. J Biomed Inform 2012 Dec;45(6):1202-1216 [FREE Full text] [doi: 10.1016/j.jbi.2012.09.002] [Medline: 22995208]
4. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. J Biomed Inform 2008 Apr;41(2):387-392 [FREE Full text] [doi: 10.1016/j.jbi.2007.09.003] [Medline: 18029232]
5. Shah N, Seger A, Seger D, Fiskio J, Kuperman G, Blumenfeld B, et al. Improving acceptance of computerized prescribing alerts in ambulatory care. J Am Med Inform Assoc 2006;13(1):5-11 [FREE Full text] [doi: 10.1197/jamia.M1868] [Medline: 16221941]
6. Horsky J, Phansalkar S, Desai A, Bell D, Middleton B. Design of decision support interventions for medication prescribing. Int J Med Inform 2013 Jun;82(6):492-503. [doi: 10.1016/j.ijmedinf.2013.02.003] [Medline: 23490305]
7. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc 2003 Nov 1;10(6):523-530. [doi: 10.1197/jamia.m1370]
8. Marcilly R, Ammenwerth E, Roehrer E, Niès J, Beuscart-Zéphir MC. Evidence-based usability design principles for medication alerting systems. BMC Med Inform Decis Mak 2018 Jul 24;18(1):69 [FREE Full text] [doi: 10.1186/s12911-018-0615-9] [Medline: 30041647]

9. Feldstein AC, Glasgow RE. A practical, robust implementation and sustainability model (PRISM) for integrating research findings into practice. Jt Comm J Qual Patient Saf 2008 Apr;34(4):228-243. [Medline: 18468362]
10. Trinkley KE, Kahn MG, Bennett TD, Glasgow RE, Haugen H, Kao DP, et al. Integrating the practical robust implementation and sustainability model with best practices in clinical decision support design: implementation science approach. J Med Internet Res 2020 Oct 29;22(10):e19676 [FREE Full text] [doi: 10.2196/19676] [Medline: 33118943]
11. Qian Q, Manning DM, Ou N, Klarich MJ, Leutink DJ, Loth AR, et al. ACEi/ARB for systolic heart failure: closing the quality gap with a sustainable intervention at an academic medical center. J Hosp Med 2011 Mar 22;6(3):156-160. [doi: 10.1002/jhm.803] [Medline: 20652962]
12. Yancy C, Jessup M, Bozkurt B, Butler J, Casey D, Drazner M, et al. 2013 ACCF/AHA guideline for the management of heart failure: executive summary: a report of the American College of Cardiology Foundation/American Heart Association Task Force on practice guidelines. Circulation 2013 Oct 15;128(16):1810-1852. [doi: 10.1161/CIR.0b013e31829e8807] [Medline: 23741057]
13. Yancy C, Januzzi J, Allen L, Butler J, Davis L, Fonarow G, et al. 2017 ACC Expert Consensus Decision Pathway for Optimization of Heart Failure Treatment: Answers to 10 Pivotal Issues About Heart Failure With Reduced Ejection Fraction: A Report of the American College of Cardiology Task Force on Expert Consensus Decision Pathways. J Am Coll Cardiol 2018 Jan 16;71(2):201-230 [FREE Full text] [doi: 10.1016/j.jacc.2017.11.025] [Medline: 29277252]
14. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Colvin MM, et al. 2017 ACC/AHA/HFSA Focused Update of the 2013 ACCF/AHA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Failure Society of America. J Card Fail 2017 Aug;23(8):628-651. [doi: 10.1016/j.cardfail.2017.04.014] [Medline: 28461259]
15. Greene SJ, Butler J, Albert NM, DeVore AD, Sharma PP, Duffy CI, et al. Medical therapy for heart failure with reduced ejection fraction: the CHAMP-HF registry. J Am Coll Cardiol 2018 Jul 24;72(4):351-366 [FREE Full text] [doi: 10.1016/j.jacc.2018.04.070] [Medline: 30025570]
16. NIH Office of Behavioral and Social Sciences. Best practices for mixed methods research in the health sciences, 2nd ed. Bethesda, MD: National Institutes of Health; 2018.
17. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. Br Med J 2000 Sep 16;321(7262):694-696 [FREE Full text] [Medline: 10987780]
18. Moher D, Schulz KF, Altman D, CONSORT Group. The CONSORT Statement: revised recommendations for improving the quality of reports of parallel-group randomized trials 2001. Explore (NY) 2005 Jan;1(1):40-45. [doi: 10.1016/j.explore.2004.11.001] [Medline: 16791967]
19. Augestad KM, Berntsen G, Lassen K, Bellika JG, Wootton R, Lindsetmo RO, et al. Standards for reporting randomized controlled trials in medical informatics: a systematic review of CONSORT adherence in RCTs on clinical decision support. J Am Med Inform Assoc 2012 Jan;19(1):13-21 [FREE Full text] [doi: 10.1136/amiajnl-2011-000411] [Medline: 21803926]
20. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. Re-aim planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64 [FREE Full text] [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
21. Gaglio B, Shoup JA, Glasgow RE. The RE-AIM framework: a systematic review of use over time. Am J Public Health 2013 Jun;103(6):e38-e46. [doi: 10.2105/AJPH.2013.301299] [Medline: 23597377]
22. Trinkley KE, Blakeslee WW, Matlock DD, Kao DP, Van Matre AG, Harrison R, et al. Clinician preferences for computerised clinical decision support for medications in primary care: a focus group study. BMJ Health Care Inform 2019 Apr 17;26(1) [FREE Full text] [doi: 10.1136/bmjhci-2019-000015] [Medline: 31039120]
23. Barbui C, Cipriani A. Cluster randomised trials. Epidemiol Psychiatr Sci 2011 Dec 28;20(4):307-309. [doi: 10.1017/s2045796011000515] [Medline: 22201207]
24. Gums T, Carter B, Foster E. Cluster randomized trials for pharmacy practice research. Int J Clin Pharm 2016 Jun 29;38(3):607-614. [doi: 10.1007/s11096-015-0205-1] [Medline: 26715549]
25. Beller EM, Gebski V, Keech AC. Randomisation in clinical trials. Med J Aust 2002 Nov 18;177(10):565-567. [doi: 10.5694/j.1326-5377.2002.tb04955.x] [Medline: 12429008]
26. Cypress BS. Rigor or reliability and validity in qualitative research. Dimensions of Critical Care Nursing 2017;36(4):253-263. [doi: 10.1097/dcc.0000000000000253]
27. System Usability Scale SUS Internet. Usability. URL: https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html [accessed 2018-01-22]
28. Chapman AL, Hadfield M, Chapman CJ. Qualitative research in healthcare: an introduction to grounded theory using thematic analysis. J R Coll Physicians Edinb 2015;45(3):201-205. [doi: 10.4997/JRCPE.2015.305] [Medline: 26517098]
29. Castleberry A, Nolen A. Thematic analysis of qualitative research data: Is it as easy as it sounds? Curr Pharm Teach Learn 2018;10(6):807-815. [doi: 10.1016/j.cptl.2018.03.019] [Medline: 30025784]
30. Measuring and Interpreting System Usability Scale (SUS) - UIUX Trend. UIUX Trend. 2017. URL: https://uiuxtrend.com/measuring-system-usability-scale-sus/ [accessed 2021-02-21]

31. Tierney WM, Overhage JM, Murray MD, Harris LE, Zhou X, Eckert GJ, et al. Effects of computerized guidelines for managing heart disease in primary care. J Gen Intern Med 2003 Dec;18(12):967-976 [FREE Full text] [doi: 10.1111/j.1525-1497.2003.30635.x] [Medline: 14687254]
32. Blecker S, Pandya R, Stork S, Mann D, Kuperman G, Shelley D, et al. Interruptive versus noninterruptive clinical decision support: usability study. JMIR Hum Factors 2019 Apr 17;6(2):e12469 [FREE Full text] [doi: 10.2196/12469] [Medline: 30994460]
33. CDS authoring tool. Agency for Healthcare Research and Quality. URL: https://cds.ahrq.gov/cdsconnect/authoring [accessed 2021-02-22]
34. McIlvennan CK, Morris MA, Guetterman TC, Matlock DD, Curry L. Qualitative Methodology in Cardiovascular Outcomes Research: A Contemporary Look. Circ Cardiovasc Qual Outcomes 2019 Sep;12(9):e005828. [doi: 10.1161/CIRCOUTCOMES.119.005828] [Medline: 31510771]
35. Glasgow RE, Huebschmann AG, Brownson RC. Expanding the consort figure: increasing transparency in reporting on external validity. Am J Prev Med 2018 Sep;55(3):422-430. [doi: 10.1016/j.amepre.2018.04.044] [Medline: 30033029]

Abbreviations

BB: beta-blocker
BP: blood pressure
CDS: clinical decision support
CONSORT: Consolidated Standards of Reporting Trials
EF: ejection fraction
EHR: electronic health record
HFrEF: heart failure with reduced ejection fraction
PCP: primary care provider
PRISM: Practical, Robust, Implementation, and Sustainability Model
RCT: randomized controlled trial
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
SUS: System Usability Scale

Edited by G Eysenbach; submitted 21.09.20; peer-reviewed by M Afzal, S Sarbadhikari; comments to author 24.10.20; revised version received 07.12.20; accepted 16.01.21; published 22.03.21.

Please cite as:
Trinkley KE, Kroehl ME, Kahn MG, Allen LA, Bennett TD, Hale G, Haugen H, Heckman S, Kao DP, Kim J, Matlock DM, Malone DC, Page 2nd RL, Stine J, Suresh K, Wells L, Lin CT
Applying Clinical Decision Support Design Best Practices With the Practical Robust Implementation and Sustainability Model Versus Reliance on Commercially Available Clinical Decision Support Tools: Randomized Controlled Trial
JMIR Med Inform 2021;9(3):e24359
URL: https://medinform.jmir.org/2021/3/e24359
doi:10.2196/24359
PMID:33749610

Original Paper

Early Prediction of Unplanned 30-Day Hospital Readmission: Model Development and Retrospective Data Analysis

Peng Zhao[1], MS, PhD; Illhoi Yoo[1,2], PhD; Syed H Naqvi[3], MD

[1]Institute for Data Science and Informatics, University of Missouri, Columbia, MO, United States

[2]Department of Health Management and Informatics, School of Medicine, University of Missouri, Columbia, MO, United States

[3]Division of Hospital Medicine, Department of Medicine, University of Missouri School of Medicine, Columbia, MO, United States

Corresponding Author:
Illhoi Yoo, PhD
Department of Health Management and Informatics
School of Medicine
University of Missouri
Five Hospital Drive
CE718 Clinical Support and Education Building (DC006.00)
Columbia, MO, 65212
United States
Phone: 1 5738827642
Fax: 1 573 882 6158
Email: YooIL@health.missouri.edu

Abstract

Background: Existing readmission reduction solutions tend to focus on complementing inpatient care with enhanced care transition and postdischarge interventions. These solutions are initiated near or after discharge, when clinicians' impact on inpatient care is ending. Preventive intervention during hospitalization is an underexplored area that holds potential for reducing readmission risk. However, it is challenging to predict readmission risk at the early stage of hospitalization because few data are available.

Objective: The objective of this study was to build an early prediction model of unplanned 30-day hospital readmission using a large and diverse sample. We were also interested in identifying novel readmission risk factors and protective factors.

Methods: We extracted the medical records of 96,550 patients in 205 participating Cerner client hospitals across four US census regions in 2016 from the Health Facts database. The model was built with index admission data that can become available within 24 hours and data from previous encounters up to 1 year before the index admission. The candidate models were evaluated for performance, timeliness, and generalizability. Multivariate logistic regression analysis was used to identify readmission risk factors and protective factors.

Results: We developed six candidate readmission models with different machine learning algorithms. The best performing model of extreme gradient boosting (XGBoost) achieved an area under the receiver operating characteristic curve of 0.753 on the development data set and 0.742 on the validation data set. By multivariate logistic regression analysis, we identified 14 risk factors and 2 protective factors of readmission that have never been reported.

Conclusions: The performance of our model is better than that of the most widely used models in US health care settings. This model can help clinicians identify readmission risk at the early stage of hospitalization so that they can pay extra attention during the care process of high-risk patients. The 14 novel risk factors and 2 novel protective factors can aid understanding of the factors associated with readmission.

(JMIR Med Inform 2021;9(3):e16306) doi:10.2196/16306

KEYWORDS

patient readmission; risk factors; unplanned; early detection; all-cause; predictive model; 30-day; machine learning

Introduction

Unplanned hospital readmission continues to attract much attention due to its negative influence on patients' quality of life and substantial contribution to health care costs. During July 2015 to June 2016, 15.2% of Medicare beneficiaries experienced unplanned readmission within 30 days after discharge [1]. It has been estimated that unplanned readmission accounts for US $17.4 billion in Medicare expenditures annually [2]. In an effort to improve health care quality and decrease unplanned hospital readmission rates, the Affordable Care Act [3] implemented the Hospital Readmission Reduction Program (HRRP) [4] in 2012 to use unplanned 30-day hospital readmission as a metric to financially penalize hospitals with excessive readmission rates. The high associated cost and penalties from the HRRP have intensified the efforts of the entire health care industry to reduce unplanned hospital readmissions.

Existing readmission reduction interventions, especially transition interventions and postdischarge interventions, focus on complementing inpatient care with enhanced services; however, the planning, implementation, and monitoring of these interventions can be resource-intensive [5]. In addition, no single intervention or bundle of interventions were found to be reliable in reducing readmissions, according to the review by Hansen et al [6]. Another disadvantage is that these interventions do not greatly impact the quality improvement of inpatient care because they are mostly initiated near or after discharge, when clinicians' impact on inpatient care is ending. Preventive intervention during hospitalization is an underexplored area that holds potential for reducing readmission risk. It has been shown that early interventions during inpatient hospitalization, such as early discharge planning [7], can reduce readmissions. However, it is impractical to deliver readmission-preventive interventions to all patients because health care resources are restricted. Predictive modeling is an efficient method to optimize the allocation of valuable clinical resources by stratifying patients' readmission risk and targeting the delivery of preventive interventions to patients at high risk [8]. Evidence has shown that focusing interventions on high-risk patients can reduce 30-day hospital readmission risk by 11%-28% [9-11]. However, the majority of reported hospital readmission predictive models have limited value in real-world health care settings because they require variables whose values only become completely available at discharge [12]. For example, the HOSPITAL score [13] and the LACE index [14] are the most widely used readmission risk calculators in US healthcare settings. They only work at the end of inpatient care because they require variables that are not available in a timely fashion, such as the length of stay and the results of some laboratory tests before discharge. It is essential to perform early risk assessments of high-risk patients to enable clinicians to deliver timely preventive interventions at the early stage of hospitalization [15].

Several 30-day hospital readmission early detection models have been reported; however, their performance and design are unsatisfactory. Wang et al [16] developed a real-time readmission model using a time series of vital signs and discrete features such as laboratory tests. However, this model was a black box, and it was unclear how the clinical factors led to the predictions. In health care applications, the interpretability of a model is as important as its performance because the attributes and the decision path must be medically rational. Horne et al [17] developed a laboratory-based model specific to heart failure patients. It can be used within 24 hours of admission; however, the performance was poor, with areas under the receiver operating characteristic curve (AUCs) [18] of 0.571 and 0.596 in female and male validation data sets. Cronin et al [19] reported an early detection model based on the information available at admission and medications used in index admission; it showed a moderate performance in the validation data set, with an AUC of 0.671. El Morr et al [20] created a modified LACE index (LACE-rt) to support real-time prediction by replacing the length of stay during the current admission in the original LACE index with that of the previous admission within the last 30 days. However, this model only showed fair performance (AUC 0.632) [20]. Shadmi et al [21] developed an early prediction model for emergency readmissions based on data available before the index admission, and they achieved an AUC of 0.69 in the validation data set. The same team further modified the model by adding risk factors accrued during index admissions; however, they obtained a similarly moderate AUC (0.68) in the validation data set [22]. Amarasingham *et al* [23] reported a real-time readmission model (AUC of 0.69 in the validation data set) for patients with heart failure using clinical and social factors available within 24 hours of admission. However, their cohort size was too small, with only 1372 index admissions.

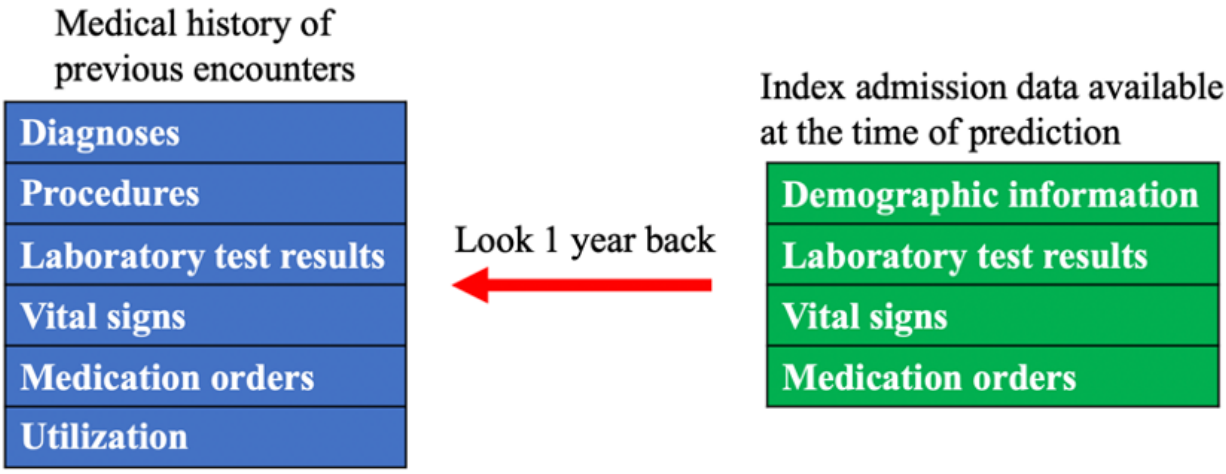
The objective of this work was to build a predictive model for early detection of unplanned 30-day hospital readmission using a large and diverse sample. We were also interested in identifying novel risk factors and protective factors of readmission. We used machine learning methods to develop a predictive model that can monitor readmission risk at the early stage of hospitalization. Unlike most models, which focus only on characteristics of index admissions, we included the detailed medical history of previous encounters up to 1 year before index admissions to construct a better readmission prediction model.

Methods

Study Design

This study was a retrospective analysis of electronic health record (EHR) data. To ensure the readmission prediction model can be accurate at the early stage of hospitalization, we only used index admission attributes whose values can become available in the EHR within 24 hours, including patient demographics, laboratory test results, vital signs, and medication orders. The patients' data were enriched by the detailed history of previous hospital encounters within one year before the current inpatient stay, including the information of diagnoses, procedures, laboratory test results, vital signs, medication orders, and health care utilization. Figure 1 shows the types of variables used for modeling.

Figure 1. The variables used to develop the models.



Data Source

The data were extracted from Health Facts [24], an EHR database curated by Cerner Corporation. This Health Insurance Portability and Accountability Act (HIPAA)–compliant database was collected from participating Cerner client hospitals and from clinics with de-identified longitudinal records of diagnoses, laboratory test results, surgeries, microbiology test results, medications, medical events, and medical histories. The version accessed by researchers at the University of Missouri contained 3.15 TB of encounter-level medical records extracted from 782 hospitals and clinics in the United States between 2000 and 2016.

Ethics

This retrospective data analysis study was not required to obtain approval from the University of Missouri Institutional Review Board because the data used in the study were already fully deidentified by the data owner (Cerner Corporation).

Data Inclusion and Exclusion Criteria

The data inclusion and exclusion criteria were based on the criteria used by the Centers for Medicare & Medicaid Services (CMS) [25] with minor modifications. (1) We captured inpatient encounters between January 1 and December 31, 2016, in acute care hospitals with a length of stay longer than 24 hours. (2) The gap between index admission discharge and readmission was between 1 and 30 days (inclusive). (3) If a patient had more than one inpatient visit within 30 days of discharge, only the first visit was considered as readmission. (4) Patients were aged older than 18 years at admission. (5) Patients were not transferred to other acute care facilities and were alive at discharge. (6) Patients were not readmitted for newborn status, labor, accident, trauma, rehabilitation services, or other scheduled care according to the CMS planned readmission identification algorithm [25]. (7) In this work, we adopted the concept of "hospital-wide all-cause readmission" used by CMS [26] to study readmissions due to medical and health care–related reasons. (8) Patients without readmissions had the same requirements for their index admissions. (9) Each patient had only one record.

Feature Engineering and Data Transformation

According to our previous literature review of readmission risk factors [27], patients' demographic and social factors as well as their previous health care utilization were strong predictors for readmission. In this work, we incorporated the patients' age at admission, sex, race, insurance payer, hospital census region, census division, rurality, and health care utilization in the previous year, including the number of inpatient visits, outpatient visits, emergency department visits, and times the patient left the hospital against medical advice. We only retained records without any missing demographic information. We also investigated the impact of the patients' medical history within the year before the index admission. We used counts to condense longitudinal medical histories into structured data so that patients with different medical histories could be represented in the same feature space. Patients with a medical history had higher counts, and patients without any medical history had counts of zero. In this way, we were able to handle the missing value problem for new patients. For example, if a patient had the same diagnosis of heart failure in two separate encounters in the previous year, this diagnosis would have a count of two. For laboratory tests and vital signs, the latest results were checked and recorded if they were abnormal. Suppose a patient's systolic blood pressure was taken twice in one encounter, and the result of the second test was abnormal. In another encounter, it was taken three times, and the latest result was normal. This patient would be determined to have had one abnormal systolic blood pressure result during the two encounters in the last year. For the index admission, we only checked the medication record and the latest results of laboratory tests and vital signs. Diagnosis codes were mapped from the *International Classification of Disease, Tenth Revision, Clinical Modification* (*ICD-10-CM*) [28] into the Clinical Classifications Software (CCS) categories [29] because the ICD codes were too granular for data mining purposes. For the same reason, procedure codes were mapped from the *International Classification of Disease, Tenth Revision, Procedure Coding System* (*ICD-10-PCS*) [28], current procedural terminology (CPT) [30], and Healthcare Common Procedure Coding System (HCPCS) [31] codes into CCS categories. Laboratory tests and vital signs were represented by

their original names. We used generic names to represent medication orders. Table 1 shows an explanation of these features. After data transformation and feature engineering, the final data set contained 432 variables.

Table 1. Feature representation and value types.

Type and category	Representation	Data type
Medical history in last year		
Diagnosis	CCS[a]	Count
Procedure	CCS	Count
Laboratory test	Name	Count
Vital sign	Name	Count
Medication	Generic name	Count
Utilization	Name	Count
Index admission		
Demographic	Name	Discretized age, race, sex, payer, region, or rurality
Medication	Generic name	Ordered or not
Laboratory test	Name	Latest result is abnormal or not
Vital sign	Name	Latest result is abnormal or not

[a]CCS: Clinical Classifications Software.

Candidate Algorithms and Baseline Models

Interpretability is an important consideration for clinical predictive models because it is crucial to ensure medical rationality in the classification process. We selected six candidate machine learning algorithms that can generate probabilistic outputs, including logistic regression, naïve Bayes, decision tree, random forest, gradient boosting tree, and artificial neural networks. Logistic regression belongs to the family of generalized linear models [32], and it predicts the log odds of the positive class as a linear combination of variables weighted by coefficients [33]. The association of a variable (factor) with the response target can be measured by the odds ratio [34], which is equal to the exponential of the coefficient of the variable. An odds ratio >1 indicates that the presence of the factor increases the odds of the outcome (eg, readmission). Naïve Bayes is a probabilistic classification algorithm based on the Bayes theorem [35] with the assumption that variables are independent [36]. Classifications are achieved by assigning the class label that can maximize the posterior probability given the features of an instance. A naïve Bayes model can be interpreted by taking the conditional probability of a variable given a class, and a higher probability indicates a stronger relationship with the class. Decision trees are a family of tree-structured predictive algorithms that iteratively split the data into disjoint subsets in a greedy manner [37]. Classifications are made by walking the tree splits until arriving at a leaf node (the class). Decision trees are self-explainable because each leaf node is represented as an if-then rule, and the decision process can be visualized. The contribution of a variable to the classification can be measured using various methods, such as information gain based on information theory and Gini importance [38]. Random forests are ensemble learning algorithms generated by bootstrap aggregation; the algorithm repeatedly selects a random sample from the training data set (with replacement) and builds a decision tree for the sample [39]. When making predictions, the outputs from different decision trees will be ensembled. Gradient boosting trees are another type of tree ensemble algorithm; they build the model in a stagewise fashion by iteratively generating new trees to improve the previous weaker trees [40]. Predictions are made by the weighted average of tree outcomes, with stronger trees having higher weights. Random forests and gradient tree boosting algorithms can be interpreted by measuring the Gini importance of the variables. Artificial neural networks are an interconnected group of computing units called artificial neurons [41]. The artificial neurons are aggregated into layers and connected by edges that have different weights to control the signals transmitted between neurons. The signals in the final output layer are used for prediction. The importance of each feature can be measured by the increase in prediction error after permuting the values of the feature.

We implemented the HOSPITAL score, LACE index, and LACE-rt index to compare their performance with that of our models. The HOSPITAL score has seven variables, including hemoglobin level at discharge, discharge from an oncology service, sodium level at discharge, any ICD procedures during the hospital stay, the type of index admission, the number of admissions 1 year before the index admission, and the length of stay [13]. Each factor level has a weighted point, and the maximum total score is 13 points. The LACE index has four variables, including length of stay, acuity of admission, the Charlson comorbidity index, and the number of emergency department visits 6 months before the index admission [14]. Its score ranges from 0 to 19 points. The LACE-rt index has the same variable weights and the same maximum score as the original LACE index [20]. The only difference is that it requires the length of stay during the previous admission within last 30 days instead of the current admission.

Model Training and Benchmark

Based on the inclusion and exclusion criteria, we identified 96,550 eligible patients. We randomly split the 96,550 records into a development data set (91,550 records) and a validation data set (5000 records). The readmission rate (11.7%) was preserved in these two data sets. The development set was used to derive and test the five candidate models in 10-fold cross-validation. The validation set was used to assess if the models can be generalized to unseen data. For the three baseline models, we extracted the required variables from the encounters in the validation set so that we could perform a fair comparison of our candidate models and the baseline models.

Because accuracy is sensitive to class imbalance, it cannot be used to evaluate readmission models (readmission rate <50%). To measure the performance of the models, we used the AUC, precision, recall, specificity, and F1 measure, which are less sensitive to data imbalance. The AUC is the probability that a model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. The AUC ranges from 0.5 to 1.0, with 1.0 indicating that the model has perfect discrimination ability and 0.5 indicating that it performs no better than random guessing. Precision is the fraction of true positives among all instances predicted to be positive. Recall is the fraction of correctly identified positives in all positive instances. Specificity is the fraction of correctly identified negatives in all negative instances. The F1 measure is the harmonic mean of precision and recall. The values of the precision, recall, specificity, and F1 measure range from 0 to 1.0. A higher value indicates better performance.

Multivariate Logistic Regression Analysis

We performed multivariate logistic regression analysis [32] to evaluate associations between the independent variables and the patients' readmission status. Features were selected by backward elimination. We chose the significance level of .05 for the statistical tests.

Software

We used Weka [42] to build and evaluate the logistic regression, naïve Bayes, decision tree, random forest, and gradient boosting tree models. The extreme gradient boosting (XGBoost) model and neural network models were developed in Python. The hyperparameters of these models were optimized. We implemented the HOSPITAL score, LACE index, and LACE-rt index in Python. The multivariate logistic regression analysis used the GLM package in R (R Project).

Results

Patient Demographics

From the 96,550 included patients, 11,294 experienced unplanned 30-day hospital readmission. The readmission rate (11.7%) is lower than the Medicare readmission rate (15.2% [1]). One possible reason was that in contrast to Medicare patients, who are normally older than 65 years, our study population included younger and less vulnerable adult patients (aged 18 to 64 years) as well as older adults (aged 65 years and above). Table 2 shows the demographic information of patients with and without readmissions. Most patients were White, female, and between 65 and 79 years of age.

Table 2. Demographic information of the 96,550 patients included in the data set. The characteristics with the highest frequencies are indicated with italic text.

Characteristic	Value, n (%)	
	Readmission=yes	Readmission=no
Total	11,294 (11.7)	85,256 (88.3)
Age (years)		
18-34	930 (8.2)	13,242 (15.5)
35-49	1525 (13.5)	12,541 (14.7)
50-64	3116 (27.6)	21,559 (25.3)
*65-79**	*3380 (29.9)*	*22,634 (26.5)*
≥80	2343 (20.7)	15,280 (17.9)
Sex		
Female	*5966 (52.8)*	*49,619 (58.2)*
Male	5328 (47.2)	35,637 (41.8)
Race		
African American	2612 (23.1)	16,248 (19.1)
White	*7750 (68.6)*	*61,685 (72.3)*
Other	932 (8.3)	7323 (8.6)

*Italicized text represents majority in the group.

Model Development and Selection

Table 3 shows the 10-fold cross-validation AUCs (mean and standard deviation) of the six candidate models on the development data set. Especially, the alternating decision tree (ADTree) algorithm [43], the XGBoost [44] algorithm, and the feedforward neural networks with three hidden layers (256 neurons, 512 neurons, and 256 neurons) had the best AUCs within the decision tree, gradient boosting tree, and artificial neural network families, respectively.

Table 3. 10-fold cross-validation AUCs of the candidate models on the development set.

Model	10-fold cross-validation AUC[a], mean (SD)
Logistic regression	0.750 (0.005)
Naïve Bayes	0.730 (0.006)
Alternating decision tree	0.730 (0.010)
Random forest	0.734 (0.006)
XGBoost[b]	0.753 (0.007)
Neural network	0.746 (0.004)

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

[b]XGBoost: extreme gradient boosting.

We further compared the performance of the six candidate models and the three baseline models (HOSPITAL score, LACE index, and LACE-rt index) on the validation data set by measuring the precision, recall, specificity, F1 measure, and AUC (Table 4). Because of the imbalanced prevalence of readmissions (eg, 11.7% in this study), it was infeasible to use 0.5 as the cutoff probability to dichotomize probabilistic outputs. We chose cutoffs that could maximize the Youden index of each model [45]. The optimal cutoffs of the three baseline models are integers because they do not generate probabilities. It can be seen that the random forest model has the best specificity and precision, while the XGBoost model has the best recall, F1 measure, and AUC. In the medical domain, recall is a more important metric because false negatives are considered more risky than false positives. Therefore, although the XGBoost and logistic regression models had similar AUC on development set (Table 3) and validation set, we chose XGBoost as the final model. It can be seen that the XGBoost model is better than the three baseline models in all performance metrics. Features of the XGBoost model and their importance ranking are shown in Multimedia Appendix 1. The optimized XGBoost model has “gbtree” as the booster, “binary:logistic” as the objective, a gamma of 0.4, a learning rate of 0.1, a maximum depth of 3, a maximum delta step of 0, a minimum child weight of 8, a reg_alpha parameter of 5, a reg_lambda parameter of 1, a subsample of 0.7, and 240 estimators.

Table 4. Performance of the candidate models and baseline models on the validation set. The best-performing parameters are indicated in italic text.

Model	Optimal cutoff	Specificity	Precision	Recall	F1 measure	AUC[a]
Logistic regression	0.157	0.642	0.857	0.729	0.773	0.741
Naïve Bayes	0.220	0.666	0.855	0.685	0.740	0.720
Alternating decision tree	0.298	0.662	0.857	0.705	0.755	0.732
Random forest	0.122	*0.747*	*0.862*	0.611	0.680	0.726
XGBoost[b]	0.175	0.611	0.856	*0.759*	*0.794*	*0.743*
Neural network	0.125	0.686	0.858	0.681	0.737	0.735
HOSPITAL score	4	0.564	0.838	0.694	0.745	0.688
LACE index	11	0.469	0.830	0.745	0.779	0.675
LACE-rt index	7	0.542	0.833	0.688	0.740	0.668

[a]AUC: area under the curve.

[b]XGBoost: extreme gradient boosting.

Risk Factors and Protective Factors of Readmission

To understand the statistical significance of the factors, we performed multivariate logistic regression analysis on all the data (96,550 records and 432 variables). By backward elimination, we reduced the feature space to 83, as shown in Multimedia Appendix 2. We reidentified 40 risk factors and significant predictors reported in previous studies. In addition, we discovered 14 risk factors and 2 protective factors that have never been reported in the literature. These 16 novel factors belong to 13 variables, and they are displayed in Table 5.

Table 5. The 14 novel risk factors and 2 novel protective factors of readmission identified in the study.

Risks or protective factors	Coefficient	*P* value	Odds ratio (95% CI)
Medical history in last year			
Diagnosis			
1 maintenance chemotherapy visit in the last year	0.390	<.001	1.476 (1.218-1.790)
Laboratory test result			
1 abnormal lymphocyte count test in the last year	0.221	<.001	1.247 (1.144-1.359)
≥2 abnormal lymphocyte count tests in the last year	0.228	.001	1.257 (1.091-1.447)
1 abnormal monocyte count test in the last year	0.182	.005	1.199 (1.056-1.362)
≥2 abnormal monocyte percent tests in the last year	0.316	<.001	1.371 (1.178- 1.596
1 abnormal serum calcium quantitative test in the last year	0.226	<.001	1.254 (1.107-1.420)
≥2 abnormal serum calcium quantitative tests in the last year	0.297	.001	1.345 (1.122-1.612)
Medication			
1 albuterol ipratropium order in the last year	0.071	.02	1.073 (1.010-1.141)
≥2 albuterol ipratropium orders in the last year	0.145	.003	1.157 (1.052-1.272)
1 cefazolin order in the last year	–0.123	.001	0.884 (0.822-0.950)
Index admission			
Demographic information			
Index admission to hospital in Northeast census region	0.365	<.001	1.441 (1.345-1.543)
Medication			
Gabapentin ordered at index admission	0.162	<.001	1.176 (1.113-1.243)
Ondansetron ordered at index admission	0.105	<.001	1.111 (1.057-1.168)
Polyethylene glycol 3350 ordered at index admission	0.073	.01	1.076 (1.017-1.139)
Cefazolin ordered at index admission	–0.147	<.001	0.863 (0.798-0.934)
Laboratory test result			
≥16 abnormal laboratory test results at index admission	0.140	.005	1.151 (1.043-1.269)

Discussion

Novel Risk Factors and Protective Factors of Readmission

The 14 novel risk factors and 2 novel protective factors of readmission are related to medical history and index admission. They belong to four categories: diagnosis, laboratory test results, medications, and demographic information.

Patients with one CCS-level diagnosis of maintenance chemotherapy in the previous year were found to be more likely to be readmitted than patients without this diagnosis. This can be explained by the linkage between chemotherapy and cancer, which has been reported as a predictor of readmission [46,47].

A blood disorder or an abnormal amount of substance in the blood can indicate certain diseases or side effects. Having an increased number of abnormal test results indicates that the patient is frailer and can be more prone to readmission.

Four medications were found to be positively linked to readmission. These medications may have side effects that are associated with readmission. Another interpretation is that conditions treated by these medications may be related to readmission. For example, albuterol ipratropium is a combination of two bronchodilators, which are used in the treatment of chronic obstructive pulmonary disease (COPD). COPD has been reported as a risk factor of readmission [47]. It is interesting that the prescriptions of cefazolin in previous encounters and at index admission were both negatively associated with readmission. One possible explanation is that cefazolin is an antibiotic that is used to treat infections caused by bacteria. The use of cefazolin may reduce patients' chance of infection and reduce their readmission risk.

The Northeast census region was found to be more positively associated with readmission than the Midwest census region. One possible reason is that geolocation is associated with socioeconomic status, which has been reported to be linked to readmission [48].

Timeliness of Prediction

Most readmission predictive models are based on index admission data. Many highly predictive variables of the index admission, such as the length of stay, diagnosis codes, procedure codes, and laboratory test results before discharge, are only available near or after discharge. To achieve good predictive performance, most studies include these variables in their

models. As a result, these models can only be used near or after discharge. They are useful for public reporting but not for clinical decision support because they are not timely.

In this work, we used the data from index admission and patients' medical history up to 1 year before the index admission. To ensure that the model could work at any time during hospitalization, we only used index admissions data that become available in the EHR within 24 hours during hospitalization, such as medication orders and laboratory test results. We used the detailed medical history from the patients' previous encounters. Although some studies include medical histories in their models, they only use high-level information from previous encounters (eg, the number of inpatient stays in the previous year) instead of detailed information such as previous laboratory test results. By using the patients' detailed medical history, we were able to add more variables to the model without sacrificing its timeliness. As a result, our model enables point-of-care prediction and can be used to continuously monitor the readmission risk during the entire episode of hospitalization.

Generalizability

In addition to the performance of the model, we considered its generalizability. From the modeling point of view, generalizability indicates if a model can achieve similar performance on new data. In other words, the model should be trained and built using a large and diverse training sample to represent the whole population. Most existing readmission prediction models were based on relatively homogenous (eg, single-center studies) and small (eg, less than 20,000 patients) samples. For example, the LACE index and the HOSPITAL score were derived from only 4812 Canadian and 9212 American patients, respectively [13,14]. To ensure good generalizability, we captured all eligible inpatient encounters in 2016 from the Health Facts database, with 96,550 patients discharged from 205 hospitals across the four US Census regions. The best performing model (XGBoost) has a validation AUC close to the mean 10-fold cross-validation AUC on the development set (0.742 vs 0.753). This indicates that the model has good generalizability.

Another consideration of generalizability is whether the model can work on various types of patients. There is no consensus on data inclusion criteria for readmission studies, and the study outcomes span condition- or procedure-specific to all-cause readmission predictive models [27]. The choice between these two types of models has long been under debate. In two systematic reviews [8,12] of 99 readmission predictive models reported between 1985 and 2015, 77% of the models were specialized for one patient subpopulation. The condition-specific design limits the adaptability of the models to other patient subpopulations and may overlook patients in some at-risk minority groups if specific models are not available [49,50]. In practice, it can be challenging for a hospital to maintain separate readmission prediction models for different patient subpopulations, and this situation will be further exacerbated if patients have comorbidities [50]. All-cause models are designed for broad patient populations without limiting diagnoses or procedures. In this work, we were interested in hospital-wide readmissions caused by medical and health care–related reasons. Our model is not specific to any conditions or procedures because we wanted to use it as an early screening tool to assess all patients' risk.

Limitations

Although our model was designed to be nonspecific to patient populations, it does not work for patients under 18 years of age. This is because infant and pediatric readmissions were reported to have different patterns from adult readmissions [12,51] and could be influenced by parental factors [51,52]. The Health Facts database is deidentified, and there is no information about the patients' families. Therefore, we removed patients aged younger than 18 years from the data. In addition, the Health Facts database only contains data collected from US health care settings. For readmissions in other countries, where patient demographics (eg, race) and medical interventions (eg, medications) are different from those in the United States, our model may not work well.

Another limitation was that the 14 novel risk factors and 2 protective factors were identified based on associations. Because this work was a retrospective study on deidentified data, we were not able to further investigate the relationship between our findings and factors reported in other studies.

Conclusions

In this work, we developed an early prediction model for unplanned 30-day hospital readmission. The model has better performance (AUC of 0.753 on the development data set and 0.742 on the validation data set) and timeliness than established readmission models such as the HOSPITAL score, LACE index, and LACE-rt index. The model was derived and validated from a large and diverse patient population (96,550 patients discharged from 205 hospitals across four US census regions), and it can be generalized in use for adult patients in the United States. We identified 14 novel risk factors and 2 novel protective factors of readmission that may shed light on the understanding of the complex readmission problem. More studies or trials are necessary to verify the relationship of these factors with readmission in the future.

Acknowledgments

We appreciate Cerner Corporation and the University of Missouri School of Medicine for providing the Health Facts data and computing resources.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Features of the XGBoost model.
[PDF File (Adobe PDF File), 894 KB - medinform_v9i3e16306_app1.pdf]

Multimedia Appendix 2
Multivariate logistic regression analysis.
[XLSX File (Microsoft Excel File), 23 KB - medinform_v9i3e16306_app2.xlsx]

References

1. Hospital Quality Initiative - Outcome Measures 2016 Chartbook Internet. The Centers for Medicare & Medicaid Services. 2016. URL: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/OutcomeMeasures.html [accessed 2021-03-15]
2. Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the Medicare fee-for-service program. N Engl J Med 2009 Apr 02;360(14):1418-1428. [doi: 10.1056/NEJMsa0803563] [Medline: 19339721]
3. The Patient Protection and Affordable Care Act Internet. The 111th United States Congress. 2010. URL: https://www.govinfo.gov/content/pkg/PLAW-111publ148/pdf/PLAW-111publ148.pdf [accessed 2019-08-22]
4. Readmissions Reduction Program (HRRP) Internet. Centers for Medicare & Medicaid Services (CMS). 2012. URL: https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program.html [accessed 2019-08-22]
5. Kripalani S, Theobald CN, Anctil B, Vasilevskis EE. Reducing hospital readmission rates: current strategies and future directions. Annu Rev Med 2014;65:471-485 [FREE Full text] [doi: 10.1146/annurev-med-022613-090415] [Medline: 24160939]
6. Hansen LO, Young RS, Hinami K, Leung A, Williams MV. Interventions to reduce 30-day rehospitalization: a systematic review. Ann Intern Med 2011 Oct 18;155(8):520-528. [doi: 10.7326/0003-4819-155-8-201110180-00008] [Medline: 22007045]
7. Fox M. Nurse-led early discharge planning for chronic disease reduces hospital readmission rates and all-cause mortality. Evid Based Nurs 2016 Apr;19(2):62. [doi: 10.1136/eb-2015-102197] [Medline: 26701748]
8. Zhou H, Della PR, Roberts P, Goh L, Dhaliwal SS. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. BMJ Open 2016 Jun 27;6(6):e011060 [FREE Full text] [doi: 10.1136/bmjopen-2016-011060] [Medline: 27354072]
9. Naylor MD, Brooten D, Campbell R, Jacobsen BS, Mezey MD, Pauly MV, et al. Comprehensive discharge planning and home follow-up of hospitalized elders: a randomized clinical trial. JAMA 1999 Feb 17;281(7):613-620. [doi: 10.1001/jama.281.7.613] [Medline: 10029122]
10. Koehler BE, Richter KM, Youngblood L, Cohen BA, Prengler ID, Cheng D, et al. Reduction of 30-day postdischarge hospital readmission or emergency department (ED) visit rates in high-risk elderly medical patients through delivery of a targeted care bundle. J Hosp Med 2009 Apr;4(4):211-218. [doi: 10.1002/jhm.427] [Medline: 19388074]
11. Evans RL, Hendricks RD. Evaluating hospital discharge planning: a randomized clinical trial. Med Care 1993 Apr;31(4):358-370. [doi: 10.1097/00005650-199304000-00007] [Medline: 8464252]
12. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. JAMA 2011 Oct 19;306(15):1688-1698 [FREE Full text] [doi: 10.1001/jama.2011.1515] [Medline: 22009101]
13. Donzé J, Aujesky D, Williams D, Schnipper JL. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. JAMA Intern Med 2013 Apr 22;173(8):632-638. [doi: 10.1001/jamainternmed.2013.3023] [Medline: 23529115]
14. van Walraven C, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. CMAJ 2010 Apr 06;182(6):551-557 [FREE Full text] [doi: 10.1503/cmaj.091117] [Medline: 20194559]
15. Stricker P. Best practice strategies and interventions to reduce hospital readmission rates. TCS Healthcare Technologies. 2018. URL: https://www.tcshealthcare.com/clinical-corner/best-practice-strategies-and-interventions-to-reduce-hospital-readmission-rates/ [accessed 2019-08-22]
16. Wang H, Cui Z, Chen Y, Avidan M, Abdallah AB, Kronzer A. Predicting hospital readmission via cost-sensitive deep learning. IEEE/ACM Trans Comput Biol Bioinform 2018;15(6):1968-1978. [doi: 10.1109/TCBB.2018.2827029] [Medline: 29993930]
17. Horne BD, Budge D, Masica AL, Savitz LA, Benuzillo J, Cantu G, et al. Early inpatient calculation of laboratory-based 30-day readmission risk scores empowers clinical risk modification during index hospitalization. Am Heart J 2017 Mar;185:101-109. [doi: 10.1016/j.ahj.2016.12.010] [Medline: 28267463]
18. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983 Sep;148(3):839-843. [doi: 10.1148/radiology.148.3.6878708] [Medline: 6878708]

19. Cronin PR, Greenwald JL, Crevensten GC, Chueh HC, Zai AH. Development and implementation of a real-time 30-day readmission predictive model. AMIA Annu Symp Proc 2014;2014:424-431 [FREE Full text] [Medline: 25954346]
20. El Morr C, Ginsburg L, Nam S, Woollard S. Assessing the performance of a modified LACE index (LACE-rt) to predict unplanned readmission after discharge in a community teaching hospital. Interact J Med Res 2017 Mar 08;6(1):e2 [FREE Full text] [doi: 10.2196/ijmr.7183] [Medline: 28274908]
21. Shadmi E, Flaks-Manov N, Hoshen M, Goldman O, Bitterman H, Balicer R. Predicting 30-day readmissions with preadmission electronic health record data. Med Care 2015 Mar;53(3):283-289. [doi: 10.1097/MLR.0000000000000315] [Medline: 25634089]
22. Flaks-Manov N, Topaz M, Hoshen M, Balicer R, Shadmi E. Identifying patients at highest-risk: the best timing to apply a readmission predictive model. BMC Med Inform Decis Mak 2019 Jun 26;19(1):118 [FREE Full text] [doi: 10.1186/s12911-019-0836-6] [Medline: 31242886]
23. Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA, Zhang S, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Med Care 2010 Nov;48(11):981-988. [doi: 10.1097/MLR.0b013e3181ef60d9] [Medline: 20940649]
24. Health Facts Database. Cerner Corporation. URL: https://www.cerner.com/ap/en/solutions/data-research [accessed 2020-01-01]
25. 2018 Condition-Specific Measures Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Readmission Measures. Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. 2018. URL: https://qualitynet.cms.gov/files/5d0d375a764be766b010141f?filename=2018_Rdmsn_Updates%26Specs_Rpts.zip [accessed 2020-01-01]
26. 2018 All-Cause Hospital Wide Measure Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Readmission Measures. Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. 2018. URL: https://qualitynet.cms.gov/files/5d0d375a764be766b010141f?filename=2018_Rdmsn_Updates%26Specs_Rpts.zip [accessed 2020-01-01]
27. Zhao P, Yoo I. A Systematic Review of Highly Generalizable Risk Factors for Unplanned 30-Day All-Cause Hospital Readmissions. J Health Med Inform 2017;08(04). [doi: 10.4172/2157-7420.1000283]
28. International Classification of Diseases. World Health Organization. URL: https://www.who.int/classifications/icd/en/ [accessed 2021-03-14]
29. Clinical classification software (CCS) for ICD10-CM/PCS. Healthcare Cost and Utilization Project. 2018. URL: https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp [accessed 2021-03-15]
30. Current procedural terminology (CPT). American Medical Association. URL: https://www.ama-assn.org/amaone/cpt-current-procedural-terminology [accessed 2021-03-15]
31. HCPCS Level II codes. The Centers for Medicare & Medicaid Services, America's Health Insurance Plans, Blue Cross and Blue Shield Association. URL: https://hcpcs.codes/ [accessed 2021-03-15]
32. Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc Ser A 1972;135(3):370. [doi: 10.2307/2344614]
33. Cox DR. The regression analysis of binary sequences. J R Stat Soc Ser B 2018 Dec 05;20(2):215-232. [doi: 10.1111/j.2517-6161.1958.tb00292.x]
34. Szumilas M. Explaining odds ratios. J Can Acad Child Adolesc Psychiatry 2010 Aug;19(3):227-229 [FREE Full text] [Medline: 20842279]
35. Bayes T, Price R. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. Phil Trans R Soc 1997 Jan 31;53:370-418. [doi: 10.1098/rstl.1763.0053]
36. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn 1997;29:103-130 [FREE Full text] [doi: 10.1023/A:1007413511361]
37. Quinlan JR. Induction of decision trees. Mach Learn 1986 Mar;1(1):81-106. [doi: 10.1007/bf00116251]
38. Breiman L. Random forests. Mach Learn 2001;45(1):5-32 [FREE Full text] [doi: 10.1023/A:1010933404324]
39. Tin KH. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. 1995 Presented at: 3rd International Conference on Document Analysis and Recognition; August 14-16, 1995; Montréal, QC p. 278-282. [doi: 10.1109/icdar.1995.598994]
40. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002 Feb;38(4):367-378. [doi: 10.1016/s0167-9473(01)00065-2]
41. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biol 1943 Dec;5(4):115-133. [doi: 10.1007/bf02478259]
42. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software. SIGKDD Explor Newsl 2009 Nov 16;11(1):10-18. [doi: 10.1145/1656274.1656278]
43. Freund Y, Mason L. The alternating decision tree learning algorithm. In: Proceedings of the Sixteenth International Conference on Machine Learning. 1999 Presented at: Sixteenth International Conference on Machine Learning; June 27-30, 1999; Bled, Slovenia p. 124-133.

44. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.: ACM Press; 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, CA; August 13-17, 2016. [doi: 10.1145/2939672.2939785]
45. Bay S, Pazzani M. Detecting change in categorical data: mining contrast sets. In: KDD '99: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, New York, USA: ACM Press; 1999 Presented at: Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 1999; San Diego, CA. [doi: 10.1145/312129.312263]
46. Tonkikh O, Shadmi E, Flaks-Manov N, Hoshen M, Balicer RD, Zisberg A. Functional status before and during acute hospitalization and readmission risk identification. J Hosp Med 2016 Sep;11(9):636-641. [doi: 10.1002/jhm.2595] [Medline: 27130176]
47. Lin K, Chen P, Huang L, Mao H, Chan DD. Predicting inpatient readmission and outpatient admission in elderly: a population-based cohort study. Medicine (Baltimore) 2016 Apr;95(16):e3484. [doi: 10.1097/MD.0000000000003484] [Medline: 27100455]
48. Kim H, Hung WW, Paik MC, Ross JS, Zhao Z, Kim G, et al. Predictors and outcomes of unplanned readmission to a different hospital. Int J Qual Health Care 2015 Dec;27(6):513-519 [FREE Full text] [doi: 10.1093/intqhc/mzv082] [Medline: 26472739]
49. Lavenberg JG, Leas B, Umscheid CA, Williams K, Goldmann DR, Kripalani S. Assessing preventability in the quest to reduce hospital readmissions. J Hosp Med 2014 Sep;9(9):598-603 [FREE Full text] [doi: 10.1002/jhm.2226] [Medline: 24961204]
50. Tanzer M, Heil E. Why majority of readmission risk assessment tools fail in practice. 2013 Presented at: 2013 IEEE International Conference on Healthcare Informatics; September 9-11, 2013; Philadelphia, PA p. 567-569. [doi: 10.1109/ichi.2013.89]
51. Vest JR, Gamm LD, Oxford BA, Gonzalez MI, Slawson KM. Determinants of preventable readmissions in the United States: a systematic review. Implement Sci 2010 Nov 17;5:88 [FREE Full text] [doi: 10.1186/1748-5908-5-88] [Medline: 21083908]
52. Berry JG, Ziniel SI, Freeman L, Kaplan W, Antonelli R, Gay J, et al. Hospital readmission and parent perceptions of their child's hospital discharge. Int J Qual Health Care 2013 Oct;25(5):573-581 [FREE Full text] [doi: 10.1093/intqhc/mzt051] [Medline: 23962990]

Abbreviations

ADTree: alternating decision tree
AUC: area under the receiver operating characteristic curve
CCS: Clinical Classifications Software
CMS: Centers for Medicare & Medicaid Services
COPD: chronic obstructive pulmonary disease
CPT: current procedural terminology
EHR: electronic health record
HCPCS: Healthcare Common Procedure Coding System
HIPAA: Health Insurance Portability and Accountability Act
HRRP: hospital readmissions reduction program
ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification
ICD-10-PCS: International Classification of Diseases, Tenth Revision, Procedure Coding System
XGBoost: extreme gradient boosting

Edited by G Eysenbach; submitted 18.09.19; peer-reviewed by E Shadmi, V Foufi, Z Zhang, TW Chien, P Roberts; comments to author 26.10.20; revised version received 06.02.21; accepted 03.03.21; published 23.03.21.

Please cite as:
Zhao P, Yoo I, Naqvi SH
Early Prediction of Unplanned 30-Day Hospital Readmission: Model Development and Retrospective Data Analysis
JMIR Med Inform 2021;9(3):e16306
URL: https://medinform.jmir.org/2021/3/e16306
doi:10.2196/16306
PMID:33755027

Original Paper

Noninvasive Real-Time Mortality Prediction in Intensive Care Units Based on Gradient Boosting Method: Model Development and Validation Study

Huizhen Jiang[1*], BSc; Longxiang Su[2*], MD; Hao Wang[2], MD; Dongkai Li[1], MD; Congpu Zhao[1], BSc; Na Hong[3], PhD; Yun Long[2], MD; Weiguo Zhu[1], MD

[1]Department of Information Center, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Science and Peking Union Medical College, Beijing, China

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

[3]Digital Health China Technologies Co., Ltd, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Weiguo Zhu, MD
Department of Information Center, State Key Laboratory of Complex Severe and Rare Diseases
Peking Union Medical College Hospital
Chinese Academy of Medical Science and Peking Union Medical College
1 Shuaifuyuan, Dongcheng District
Beijing, 100730
China
Phone: 86 01069154149
Email: Zhuwg@pumch.cn

Abstract

Background: Monitoring critically ill patients in intensive care units (ICUs) in real time is vitally important. Although scoring systems are most often used in risk prediction of mortality, they are usually not highly precise, and the clinical data are often simply weighted. This method is inefficient and time-consuming in the clinical setting.

Objective: The objective of this study was to integrate all medical data and noninvasively predict the real-time mortality of ICU patients using a gradient boosting method. Specifically, our goal was to predict mortality using a noninvasive method to minimize the discomfort to patients.

Methods: In this study, we established five models to predict mortality in real time based on different features. According to the monitoring, laboratory, and scoring data, we constructed the feature engineering. The five real-time mortality prediction models were RMM (based on monitoring features), RMA (based on monitoring features and the Acute Physiology and Chronic Health Evaluation [APACHE]), RMS (based on monitoring features and Sequential Organ Failure Assessment [SOFA]), RMML (based on monitoring and laboratory features), and RM (based on all monitoring, laboratory, and scoring features). All models were built using LightGBM and tested with XGBoost. We then compared the performance of all models, with particular focus on the noninvasive method, the RMM model.

Results: After extensive experiments, the area under the curve of the RMM model was 0.8264, which was superior to that of the RMA and RMS models. Therefore, predicting mortality using the noninvasive method was both efficient and practical, as it eliminated the need for extra physical interventions on patients, such as the drawing of blood. In addition, we explored the top nine features relevant to real-time mortality prediction: invasive mean blood pressure, heart rate, invasive systolic blood pressure, oxygen concentration, oxygen saturation, balance of input and output, total input, invasive diastolic blood pressure, and noninvasive mean blood pressure. These nine features should be given more focus in routine clinical practice.

Conclusions: The results of this study may be helpful in real-time mortality prediction in patients in the ICU, especially the noninvasive method. It is efficient and favorable to patients, which offers a strong practical significance.

(JMIR Med Inform 2021;9(3):e23888) doi:10.2196/23888

KEYWORDS

real time; mortality prediction; intensive care unit; noninvasive

Introduction

Patients in intensive care units (ICUs) are usually suffering from the most severe and complicated diseases. Thus, they require more intensive care and hospital resources [1]. Research shows that the cost of ICUs accounts for 22% of total hospital costs [2]. The cost of doctors and nurses in the ICU is also a massive burden. Therefore, hospitals usually use scoring systems to help assess patients' risks and then place more efforts on improving the patient care and management. Scoring systems such as the Acute Physiology and Chronic Health Evaluation (APACHE) [3] systems II, III, and IV; the Simplified Acute Physiology Score II (SAPS II) [4]; and the Sequential Organ Failure Assessment (SOFA) score [5] are commonly used to estimate the illness severity of patients in the ICU [6,7]. However, the scoring systems cannot reflect the condition of patients in real time, and clinical staff must spend plenty of time calculating the scores to make decisions. Further, the scores alone are insufficient for the needs of the clinical staff. Johnson and Mark [8] found that the gradient boosting method outperformed the scoring systems on predicting mortality, which provided inspiration for our study.

Meanwhile, the severity and mortality of ICU patients can be specifically assessed in real time using machine learning methods. This would allow doctors and nurses to prepare lifesaving interventions ahead of time and provide families with more time to make decisions [9]. Hence, precisely predicting the mortality of ICU patients is significant. Machine learning technology has significantly changed lives in many aspects in recent years, even in the health care field [10,11]. Usually, the shortest time period for predicting mortality is 24 h [12,13], which is not sufficient for the ICU staff to obtain the real-time condition of patients. Kim et al [14] presented a deep learning method to predict the mortality of patients 6 h to 60 h prior to death, where the time period was a little longer than the real time. With regard to machine learning techniques, the ensemble and neural network models demonstrate better performance in predicting mortality [2]. Brand et al [15] proposed a deep learning method to predict mortality based only on heart rate, respiratory rate, and blood pressure, which had an accuracy of 76.3%, but its performance was not as good as that of other methods. Besides, the neural network model cannot interpret the gap between the input and the output. Further, it is vulnerable to attack when the training set is inadvertently being modified [16].

In this study, we established a real-time mortality prediction model based on clinical data where we explored a noninvasive method to predict mortality by only monitoring features. Because frequent laboratory examination can cause physical trauma to patients whose bodies are already weak, using a model that can show general performance and is noninvasive is clinically meaningful.

Methods

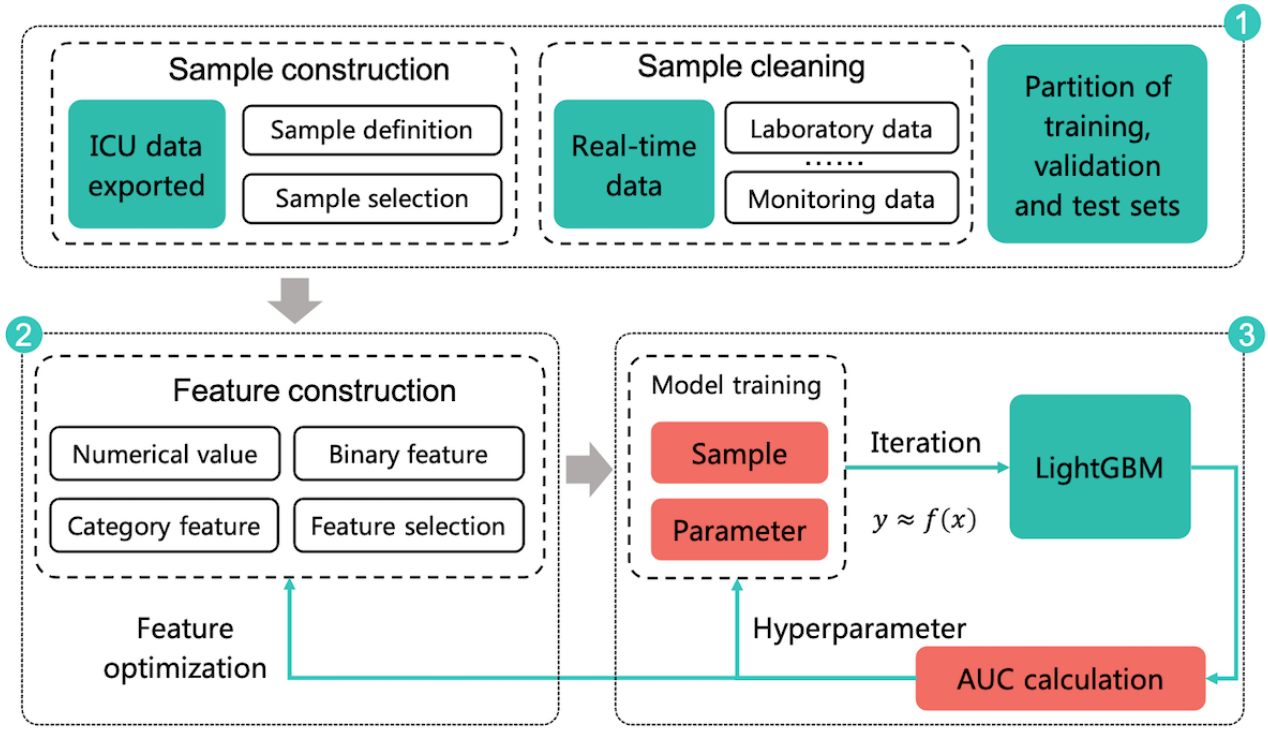
Data Sources

We used the ICU data from Peking Union Medical College Hospital from 2013 to 2018. A total of 13,649 patients were investigated in our experiments with the privacy information filtered out. We mined features from three types of data: real-time monitoring, laboratory, and scoring data. The main features from the monitoring data are listed in Multimedia Appendix 1.

Prediction Model

In this study, we constructed the real-time mortality prediction model based on the clinical data. The data of the patients in the ICU were updated all the time, and the model could predict each patient's mortality once the data were updated; the model could predict the mortality after 2 h at any time if the data were not updated during the 2 h period. Therefore, it is a real-time prediction model. The modeling process involved three steps, which are shown in Figure 1. First, we constructed and cleaned the sample data according to the clinical data. Second, after dividing the data into training and test sets, we normalized all types of data as features. Third, we used the LightGBM method to train the data and optimized the model by adjusting the parameters. LightGBM and XGBoost are both gradient boosting decision tree methods, and LightGBM has good performance and high training efficiency [17]. In this paper, we also compared the performance of the LightGBM and XGBoost methods.

Figure 1. Modeling process, including sample construction and cleaning, feature engineering, model training, and optimization. AUC: area under the curve; ICU: intensive care unit.

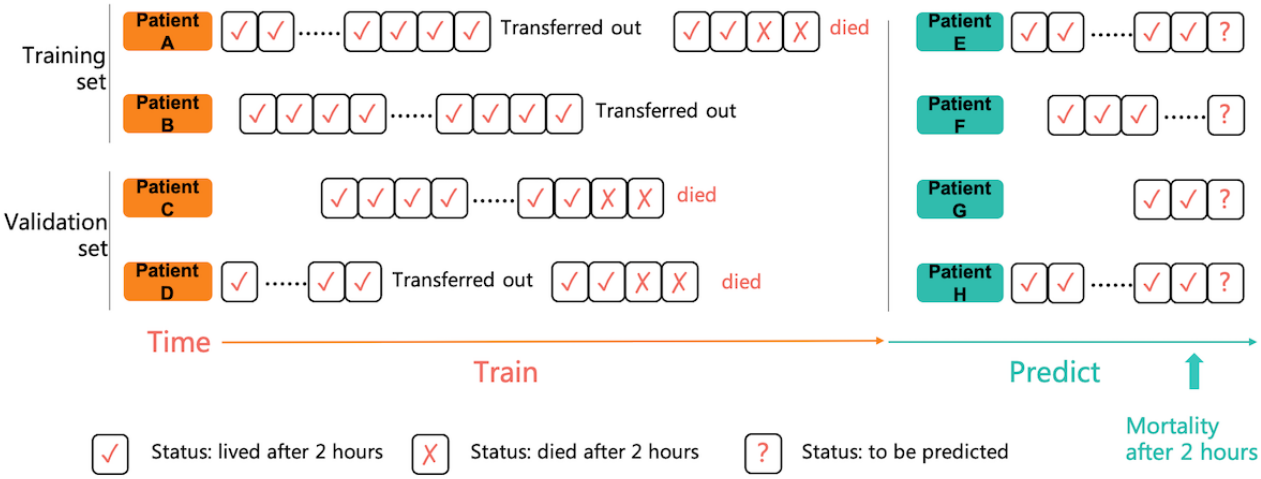


Sample Construction

Usually, patients in ICUs are weak and at high risk. Therefore, focusing on the real-time condition of an ICU patient by the clinical staff is meaningful. In this research, we predicted the mortality of a patient after 2 h based on the clinical data.

Figure 2 shows how we constructed the data by the hour. One record of a patient was captured in each hour, and each patient may have a sample sequence based on the timeline. As a result, there might be several samples for one patient. For example, patient A had been admitted to the ICU twice, and patients B and C had each been admitted to the ICU once. There were 2 samples for patient A. During patient A's second stay in the ICU, the third box contained a cross, representing a status of "died after 2 h," so he/she died 2 h after he/she was admitted into the ICU because one box meant one sample in 1 h. The data of patients A and B were the training data, and the data of patients C and D were the validation data. Samples of patients E, F, G, and H were the test data set. Then, the samples were constructed according to the process shown in Figure 2. For the 13,649 patients, we constructed 1,172,652 samples in all.

Figure 2. Feature engineering process. Each square represents a 1 h record in the intensive care unit (ie, one sample). The symbols in the squares indicate the status of the patient.



Feature Engineering

Feature engineering is the key process in machine learning. The modeling performance depends on the feature engineering quality to a large extent.

In this study, two data types existed: numerical and categorical data from the monitoring, laboratory, and scoring data. For the numerical data, we directly considered the numerical value as the feature, such as the heart rate and temperature. The categorical data included gender and positive or negative status. We used the LabelEncoder method [18] to normalize these categorical data. LabelEncoder is a method that converts text data into multinumeric values. It can convert two-class and multiclass features. For example, the positive and negative states were represented by 0 and 1, respectively. We left the missing value blank to ensure the authenticity of the data.

Model Training

In this research, we needed to predict the real-time condition of a patient 2 h after each moment. Actually, this process was a binary classification problem (ie, life or death). LightGBM is a gradient boosting method that is superior in dealing with the binary classification problem and has high efficiency and performance, especially in dealing with structured data. The 1,172,652 samples were randomly divided into three parts. One-third of the samples was set as the training set, one-third was set as the validation set, and the rest was set as the test set. The area under the curve (AUC) was used to evaluate the model's performance.

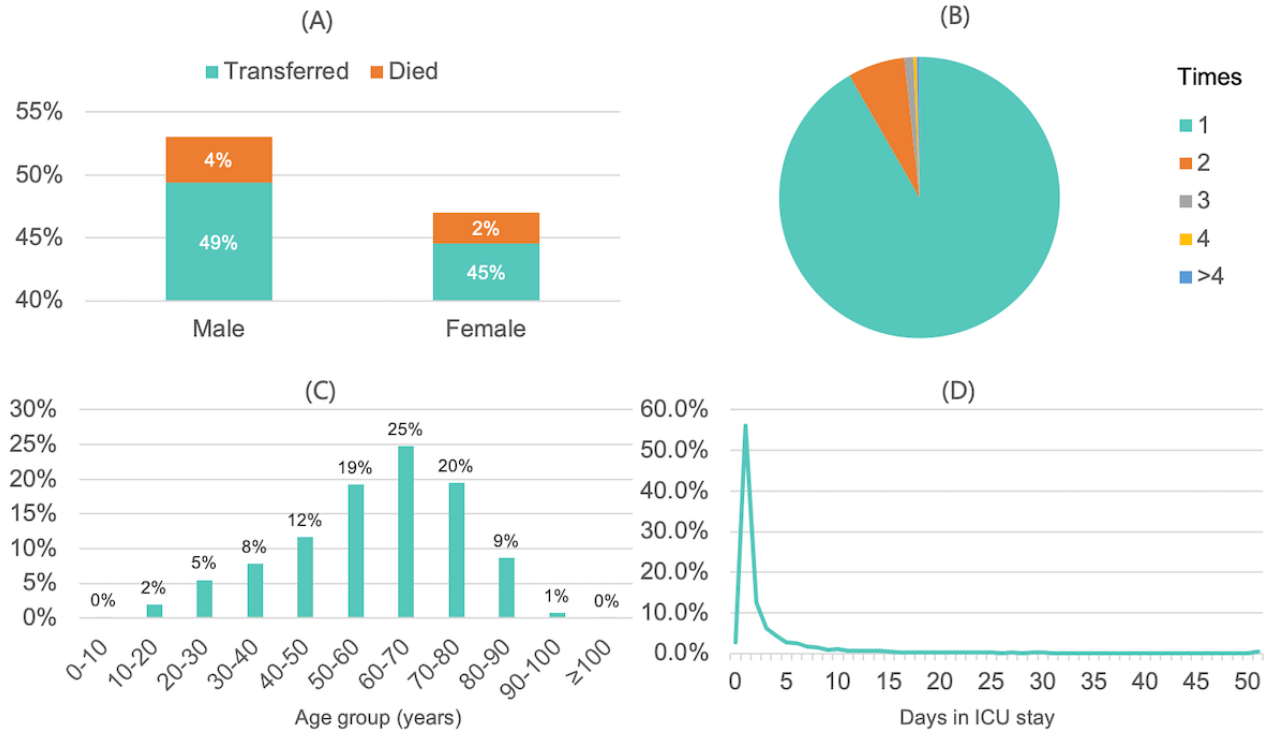
Based on different features, we constructed five real-time mortality prediction models:

- RMM: based on monitoring features;
- RMA: based on monitoring features and APACHE;
- RMS: based on monitoring features and SOFA;
- RMML: based on monitoring and laboratory features; and
- RM: based on all monitoring, laboratory, and scoring features.

Results

In presenting the results of our study, we will focus on the results of the models in the test set. Figure 3 shows the distributions and proportions of patients in the ICU in the data set. Figure 3A shows that male patients in the ICU outnumbered female patients irrespective of whether they were transferred out or died. Figure 3B shows that more than 12,510 patients were transferred into the ICU only once, and 898 patients stayed in the ICU twice. Patients between the ages of 50 and 80 years accounted for 8700 of the total number of patients, as shown in Figure 3C. In addition, patients between the ages of 60 and 70 years represented the largest group, accounting for one-quarter of the total number of patients. Figure 3D shows the length of stay of patients in the ICU in a single visit; we can observe that most patients stayed in the ICU for fewer than 5 days.

Figure 3. Distribution of the data set. (A) Proportion of patients that were transferred out of the intensive care unit (ICU) or died, according to gender. (B) Distribution of the number of times patients transferred into the ICU. (C) Age group distribution of ICU patients. (D) Length of stay of patients in the ICU.



First, we evaluated the influence of the scoring systems through extensive experiments; the results are shown in Figures 4 and 5. Figure 4 shows that the RMM model outperformed the RMA and RMS models. Overall, all three models showed an upward trend with the increase in tree number and became stable after the tree number reached 200. The RMS model demonstrated better performance than the RMA model. Therefore, the SOFA scoring system was more valuable than the APACHE scoring

system in predicting mortality. Compared with the RMA and RMS models, the RMM model was superior and demonstrated the best performance (AUC 0.8264) when the tree number was 299.

Figure 4. Performance of the RMM, RMA, and RMS models with parameter variation. Each point on the line represents one experiment. AUC: area under the curve.

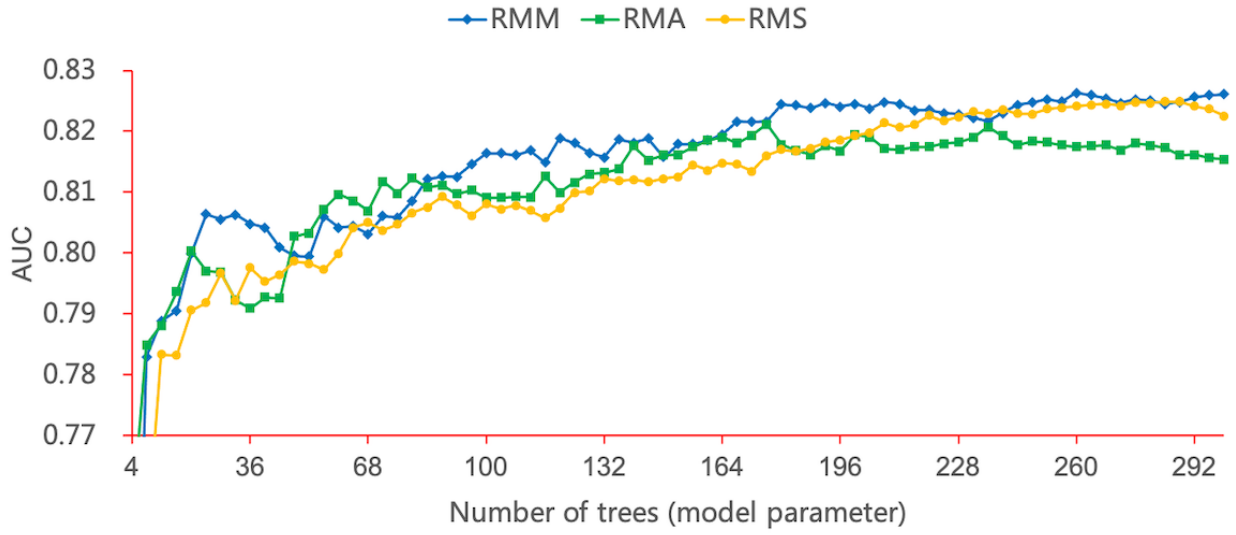
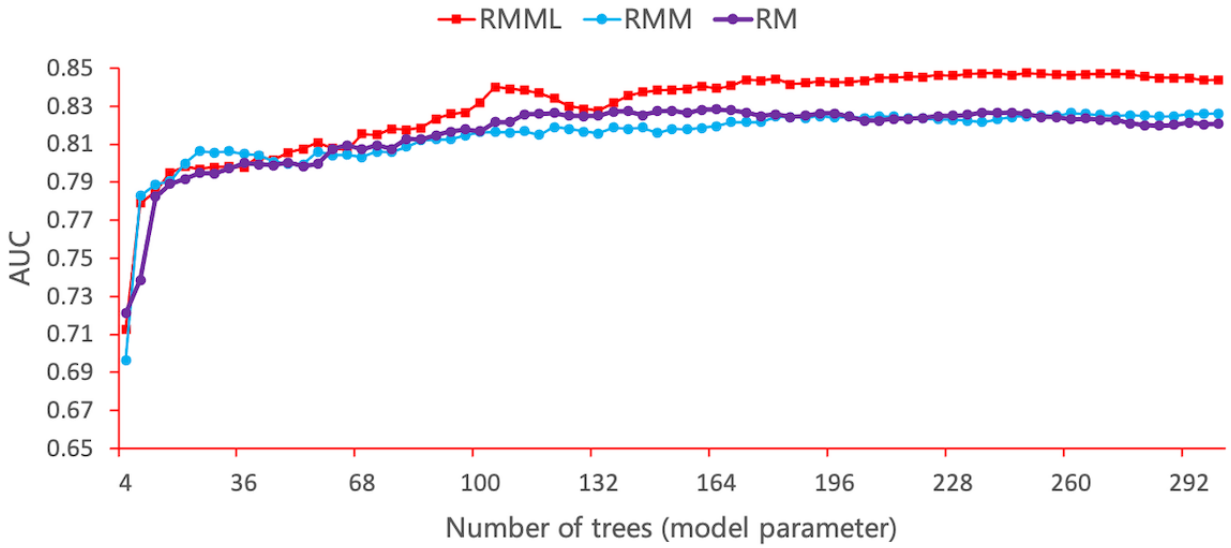


Figure 5 shows the results of the experiments that we conducted on the RMML, RMM, and RM models to compare their performance in terms of the monitoring, laboratory, and scoring features. The RMML model exhibited the best performance based on the monitoring and laboratory features than the other two models. When the tree number was 234, RMML obtained the best AUC (0.8476). In the RM model, monitoring, laboratory, and all scoring features were considered. The RM model exhibited worse performance than the RMM model.

Figure 5. Performance of the RMML, RMM, and RM models with parameter variation. Each point on the line represents one experiment. AUC: area under the curve.



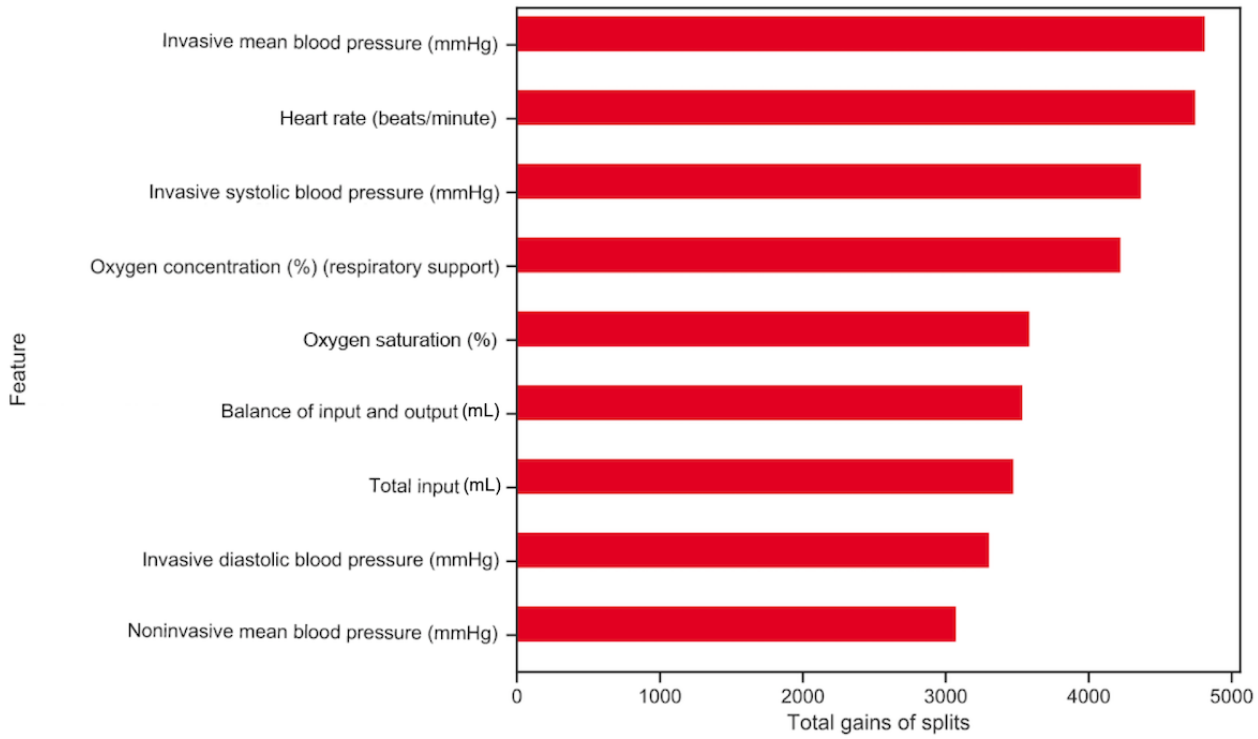
In addition, we repeated the experiments above using XGBoost. The AUCs of XGBoost were relatively lower than those of LightGBM, as shown in Table 1. The best performance with XGBoost was 0.8452 on the RMML model and 0.8154 on the RMM model. As well, we showed the RMML and RMM models using LightGBM and XGBoost on the validation set, and the results are shown in Table 1. Therefore, LightGBM outperformed XGBoost in these experiments.

Table 1. Performance of the RMML and RMM models using LightGBM and XGBoost.

Model and method	Area under the curve	
	Test set	Validation set
RMML		
LightGBM	0.8476	0.8483
XGBoost	0.8452	0.8466
RMM		
LightGBM	0.8264	0.8269
XGBoost	0.8154	0.8167

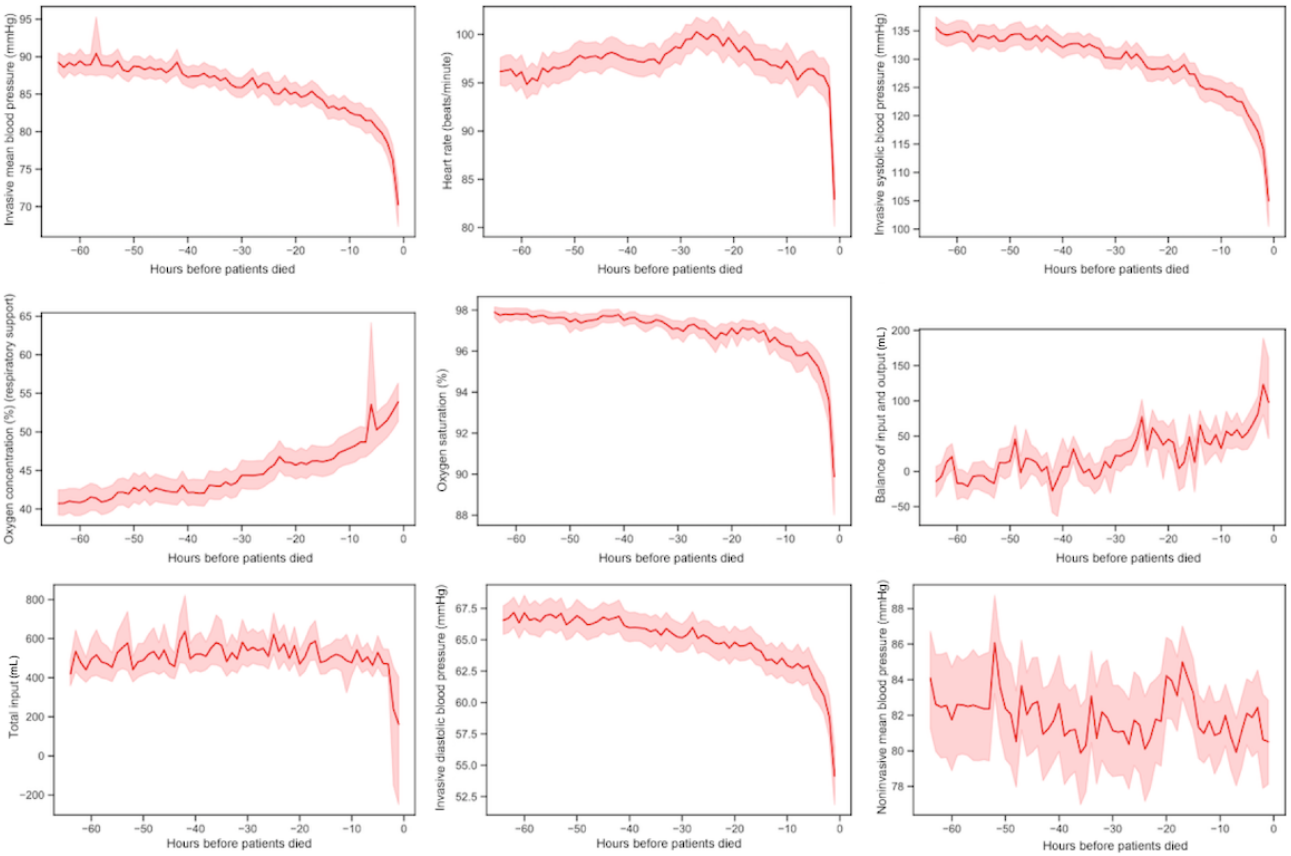
Because the RMML model demonstrated the best performance, we analyzed the relevant features that predicted mortality in that model. Figure 6 shows the top nine features relevant to the mortality prediction. The "gain" of the feature splitting implies the importance of the feature in the model, which was computed during the model training. Thus, the bigger the gains of the feature, the more important the feature was in the model. It was shown that invasive mean blood pressure was the most important feature related to mortality prediction. Among the top nine features, heart rate, invasive systolic blood pressure, oxygen concentration, oxygen saturation, balance of input and output, total input, invasive diastolic blood pressure, and noninvasive mean blood pressure were all vital sign features in the monitoring. "Balance of input and output" was the difference between input and output data, while "total input" was the input data only. They all demonstrated a relatively strong correlation with the mortality prediction.

Figure 6. Top nine features relevant to mortality prediction. The horizontal bar represents the gain of each feature in the model; a bigger gain means more relevance and importance in the mortality prediction.



In addition, we exploited the variation in each of the top nine features with time during the last 64 h before a patient died. Figure 7 shows that all nine features showed an obvious trend with the time variation. "Oxygen concentration" and "balance of input and output" exhibited an upward trend during the final 64 h before the patient died. The other seven features all decreased with time during the final 64 h before the patient died. For example, invasive diastolic blood pressure exhibited a downward trend and sharply declined in the last 5 h. Similarly, the top eight features all rapidly changed in the last 5 h. The "noninvasive mean blood pressure" exhibited dithering but an overall decreasing trend.

Figure 7. Variation in the top nine features relevant to mortality prediction with time. The abscissa represents 64 h before the patients died, and the ordinate represents the value of the feature.



Discussion

In this paper, we used clinical data to predict the real-time mortality of ICU patients. Several models were established based on different features. Extensive experiments showed that the models that used the machine learning method were superior to the scoring systems. More importantly, they can be employed to predict real-time mortality in a noninvasive manner.

Constant care of ICU patients is necessary against their life-threatening conditions. Intensive care is based on more financial support and more professional hospital staff [19,20]. The US health care spending was approximately 17% of the gross domestic product (GDP) in 2011 and may reach 26% of the GDP by 2035 [21]. In addition to the cost, the mortality rate in ICUs cannot be ignored. Studies show that ICUs have the highest mortality rate of all hospital units (16.2% [22] and 22.4% [23]). Therefore, helping predict patient mortality in ICUs is significant, as it could save time for nurses and doctors by more efficiently measuring the risk of ICU patients. It would be better if there was less trauma to patients in the clinical process.

Commonly, hospital staff use scoring systems to help predict the severity status of ICU patients. Most of these scoring systems calculate the scores based on the worst values during the first 24 h after ICU admission [24]. The SAPS score only uses the data in the first hour after ICU admission, which are more robust because the missing data have a lesser effect on specificity [25]. Saleh et al [24] compared APACHE II and III, SAPS II, and SOFA and showed that APACHE II and III demonstrated better performance than the others. However, Yap et al [26] verified that the National Early Warning Score demonstrated the best performance for predicting the severity status of patients with emphysematous pyelonephritis patients. Tan et al [27] explored the ability of the scoring systems to predict sepsis mortality in the short term (less than 30 days in the hospital) and long term (more than 30 days). They discovered that the sensitivity and specificity were similar in both factors, whereas geographical region had a significant effect on the short-term mortality prediction. Therefore, the scoring systems can show different performance on different diseases and under different situations [28,29]. In addition, Nielsen et al [30] compared the APACHE II and SAPS II with the aggregation of the APACHE II and SAPS II, and the aggregation of APACHE II and SAPS II outperformed each single model. Similarly, Fei et al [31] presented the use of the fibrin degradation product level and APACHE II scores in parallel to improve the prediction performance.

Machine learning technologies have been increasingly used in the health care field because of their excellent performance [20,32]. Further, machine learning models have been confirmed to perform better in predicting the severity status of ICU patients than the scoring systems. Henry et al [33] proposed a supervised learning model to predict the risk of patients getting septic shock, and machine learning was found to have higher sensitivity and specificity than the scoring systems. An ensemble machine learning model was investigated by Pirracchio et al

[2], and the results showed better performance for the machine learning model than for the common scoring systems. In recent years, many studies have focused on using the neural network model to predict mortality [34,35]. Most of the experiments demonstrated that the neural network model outperformed the other models. Norrie [36] innovatively proposed a prespecified library of models and established an optimum model. However, the neural network model is difficult to explain in terms of the black box principle [32,37], which is not clear for high recursion [38]. Using the inherently interpretable models is vital and important in health care because decisions in health care involve high stakes [39]. Awad et al [20] demonstrated that the decision tree model is interpretable and better than the neural network model in predicting mortality. Similarly, Blanco-Justicia et al [40] used a depth-limited decision tree model to avoid the black box problem. In reality, the gradient boosting methods usually perform better than the deep learning method on structured data, especially on a small data set.

The limitation of this study is that the data were obtained from one hospital only. The structure and quality of data may vary in different hospitals. In the future, we would try to improve our model based on multicenter data.

In the present study, we constructed the real-time mortality prediction model based on the monitoring, laboratory, and scoring data. Compared with the RMM, RMA, RMS, and RM models, the RMML model demonstrated the best performance. Moreover, we found that the invasive mean blood pressure, heart rate, and invasive systolic blood pressure were the top three features relevant to the mortality prediction. In addition, the RMM model performed better than the RMA and RMS models. Therefore, noninvasively predicting real-time mortality would be meaningful. Not only can the results of our research provide support for decision making by clinical staff, but our method is also better for patients because the real-time mortality prediction is noninvasive.

Acknowledgments

This study was supported by the National Key Research & Development Program of China (project 2018YFC0116905) and the CAMS Innovation Fund for Medical Sciences (CIFMS; project 2016-I2 M-2-004).

Authors' Contributions

HJ, LS, YL, and WZ contributed to the study concept and design. DL, HW, and CZ contributed to the acquisition of the data set, and HJ created the figures. LS, HJ, and NH consulted on the analyses. All authors interpreted the results, contributed to the manuscript, and approved the final draft.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The main monitoring features.
[DOCX File , 17 KB - medinform_v9i3e23888_app1.docx]

References

1. Angus DC, Black N. Improving care of the critically ill: institutional and health-care system approaches. Lancet 2004 Apr 17;363(9417):1314-1320 [FREE Full text] [doi: 10.1016/S0140-6736(04)16007-8] [Medline: 15094279]
2. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med 2015 Jan;3(1):42-52 [FREE Full text] [doi: 10.1016/S2213-2600(14)70239-5] [Medline: 25466337]
3. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 2006 May;34(5):1297-1310. [doi: 10.1097/01.CCM.0000215112.84523.F0] [Medline: 16540951]
4. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 1993;270(24):2957-2963. [doi: 10.1001/jama.270.24.2957] [Medline: 8254858]
5. Vincent J, de Mendonça A, Cantraine F, Moreno R, Takala J, Suter PM, et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. Working group on "sepsis-related problems" of the European Society of Intensive Care Medicine. Crit Care Med 1998 Nov;26(11):1793-1800. [doi: 10.1097/00003246-199811000-00016] [Medline: 9824069]
6. Rothwell P. Physiological scoring systems and audit. Lancet 1993 Jul 31;342(8866):306. [Medline: 8101332]
7. Platon L, Amigues L, Ceballos P, Fegueux N, Daubin D, Besnard N, et al. A reappraisal of ICU and long-term outcome of allogeneic hematopoietic stem cell transplantation patients and reassessment of prognosis factors: results of a 5-year cohort study (2009-2013). Bone Marrow Transplant 2016 Feb;51(2):256-261. [doi: 10.1038/bmt.2015.269] [Medline: 26569092]

8. Johnson AEW, Mark RG. Real-time mortality prediction in the Intensive Care UnitC//AMIA Annual Symposium Proceedings. 2018 Apr 16 Presented at: AMIA Annu Symp Proc. 2017; 2017; Washington p. 994-1003.
9. Shickel B, Loftus TJ, Ozrazgat-Baslanti L, Ebadi T, Bihorac A, Rashidi P. DeepSOFA: A Real-Time Continuous Acuity Score Framework using Deep Learning. ArXiv e-prints 2018 [FREE Full text]
10. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
11. Alber M, Buganza Tepole A, Cannon WR, De S, Dura-Bernal S, Garikipati K, et al. Integrating machine learning and multiscale modeling-perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. NPJ Digit Med 2019;2:115 [FREE Full text] [doi: 10.1038/s41746-019-0193-y] [Medline: 31799423]
12. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data 2019 Jun 17;6(1):96 [FREE Full text] [doi: 10.1038/s41597-019-0103-9] [Medline: 31209213]
13. Kim MJ, Kim YH, Sol IS, Kim SY, Kim JD, Kim HY, et al. Serum anion gap at admission as a predictor of mortality in the pediatric intensive care unit. Sci Rep 2017 May 03;7(1):1456 [FREE Full text] [doi: 10.1038/s41598-017-01681-9] [Medline: 28469150]
14. Kim SY, Kim S, Cho J, Kim YS, Sol IS, Sung Y, et al. A deep learning model for real-time mortality prediction in critically ill children. Crit Care 2019 Aug 14;23(1):279 [FREE Full text] [doi: 10.1186/s13054-019-2561-z] [Medline: 31412949]
15. Brand L, Patel A, Singh I, Brand C. Real Time Mortality Risk Prediction: A Convolutional Neural Network Approach. In: Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018). 2018 Presented at: Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies; January 19-21, 2018; Funchal, Madeira, Portugal p. 463-470. [doi: 10.5220/0006596204630470]
16. Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami. Practical Black-Box Attacks against Machine Learning. In: McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, A. Swami, Practical black-box attacks against machine learning, pp. 506?. New York: Association for Computing Machinery; 2017 Apr Presented at: ASIA CCS '17: ACM Asia Conference on Computer and Communications Security; April, 2017; Abu Dhabi United Arab Emirates p. 506-519. [doi: 10.1145/3052973.3053009]
17. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision treeC//Advances in neural information processing systems. In: Advances in neural information processing systems. Red Hook, NY: Curran Associates Inc; 2017 Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing System; December 2017; Long Beach, CA, USA p. 3149-3157.
18. Bisong E. Introduction to Scikit-learn. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Berkeley, CA: Apress; 2019:215-229.
19. Wang YC, McPherson K, Marsh T, Gortmaker SL, Brown M. Health and economic burden of the projected obesity trends in the USA and the UK. Lancet 2011 Aug 27;378(9793):815-825. [doi: 10.1016/S0140-6736(11)60814-3] [Medline: 21872750]
20. Awad A, Bader-El-Den M, McNicholas J. Patient length of stay and mortality prediction: A survey. Health Serv Manage Res 2017 May;30(2):105-120. [doi: 10.1177/0951484817696212] [Medline: 28539083]
21. Baicker K, Goldman D. Patient cost-sharing and healthcare spending growth. J Econ Perspect 2011;25(2):47-68. [doi: 10.1257/jep.25.2.47] [Medline: 21595325]
22. Vincent J, Marshall JC, Namendys-Silva SA, François B, Martin-Loeches I, Lipman J, ICON investigators. Assessment of the worldwide burden of critical illness: the intensive care over nations (ICON) audit. Lancet Respir Med 2014 May;2(5):380-386. [doi: 10.1016/S2213-2600(14)70061-X] [Medline: 24740011]
23. Machado FR, Cavalcanti AB, Bozza FA, Ferreira EM, Angotti Carrara FS, Sousa JL, SPREAD Investigators, Latin American Sepsis Institute Network. The epidemiology of sepsis in Brazilian intensive care units (the Sepsis PREvalence Assessment Database, SPREAD): an observational study. Lancet Infect Dis 2017 Nov;17(11):1180-1189. [doi: 10.1016/S1473-3099(17)30322-5] [Medline: 28826588]
24. Saleh A, Ahmed M, Sultan I, Abdel-lateif A. Comparison of the mortality prediction of different ICU scoring systems (APACHE II and III, SAPS II, and SOFA) in a single-center ICU subpopulation with acute respiratory distress syndrome. Egypt J Chest Dis Tuberc 2015 Oct;64(4):843-848. [doi: 10.1016/j.ejcdt.2015.05.012]
25. Nagrebetsky A, Bittner EA. Missing Data and ICU Mortality Prediction: Gone But Not to Be Forgotten. Crit Care Med 2017 Dec;45(12):2108-2109. [doi: 10.1097/CCM.0000000000002780] [Medline: 29148991]
26. Yap X, Ng C, Hsu K, Chien C, Goh ZNL, Li C, et al. Predicting need for intensive care unit admission in adult emphysematous pyelonephritis patients at emergency departments: comparison of five scoring systems. Sci Rep 2019 Nov 12;9(1):16618 [FREE Full text] [doi: 10.1038/s41598-019-52989-7] [Medline: 31719593]
27. Tan TL, Tang YJ, Ching LJ, Abdullah N, Neoh H. Scientific reports 2018 Nov 12;8(1):16698 [FREE Full text] [doi: 10.1038/s41598-018-35144-6] [Medline: 30420768]
28. Lee H, Yoon S, Oh S, Shin J, Kim J, Jung C, et al. Comparison of APACHE IV with APACHE II, SAPS 3, MELD, MELD-Na, and CTP scores in predicting mortality after liver transplantation. Sci Rep 2017 Sep 07;7(1):10884 [FREE Full text] [doi: 10.1038/s41598-017-07797-2] [Medline: 28883401]

29. Cui Y, Wang T, Bao J, Tian Z, Lin Z, Chen D. Comparison of Charlson's weighted index of comorbidities with the chronic health score for the prediction of mortality in septic patients. Chin Med J (Engl) 2014;127(14):2623-2627. [Medline: 25043078]
30. Nielsen AB, Thorsen-Meyer H, Belling K, Nielsen AP, Thomas CE, Chmura PJ, et al. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. Lancet Digit Health 2019 Jun;1(2):e78-e89 [FREE Full text] [doi: 10.1016/S2589-7500(19)30024-X] [Medline: 33323232]
31. Fei A, Lin Q, Liu J, Wang F, Wang H, Pan S. The relationship between coagulation abnormality and mortality in ICU patients: a prospective, observational study. Sci Rep 2015 Mar 23;5:9391 [FREE Full text] [doi: 10.1038/srep09391] [Medline: 25797521]
32. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep Learning for Health Informatics. IEEE J Biomed Health Inform 2016 Dec 29;21(1):4-21. [doi: 10.1109/JBHI.2016.2636665] [Medline: 28055930]
33. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015 Aug 05;7(299):299ra122. [doi: 10.1126/scitranslmed.aab3719] [Medline: 26246167]
34. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. Sci Rep 2019 Feb 12;9(1):1879 [FREE Full text] [doi: 10.1038/s41598-019-38491-0] [Medline: 30755689]
35. Davoodi R, Moradi MH. Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier. J Biomed Inform 2018 Mar;79:48-59 [FREE Full text] [doi: 10.1016/j.jbi.2018.02.008] [Medline: 29471111]
36. Norrie J. Mortality prediction in ICU: a methodological advance. Lancet Respir Med 2015 Jan;3(1):5-6. [doi: 10.1016/S2213-2600(14)70268-1] [Medline: 25466334]
37. Zhengping Che, Sanjay Purushotham, Robinder Khemani, Yan Liu. Interpretable deep models for icu outcome prediction. In: AMIA Annu Symp Proc. 2017 Feb 10 Presented at: Purushotham, R. Khemani, Y. Liu, Interpretable deep models for icu outcome prediction, Amia Annu Symp Proc () 371?380; 2016; Chicago p. 371-380.
38. The Lancet Respiratory Medicine. Opening the black box of machine learning. Lancet Respir Med 2018 Nov;6(11):801. [doi: 10.1016/S2213-2600(18)30425-9] [Medline: 30343029]
39. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019 May 13;1(5):206-215. [doi: 10.1038/s42256-019-0048-x]
40. Blanco-Justicia A, Domingo-Ferrer J, Martínez S, Sánchez D. Machine learning explainability via microaggregation and shallow decision trees. Knowl-Based Syst 2020 Apr;194:105532. [doi: 10.1016/j.knosys.2020.105532]

Abbreviations

APACHE: Acute Physiology and Chronic Health Evaluation
AUC: area under the curve
GDP: gross domestic product
ICU: intensive care unit
SAPS II: Simplified Acute Physiology Score II
SOFA: Sequential Organ Failure Assessment

Edited by C Lovis; submitted 26.08.20; peer-reviewed by B Qian, P Karsmakers, X Li; comments to author 24.10.20; revised version received 17.12.20; accepted 25.01.21; published 25.03.21.

Please cite as:
Jiang H, Su L, Wang H, Li D, Zhao C, Hong N, Long Y, Zhu W
Noninvasive Real-Time Mortality Prediction in Intensive Care Units Based on Gradient Boosting Method: Model Development and Validation Study
JMIR Med Inform 2021;9(3):e23888
URL: https://medinform.jmir.org/2021/3/e23888
doi:10.2196/23888
PMID:33764311

Original Paper

Accuracy of an Artificial Intelligence System for Cancer Clinical Trial Eligibility Screening: Retrospective Pilot Study

Tufia Haddad[1], MD; Jane M Helgeson[1], BSc; Katharine E Pomerleau[1], MBA; Anita M Preininger[2], PhD; M Christopher Roebuck[3], PhD, MBA; Irene Dankwa-Mullan[2], MD, MPH; Gretchen Purcell Jackson[2,4], MD, PhD, FACS; Matthew P Goetz[1], MD

[1]Mayo Clinic, Rochester, MN, United States

[2]IBM Watson Health, Cambridge, ME, United States

[3]RxEconomics LLC, Hunt Valley, MD, United States

[4]Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Tufia Haddad, MD
Mayo Clinic
200 First Street SW
Rochester, MN, 55905
United States
Phone: 1 507 284 3731
Fax: 1 507 284 1803
Email: haddad.tufia@mayo.edu

Abstract

Background: Screening patients for eligibility for clinical trials is labor intensive. It requires abstraction of data elements from multiple components of the longitudinal health record and matching them to inclusion and exclusion criteria for each trial. Artificial intelligence (AI) systems have been developed to improve the efficiency and accuracy of this process.

Objective: This study aims to evaluate the ability of an AI clinical decision support system (CDSS) to identify eligible patients for a set of clinical trials.

Methods: This study included the deidentified data from a cohort of patients with breast cancer seen at the medical oncology clinic of an academic medical center between May and July 2017 and assessed patient eligibility for 4 breast cancer clinical trials. CDSS eligibility screening performance was validated against manual screening. Accuracy, sensitivity, specificity, positive predictive value, and negative predictive value for eligibility determinations were calculated. Disagreements between manual screeners and the CDSS were examined to identify sources of discrepancies. Interrater reliability between manual reviewers was analyzed using Cohen (pairwise) and Fleiss (three-way) κ, and the significance of differences was determined by Wilcoxon signed-rank test.

Results: In total, 318 patients with breast cancer were included. Interrater reliability for manual screening ranged from 0.60-0.77, indicating substantial agreement. The overall accuracy of breast cancer trial eligibility determinations by the CDSS was 87.6%. CDSS sensitivity was 81.1% and specificity was 89%.

Conclusions: The AI CDSS in this study demonstrated accuracy, sensitivity, and specificity of greater than 80% in determining the eligibility of patients for breast cancer clinical trials. CDSSs can accurately exclude ineligible patients for clinical trials and offer the potential to increase screening efficiency and accuracy. Additional research is needed to explore whether increased efficiency in screening and trial matching translates to improvements in trial enrollment, accruals, feasibility assessments, and cost.

(JMIR Med Inform 2021;9(3):e27767) doi:10.2196/27767

KEYWORDS

clinical trial matching; clinical decision support system; machine learning; artificial intelligence; screening; clinical trials; eligibility; breast cancer

Introduction

Patients with cancer treated in multispecialty clinical settings with access to clinical trials may experience better survival and quality of life [1-5]. Cancer research involving clinical trials is essential to bring new drugs, combination therapies, devices, and procedures into clinical practice, with the ultimate goal of decreasing cancer morbidity and mortality. Implementing a program to systematically screen patients for clinical trials can improve accruals but requires dedicated and skilled staff to complete a demanding and often tedious task [6]. Identifying patients that fit complex protocol eligibility criteria is key to successful trial recruitment and enrollment [7]; however, most clinics are not optimally staffed for the time-intensive nature of manual patient screening. Emerging health information technologies leveraging artificial intelligence (AI) techniques, such as natural language processing (NLP) and machine learning (ML), can play important roles in the clinical trial-matching and enrollment processes. Matching of eligible patients to relevant trials requires retrieval of patient information buried in the electronic health record (EHR) and extensive knowledge of complex exclusion and inclusion criteria for each trial protocol. Therefore, an automated technology that enhances efficiencies of eligibility screening for diverse cohorts of patients and large portfolios of clinical trials holds great promise for advancing cancer translational and research activities.

In oncology practices, clinical decision support systems (CDSSs) designed for cancer clinical trial matching have the potential to assist research program managers, trial coordinators, principal investigators, and cancer care providers with the eligibility screening process [8]. This assistance is needed, as the time and effort required to identify trials for individual patients increases the burden on already overstretched research and clinical care teams, but also poses a potential barrier to trials being offered to eligible patients with cancer. Protocols often include numerous complex inclusion and exclusion criteria that must be evaluated for each patient, and depending on the number of active clinical trials, research teams may need to screen and evaluate patients against a long list of possible trials. Automation of the screening process with trial-matching tools can reduce screening time and research team fatigue, thereby increasing coordinator availability to address other patient and provider barriers to clinical trial enrollment [9-12].

Sponsors of clinical trials typically seek to open new clinical trials at sites based on the expertise of the principal investigator, his or her track record of trial enrollments, as well as results of site feasibility questionnaires that may or may not accurately reflect the potential for enrollment of patients who meet eligibility criteria for the trial. An automated trial-matching system can help identify factors associated with accrual rates, including common reasons for patient exclusion, and inform discussions regarding eligibility criteria.

Watson for Clinical Trial Matching is a CDSS designed to interpret clinical trial protocols written in natural language and patient information from EHRs and provide just-in-time information to determine patient eligibility for clinical trials. The CDSS integration with an EHR facilitates intake of structured data (eg, laboratory values, demographics) and processing of unstructured information (eg, pathology reports, clinical notes) with NLP. This enables an assessment of patient eligibility across studies that have been ingested into its trial corpus (ie, ClinicalTrials.gov trials, sponsor- or investigator-initiated trials).

There are two types of approaches for the CDSS to screen and match patients to clinical trials. The first approach involves identifying the cohort of potentially eligible patients for a trial, referred to as trial-centered matching; specifically, for an individual trial, which patients among a cohort match to the trial inclusion and exclusion criteria. Trial-centered analysis by the CDSS can provide feedback to a study team on trial feasibility, including recognition of criteria that commonly lead to patient exclusion. This information can help estimate the projected site enrollment or generate protocol modifications to eligibility to optimize patient inclusion. The second approach involves identifying appropriate clinical trials for a patient, or patient-centered matching; specifically, for an individual patient, which trials among a portfolio of options match to the patient and his/her tumor characteristics. Patient-centered analysis by the CDSS can provide a ranked list of trials to clinicians or research teams at point-of-care or be used for just-in-time screening when patients contact cancer centers with interest for clinical trial opportunities.

The CDSS used in this study was initially designed to support patient-centered matching. In the current study, however, we report the evaluation of a trial-centered matching approach by the CDSS to identify eligible patients for each of 4 different clinical trials from a pool of patients with breast cancer treated at Mayo Clinic (Rochester, MN), a National Cancer Institute–designated comprehensive cancer center. The purpose of this pilot study was to determine the accuracy, efficiency, feasibility, clinical validity, and performance of the CDSS using a trial-centered matching approach.

Methods

Institutional Review Board Review

This study was conducted under an exemption from the Western Institutional Review Board (WIRB) as a technology pilot for epidemiologic research (Protocol 20152322). The WIRB determined that this research met requirements for waiver of consent. This pilot study was also approved by the Mayo Clinic Institutional Review Board. This pilot was not intended to direct patient care or recruitment of patients into trials. All evaluations on patient data were performed in a retrospective manner. For actual trial participation, patients were evaluated via the standard manual screening process at Mayo Clinic.

CDSS Description and Training

This study evaluated a trial-centered approach by the Watson for Clinical Trial Matching CDSS system in a research setting. The core NLP and ML technologies designed for patient-centered matching within the system have been described elsewhere and will not be detailed in this manuscript focused on performance evaluation [13-15]. The CDSS uses NLP to determine cancer-specific attribute values from structured and

unstructured deidentified data sources. In developing the CDSS, specific attributes for cancers, such as cancer stage, cancer subtype, genetic markers, prior cancer therapy, surgical status, and pathology, as well as attributes needed for trial consideration, such as therapy-related characteristics, were defined by subject matter experts (SMEs), including clinical specialists and PhD-level nurses. The NLP was trained through iterative teaching cycles with clinical information obtained from patient cases [16,17]. During these training cycles, SMEs reviewed partially trained outputs from the CDSS to identify initial attributes and values, as well as correct system outputs by providing supporting information and evidence as needed. Corrected outputs were given to developers for additional system training. This process was used to iteratively create a ground truth and allowed for greater scalability and agility during the system development process. Medical logic algorithms allowed the CDSS to prioritize clinical information when determining attribute values when two values for the same attribute were available (eg, mastectomy was prioritized over lumpectomy) [18,19]. The CDSS's learning was continual as patient cases were processed and as medical knowledge advanced.

For the NLP process of trial ingestion, the CDSS used protocol inclusion and exclusion criteria from analysis of several thousand trials available from ClinicalTrials.gov. In this study, the Novartis protocol library was made available to Mayo Clinic, and the full inclusion and exclusion criteria from 4 breast cancer trials (ie, NCT02069093, NCT01633060, NT02437318, and NCT01923168) were ingested into the CDSS. NLP training at the protocol level was conducted by processing PDFs of the final readable trial protocols, including amendments approved by the WIRB. The CDSS applied NLP against the full protocol criteria and an evaluation file was provided to Novartis for protocol disambiguation. The disambiguation file indicated which criteria required clarification; clarification was provided with input from Novartis. Therefore, trial ingestion errors were corrected and not evaluated as part of this study.

Study Population and Analytic Methods

Based on binomial distribution to detect accuracy of at least 80% with a power of 90% and a probability of error at the α=.05 level, a minimum sample of 172 individuals was required. We identified patient records suitable for inclusion in this retrospective pilot study from a population of patients with breast cancer treated in the medical oncology clinic at Mayo Clinic in Rochester, Minnesota, between May and June of 2017 with at least one unstructured health record note for processing by the CDSS.

The CDSS processes structured and unstructured patient data contained in the EHR (including medical oncology progress notes, pathology records, surgical reports, and laboratory values) to derive patient- and tumor-specific attributes. In this study, two groups of patient records were evaluated. Group 1 was comprised of a subset of patients that had been previously processed by the CDSS with a patient-centered approach, during which time any missing or conflicting attributes were resolved through human intervention. Group 2 patient records were processed solely by CDSS without additional human verification; any missing or conflicting attributes were marked as "unknown" by the system.

Gold Standard

To establish a gold standard for eligibility determinations, attribute filters for the 4 preselected clinical trials were established according to tumor stage (metastatic or nonmetastatic breast cancer), patient setting (neoadjuvant/preoperative or adjuvant/postoperative setting), tumor HER2 status (positive or negative), and tumor hormone receptor status (positive or negative). One or more qualified staff (nurse abstractors) then manually reviewed patient EHRs, screened trial eligibility based on the attribute filters, and made determinations to "include" or "exclude" patients if their data attributes matched those of each individual trial.

To measure the reproducibility of manual review, a random subset of 38 breast cancer cases from Group 2 (those whose attributes were determined solely by the CDSS) were rereviewed by two additional reviewers, and interrater reliability between these additional reviewers and the gold standard manual review was calculated. The additional reviews were performed by trained breast oncology clinical research coordinators from the Mayo Clinic. For all 38 cases, the two additional reviewers repeated the same sequence of steps performed by the initial reviewer, using the same set of relevant data from filter parameters described above. Interrater reliability was assessed using Cohen κ for each of the additional reviewers compared to manual review (ie, two pairs). A single Fleiss κ coefficient was also calculated for three raters (ie, the two additional reviewers and the gold standard manual). Significance of differences was analyzed using the Wilcoxon signed-rank test.

CDSS Performance Evaluation

The CDSS abstracted the same attributes from the EHR and determined patient eligibility by matching them with those attribute filters assigned to each trial, including tumor stage, patient setting, and tumor HER2 and hormone receptor status. The CDSS eligibility determinations were next compared to manual classifications using confusion matrices constructed by cross tabulation of inclusion and exclusion determinations from the CDSS versus manual reviewers.

The CDSS clinical performance was assessed for its predictive accuracy, sensitivity, specificity, positive predictive value, and negative predictive value using manual review as the gold standard. Discrepancies in clinical trial matches (inclusion/exclusion determinations) between the CDSS and the manual review were identified and evaluated by independent SMEs with breast cancer clinical expertise and knowledge of the CDSS's trial-matching processes. All discrepant determinations were resolved by SMEs, categorized by type, and recorded. Types of discrepancies include the following: manual screening errors (human error), incorrectly derived by the CDSS (machine error), unsupported CDSS functionality (CDSS untrained for all breast cancer clinical scenarios, such as multiple primary tumors), filter parameters (differences in patient setting or tumor attribute without error, such as when the CDSS system's medical logic did not include all variations of reasoning used in practice for estrogen receptor/progesterone

receptor interpretation), limited records provided (insufficient data available), and project design errors (due to patient tumor attributes or setting changing over the time frame for which the patient EHR was used for the study).

Results

Study Population

The study sample included 327 patients with breast cancer. From the original sample, 4 patients were removed due to an unsupported disease type (noninvasive breast cancer) and 5 patients were removed as duplicates, resulting in a total of 318 patients with breast cancer.

Reproducibility of Manual Screening

Manual review of breast cancer cases was employed to create the gold standard for this evaluation. Interrater reliability for manual assignment was substantial as Cohen κ between the gold standard and each of the two additional reviewers was 0.60 and 0.77, respectively. Fleiss κ coefficient across all three reviewers was 0.64. No statistically significant differences in assignment were detected (P=.16).

CDSS Accuracy in Eligibility Determination

Group 1

Group 1, with attributes verified by humans as described in the Methods section, included 117 breast cancer cases. The CDSS accuracy of trial eligibility determinations for Group 1 (included/excluded) was 90.6% overall. Sensitivity (true positive rate) was 82.1%, and specificity (true negative rate) was 93.3%. The mean accuracy for this group was determined for the following filters: metastatic stage (95.4%), neoadjuvant setting (100%), HER2 status (88.9%), and hormone receptor status (93.5%; all results shown in Table 1). Discrepancies (Table 2) included 5 false positive values (originating from 4 filter parameter errors and 1 manual screening error) and 6 false negative values (3 from filter parameters, 1 due to manual screening error, 1 incorrectly derived by the CDSS, and 1 error from an unsupported CDSS functionality).

Table 1. Clinical decision support system accuracy by cohort.

Cohort	Overall eligibility (%)	Metastatic breast cancer (%)	Neoadjuvant setting (%)	HER2 status (%)	Hormone receptor status (%)
Group 1	90.6	95.4	100	88.9	93.5
Group 2	87.6	90.4	87.2	93	88.6

Table 2. False positive, false negative, and discrepancy type by cohort.

Cohort and discrepancy type(s)	Counts
Group 1, false positive (n=5)	
Project design errors	4
Manual screening errors	1
Group 1, false negative (n=6)	
Filter parameters	3
Manual screening errors	1
Incorrectly derived by the CDSS[a]	1
Unsupported CDSS functionality	1
Group 2, false positive (n=7)	
Manual screening errors	5
Incorrectly derived by the CDSS	2
Group 2, false negative (n=18)	
Incorrectly derived by the CDSS	9
Limited records provided	3
Unsupported CDSS functionality	3
Manual screening errors	2
Filter parameters	1

[a]CDSS: clinical decision support system.

Group 2

In Group 2, a total of 201 cases were processed without human reconciliation of attributes. Inclusion/exclusion determinations were based solely on attribute abstraction and trial matching by the CDSS. The mean system accuracy of trial eligibility determinations of Group 2 (included/excluded) was 87.6%. Sensitivity (true positive rate) was 81.1%, and specificity (true

negative rate) was 89%. The mean accuracy was determined for the following filters: metastatic stage (90.4%), neoadjuvant setting (87.2%), HER2 status (93%), and hormone receptor status (88.6%; all results shown in Table 1). Since the system processed information without human verification of attributes, unresolved attribute conflicts that led to discrepancies in trial inclusion/exclusion determinations were classified as filter parameter errors. Discrepancies (Table 2) included 7 false positive values (originating from 5 manual screening errors and 2 incorrectly derived by the CDSS) and 18 false negative values (9 values incorrectly derived by the CDSS, 3 with limited records provided, 3 related to an unsupported system functionality, 2 manual screening errors, and 1 value related to filter parameters).

Discussion

Principal Results

Screening for clinical trials is a complex and laborious process. This study demonstrated that an AI CDSS can automate eligibility screening accurately and identify potentially eligible patients with breast cancer with a wide variety of clinical characteristics for clinical trials. The clinical trial eligibility screening tool had a mean accuracy of 90.6% after attribute validation by research staff, which is part of the normal clinical workflow when this CDSS is used as a patient-centered solution in the practice. CDSSs such as the one used in this study can aid humans in the process of finding clinical trial matches for patients and replace the slower manual process of screening by search of EHR and eligibility criteria for each of many trial protocols with automation. The system facilitates and engages clinical staff in the completion of tasks that require human intervention, such as attribute verification and resolution of conflicting attributes. Such conflicts can arise from abstraction of attribute values from different sources within the EHR that lack consistency. This CDSS was not intended to make eligibility determinations without human interaction, but it nonetheless exhibited an accuracy of 87.6% without attribute validation by humans.

Interrater reliability of manual eligibility determination demonstrated substantial but not perfect agreement, illustrating a gap that might be filled by a combination of human and machine. Some of the manual screening errors identified in this study included marking a tumor HER2 status as unknown when the information was available in the EHR, recording incorrect hormone receptor status (estrogen receptor and/or progesterone receptor values), or failing to include attributes that changed with subsequent testing. These errors would most likely have been corrected by the combination of CDSS attribute ingestion and human verification. Errors in attribute abstraction by the system included labeling a patient as metastatic based only on disease in regional lymph nodes or annotating T3 as bone metastases when T3 referred to another clinical test or reference. In a few cases, the actual content of the unstructured notes was insufficient for the system to determine trial eligibility. In addition, the system's medical logic did not include all variations of reasoning used in practice. For example, weakly positive hormone receptor values scored as positive may be interpreted as negative in clinical practice based on the tumor biology or behavior.

Several sources of discrepancies in trial eligibility determinations were artifacts of the study design that are unlikely to be seen in practice. For example, manual reviewers were instructed to rely solely on information explicitly documented in the medical record, without data that might be obtained from clinical inference. For patients in Group 1 with human verification of attributes, the CDSS required the end user to select a correct value when two different values for the same attribute were found within the same source document with the same date. Any attributes that might have conflicting values in Group 2 (lacking human verification) were marked as unknown by the CDSS. Overall accuracy of the system as typically used in the clinic would be expected to be closer to that obtained for Group 1 than Group 2, as verification by humans is recommended for use of the CDSS in practice.

Limitations

There were several limitations of this study. First, the study included a relatively small number of patients with cancer from a single academic medical center and a small number of trials for breast cancer. The findings may not generalize to other settings or cancer types. Patients with multiple primary cancers, including patients with bilateral breast cancer, were not supported by the CDSS at the time of the study. Conflicting values in the CDSS were not used by the system to determine eligibility, although all data were available to manual reviewers in the EHR. The relatively fewer patients in the cohort processed by the system in Group 1 (with human verification) lacked the statistical power of the cohort of patients in Group 2 (without human verification).

Future Work

Research is underway to evaluate system performance related to other cancer types, and this is anticipated to be successful given patient-centered matching across multiple cancers has been demonstrated [14,15]. Additionally, research to evaluate trial-centered matching scalability toward a larger volume of trials is in progress. There are also opportunities to expand patient and tumor attribute training to reflect other common and more nuanced eligibility criteria, such as prior therapies and medical comorbidities. Additional studies will be necessary to evaluate the effectiveness in translating enhanced screening into increased enrollment in clinical trials. Although this work provides evidence of the ability of technologies to expedite the trial-matching process, and automation of this process can facilitate unbiased patient screening for clinical trials, multifactorial barriers to trial recruitment remain, including racial and ethnic disparities. Further innovation and research are needed to identify strategies to address such inequities.

Conclusions

In this study, we demonstrated the ability of an AI CDSS to screen a cohort of patients with breast cancer and determine eligibility for 4 clinical trials with very good accuracy. AI-based CDSSs have the potential to optimize the efficiency and accuracy of the trial-matching process, with the overall goal of increasing clinical trial enrollment and completion of trial

objectives. This may ultimately expedite the approval of lifesaving drugs to improve cancer outcomes.

Acknowledgments

The clinical research study reported in this publication was funded by a research grant from Novartis Pharmaceuticals Corporation to Mayo Clinic.

Technical support was provided by Melissa Rammage, Sadie E Coverdill, Brett South, Steven Hancock, and Kyungae Lim of IBM Watson Health. Project guidance and clinical expertise were provided by Ryad A Ali, Paul Williamson, MSc, Quincy Chau, Kenneth W Culver, Robert W Sweetman, and Michael Vinegra of Novartis Pharmaceuticals Corporation (East Hanover, NJ).

Authors' Contributions

MPG and JMH contributed to the study conception and design. Study execution and data collection were performed by TCH, JMH, and MPG. Analysis and interpretation of results were performed by all authors. The first draft of the manuscript was written by TCH and JMH with critical review and support by MPG and GPJ. MPG supervised all aspects of the study. All authors provided review of the manuscript and approved the final version.

Conflicts of Interest

IDM, AMP, and GPJ are employed by IBM Watson. Mayo Clinic has a business collaboration with IBM Watson for Clinical Trial Matching. This activity is not undertaken to allow the company to indicate Mayo Clinic endorsement of a product or service. MPG declares funding acknowledgment to a named professorship (Erivan K Haub Family Professor of Cancer Research Honoring Richard F Emslander, MD) and consulting fees to institution from Eagle Pharmaceuticals, Lilly, Biovica, Novartis, Sermonix, Context Pharm, Pfizer, and Biotheranostics, and grant funding to institution from Pfizer, Sermonix, and Lilly. TCH declares grant funding to Mayo Clinic from Takeda Oncology. There are no conflicts of interest declared by JMH, KEP, and MCR.

References

1. Trogdon JG, Chang Y, Shai S, Mucha PJ, Kuo T, Meyer AM, et al. Care Coordination and Multispecialty Teams in the Care of Colorectal Cancer Patients. Medical Care 2018 May 1;56(5):430-435. [doi: 10.1097/MLR.0000000000000906] [Medline: 29578953]
2. Hussain T, Chang H, Veenstra CM, Pollack CE. Collaboration Between Surgeons and Medical Oncologists and Outcomes for Patients With Stage III Colon Cancer. J Oncol Pract 2015 May;11(3):e388-e397 [FREE Full text] [doi: 10.1200/JOP.2014.003293] [Medline: 25873063]
3. Vinciguerra V, Degnan TJ, Sciortino A, O'Connell M, Moore T, Brody R, et al. A comparative assessment of home versus hospital comprehensive treatment for advanced cancer patients. J Clin Oncol 1986 Oct;4(10):1521-1528. [doi: 10.1200/JCO.1986.4.10.1521] [Medline: 3760919]
4. Hess LM, Pohl G. Perspectives of quality care in cancer treatment: a review of the literature. Am Health Drug Benefits 2013 Jul;6(6):321-329 [FREE Full text] [Medline: 24991367]
5. Clauser SB, Johnson MR, O'Brien DM, Beveridge JM, Fennell ML, Kaluzny AD. Improving clinical research and cancer care delivery in community settings: evaluating the NCI community cancer centers program. Implement Sci 2009 Sep 26;4:63 [FREE Full text] [doi: 10.1186/1748-5908-4-63] [Medline: 19781094]
6. Chen L, Grant J, Cheung WY, Kennecke HF. Screening intervention to identify eligible patients and improve accrual to phase II-IV oncology clinical trials. J Oncol Pract 2013 Jul;9(4):e174-e181 [FREE Full text] [doi: 10.1200/JOP.2012.000763] [Medline: 23942936]
7. Afrin LB, Oates JC, Kamen DL. Improving clinical trial accrual by streamlining the referral process. Int J Med Inform 2015 Jan;84(1):15-23 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.09.001] [Medline: 25256066]
8. Etchells E, Adhikari NKJ, Wu R, Cheung M, Quan S, Mraz R, et al. Real-time automated paging and decision support for critical laboratory abnormalities. BMJ Qual Saf 2011 Nov;20(11):924-930. [doi: 10.1136/bmjqs.2010.051110] [Medline: 21725046]
9. Butte AJ, Weinstein DA, Kohane IS. Enrolling patients into clinical trials faster using RealTime Recuiting. Proc AMIA Symp 2000:111-115 [FREE Full text] [Medline: 11079855]
10. Heinemann S, Thüring S, Wedeken S, Schäfer T, Scheidt-Nave C, Ketterer M, et al. A clinical trial alert tool to recruit large patient samples and assess selection bias in general practice research. BMC Med Res Methodol 2011 Feb 15;11(1):16 [FREE Full text] [doi: 10.1186/1471-2288-11-16] [Medline: 21320358]
11. Treweek S, Pearson E, Smith N, Neville R, Sargeant P, Boswell B, et al. Desktop software to identify patients eligible for recruitment into a clinical trial: using SARMA to recruit to the ROAD feasibility trial. Inform Prim Care 2010;18(1):51-58 [FREE Full text] [doi: 10.14236/jhi.v18i1.753] [Medline: 20429978]

12. Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc 2015 Jan;22(1):166-178 [FREE Full text] [doi: 10.1136/amiajnl-2014-002887] [Medline: 25030032]
13. Helgeson J, Rammage M, Urman A, Roebuck MC, Coverdill S, Pomerleau K, et al. Clinical performance pilot using cognitive computing for clinical trial matching at Mayo Clinic. JCO 2018 May 20;36(15_suppl):e18598-e18598. [doi: 10.1200/jco.2018.36.15_suppl.e18598]
14. Beck JT, Rammage M, Jackson GP, Preininger AM, Dankwa-Mullan I, Roebuck MC, et al. Artificial Intelligence Tool for Optimizing Eligibility Screening for Clinical Trials in a Large Community Cancer Center. JCO Clin Cancer Inform 2020 Jan;4:50-59 [FREE Full text] [doi: 10.1200/CCI.19.00079] [Medline: 31977254]
15. Alexander M, Solomon B, Ball DL, Sheerin M, Dankwa-Mullan I, Preininger AM, et al. Evaluation of an artificial intelligence clinical trial matching system in Australian lung cancer patients. JAMIA Open 2020 Jul;3(2):209-215 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa002] [Medline: 32734161]
16. Patel NM, Michelini VV, Snell JM, Balu S, Hoyle AP, Parker JS, et al. Enhancing Next-Generation Sequencing-Guided Cancer Care Through Cognitive Computing. Oncologist 2018 Feb;23(2):179-185 [FREE Full text] [doi: 10.1634/theoncologist.2017-0170] [Medline: 29158372]
17. Liang JJ, Tsou C, Devarakonda MV. Ground Truth Creation for Complex Clinical NLP Tasks - an Iterative Vetting Approach and Lessons Learned. AMIA Jt Summits Transl Sci Proc 2017;2017:203-212 [FREE Full text] [Medline: 28815130]
18. Bang Y, Van Cutsem E, Feyereislova A, Chung HC, Shen L, Sawaki A, et al. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. The Lancet 2010 Aug 28;376(9742):687-697. [doi: 10.1016/S0140-6736(10)61121-X] [Medline: 20728210]
19. Lal P, Salazar PA, Hudis CA, Ladanyi M, Chen B. HER-2 Testing in Breast Cancer Using Immunohistochemical Analysis and Fluorescence In Situ Hybridization : A Single-Institution Experience of 2,279 Cases and Comparison of Dual-Color and Single-Color Scoring. American Journal of Clinical Pathology 2004 May 1;121(5):631-636. [doi: 10.1309/VE78-62V2-646B-R6EX] [Medline: 15151202]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
EHR: electronic health record
ML: machine learning
NLP: natural language processing
SME: subject matter expert

Edited by C Lovis; submitted 08.02.21; peer-reviewed by R Pozzar, K Turner; comments to author 20.02.21; revised version received 05.03.21; accepted 07.03.21; published 26.03.21.

Please cite as:
Haddad T, Helgeson JM, Pomerleau KE, Preininger AM, Roebuck MC, Dankwa-Mullan I, Jackson GP, Goetz MP
Accuracy of an Artificial Intelligence System for Cancer Clinical Trial Eligibility Screening: Retrospective Pilot Study
JMIR Med Inform 2021;9(3):e27767
URL: https://medinform.jmir.org/2021/3/e27767
doi:10.2196/27767
PMID:33769304

Review

Physicians' Use of the Computerized Physician Order Entry System for Medication Prescribing: Systematic Review

Asra Mogharbel[1], BA, MSc; Dawn Dowding[2], PhD; John Ainsworth[1], MSc, PhD

[1]Division of Informatics Imaging and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, Centre for Health Informatics, The University of Manchester, Manchester, United Kingdom

[2]Division of Nursing, Midwifery and Social Work, School of Health Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, United Kingdom

Corresponding Author:
Asra Mogharbel, BA, MSc
Division of Informatics Imaging and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre
Centre for Health Informatics
The University of Manchester
Vaughan House
Portsmouth St
Manchester, M13 9GB
United Kingdom
Phone: 44 161 275 1274
Email: asra.mogharbel@postgrad.manchester.ac.uk

Abstract

Background: Computerized physician order entry (CPOE) systems in health care settings have many benefits for prescribing medication, such as improved quality of patient care and patient safety. However, to achieve their full potential, the factors influencing the usage of CPOE systems by physicians must be identified and understood.

Objective: The aim of this study is to identify the factors influencing the usage of CPOE systems by physicians for medication prescribing in their clinical practice.

Methods: We conducted a systematic search of the literature on this topic using four databases: PubMed, CINAHL, Ovid MEDLINE, and Embase. Searches were performed from September 2019 to December 2019. The retrieved papers were screened by examining the titles and abstracts of relevant studies; two reviewers screened the full text of potentially relevant papers for inclusion in the review. Qualitative, quantitative, and mixed methods studies with the aim of conducting assessments or investigations of factors influencing the use of CPOE for medication prescribing among physicians were included. The identified factors were grouped based on constructs from two models: the unified theory of acceptance and use of technology model and the Delone and McLean Information System Success Model. We used the Mixed Method Appraisal Tool to assess the quality of the included studies and narrative synthesis to report the results.

Results: A total of 11 articles were included in the review, and 37 factors related to the usage of CPOE systems were identified as the factors influencing how physicians used CPOE for medication prescribing. These factors represented three main themes: individual, technological, and organizational.

Conclusions: This study identified the common factors that influenced the usage of CPOE systems by physicians for medication prescribing regardless of the type of setting or the duration of the use of a system by participants. Our findings can be used to inform implementation and support the usage of the CPOE system by physicians.

(JMIR Med Inform 2021;9(3):e22923) doi:10.2196/22923

KEYWORDS

computerized physician order entry; CPOE; e-prescribing; system use; actual usage; systematic review

Introduction

Background

Computerized physician order entry (CPOE) systems for medication prescribing allow health care professionals to enter accurate and complete medication orders electronically [1]. The CPOE system has clinical decision support (CDS) features that help reduce medication errors and increase safety, such as an alert system, to warn a physician of drug allergies and drug-drug interactions and a feature offering advice regarding medication dosages and frequencies [1]. CPOE for prescribing medication has been reported to be helpful to clinicians by providing them with easy access to patient data, a faster prescribing process [2], and guidelines to enhance compliance with best practices; it also reduces medical costs and improves organizational efficiency [3].

In addition to being beneficial for clinicians, CPOE for medication prescribing also has drawbacks that affect its usage by clinicians. Issues such as excessive alerting can lead physicians to ignore these safety warnings, which might be harmful for patients [4]. In addition, owing to the expense associated with continuous training required for such a system, physicians may lack adequate skills to use CPOE, which leads to underutilization [5].

The adoption and use of CPOE usually starts at the organizational level, where health organizations decide to implement such a system. Studies have shown that the adoption of CPOE for medication prescribing by health care organizations is associated with the high cost of installing a CPOE system. This may hinder many health care organizations from having a system within their practice. However, the benefits offered by the system in the long run can compensate for these costs [6].

For example, in 2013, a CPOE was implemented in 2 groups of 4 community hospitals in the United States at a cost of US $7,130,894 and US $19,293,379, respectively. After adopting the CPOE, the avoided financial cost of adverse drug events alone saves the hospital about US $7,937,651 and US $16,557,056 [7]. The organization makes the decision to implement the CPOE system; however, to achieve benefits and reach its full potential, CPOE depends on effective use by individual clinicians. There is a need to understand the factors influencing the usage of this system by physicians after it has been implemented. The aim of this review is to identify the factors that influence actual use of CPOE by physicians for medication prescribing.

The rationale for this systematic review was based on the results of previous studies, which suggested that the use of CPOE at the international level appears to be low [8-10]. The adoption of CPOE as a computerized ordering system for all types of medical orders (not only medication prescriptions) has international relevance [8,9]; however, evidence from studies conducted in several countries has shown a low rate of acceptance and adoption of these systems by health care providers [8,9]. For example, in some developing countries, despite the availability of several types of computerized health systems, such as electronic medical records, CDS systems, CPOE, and telemedicine, these systems are not properly used [9]. Although little has been reported in recent years about the proportion of CPOE users, in 2009 [8], the proportion of hospitals that implemented and adopted CPOE as an ordering system, including medication prescribing, in 7 western countries was reported. The study indicated that 15% of the hospitals in the United States, 2% in the United Kingdom, and 20% in the Netherlands had CPOE, with very few in Germany, France, and Australia. This shows a significantly low adoption rate [8], which was related to financial, organizational, and technological factors and attitudes of users [8].

In the United Kingdom, for example, vendors of CPOE systems for electronic prescribing have challenges related to implementation because of the factors related to policies [10]. In other countries with different health care systems and policies, the factors affecting the adoption and use of CPOE might vary.

Objectives

The first rationale for conducting this study was to identify the factors influencing the underutilization of CPOE by physicians for medication prescribing and understand their reasons.

Second, we identified only 4 reviews with a main focus on CPOE as a medication-prescribing system [11-14]. The evidence from these reviews focused on the factors affecting health care providers during the implementation and adoption phases, rather than their actual use of CPOE postimplementation. The implementation phase refers to the time between deciding to introduce a new system and the activities involved in this decision by the hospital, up to the point the system is ready to be used [11]. In this study, we aim to identify the factors affecting the actual use of CPOE.

The actual usage of a system follows the implementation process [15]: actual usage is defined as a behavior that can be measured through indicators, such as an individual's frequency or duration of usage [16]. The term system usage consists of 3 fundamental components: the subject using the system (user), the system itself, and the task to be accomplished through the system [17]. Although one of the reviews [14] focused on medication-related CDS after it was fully implemented, it included evidence only from qualitative studies, and there was no indication that the actual usage, as defined here, was the main focus of that review.

Two of the reviews [11,12] identified factors influencing different types of health care providers as users (eg, physicians, nurses, pharmacists), whereas the other 2 reviews [13,14] identified their targeted users. This study focused entirely on physicians as users and the factors that were likely to affect their usage, as professionals from different disciplines might be influenced by different factors in their decisions to use CPOE for prescribing medication. Hence, the second rationale for conducting this study was to fill the gap in the evidence found in prior reviews.

Third, most of the studies included in these reviews were conducted in industrialized western countries (the United States, the United Kingdom, Sweden, the Netherlands, Australia, and Canada); only 1 study was conducted in a developing country. There is a huge gap in the literature on the factors affecting the

usage of CPOE for prescribing medication among developing countries [9]. This study was part of a research project conducted in Saudi Arabia (a developing country) to investigate the factors that influence the actual usage of CPOE by physicians for medication prescribing.

In summary, the aforementioned gap in the literature regarding the factors influencing the actual use of CPOE for medication prescribing by physicians is the reason for carrying out this systematic review. In this study, we used the unified theory of acceptance and use of technology (UTAUT) model [18] and the Delone and McLean Information System Success Model [19] as frameworks to classify the evidence on the actual use of CPOE by physicians for medication prescribing. To the best of our knowledge, there is no published analysis of the factors affecting the actual use of CPOE in particular by physicians for medication prescribing using this theoretical approach.

Methods

Search Strategy

This study was based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) guidelines [20]. The following databases were searched from September 2019 to December 2019: PubMed, Embase, Ovid MEDLINE, and CINAHL. The search was performed without any restrictions on dates; however, it was limited to English language papers. Reference lists in the identified reviews and included studies were checked to retrieve relevant papers. We combined medical subject headings (MeSH terms) related to CPOE retrieved from PubMed and keywords from the relevant research literature (Textbox 1).

Textbox 1. Medical subject headings (MeSH) terms and keywords used in the searches of PubMed, Embase, Ovid MEDLINE, and CINAHL. The final search strategy (A10, B8, and C3) was applied to all 4 databases.

Group A: type of system

1. Medication alert systems
2. Computerized provider order entry
3. Computerized physician order entry
4. CPOE
5. Electronic prescription
6. Prescription decision support system
7. Computerized prescriber order entry
8. Pharmaceutical decision-support systems
9. Pharmacy information system
10. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9

Group B: usage

1. Use
2. Actual usage
3. System use
4. Utilization
5. Acceptance
6. Adoption
7. Usage
8. 1 or 2 or 3 or 4 or 5 or 6 or 7

Group C: factors

1. Factors
2. Determinants
3. 1 or 2

A draft of the search strategies used in three of the databases is presented in Multimedia Appendix 1.

Eligibility Criteria

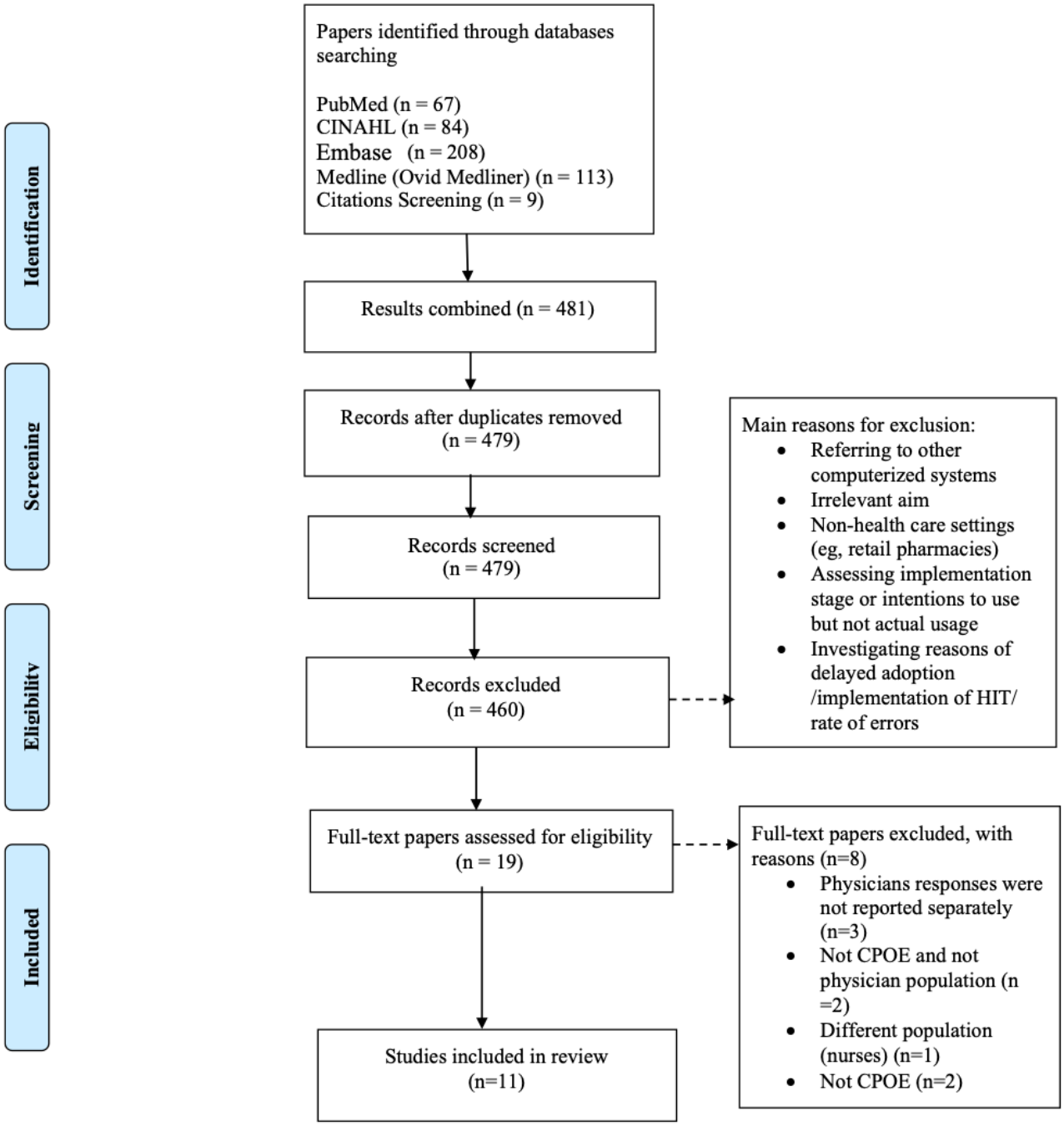
The included studies were peer-reviewed research reports written in English, with the stated aim of exploring, investigating, or assessing factors that influence the use of medication-related CPOE systems as our target intervention. The population of

interest was physicians, with the included studies reporting the results of physicians only or papers in which physicians' responses were reported separately. The included studies also had to be conducted in clinical settings, that is, inpatient and outpatient departments of hospitals, health care centers, primary care centers, and polyclinics. Quantitative, qualitative, and mixed methods designs were considered eligible for inclusion. Studies were excluded if the CPOE system had not been implemented at the time of this study or if the study assessed the influence of factors on intentions to use the CPOE system rather than on its actual use. Papers with a population of nurses, pharmacists, information technology (IT) personnel, managers, or patients and those with interventions that were not strictly CPOE, as defined earlier, were excluded from the review. Studies that were conducted in nonclinical settings (eg, retail pharmacies, community pharmacies, nursing homes) were excluded from this review.

Selection Process

The primary researcher (AM) independently screened the titles and abstracts of all papers retrieved from the search using the inclusion criteria. The full-text articles of all potentially relevant studies were assessed independently by all 3 authors for eligibility. A calibration exercise was conducted to cross-check the results obtained by the authors. All disagreements were resolved through discussion. The details of the exclusion criteria are shown in Figure 1.

Figure 1. Flow diagram of the selection process for the included papers. CPOE: computerized physician order entry; HIT: health information technology.



Data Collection Process and Data Items

The primary researcher performed the data extraction. The data included names of the authors, publication year, country, objective, study design, data collection method, type of intervention, setting, population and sample, factors associated with CPOE use, how actual use was assessed, and the duration of the system's use before the data were collected.

Risk of Bias of the Included Studies (Quality Assessment)

The Mixed Methods Appraisal Tool (MMAT) was used to assess the quality of the included studies [21]. The MMAT is a comprehensive tool designed to evaluate reviews, including quantitative, qualitative, and mixed methods studies [21]. All the 3 authors independently appraised the included studies. The primary researcher (AM) reviewed all of the studies, and each of the other 2 researchers (JA and DD) reviewed half of the studies. Any disagreements were resolved through discussion. MMAT does not recommend assigning a single score based on the assessment [21]. However, in this review, we used a specific metric derived from a previous study [22]. To rate the quality of each of the studies to justify the reasons for the final inclusions and exclusions. Studies were classified as high, medium, or low quality, depending on the number of criteria that were met. A study was considered high quality if all 5 MMAT criteria were met, medium if 3 or 4 criteria were met, and low when a study met 1 or 2 criteria [22].

Data Synthesis

Narrative synthesis was used to summarize the evidence from the included studies. Narrative synthesis is appropriate when a review includes both qualitative and quantitative findings [23].

Results

Study Selection

The electronic database search retrieved 67 records from PubMed, 84 from CINAHL, 208 from Embase, 113 from Ovid MEDLINE, and 9 from the reference lists of the included studies. After duplicates were removed, the titles and abstracts of the remaining 479 studies were assessed for eligibility. Of these, 460 studies were excluded because they were ineligible and 19 articles were selected for in-depth analyses. A total of 11 studies were included in the final review. The study selection process and reasons for exclusion are shown in Figure 1.

Characteristics of the Included Studies

Multimedia Appendix 2 [24-34] summarizes the characteristics of the included studies. The 11 studies included in the review were from different regions of the world: 4 are from the United States [24-27], 3 are from Sweden [28-30], 1 is from the Netherlands [31], 1 from Saudi Arabia [32], 1 from Australia [33], and 1 from Singapore [34]. Of the total number of studies, 4 used qualitative methods (interviews) [24,25,29,33], 6 used quantitative methods (surveys or questionnaires) [26-28,30,32,34], and 1 used a mixed methods approach [31]. Among the 11 included studies, the factors associated with the use of CPOE for medication prescribing were mainly related to technical, organizational, or individual characteristics. All the included studies were conducted in either a hospital or a primary care center. Of the total number of studies, 7 were conducted in a hospital setting [24-27,29,32,33], 2 in a hospital and a primary care center [28,30], 1 in a primary care center [31], and another in a group of polyclinics [34].

Quality of the Included Studies

Multimedia Appendix 3 [24-34] summarizes the results of the quality assessment of the included studies. Of the total number of studies, 3 (all qualitative) were rated as *high* quality because they met all 5 MMAT criteria [24,25,29]. Of the total number of studies, 5 (all quantitative) were rated as *medium* quality, as they met 3 or 4 of the MMAT criteria [26,28,30,32,34] and 3 studies were evaluated as having low quality because they met either 1 or none of the MMAT criteria. Of these, 1 was a quantitative study [27], 1 study used a mixed methods design [31], and 1 was a qualitative study [33]. We chose not to exclude these studies from the final synthesis based on their quality because of the exploratory nature of the review.

Synthesis of the Results

The factors that influenced physicians' usage of CPOE for medication prescribing are presented in Table 1. On the basis of the perceived commonality among the reported factors, we organized them according to the definitions of the constructs from the UTAUT [18] and the Delone and McLean Information System Success Model [19].

Table 1. Factors influencing the frequency of use of the computerized physician order entry system by physicians.

Theme, construct, and factor	Studies, n	Study
Individual factors		
Performance expectancy: perception that using CPOE[a] will improve the physician's job performance [18]		
Perceived usefulness	1	[29]
Relative advantage	1	[30]
Effect on quality of care and/or patient outcomes	3	[25,26,32]
Effects on productivity	2	[25,34]
Effects on safety	1	[24]
Performance outcomes	1	[25]
Effort expectancy: belief that the CPOE is easy to use [18]		
Ease of use	3	[28,29,32]
User-friendliness	1	[31]
Difficult to use	2	[24,25]
Complexity	1	[30]
Social influence: perceived importance of others' (eg, leaders, colleagues) opinions that the physician should or should not use the system [18]		
External normative beliefs	1	[25]
Organizational factors		
Facilitating conditions: available resources, facilities, and infrastructure that facilitate using CPOE [18]		
Training	4	[24,25,33,34]
Availability of technical support	4	[25,27,31,32]
Compatibility	1	[30]
Computer skills	1	[34]
Time constraints	3	[24,25,27]
Availability of hardware	2	[25,27]
Lack of awareness of the availability of certain features	1	[33]
Management support	1	[25]
User involvement	1	[25]
Technological factors		
Information quality: relevance, accuracy, comprehensiveness, understandability, prevalence, timeliness, and usability of the outputs or content [19]		
Usefulness of error messages	1	[32]
Clarity and brevity of the reminders	1	[31]
Confidentiality, privacy, and security of patients' records	1	[25]
System quality: reliability, functionality, flexibility, ease of use, integration, and response time of the system [19]		
Clarity	2	[28,32]
Layout	1	[31]
Technical problems causing delays during prescribing	1	[31]
System's speed	3	[31,32,34]
Software barriers	1	[25]
Reliability	1	[32]

Theme, construct, and factor	Studies, n	Study
Customization to individual departments	2	[25,33]
Functionality of the tools in the system	1	[34]
Locating items on the system	1	[32]
Retrieval of radiology data	1	[32]
Usability	1	[24]
System's efficiency	2	[24,26]
Availability of reference materials	1	[32]
Alert fatigue	2	[24,33]

[a]CPOE: computerized physician order entry.

UTAUT is a theoretical model that can explain about 70% of the variance in a user's behavior in relation to technology acceptance and use [18]. It consists of 4 main constructs: performance expectancy, effort expectancy, social influence, and facilitating conditions [18]. Performance expectancy refers to physicians' perceptions that using CPOE will improve their job performance [18]. Effort expectancy refers to physicians' beliefs that using CPOE is effortless and easy [18]. Social influence pertains to physicians' perceptions of the importance of others' (eg, leaders' and colleagues') opinions about whether physicians should or should not use the system [18]. Facilitating conditions refers to the existence of resources, facilities, and infrastructure that are helpful to physicians when using CPOE [18].

The Delone and McLean Information System Success Model is used to assess and understand the success of any information system and its impact on the individual and the organization [19]. It consists of 6 components: system quality, information quality, use, user satisfaction, individual impact, and organizational impact [19]. However, we assessed only system quality and information quality. Information quality refers to the system's outputs or content in terms of relevance, accuracy, comprehensiveness, understandability, prevalence, timeliness, and usability [19]. System quality refers to the quality of the system, in particular, the system's reliability, functionality, flexibility, ease of use, integration, and response time [19]. We assessed these 2 constructs because the identified factors that are mainly related to the technological aspects of the CPOE system are also related to the quality of the information and the system. The other 4 constructs were addressed in the UTAUT model.

The results of the included studies were synthesized under 3 themes: individual, organizational, and technological factors. Individual factors are related to the constructs of performance expectancy, effort expectancy, and social influence. Organizational factors are related to the construct of facilitating conditions, and technological factors are related to the constructs of information quality and system quality (Table 1).

Individual Factors

Individual factors refer to issues related to physicians' perceptions of the possible effects of using CPOE for medication prescribing [35]. A total of 11 factors related to physicians' perceptions were identified. The most cited factors were the effect on the quality of patient care [25,26,32] and ease of use [28,29,32]. Physicians perceived that using CPOE enhanced patient care. In one study [26], the features of the CPOE system were associated with better quality of patient care by providing easy and direct access to patient records and reminders and alerts for physicians, which led to a reduction in duplicate tests and expediting the ordering process. Ease of use refers to physicians' belief that using the system is easy and effortless [18,28,29]. In another study [32], physicians agreed that their satisfaction with the system was greater because it was easy to use, which led to their usage of the system. Three studies reported limited use of CPOE by physicians because they found it difficult to use and complex in terms of navigating, accessing, and finding information [24,29,30].

Organizational Factors

Organizational factors include resources (eg, materials, humans, circumstances) provided by the organization that facilitate usage of the CPOE system by physicians [12]. In total, 8 studies identified 9 organizational factors that affected the use of CPOE. Training [24,25,33,34], availability of technical support (such as a help desk) [25,27,31,32], and time constraints [24,25,27] were the most cited factors. Training issues reported by physicians included either the need for retraining because of new features [24] or lack of training [33]. The availability of technical support means the physicians need to have IT staff accessible to help them in case of any technical issues while using the CPOE system [25,27,32] or the extent of the physician's awareness that there is a designated help desk to assist them [31].

The timing of the reporting of these factors in the included studies suggests that the factors related to the organization were critical for the usage of the CPOE system by physicians, regardless of whether the physicians recently began using the system or have been using it for a longer time. For example, studies that reported training [24,25,33,34] were conducted at different time points after the implementation of CPOE. One study conducted its assessment after 2 years of CPOE usage [24], while 3 other studies investigated the factors affecting usage after only months of use [25,33,34]. Technical support availability was reported in studies after weeks [25,31,32] and after 1 year of usage [27].

Time constraints were the second most cited factor influencing physicians' CPOE usage [24,25,27]. The complexity of CPOE

[24], its slowness [25], and physicians' unfamiliarity with its features [27] were reasons why it was so time-consuming for physicians to use it.

Technological Factors

Technological factors included the technical and design aspects of CPOE in terms of the system's quality; information quality; and its reliability, functionality, flexibility, ease of use, integration, and response time [19]. Evidence from 8 of the included studies [24-26,28,31-34] indicated that the factors related to CPOE were the most relevant for affecting its use by physicians. A total of 17 factors were reported (Table 1). The system's efficiency was the most cited factor [31,32,34], specifically the quick prescribing process [31], fast data retrieval, response time [32], and the system's speed, in terms of entering patient data [34]. Furthermore, studies that reported the system's speed as an influential factor in its use by physicians were conducted shortly after the implementation phase, that is, halfway through the intervention year (about 6 months later), shortly after implementation (not clear), and 3 months after implementation. This finding suggests that because the system was newly implemented, the processing speed was significant for physicians' performance of tasks.

The findings indicate that ease of use, the effect of using CPOE on quality of care, training, availability of technical support, time, and the system's speed were the factors with the strongest influence on the use of CPOE for medication prescribing among all the studies.

Discussion

Principal Findings and Comparisons With Other Works

CPOE for medication prescribing can serve physicians as a tool to enhance patient quality of care. However, this has not led to a rapid uptake of the system by health organizations and clinicians to use it [6,14]. A key factor in the slow adoption of CPOE by health care organizations is attributed to the costs associated with installing the system and the costs of sustaining it [6]. The first CPOE was installed in the United States in 1971 [36]. Although that was long ago, the adoption rate in health organizations is still rare to moderate, with a percentage of 15.7% [13]. This low adoption rate has been reported in other countries [8,9].

Despite many years of implementation of CPOE for medication prescription, development, and research, the issue of low adoption postimplementation remains. This study focuses on the usage of the user—the physician—after the system has been implemented. We identified factors that were related to the users (physicians), organization, and technological aspects of CPOE that influence the actual use of CPOE by physicians for medication prescribing, rather than intention to use a CPOE system.

The findings of this study are consistent with those of Van Dort et al [14] and Gagnon et al [12]. Nevertheless, these reviews identified other factors that were not found in this study. Resistance to use was reported in both reviews [12,14], as a factor that negatively affected the usage of the system by physicians for medication prescribing. CDS systems embedded in the CPOE system for medication prescribing were examined in Van Dort et al [14]. As CDS systems are known to offer suggestions and recommendations, user resistance was present as the physicians reported concerns that the information presented might not be reliable [14].

In addition to resistance to using CPOE, Gagnon et al [12] described how the system could negatively affect the patient-clinician relationship and identified financial issues as another influential factor, neither of which was detected in this study. This inconsistency might be because of the focus of this study on the actual use of CPOE after the system had been installed and used and resistance is no longer an issue.

This study showed that technological factors related to the system were the most frequently reported factors that influenced how a physician used the CPOE system for medication prescribing. This finding is consistent with the results reported by Gagnon et al [12]. As their findings suggest, technical and design concerns were the most frequently identified factors limiting the system's use [12].

One of the principal findings of this study is that among the 3 main themes, 5 factors were cited most frequently (any factor cited 3 or more times was considered frequently cited), indicating that it was significant in the physicians' decisions about using the CPOE system. Quality of care, ease of use, training, availability of technical support, time constraints, and system speed were key factors in the use of CPOE by physicians. A similar pattern of results has been reported in an extensive body of literature [12,14,37,38]. One unexpected finding was that the effect of alert fatigue, as a factor in the use of CPOE, was identified in only 2 studies [24,33]. Alert fatigue is the receipt of a massive amount of reminders or warnings that cost time and effort and is eventually ignored [39].

This finding contradicts the observation that alert fatigue has previously been found to be associated with the usage of CPOE for medication prescribing. In their review, Gagnon et al [12] showed that alert fatigue was associated with the use of an electronic prescription system in 5 studies. In addition, Van Dort et al [14] showed that too many irrelevant alerts were related to the uptake of medication-related CDS systems in 10 studies.

In these 2 studies [24,33], alert fatigue affected physicians' use. In the first study [24], physicians' perception of the alerts was that after transitioning to a more advanced new system, the alerts were more sensitive than those of the older system. In the second study [33], the ratings of the alerts were higher when the study's setting was an intensive care unit (ICU), compared with their ratings by other departments in the hospital.

All factors identified in this study are similar to those of other reviews related to the implementation [12], adoption [37], or acceptance [38] of CPOE.

However, a factor not discussed in previous CPOE for e-prescription studies and detected in this study was customization of the CPOE system's features for medication prescribing to each department. Customize refers to tailoring

the features of a CPOE system to the preferences and needs of a specific department. For example, ICU physicians reported that some alerts were irrelevant to ICU patients and more suitable for other departments in the hospital [33]. This finding is in line with that reported in the review by Li et al [40], who suggested the importance of customization of the system's features according to different specialties and emphasized its significance for the provider's workflow.

We have used constructs from the UTAUT [18] and Delone and McLean Information System Success Models [19] to organize the identified factors to provide a better understanding of what each factor means to the user and how it may influence physicians' attitudes toward the actual use of the CPOE for medication prescribing. The UTAUT model is a combination of 8 technology acceptance models, which covers almost all the factors identified in the literature [18]. All the factors reported in the included literature in this study were aligned with the constructs of the UTAUT and Delone and McLean Information System Success Models. The examination of factors using these 2 models provides a useful framework for this systematic review.

Two of the constructs (system quality and information quality) from the Delone and McLean Information System Success Model were found to be highly relevant, as the most frequently reported factors were the technological ones [19]. These factors were mainly related to the quality of the system or information. Both models have been extensively used in research related to health care technology assessment [41,42].

Limitations and Strengths

The limitations of this study should be acknowledged. First, we searched only 4 databases. Although these databases are the most relevant for health care publications, there is a possibility that relevant studies could have been missed. Second, the first step of the database search—checking every single title and abstract—was performed by a single author. However, we believe that this does not affect the quality of this paper as the results of the selection and screening were revised in regular meetings with the other reviewers who are experts in the field and no issues were raised by them during the review process. In addition, all the assessment steps for article eligibility were conducted by all 3 authors in parallel. We systematically discussed any disputes between all the reviewers to ensure consistency.

Third, we acknowledge the fact that our search resulted in only 11 articles that could be viewed as a small sample for a system that has been in use for a number of years. However, this study focused on the medication ordering aspect of the CPOE and did not evaluate the CPOE as a whole system. In addition, we also focused on physicians as our target population and studies that indicated that the system is being actually used and not the intention to use (installation phase or implementation phase). The strength of this study lies in the presentation of 4 elements that are absent from previous attempts to synthesize primary research on this topic: (1) it evaluated research that used major study designs (quantitative, qualitative, and mixed methods); (2) it drew on the perspectives of physicians only; and (3) it included research on the period of actual usage of CPOE for e-prescribing in particular (while the physicians were using the system) and not the intention to use. (4) Factors that are unique to the physician's actual usage were explained using a framework that consists of a combination of 2 theoretical approaches. To the best of our knowledge, no previous systematic reviews have explored specific factors influencing physicians' actual usage of CPOE or e-prescriptions according to the presented framework.

Conclusions

This study suggests that an individual's perceptions, technical factors, and organizational factors are all significant influences on the usage of CPOE by physicians for medication prescribing. Although most of the identified factors are similar to those reported in previous reviews related to CPOE, the results of our work have allowed us to identify an additional factor that was not discussed in earlier reviews, namely, the preference of physicians to customize the CPOE system to the needs of the medical department. Finally, as much as there are issues at the organizational level during the implementation process, it is important to focus on the individual physicians after the implementation is completed. The outcomes of this study provide a source of knowledge for health care decision makers, managers, and staff and a clear understanding of the factors influencing the usage of CPOE by physicians for medication prescribing, which can inform future system designs and implementation.

Acknowledgments

This systematic review was conducted as a first phase of doctoral study sponsored by the Ministry of Education, Saudi Arabia.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Results of the search strategies used in the PubMed, EMBASE, and CINAHL databases.
[DOCX File , 16 KB - medinform_v9i3e22923_app1.docx]

Multimedia Appendix 2
Characteristics of the included studies.

[DOCX File , 24 KB - medinform_v9i3e22923_app2.docx]

Multimedia Appendix 3
Quality assessment of the included studies using the Mixed Methods Appraisal Tool (2018).
[DOCX File , 17 KB - medinform_v9i3e22923_app3.docx]

References

1. Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Arch Intern Med 2003 Jun 23;163(12):1409-1416. [doi: 10.1001/archinte.163.12.1409] [Medline: 12824090]
2. Cresswell KM, Bates DW, Williams R, Morrison Z, Slee A, Coleman J, et al. Evaluation of medium-term consequences of implementing commercial computerized physician order entry and clinical decision support prescribing systems in two 'early adopter' hospitals. J Am Med Inform Assoc 2014 Oct;21(e2):e194-e202 [FREE Full text] [doi: 10.1136/amiajnl-2013-002252] [Medline: 24431334]
3. Eslami S, de Keizer NF, Abu-Hanna A. The impact of computerized physician medication order entry in hospitalized patients--a systematic review. Int J Med Inform 2008 Jun;77(6):365-376. [doi: 10.1016/j.ijmedinf.2007.10.001] [Medline: 18023611]
4. Magid S, Forrer C, Shaha S. Duplicate orders: an unintended consequence of computerized provider/physician order entry (CPOE) implementation: analysis and mitigation strategies. Appl Clin Inform 2012;3(4):377-391 [FREE Full text] [doi: 10.4338/ACI-2012-01-RA-0002] [Medline: 23646085]
5. Kuperman GJ, Gibson RF. Computer physician order entry: benefits, costs, and issues. Ann Intern Med 2003 Jul 01;139(1):31-39. [doi: 10.7326/0003-4819-139-1-200307010-00010] [Medline: 12834316]
6. Coustasse A, Shaffer J, Conley D, Coliflower J, Deslich S, Sikula SA. Computer Physician Order Entry (CPOE) : benefits and concerns - a status report. Healthc Admin 2015:2-742. [doi: 10.4018/978-1-4666-6339-8.ch036]
7. Zimlichman E, Keohane C, Franz C, Everett WL, Seger DL, Yoon C, et al. Return on investment for vendor computerized physician order entry in four community hospitals: the importance of decision support. Jt Comm J Qual Patient Saf 2013 Jul;39(7):312-318. [doi: 10.1016/s1553-7250(13)39044-8] [Medline: 23888641]
8. Aarts J, Koppel R. Implementation of computerized physician order entry in seven countries. Health Aff (Millwood) 2009;28(2):404-414. [doi: 10.1377/hlthaff.28.2.404] [Medline: 19275996]
9. Ahlan AR, Ahmad BI. User acceptance of health information technology (hit) in developing countries: a conceptual model. Procedia Technology 2014;16:1287-1296. [doi: 10.1016/j.protcy.2014.10.145]
10. Mozaffar H, Williams R, Cresswell K, Morrison Z, Bates DW, Sheikh A. The evolution of the market for commercial computerized physician order entry and computerized decision support systems for prescribing. J Am Med Inform Assoc 2016 Mar;23(2):349-355. [doi: 10.1093/jamia/ocv095] [Medline: 26338217]
11. Farre A, Heath G, Shaw K, Bem D, Cummins C. How do stakeholders experience the adoption of electronic prescribing systems in hospitals? A systematic review and thematic synthesis of qualitative studies. BMJ Qual Saf 2019 Dec;28(12):1021-1031 [FREE Full text] [doi: 10.1136/bmjqs-2018-009082] [Medline: 31358686]
12. Gagnon MP, Nsangou ER, Payne-Gagnon J, Grenier S, Sicotte C. Barriers and facilitators to implementing electronic prescription: a systematic review of user groups' perceptions. J Am Med Inform Assoc 2014;21(3):535-541 [FREE Full text] [doi: 10.1136/amiajnl-2013-002203] [Medline: 24130232]
13. Kruse CS, Goetz K. Summary and frequency of barriers to adoption of CPOE in the U.S. J Med Syst 2015 Feb;39(2):15 [FREE Full text] [doi: 10.1007/s10916-015-0198-2] [Medline: 25638719]
14. Van Dort BA, Zheng WY, Baysari MT. Prescriber perceptions of medication-related computerized decision support systems in hospitals: a synthesis of qualitative research. Int J Med Inform 2019 Sep;129:285-295. [doi: 10.1016/j.ijmedinf.2019.06.024] [Medline: 31445268]
15. Ahmad N. Adoption, implementation and usage of enterprise systems: an empirical study. Researchgate 2012. [doi: 10.13140/2.1.2010.9762]
16. Trice AW, Treacy ME. Utilization as a dependent variable in MIS research. SIGMIS Database 1988 Oct;19(3-4):33-41. [doi: 10.1145/65766.65771]
17. Burton-Jones A, Gallivan MJ. Toward a deeper understanding of system usage in organizations: a multilevel perspective. MIS Quarterly 2007;31(4):657. [doi: 10.2307/25148815]
18. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
19. Delone WH, McLean ER. The Delone and Mclean model of information systems success: a ten-year update. J Manage Info Sys 2014 Dec 23;19(4):9-30. [doi: 10.1080/07421222.2003.11045748]
20. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. Br Med J 2015 Jan 02;350:g7647 [FREE Full text] [doi: 10.1136/bmj.g7647] [Medline: 25555855]

21. Mixed Methods Appraisal Tool (MMAT) Version 2018 user guide. Department of Family Medicine. 2018. URL: http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_criteria%20manual_2018%202008%202001_ENG.pdf [accessed 2021-02-11]
22. Vusio F, Thompson A, Birchwood M, Clarke L. Experiences and satisfaction of children, young people and their parents with alternative mental health models to inpatient settings: a systematic review. Eur Child Adolesc Psychiatry 2020 Dec;29(12):1621-1633 [FREE Full text] [doi: 10.1007/s00787-019-01420-7] [Medline: 31637520]
23. Mays N, Pope C, Popay J. Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. J Health Serv Res Policy 2005 Jul;10 Suppl 1:6-20. [doi: 10.1258/1355819054308576] [Medline: 16053580]
24. Abramson EL, Patel V, Pfoh ER, Kaushal R. How physician perspectives on e-Prescribing evolve over time. A case study following the transition between EHRs in an outpatient clinic. Appl Clin Inform 2016 Oct 26;7(4):994-1006 [FREE Full text] [doi: 10.4338/ACI-2016-04-RA-0069] [Medline: 27786335]
25. Holden RJ. Physicians' beliefs about using EMR and CPOE: in pursuit of a contextualized understanding of health IT use behavior. Int J Med Inform 2010 Feb;79(2):71-80 [FREE Full text] [doi: 10.1016/j.ijmedinf.2009.12.003] [Medline: 20071219]
26. Schectman JM, Schorling JB, Nadkarni MM, Voss JD. Determinants of physician use of an ambulatory prescription expert system. Int J Med Inform 2005 Sep;74(9):711-717. [doi: 10.1016/j.ijmedinf.2005.05.011] [Medline: 15985385]
27. Shriner AR, Webber EC. Attitudes and perceptions of pediatric residents on transitioning to CPOE. Appl Clin Inform 2014;5(3):721-730 [FREE Full text] [doi: 10.4338/ACI-2014-04-RA-0045] [Medline: 25298812]
28. Hellström L, Waern K, Montelius E, Astrand B, Rydberg T, Petersson G. Physicians' attitudes towards ePrescribing--evaluation of a Swedish full-scale implementation. BMC Med Inform Decis Mak 2009 Aug 07;9:37 [FREE Full text] [doi: 10.1186/1472-6947-9-37] [Medline: 19664219]
29. Omar A, Ellenius J, Lindemalm S. Evaluation of electronic prescribing decision support system at a tertiary care pediatric hospital: the user acceptance perspective. Stud Health Technol Inform 2017;234:256-261. [Medline: 28186051]
30. Rahimi B, Timpka T, Vimarlund V, Uppugunduri S, Svensson M. Organization-wide adoption of computerized provider order entry systems: a study based on diffusion of innovations theory. BMC Med Inform Decis Mak 2009 Dec 31;9:52 [FREE Full text] [doi: 10.1186/1472-6947-9-52] [Medline: 20043843]
31. Martens JD, van der Weijden T, Winkens RAG, Kester ADM, Geerts PJH, Evers SMAA, et al. Feasibility and acceptability of a computerised system with automated reminders for prescribing behaviour in primary care. Int J Med Inform 2008 Mar;77(3):199-207. [doi: 10.1016/j.ijmedinf.2007.05.013] [Medline: 17631412]
32. Saddik B, Al-Fridan MM. Physicians' satisfaction with computerised physician order entry (CPOE) at the National Guard Health Affairs: a preliminary study. Stud Health Technol Inform 2012;178:199-206. [Medline: 22797042]
33. Santucci W, Day RO, Baysari MT. Evaluation of hospital-wide computerised decision support in an intensive care unit: an observational study. Anaesth Intensive Care 2016 Jul;44(4):507-512 [FREE Full text] [doi: 10.1177/0310057X1604400403] [Medline: 27456183]
34. Tan WS, Phang JS, Tan LK. Evaluating user satisfaction with an electronic prescription system in a primary care group. Ann Acad Med Singap 2009 Jun;38(6):494-497 [FREE Full text] [Medline: 19565099]
35. Chau PY, Hu PJ. Examining a model of information technology acceptance by individual professionals: an exploratory study. J Manage Info Sys 2014 Dec 23;18(4):191-229. [doi: 10.1080/07421222.2002.11045699]
36. Schneider EC, Timbie JW, Fox DS, Van Busum KR, Caloyeras JP. Dissemination and adoption of comparative effectiveness research findings when findings challenge current practices. CPOE Case-Study Report.: RAND Corporation; 2011. URL: https://www.rand.org/pubs/technical_reports/TR924.html [accessed 2021-02-11]
37. Gagnon MP, Desmartis M, Labrecque M, Car J, Pagliari C, Pluye P, et al. Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals. J Med Syst 2012 Feb;36(1):241-277 [FREE Full text] [doi: 10.1007/s10916-010-9473-4] [Medline: 20703721]
38. Handayani PW, Hidayanto AN, Budi I. User acceptance factors of hospital information systems and related technologies: systematic review. Inform Health Soc Care 2018 Dec;43(4):401-426. [doi: 10.1080/17538157.2017.1353999] [Medline: 28829650]
39. Phansalkar S, van der Sijs H, Tucker AD, Desai AA, Bell DS, Teich JM, et al. Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. J Am Med Inform Assoc 2013 May 01;20(3):489-493 [FREE Full text] [doi: 10.1136/amiajnl-2012-001089] [Medline: 23011124]
40. Li J, Talaei-Khoei A, Seale H, Ray P, Macintyre CR. Health care provider adoption of eHealth: systematic literature review. Interact J Med Res 2013 Apr 16;2(1):e7 [FREE Full text] [doi: 10.2196/ijmr.2468] [Medline: 23608679]
41. Ojo AI. Validation of the DeLone and McLean information systems success model. Healthc Inform Res 2017 Jan;23(1):60-66 [FREE Full text] [doi: 10.4258/hir.2017.23.1.60] [Medline: 28261532]
42. Williams MD, Rana NP, Dwivedi YK. The unified theory of acceptance and use of technology (UTAUT): a literature review. Journal of Ent Info Management 2015 Apr 13;28(3):443-488. [doi: 10.1108/JEIM-09-2014-0088]

Abbreviations

CDS: clinical decision support
CPOE: computerized physician order entry
ICU: intensive care unit
IT: information technology
MMAT: Mixed Methods Appraisal Tool
UTAUT: unified theory of acceptance and use of technology

Edited by G Eysenbach; submitted 30.07.20; peer-reviewed by S Santos, M Lavin; comments to author 22.09.20; revised version received 17.11.20; accepted 07.12.20; published 04.03.21.

Please cite as:
Mogharbel A, Dowding D, Ainsworth J
Physicians' Use of the Computerized Physician Order Entry System for Medication Prescribing: Systematic Review
JMIR Med Inform 2021;9(3):e22923
URL: https://medinform.jmir.org/2021/3/e22923
doi:10.2196/22923
PMID:33661126

Original Paper

Comparative Analysis of Paper-Based and Web-Based Versions of the National Comprehensive Cancer Network-Functional Assessment of Cancer Therapy-Breast Cancer Symptom Index (NFBSI-16) Questionnaire in Breast Cancer Patients: Randomized Crossover Study

Jinfei Ma[1*], MD, PhD; Zihao Zou[1*], MD, MMed; Emmanuel Eric Pazo[2], MD, MSc, PhD; Salissou Moutari[3], PhD; Ye Liu[1], BSc; Feng Jin[1], MD, PhD

[1]Department of Breast Surgery, The First Affiliated Hospital of China Medical University, Shenyang, China

[2]Department of Ophthalmology, He Eye Hospital, Shenyang, China

[3]Mathematical Science Research Centre, Queen's University Belfast, Belfast, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Feng Jin, MD, PhD
Department of Breast Surgery
The First Affiliated Hospital of China Medical University
No 155 Nanjing Road, Heping District
Shenyang, 110001
China
Phone: 86 18040031101
Email: jinfeng@cmu.edu.cn

Abstract

Background: Breast cancer remains the most common neoplasm diagnosed among women in China and globally. Health-related questionnaire assessments in research and clinical oncology settings have gained prominence. The National Comprehensive Cancer Network–Functional Assessment of Cancer Therapy–Breast Cancer Symptom Index (NFBSI-16) is a rapid and powerful tool to help evaluate disease- or treatment-related symptoms, both physical and emotional, in patients with breast cancer for clinical and research purposes. Prevalence of individual smartphones provides a potential web-based approach to administrating the questionnaire; however, the reliability of the NFBSI-16 in electronic format has not been assessed.

Objective: This study aimed to assess the reliability of a web-based NFBSI-16 questionnaire in breast cancer patients undergoing systematic treatment with a prospective open-label randomized crossover study design.

Methods: We recruited random patients with breast cancer under systematic treatment from the central hospital registry to complete both paper- and web-based versions of the questionnaires. Both versions of the questionnaires were self-assessed. Patients were randomly assigned to group A (paper-based first and web-based second) or group B (web-based first and paper-based second). A total of 354 patients were included in the analysis (group A: n=177, group B: n=177). Descriptive sociodemographic characteristics, reliability and agreement rates for single items, subscales, and total score were analyzed using the Wilcoxon test. The Lin concordance correlation coefficient (CCC) and Spearman and Kendall τ rank correlations were used to assess test-retest reliability.

Results: Test-retest reliability measured with CCCs was 0.94 for the total NFBSI-16 score. Significant correlations (Spearman ρ) were documented for all 4 subscales—Disease-Related Symptoms Subscale–Physical (ρ=0.93), Disease-Related Symptoms Subscale–Emotional (ρ=0.85), Treatment Side Effects Subscale (ρ=0.95), and Function and Well-Being Subscale (ρ=0.91)—and total NFBSI-16 score (ρ=0.94). Mean differences of the test and retest were all close to zero (≤0.06). The parallel test-retest reliability of subscales with the Wilcoxon test comparing individual items found GP3 (item 5) to be significantly different (*P*=.02). A majority of the participants in this study (255/354, 72.0%) preferred the web-based over the paper-based version.

Conclusions: The web-based version of the NFBSI-16 questionnaire is an excellent tool for monitoring individual breast cancer patients under treatment, with the majority of participants preferring it over the paper-based version.

(JMIR Med Inform 2021;9(3):e18269) doi:10.2196/18269

KEYWORDS

breast cancer; NFBSI-16; patient-reported outcome; reproducibility; test-retest reliability; web-based questionnaire

Introduction

Breast cancer accounts for the highest proportion of malignant tumors among women (excluding skin cancers) globally. According to an International Agency for Research on Cancer report [1], the worldwide burden for breast cancer was 2.1 million cases in the year 2018, accounting for 1 in 4 cancer cases among women. Advancements in breast cancer screening, detection, and treatment over the last few decades have produced an increased chance of cure for early-stage breast cancer patients, while advanced (metastatic) disease patients now have prolonged survival and varying degrees of controlled symptoms [2,3]. However, full-aspect and long-term treatment can impact patients' and survivors' quality of life and therefore require continual health management during and after the process of recovery [4].

Breast cancer and its treatment have been documented to significantly disrupt patients' health-related quality of life, which has been found to predict survival time and additionally showed more significance for noncurative patients [5-10]. To assess treatment benefits, patient-reported outcome measures (PROMs) provide unique perspectives on cancer symptoms from patients' experience, some of which can be neglected by clinicians and laboratory tests [11-13]. The National Comprehensive Cancer Network–Functional Assessment of Cancer Therapy–Breast Cancer Symptom Index (NFBSI-16) PROMs were regulated on the foundation of the Functional Assessment of Chronic Illness Therapy (FACIT) measurement system to assess high-priority symptoms of breast cancer, emphasizing patients' input, which can be applied to help evaluate the effectiveness of treatments for breast cancer in clinical practice and research [14-16].

The migration from paper-based to web-based versions does not guarantee preservation of psychometric properties of the scale since various factors have the potential to impact the performance of the questionnaire scale when adapted for web-based administration, such as layout, instructions, or restructuring of item and response. Researchers have investigated methods of validation, routes of administration, practical considerations, and reliability of electronic PROMs [17-27]. Gwaltney et al's meta-analysis on assessing the equivalence of computer versus paper versions of PROMs showed "a high overall level of agreement between paper and computerized measures" [28]. The review encompassed the fields of rheumatology, cardiology, psychiatry, asthma, alcoholism, pain assessment, gastrointestinal disease, diabetes, and allergies. In contrast, a study of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30 found small but statistically significant differences in scale mean scores (3 to 7 points on a 100-point scale) associated with mode of administration [29]. Various validated web-based questionnaires in oncology have been demonstrated to be reliable and effective tools for assessing PROMs in therapeutic clinical and research settings [30-33]. Currently in China, web-based versions of clinical research questionnaires using WeChat are rapidly growing in number, and various studies have validated the WeChat-based administration of health-related questionnaires [34-36]. To cover the large growing patient base in China, we expected web-based administration of the NFBSI-16 to be a reliable methodology to assess the impact of disease, treatment, and well-being status among patients with breast cancer. Additionally, it could be a more cost-effective and efficient method to apply in the growing number of patients in certain demographics.

The aim of this study was to analyze the reliability of a web-based NFBSI-16 questionnaire (Chinese language) for measuring disease- and treatment-related symptoms and concerns in breast cancer patients, comparing it with the validated paper-based version.

Methods

Study Design and Patient Enrolment

Patients were recruited from the Department of Breast Surgery of the First Affiliated Hospital of China Medical University, Shenyang, China, between October 2019 and January 2020. The inclusion criteria were female gender, full legal age, proven diagnosis of breast cancer, being under systemic anticancer treatment, ability to follow study instructions, sufficient literacy and fluency in Chinese to comprehend the questionnaires, ability and willingness to complete the study protocol, and signed declaration of consent. Potential participants were excluded if they could not provide informed consent or participated in other studies (burden of participation). Participants had an initial clinic visit at which eligibility was assessed. All eligible participants were randomly chosen from the hospital's central registry and invited to volunteer for the study via face-to-face interview with a trained research clinician. Written informed consents were obtained. The study protocol was approved by the First Affiliated Hospital of China Medical University ethics committee.

The study was a randomized crossover design in which all participants completed both a standard paper questionnaire and a web-based version of the NFBSI-16 (Multimedia Appendix 1). Patients in group A were assigned to start with the paper-based version followed by the web-based version on their smartphone in the same session. Patients in group B completed the web-based version followed by the paper-based version. Participants were randomized immediately after enrolment to group A or B in a 1:1 ratio using a computer-generated randomization list with a specified seed and block size of 6, based on the mode of administration to be completed first. Between each session from paper-based to web-based and web-based to paper-based, participants were given a break of 15 minutes during which they were invited into a quick patient education seminar, which was also a routine activity in our

department as a distractor task to lower the potential carryover effect. All participants were provided with written instructions for completion of the paper- or web-based questionnaires prior to their questionnaires being administered. After completing both versions of the NFBSI-16 questionnaire, participants were invited to state their preference for either the paper- or web-based NFBSI-16 questionnaire.

Questionnaire

The NFBSI-16 contains items from the original FBSI and FACIT measures selected by patients and clinicians according to their priority concern [15], which presented as a more direct tool to reflect the effectiveness of treatments for advanced breast cancer. The NFBSI-16 comprises 16 items with 4 dimensions for ease of use and scoring: Disease-Related Symptoms Subscale–Physical (DRS-P), Disease-Related Symptoms Subscale–Emotional (DRS-E), Treatment Side Effects Subscale (TSE), and Function and Well-Being Subscale (FWB). Therefore, clinicians and researchers can individually view and assess subscale scores when concerned about a particular class of symptoms. The questionnaires were self-completed, and careful attention was paid to the design and layout of the web-based version. In order to reduce the risk of errors in posing, interpretation, recording, and coding responses and potential interrater variability, the theory-based guidelines for self-administered questionnaire design were followed by the authors (Multimedia Appendix 1) [37]. The web-based user interface and paper for the paper-based questionnaires were free from all other information such as logos, slogans, advertisement, etc. The instructions for completing the web-based and paper-based questionnaires were included at the beginning of the web-based interface and header of the paper, respectively. In brief, while participating in the web-based assessment, patients had to scan a redesignated Quick Response code using their smartphone. This action automatically took them to a web-based test, and the user had to select the intensity or severity of the 16 items. After completing the 16th question, the interface turned into a blank screen indicating the test was over. On the other hand, the paper-based questionnaire test was conducted using white paper and pencil. The text was printed using clear 12-point font.

Testing of the Instrument

During pretesting and pilot testing, 3 colleagues specializing in oncology and 3 nonexperts evaluated the web-based questionnaire's usability, accessibility, and clarity of the user interface. This testing was only conducted on the functionality of the web-based questionnaire since the format, structure, and sequence of items in the web-based questionnaire were the same as in the validated paper-based questionnaire.

Computation of Subscale Scores

Data from the paper questionnaires were entered manually into an electronic patient management system by the authors, and data from the web-based questionnaires were automatically captured after the participant completed the online questionnaire and downloaded to the electronic patient management system. All data was anonymized. We assessed the completeness of the data on a per-item basis and questionnaire basis. The total scores were obtained by taking the mean score across completed items and multiplying by 16, the number of items (following official guidelines) [15]. All subscale totals ranged from 0 to 4, with a score of 0 representing that the patient agrees with the item "not at all" and 4 representing "very much". Subscale scores and total scores were computed for each participant and each mode of administration separately. Comparative analyses of individual items, subscales, and total score were the primary goal of the study.

Statistical Analysis

All statistical analyses were conducted using SPSS Statistics, version 25 (IBM Corp). Frequency analysis was performed to determine the descriptive sociodemographic characteristics of patients. Referring to ISPOR ePRO Good Research Practices Task Force recommendations [21], we conducted the evaluation of measurement equivalence. Reliability, internal consistency, disparity of responses, and the rate of consistency between paper- and web-based responses were assessed. Reliability was calculated for the 16 individual items as well as for scores of the 4 subscales (DRS-P, DRS-E, TSE, and FWB) and the NFBSI-16 total score in accordance with the NFBSI-16 guidelines [15]. The primary study outcome was to assess the reliability of single items and total score of the web-based questionnaire. The Wilcoxon test was used to identify possible statistically significant differences in the test of parallel forms reliability, both between the single items and the scores due to the ordinal nature of the data. The secondary outcome measure was to assess the consistency and agreement of the web-based questionnaire with the paper-based questionnaire. The mean values of the paper- and web-based measures were calculated, consistency analyses were performed by calculation of the Spearman rank correlation coefficient (Spearman ρ), and agreement rates for each item were assessed using rank correlation (Kendall τ) for each scale. As a second measure of test-retest reliability, we calculated the Lin concordance correlation coefficient (CCC) [38]. Finally, all answers to the ''preference'' questionnaire were compared between the web-based and the paper version of the NFBSI-16 using χ^2 tests. In all analyses, $P<.05$ (2-tailed) was considered indicative of statistically significant differences ($\alpha=.05$). As such an analysis is considered an explorative study, all reported P values can be taken as purely descriptive. All figures (box plot and correlation diagram) were generated in SPSS Statistics.

Results

Enrolment of Patients

The final analysis included 354 patients with breast cancer receiving systematic treatment who completed both the paper- and web-based versions of NFBSI-16 questionnaire. Initially, 380 patients were assessed for eligibility. 26 patients were excluded, as shown in the study flow diagram (Figure 1). Since there was no internal difference between group A and group B, demographically, two groups were combined in the final analysis. The mean age was 49.5 years (SD 10.44). Other basic characteristics of patients are shown in Table 1.

Figure 1. Study flow diagram.

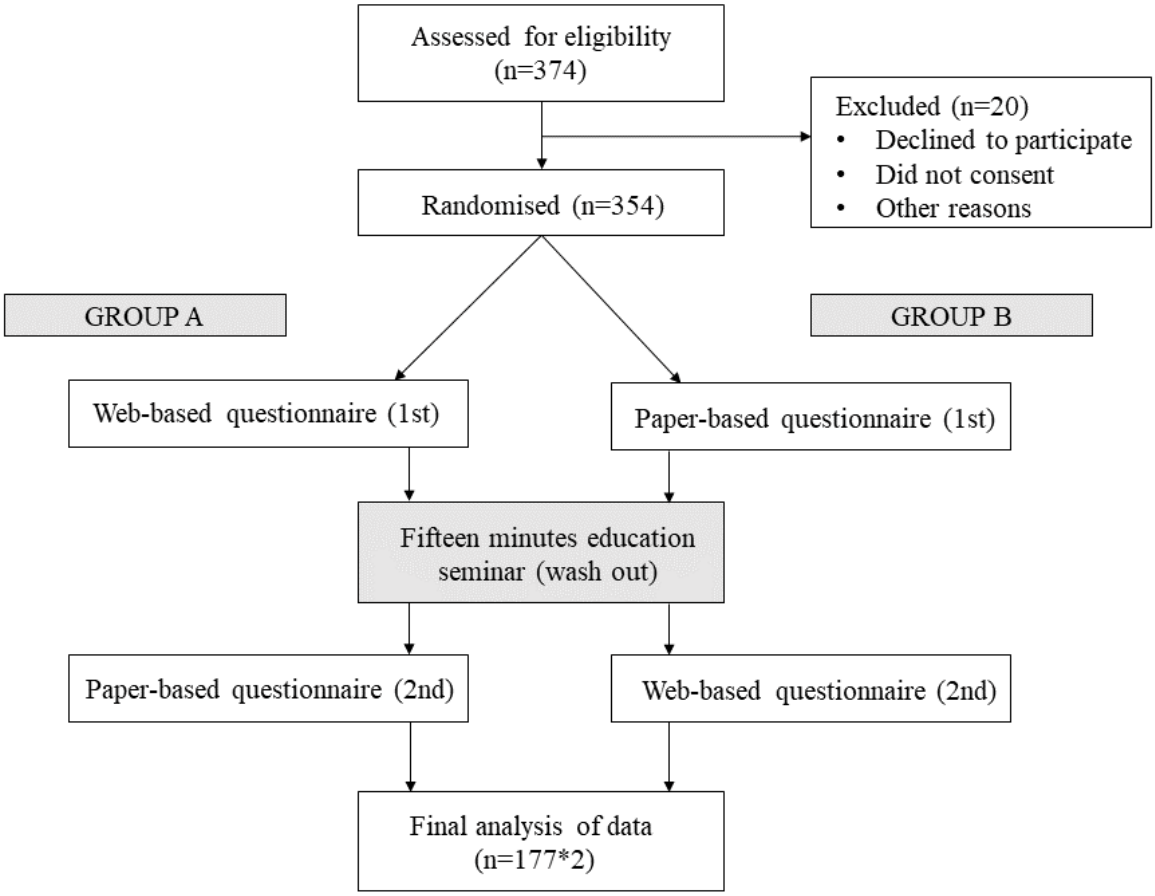


Table 1. Basic characteristics of study participants.

Patient characteristics	n (%)
Menstrual status	
Premenopause	133 (37.6)
Perimenopause	107 (30.2)
Postmenopause	114 (32.2)
Level of education completed	
Primary	59 (16.7)
Secondary	161 (45.5)
Tertiary	134 (37.9)
Marital status	
Single	16 (4.5)
Married	338 (95.5)
Region	
Rural	165 (46.6)
Urban	189 (53.4)
Treatment	
Neoadjuvant therapy	244 (68.9)
Adjuvant therapy	110 (31.1)

Parallel Forms Reliability

The Wilcoxon signed rank test analyzed parallel reliability in the single items of the NFBSI-16, shown in Table 2. No systematic location difference between the two versions of questionnaires (paper- and web-based versions) was observed for continuous variables except for item 5 (GP3 question). A very large proportion of the items answered by the patients had the same response (ties) in both versions of the questionnaire, suggesting high parallel reliability as only one significant difference (out of 16 in total) could be found in the single-item comparison. A statistically significant difference could only be identified in question GP3, "Because of my physical condition, I have trouble meeting the needs of my family." GP3 was reported slightly higher in the paper-based questionnaire (mean 2.07, SD 0.98), while in the web-based version the same participants scored it at a mean of 2.00 (SD 0.91). Additionally, the medians of the item GP3 for the paper- and web-based questionnaires were the same (median 2; IQR 1-3). While the web-based total mean score was slightly higher than the paper-based score by 0.08 points, they had no statistically significant difference between them. Figure 2 illustrates the distribution of the paper-based and web-based total scores in a box plot. The slightly higher total web-based total score can be attributed to a few outliers shown in the box plot. The web-based whisker of the box plot IQR was within the broader IQR of the paper-based version. In addition, slight differences of less than 0.50 points were found between the paper-based and web-based questionnaires when the item scores of the 4 dimensions (DRS-P, DRS-E, TSE, and FWB) were calculated and compared. However, all 4 dimensions' scores showed no statistically significant differences when compared (Table 3).

Table 2. Parallel test-retest reliability of single items and total score (Wilcoxon test).

NFBSI-16[a] items	Paper-based patient score		Web-based patient score		*P* value	Δ
	Mean (SD)	Median (IQR)	Mean' (SD)	Median (IQR)		|Mean−Mean'|
Disease-Related Symptoms Subscale – Physical (DRS-P)						
GP1 (item 1)	2.31 (0.92)	2 (2-3)	2.32 (0.90)	2 (2-3)	.58	0.01
GP4 (item 2)	2.19 (0.90)	2 (2-3)	2.17 (0.88)	2 (2-3)	.36	0.02
GP6 (item 3)	2.29 (1.13)	2 (1-3)	2.30 (1.15)	2 (1-3)	.88	0.01
B1 (item 4)	2.01 (0.89)	2 (1-3)	2.00 (0.88)	2 (1-3)	.79	0.01
GP3 (item 5)	2.07 (0.98)	2 (1-3)	2.00 (0.91)	2 (1-3)	.02[b]	0.07
HI7 (item 6)	2.59 (1.02)	2 (2-3)	2.59 (1.06)	2 (2-3)	.91	0.00
BP1 (item 7)	1.88 (0.93)	2 (1-2)	1.90 (0.93)	2 (1-2)	.27	0.02
GF5 (item 8)	2.59 (1.18)	2 (2-3)	2.55 (1.17)	2 (2-3)	.38	0.04
Disease-Related Symptoms Subscale – Emotional (DRS-E)						
GE6 (item 9)	2.00 (1.04)	2 (1-2)	2.01 (1.05)	2 (1-2)	.73	0.01
Treatment Side Effects Subscale (TSE)						
GP2 (item 10)	2.20 (1.15)	2 (1-3)	2.25 (1.10)	2 (1-3)	.16	0.05
N6 (item 11)	1.87 (0.98)	2 (1-2)	1.85 (0.93)	2 (1-2)	.42	0.02
GP5 (item 12)	2.77 (1.01)	3(2-3)	2.75 (1.00)	3 (2-3)	.45	0.02
B5 (item 13)	2.98 (1.35)	3 (2-4)	2.98 (1.33)	3 (2-4)	.89	0.00
Function and Well-Being Subscale (FWB)						
GF1 (item 14)	2.52 (1.04)	2 (2-3)	2.55 (1.01)	2.5 (2-3)	.14	0.03
GF3 (item 15)	2.82 (1.12)	3 (2-4)	2.81 (1.08)	3 (2-4)	.83	0.01
GF7 (item 16)	2.82 (1.19)	3 (2-4)	2.85 (1.21)	3 (2-4)	.67	0.03
Total score						
NFBSI-16 score	37.92 (7.79)	38 (32-42.5)	37.88 (7.71)	38 (32.75-42)	.98	0.04

[a]NFBSI-16: National Comprehensive Cancer Network–Functional Assessment of Cancer Therapy–Breast Cancer Symptom Index.

[b]Statistically significant difference.

Figure 2. Box plot comparison of paper-based and web-based distribution of total scores.

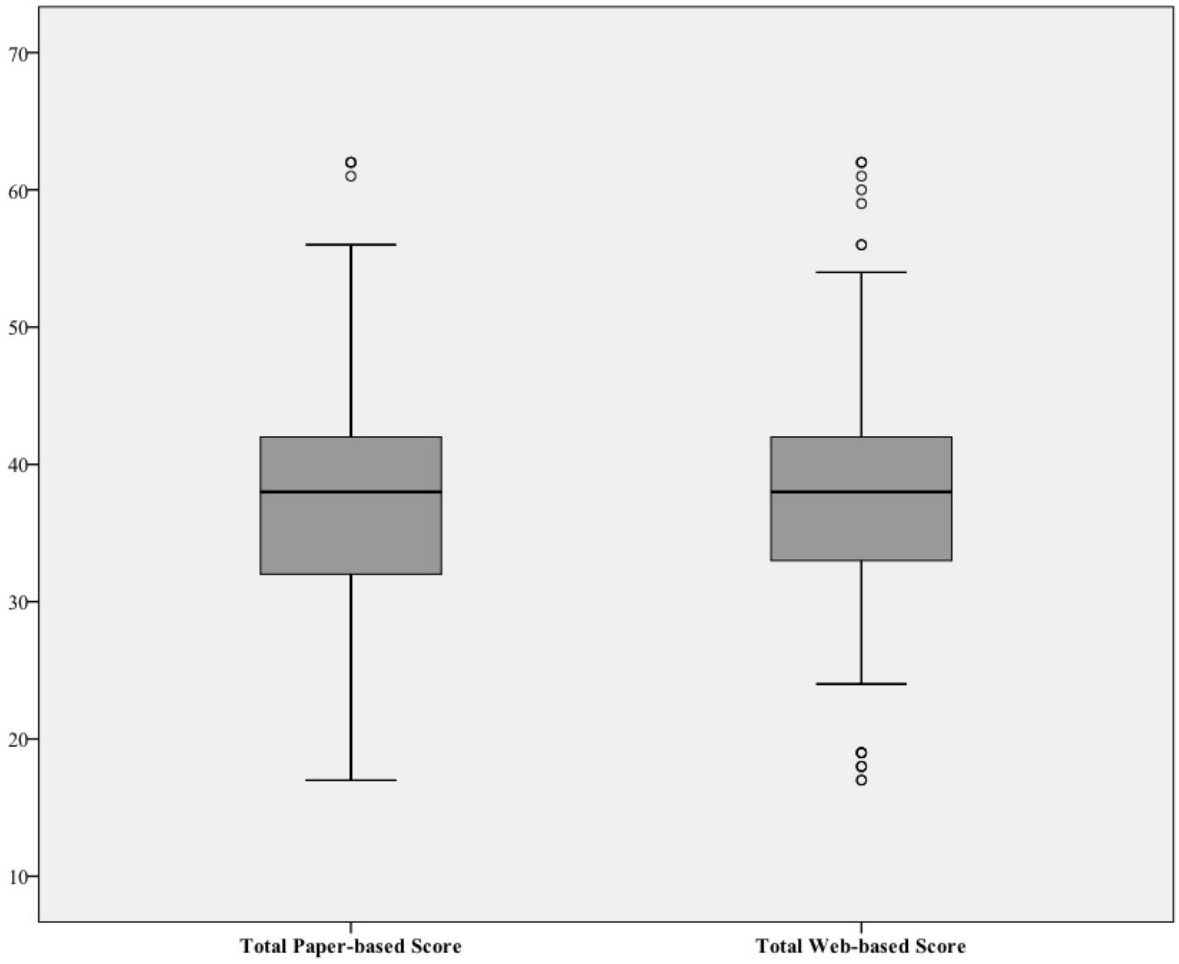


Table 3. Parallel test-retest reliability of subscales (Wilcoxon test).

NFBSI-16[a] subscale	Paper-based patient outcome		Web-based patient outcome		*P* value	Δ
	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)		|Mean−Mean'|
Disease-Related Symptoms Subscale–Physical	35.90 (9.59)	36 (30-42)	35.64 (9.58)	34 (10-42)	.43	0.26
Disease-Related Symptoms Subscale–Emotional	32.00 (16.54)	32 (16-32)	32.05 (16.84)	32 (16-32)	.98	0.05
Treatment Side Effects Subscale	39.20 (13.39)	40 (28-48)	39.20 (12.86)	36 (32-48)	.62	0.00
Function and Well-Being Subscale	43.37 (13.37)	42.67 (32-53.33)	43.62 (13.72)	42.67 (32-53.33)	.32	0.25

[a]NFBSI-16: National Comprehensive Cancer Network–Functional Assessment of Cancer Therapy–Breast Cancer Symptom Index.

Test of Internal Consistency

Table 4 shows the Spearman ρ correlation values between the individual items from the paper- and web-based questionnaires. All 16 items demonstrated a high correlation (>0.8) between paper- and web-based items. Individual item internal consistency test was performed by Kendall τ analysis between the two versions. In all items, the rank correlation was high as the Kendall τ coefficients ranged between 0.787 and 0.877 and were all statistically significant. With each data point reflecting an individual patient's total NFBSI-16 score, Figure 3 depicts a positive correlation between total paper-based and web-based scores. Overall, CCC agreement between paper-based and web-based questionnaires' item scores were all comparably high at 0.94 (fair: 0.21-0.40; moderate: 0.41-0.60; substantial: 0.61-0.80; almost perfect: 0.81-1.00), as represented in Table 5.

Table 4. Correlation between test-retest in individual items and subscale (Spearman ρ and Kendall τ analysis).

Items	Spearman ρ	*P* value	Kendall τ	*P* value
Disease-Related Symptoms Subscale – Physical (DRS-P)				
GP1 (item 1)	0.89	<.001	0.877	<.001
GP4 (item 2)	0.84	<.001	0.810	<.001
GP6 (item 3)	0.86	<.001	0.804	<.001
B1 (item 4)	0.90	<.001	0.87	<.001
GP3 (item 5)	0.85	<.001	0.825	<.001
HI7 (item 6)	0.85	<.001	0.813	<.001
BP1 (item 7)	0.89	<.001	0.856	<.001
GF5 (item 8)	0.84	<.001	0.796	<.001
Subscale total	0.93	<.001	0.827	<.001
Disease-Related Symptoms Subscale – Emotional (DRS-E)				
GE6 (item 9)	0.85	<.001	0.826	<.001
Subscale total	0.85	<.001	0.882	<.001
Treatment Side Effects Subscale (TSE)				
GP2 (item 10)	0.88	<.001	0.830	<.001
N6 (item 11)	0.89	<.001	0.857	<.001
GP5 (item 12)	0.83	<.001	0.795	<.001
B5 (item 13)	0.84	<.001	0.788	<.001
Subscale total	0.95	<.001	0.882	<.001
Function and Well-Being Subscale (FWB)				
GF1 (item 14)	0.82	<.001	0.787	<.001
GF3 (item 15)	0.86	<.001	0.821	<.001
GF7 (item 16)	0.83	<.001	0.79	<.001
Subscale total	0.91	<.001	0.825	<.001
Total score				
Score	0.94	<.001	0.823	<.001

Figure 3. Correlation between total paper-based and web-based scores.

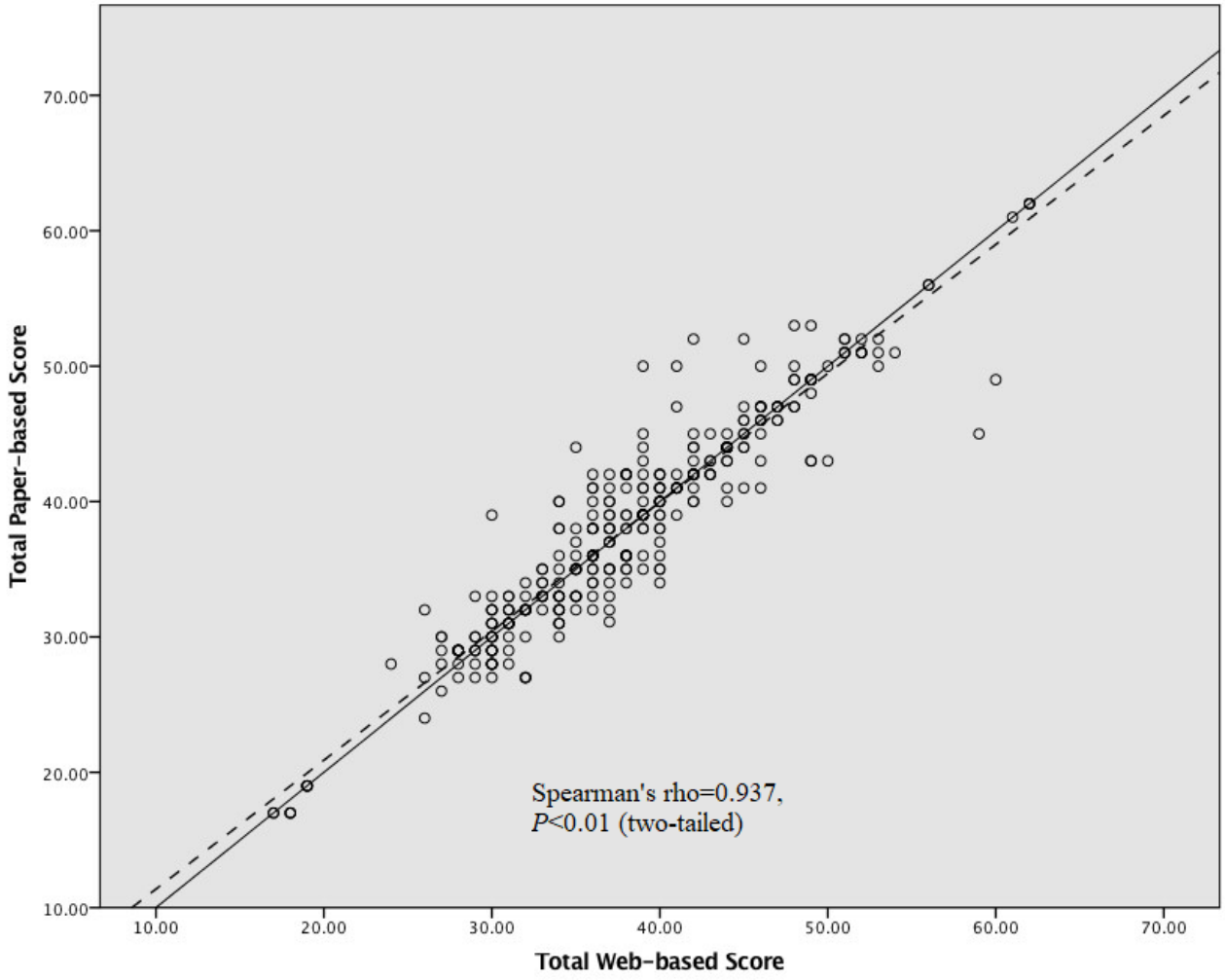


Table 5. Agreement between paper-based and web-based questionnaires scores (Lin concordance correlation coefficient analysis).

Items	R_c[a]	95% CI
Disease-Related Symptoms Subscale – Physical (DRS-P)		
GP1 (item 1)	0.92	0.90-0.94
GP4 (item 2)	0.85	0.82-0.88
GP6 (item 3)	0.86	0.83-0.88
B1 (item 4)	0.9	0.88-0.71
GP3 (item 5)	0.86	0.83-0.89
HI7 (item 6)	0.86	0.83-0.89
BP1 (item 7)	0.88	0.87-0.91
GF5 (item 8)	0.85	0.82-0.88
Subscale total	0.94	0.93-0.95
Disease-Related Symptoms Subscale – Emotional (DRS-E)		
GE6 (item 9)	0.84	0.81-0.87
Subscale total	0.84	0.81-0.87
Treatment Side Effects Subscale (TSE)		
GP2 (item 10)	0.87	0.85-0.90
N6 (item 11)	0.88	0.86-0.91
GP5 (item 12)	0.86	0.83-0.89
B5 (item 13)	0.83	0.80-0.86
Subscale total	0.96	0.95-0.97
Function and Well-Being Subscale (FWB)		
GF1 (item 14)	0.85	0.82-0.88
GF3 (item 15)	0.86	0.83-0.89
GF7 (item 16)	0.84	0.81-0.87
Subscale total	0.91	0.89-0.93
Total score		
Score	0.94	0.93-0.95

[a]R_c: concordance correlation coefficient.

Patient Preference

Table 6 shows a majority of the participants preferred answering the same questions in a web-based format rather than paper-based format. The difference in preference was statistically different.

Table 6. Analysis of participant preference.

Patient preference	Observed, n	Expected, n	Residual	Chi-square (*df*)	Asymptotic significance
Preferred paper-based questionnaire	98	177	−79		
Preferred web-based questionnaire	256	177	79		
Total	354			70.5[a] (1)	.001[b]

[a]0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 177.0.

[b]Statistically significant difference.

Estimation of the Carryover Effect

To assess the carryover effect, we let s_A denote the sum (total scores from web-based items plus the total scores from paper-based items for each respondent) from group A and let s_B denote the sum from group B. We estimate the carryover effect in both groups (A and B) using the Wilcoxon test on the sum values s_A and s_B, and at a level of significance of 5%, the

possible carryover effect is not significantly different between the different sequences (P=0.84).

Discussion

Principal Results

Overall, reliability was considered to be excellent for the web-based version as measured with the Wilcoxon signed rank test and CCC. Additionally, Spearman ρ correlation and Kendall τ analysis showed that mean differences were all close to zero, supporting good reliability of the web-based version of the NFBSI-16 self-administered questionnaire. In this study, we used the Wilcoxon signed rank test and CCC to assess test reliability. However, different methods can be used to assess test-retest reliability, and there is much discussion in the literature on the best possible methodology [39]. Intraclass correlation coefficient (ICC) was first introduced in 1954 and is a modification of the Pearson correlation coefficient. However, modern ICC is calculated by mean squares (ie, estimates of the population variances based on the variability among a given set of measures) obtained through analysis of variance (ANOVA). The disadvantage of ICC in patient group analysis is that if the groups are mainly homogeneous, the ICC tends to be low, because the ICC compares variance among patients to total variance. If patient groups are mainly heterogeneous, the ICC tends to be high. Thus, ICC would only generalize to similar populations. Additionally, the 1-way ICC does not consider the order in which observations were made [40]. Therefore, the CCC is a useful measure as it not only covers mean differences between the first and second measurements, such as ICCs calculated by a 1-way ANOVA, but also takes the variance differences between the first (paper-based) and second (web-based) measurements into consideration by reducing the magnitude of the resulting test-retest reliability estimate. In conclusion, CCC is a better tool that distinguishes bias between imprecision [39,40].

Limitations

This study may also have some limitations. First, the significant difference in item 5 (GP3) between paper- and web-based measurement of the NFBSI-16 (Table 3) was an unexpected finding. We think this significant difference might be due to an outlier. This assumption was supported by the fact that even though 293 out of 354 (total) patients had the same answer for the paper- and web-based for item 5 (high number of similarities), a significant difference in the mean was detected. Second, according to the nature of this study, it is difficult to generalize some of our findings as its limited by demographic settings.

Comparison With Prior Work

NFBSI-16 includes all 8 items from the original FBSI and 8 additional items from FACIT measures, which cover most essential breast cancer–related symptoms and concerns endorsed by both oncology patients and clinicians [15]. Compared to the previous version (FBSI), it emphasizes patient input following Food and Drug Administration guidance for PROMs [41] and has been validated as a comprehensive and powerful tool to evaluate the effectiveness of treatments for breast cancer in clinical practice and research. In addition, the layout of 4 clear separated subscales benefits any clinicians, patients, or researchers by allowing them to view particular domains they are concerned about. However, the reliability of an electronic version in Chinese language has not been tested. This paper describes the evaluation of the test-retest reliability of the web-based version of the NFBSI-16 self-administered questionnaire. When designing a web-based version of a validated paper-based questionnaire, one has to take into consideration variables such as text size, column formatting, contrast, layout, use of corrective lenses, etc. We created the web-based NFBSI-16 to be consistent with the original as far as possible. In addition, technology skills required to complete a web-based questionnaire can differ from those needed to complete a paper-based questionnaire. However, our study found no clinically significant differences between scores obtained from the paper-and web-based versions. Gwaltney et al's [28] meta-analysis reported the average correlation between paper-based and electronic assessment was 0.90 (95% CI 0.87-0.92; n=32). Our findings suggest that the NFBSI-16 questionnaire achieved a good test-retest reliability, with the total NFBSI-16 score correlation equal to 0.94.

Conclusions

In summary, the web-based version of the NFBSI-16 clearly showed comparable reliability and is thus a promising measure in evaluating studies in patients undergoing treatment for breast cancer and in monitoring individuals. The test-retest reliability supports the value of the web-based version of the NFBSI-16 for clinical studies with relatively moderate sample sizes. Furthermore, the majority of participants in our study preferred it over the paper-based version; we recommend using the web-based version of the NFBSI-16 in clinical studies. Currently, the longitudinal validity of the web-based version of the NFBSI-16 and the validity of several other demographic groups in China are being investigated, giving clinicians more choice when evaluating health-related symptoms and quality of life in patients with breast cancer and other malignant tumors.

Acknowledgments

This study was sponsored by National Natural Science Foundation of China (No. 81773163) and Science and Technology Plan Project of Liaoning Province (No. 2013225585).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screenshot of web-based version National Comprehensive Cancer Network–Functional Assessment of Cancer Therapy–Breast Cancer Symptom Index (NFBSI-16) questionnaire.
[PNG File , 231 KB - medinform_v9i3e18269_app1.png]

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018 Sep 12 [FREE Full text] [doi: 10.3322/caac.21492] [Medline: 30207593]
2. DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, et al. Breast cancer statistics, 2019. CA Cancer J Clin 2019 Nov;69(6):438-451 [FREE Full text] [doi: 10.3322/caac.21583] [Medline: 31577379]
3. Sledge GW. Curing Metastatic Breast Cancer. J Oncol Pract 2016 Jan;12(1):6-10. [doi: 10.1200/JOP.2015.008953] [Medline: 26759458]
4. Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA 2019 Jan 22;321(3):288-300. [doi: 10.1001/jama.2018.19323] [Medline: 30667505]
5. Zhang Q, Zhang L, Yin R, Fu T, Chen H, Shen B. Effectiveness of telephone-based interventions on health-related quality of life and prognostic outcomes in breast cancer patients and survivors-A meta-analysis. Eur J Cancer Care (Engl) 2018 Jan;27(1). [doi: 10.1111/ecc.12632] [Medline: 28090704]
6. Montazeri A. Health-related quality of life in breast cancer patients: a bibliographic review of the literature from 1974 to 2007. J Exp Clin Cancer Res 2008 Aug 29;27:32 [FREE Full text] [doi: 10.1186/1756-9966-27-32] [Medline: 18759983]
7. Yeo W, Kwan W, Teo P, Nip S, Wong E, Hin L, et al. Psychosocial impact of breast cancer surgeries in Chinese patients and their spouses. Psychooncology 2004 Feb;13(2):132-139. [doi: 10.1002/pon.777] [Medline: 14872532]
8. Efficace F, Therasse P, Piccart MJ, Coens C, van Steen K, Welnicka-Jaskiewicz M, et al. Health-related quality of life parameters as prognostic factors in a nonmetastatic breast cancer population: an international multicenter study. J Clin Oncol 2004 Aug 15;22(16):3381-3388. [doi: 10.1200/JCO.2004.02.060] [Medline: 15310784]
9. Lee CK, Hudson M, Simes J, Ribi K, Bernhard J, Coates AS. When do patient reported quality of life indicators become prognostic in breast cancer? Health Qual Life Outcomes 2018 Jan 12;16(1):13 [FREE Full text] [doi: 10.1186/s12955-017-0834-2] [Medline: 29329582]
10. Quinten C, Martinelli F, Coens C, Sprangers MAG, Ringash J, Gotay C, Patient Reported Outcomes and Behavioral Evidence (PROBE) and the European Organization for Research and Treatment of Cancer (EORTC) Clinical Groups. A global analysis of multitrial data investigating quality of life and symptoms as prognostic factors for survival in different tumor sites. Cancer 2014 Jan 15;120(2):302-311 [FREE Full text] [doi: 10.1002/cncr.28382] [Medline: 24127333]
11. van Egdom LS, Oemrawsingh A, Verweij LM, Lingsma HF, Koppert LB, Verhoef C, et al. Implementing Patient-Reported Outcome Measures in Clinical Breast Cancer Care: A Systematic Review. Value Health 2019 Oct;22(10):1197-1226 [FREE Full text] [doi: 10.1016/j.jval.2019.04.1927] [Medline: 31563263]
12. Rock E, Kennedy D, Furness M, Pierce W, Pazdur R, Burke L. Patient-reported outcomes supporting anticancer product approvals. J Clin Oncol 2007 Nov 10;25(32):5094-5099. [doi: 10.1200/JCO.2007.11.3803] [Medline: 17991927]
13. Yost KJ, Yount SE, Eton DT, Silberman C, Broughton-Heyes A, Cella D. Validation of the Functional Assessment of Cancer Therapy-Breast Symptom Index (FBSI). Breast Cancer Res Treat 2005 Apr;90(3):295-298. [doi: 10.1007/s10549-004-5024-3] [Medline: 15830143]
14. Krohe M, Tang DH, Klooster B, Revicki D, Galipeau N, Cella D. Content validity of the National Comprehensive Cancer Network - Functional Assessment of Cancer Therapy - Breast Cancer Symptom Index (NFBSI-16) and Patient-Reported Outcomes Measurement Information System (PROMIS) Physical Function Short Form with advanced breast cancer patients. Health Qual Life Outcomes 2019 May 29;17(1):92 [FREE Full text] [doi: 10.1186/s12955-019-1162-5] [Medline: 31142325]
15. Garcia SF, Rosenbloom SK, Beaumont JL, Merkel D, Von Roenn JH, Rao D, et al. Priority symptoms in advanced breast cancer: development and initial validation of the National Comprehensive Cancer Network-Functional Assessment of Cancer Therapy-Breast Cancer Symptom Index (NFBSI-16). Value Health 2012 Jan;15(1):183-190 [FREE Full text] [doi: 10.1016/j.jval.2011.08.1739] [Medline: 22264987]
16. Ma J, Pazo EE, Zou Z, Jin F. Prevalence of symptomatic dry eye in breast cancer patients undergoing systemic adjuvant treatment: A cross-sectional study. Breast 2020 Oct;53:164-171 [FREE Full text] [doi: 10.1016/j.breast.2020.07.009] [Medline: 32836200]
17. De Castro A, Macías JA. SUSApp: A mobile app for measuring and comparing questionnaire-based usability assessments. : Association for Computing Machinery; 2016 Presented at: ACM International Conference Proceeding Series; 2016; New York. [doi: 10.1145/2998626.2998667]
18. Schleyer TKL, Forrest JL. Methods for the design and administration of web-based surveys. J Am Med Inform Assoc 2000;7(4):416-425 [FREE Full text] [doi: 10.1136/jamia.2000.0070416] [Medline: 10887169]
19. Bateman H, Goh S, Doyle SA. Internet-based surveys of health professionals. Fam Pract 2004 Jun;21(3):329. [doi: 10.1093/fampra/cmh320] [Medline: 15128699]

20. Swoboda WJ, Mühlberger N, Weitkunat R, Schneeweiß S. Internet Surveys by Direct Mailing. Social Science Computer Review 2016 Aug 18;15(3):242-255. [doi: 10.1177/089443939701500302]
21. Coons SJ, Gwaltney CJ, Hays RD, Lundy JJ, Sloan JA, Revicki DA, ISPOR ePRO Task Force. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. Value Health 2009 Jun;12(4):419-429 [FREE Full text] [doi: 10.1111/j.1524-4733.2008.00470.x] [Medline: 19900250]
22. Thumboo J, Wee H, Cheung Y, Machin D, Luo N, Feeny D, et al. Computerized administration of health-related quality of life instruments compared to interviewer administration may reduce sample size requirements in clinical research: a pilot randomized controlled trial among rheumatology patients. Clin Exp Rheumatol 2007;25(4):577-583. [Medline: 17888214]
23. Hays RD, Bode R, Rothrock N, Riley W, Cella D, Gershon R. The impact of next and back buttons on time to complete and measurement reliability in computer-based surveys. Qual Life Res 2010 Oct;19(8):1181-1184 [FREE Full text] [doi: 10.1007/s11136-010-9682-9] [Medline: 20552282]
24. Tiplady B. ePROs: Practical Issues in Pen and Touchscreen Systems. Applied Clinical Trials. URL: https://www.appliedclinicaltrialsonline.com/view/epros-practical-issues-pen-and-touchscreen-systems [accessed 2007-02-03]
25. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
26. Barentsz MW, Wessels H, van Diest PJ, Pijnappel RM, Haaring C, van der Pol CC, et al. Tablet, web-based, or paper questionnaires for measuring anxiety in patients suspected of breast cancer: patients' preferences and quality of collected data. J Med Internet Res 2014 Oct 31;16(10):e239 [FREE Full text] [doi: 10.2196/jmir.3578] [Medline: 25364951]
27. Steele GC, Gill A, Khan AI, Hans PK, Kuluski K, Cott C. The Electronic Patient Reported Outcome Tool: Testing Usability and Feasibility of a Mobile App and Portal to Support Care for Patients With Complex Chronic Disease and Disability in Primary Care Settings. JMIR Mhealth Uhealth 2016 Jun 02;4(2):e58 [FREE Full text] [doi: 10.2196/mhealth.5331] [Medline: 27256035]
28. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. Value Health 2008 Mar;11(2):322-333 [FREE Full text] [doi: 10.1111/j.1524-4733.2007.00231.x] [Medline: 18380645]
29. Cheung YB, Goh C, Thumboo J, Khoo K, Wee J. Quality of life scores differed according to mode of administration in a review of three major oncology questionnaires. J Clin Epidemiol 2006 Feb;59(2):185-191. [doi: 10.1016/j.jclinepi.2005.06.011] [Medline: 16426954]
30. Triberti S, Savioni L, Sebri V, Pravettoni G. eHealth for improving quality of life in breast cancer patients: A systematic review. Cancer Treat Rev 2019 Mar;74:1-14. [doi: 10.1016/j.ctrv.2019.01.003] [Medline: 30658289]
31. Matthies LM, Taran F, Keilmann L, Schneeweiss A, Simoes E, Hartkopf AD, et al. An Electronic Patient-Reported Outcome Tool for the FACT-B (Functional Assessment of Cancer Therapy-Breast) Questionnaire for Measuring the Health-Related Quality of Life in Patients With Breast Cancer: Reliability Study. J Med Internet Res 2019 Jan 22;21(1):e10004 [FREE Full text] [doi: 10.2196/10004] [Medline: 30668517]
32. Wallwiener M, Matthies L, Simoes E, Keilmann L, Hartkopf AD, Sokolov AN, et al. Reliability of an e-PRO Tool of EORTC QLQ-C30 for Measurement of Health-Related Quality of Life in Patients With Breast Cancer: Prospective Randomized Trial. J Med Internet Res 2017 Sep 14;19(9):e322 [FREE Full text] [doi: 10.2196/jmir.8210] [Medline: 28912116]
33. Hartkopf AD, Graf J, Simoes E, Keilmann L, Sickenberger N, Gass P, et al. Electronic-Based Patient-Reported Outcomes: Willingness, Needs, and Barriers in Adjuvant and Metastatic Breast Cancer Patients. JMIR Cancer 2017 Aug 07;3(2):e11 [FREE Full text] [doi: 10.2196/cancer.6996] [Medline: 28784595]
34. Sun Z, Zhu L, Liang M, Xu T, Lang J. The usability of a WeChat-based electronic questionnaire for collecting participant-reported data in female pelvic floor disorders: a comparison with the traditional paper-administered format. Menopause 2016 Aug;23(8):856-862. [doi: 10.1097/GME.0000000000000690] [Medline: 27326820]
35. Wen Z, Geng X, Ye Y. Does the Use of WeChat Lead to Subjective Well-Being?: The Effect of Use Intensity and Motivations. Cyberpsychol Behav Soc Netw 2016 Oct;19(10):587-592. [doi: 10.1089/cyber.2016.0154] [Medline: 27732075]
36. Li W, Han LQ, Guo YJ, Sun J. Using WeChat official accounts to improve malaria health literacy among Chinese expatriates in Niger: an intervention study. Malar J 2016 Nov 24;15(1):567 [FREE Full text] [doi: 10.1186/s12936-016-1621-y] [Medline: 27881122]
37. Jenkins CR, Dillman DA. Towards a Theory of Self‐Administered Questionnaire Design. In: Survey Measurement and Process Quality. New Jersey: Wiley Series in Probability and Statistics; 1997:165-196.
38. Lin L, Torbeck LD. Coefficient of accuracy and concordance correlation coefficient: new statistics for methods comparison. PDA J Pharm Sci Technol 1998;52(2):55-59. [Medline: 9610168]
39. Schuck P. Assessing reproducibility for interval data in health-related quality of life questionnaires: which coefficient should be used? Qual Life Res 2004 Apr;13(3):571-586. [doi: 10.1023/B:QURE.0000021318.92272.2a] [Medline: 15130022]
40. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989 Mar;45(1):255-268. [Medline: 2720055]

41. U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes 2006 Oct 11;4:79 [FREE Full text] [doi: 10.1186/1477-7525-4-79] [Medline: 17034633]

Abbreviations

ANOVA: analysis of variance
CCC: concordance correlation coefficient
DRS-E: Disease-Related Symptoms Subscale–Emotional
DRS-P: Disease-Related Symptoms Subscale–Physical
FACIT: Functional Assessment of Chronic Illness Therapy
FWB: Function and Well-Being Subscale
ICC: intraclass correlation coefficient
NFBSI-16: National Comprehensive Cancer Network–Functional Assessment of Cancer Therapy–Breast Cancer Symptom Index
PROMs: patient-reported outcome measures
TSE: Treatment Side Effects Subscale

Edited by C Lovis; submitted 16.02.20; peer-reviewed by R Fox, PC Rassu, L Guo; comments to author 22.07.20; revised version received 15.09.20; accepted 31.01.21; published 02.03.21.

Please cite as:
Ma J, Zou Z, Pazo EE, Moutari S, Liu Y, Jin F
Comparative Analysis of Paper-Based and Web-Based Versions of the National Comprehensive Cancer Network-Functional Assessment of Cancer Therapy-Breast Cancer Symptom Index (NFBSI-16) Questionnaire in Breast Cancer Patients: Randomized Crossover Study
JMIR Med Inform 2021;9(3):e18269
URL: https://medinform.jmir.org/2021/3/e18269
doi:10.2196/18269
PMID:33650978

Original Paper

Antibiotic Prescription Rates After eVisits Versus Office Visits in Primary Care: Observational Study

Artin Entezarjou[1], MD; Susanna Calling[1], MD, PhD; Tapomita Bhattacharyya[1], MD; Veronica Milos Nymberg[1], MD, PhD; Lina Vigren[2], MD, PhD; Ashkan Labaf[3], MD, PhD; Ulf Jakobsson[1], PhD; Patrik Midlöv[1], MD, PhD

[1]Center for Primary Health Care Research, Department of Clinical Sciences in Malmö/Family Medicine, Lund University, Malmö, Sweden

[2]Capio Go AB, Gothenburg, Sweden

[3]Department of Clinical Sciences in Lund, Lund University, Lund, Sweden

Corresponding Author:
Artin Entezarjou, MD
Center for Primary Health Care Research
Department of Clinical Sciences in Malmö/Family Medicine
Lund University
Box 50332
Malmö
Sweden
Phone: 46 40 391400
Email: artin.entezarjou@med.lu.se

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2021/11/e34529

Abstract

Background: Direct-to-consumer telemedicine is an increasingly used modality to access primary care. Previous research on assessment using synchronous virtual visits showed mixed results regarding antibiotic prescription rates, and research on assessment using asynchronous chat-based eVisits is lacking.

Objective: The goal of the research was to investigate if eVisit management of sore throat, other respiratory symptoms, or dysuria leads to higher rates of antibiotic prescription compared with usual management using physical office visits.

Methods: Data from 3847 eVisits and 759 office visits for sore throat, dysuria, or respiratory symptoms were acquired from a large private health care provider in Sweden. Data were analyzed to compare antibiotic prescription rates within 3 days, antibiotic type, and diagnoses made. For a subset of sore throat visits (n=160 eVisits, n=125 office visits), Centor criteria data were manually extracted and validated.

Results: Antibiotic prescription rates were lower following eVisits compared with office visits for sore throat (169/798, 21.2%, vs 124/312, 39.7%; *P*<.001) and respiratory symptoms (27/1724, 1.6%, vs 50/251, 19.9%; *P*<.001), while no significant differences were noted comparing eVisits to office visits for dysuria (1016/1325, 76.7%, vs 143/196, 73.0%; *P*=.25). Guideline-recommended antibiotics were prescribed similarly following sore throat eVisits and office visits (163/169, 96.4%, vs 117/124, 94.4%; *P*=.39). eVisits for respiratory symptoms and dysuria were more often prescribed guideline-recommended antibiotics (26/27, 96.3%, vs 37/50, 74.0%; *P*=.02 and 1009/1016, 99.3%, vs 135/143, 94.4%; *P*<.001, respectively). Odds ratios of antibiotic prescription following office visits compared with eVisits after adjusting for age and differences in set diagnoses were 2.94 (95% CI 1.99-4.33), 11.57 (95% CI 5.50-24.32), 1.01 (95% CI 0.66-1.53), for sore throat, respiratory symptoms, and dysuria, respectively.

Conclusions: The use of asynchronous eVisits for the management of sore throat, dysuria, and respiratory symptoms is not associated with an inherent overprescription of antibiotics compared with office visits.

Trial Registration: ClinicalTrials.gov NCT03474887; https://clinicaltrials.gov/ct2/show/NCT03474887

(JMIR Med Inform 2021;9(3):e25473) doi:10.2196/25473

KEYWORDS

telemedicine; antibiotics; streptococcal tonsillitis; cystitis; respiratory tract infection; virtual visit; virtual; eVisit

Introduction

Direct-to-consumer telemedicine is an increasingly used modality to access primary care in Sweden [1]. Such visits can take the form of asynchronous chat-based visits (eVisits) or synchronous video-based visits (virtual visits). While telemedicine has the potential to address many challenges facing primary care [2] and provide an appropriate alternative for minimizing risk of COVID-19 during the current pandemic [3], concerns have been raised regarding overprescription of antibiotics [4] and potential ramifications to increasing widespread antibiotic resistance. Antibiotic resistance is already predicted to cause more deaths than cancer by the year 2050 [5].

Most research has been conducted on data derived from synchronous virtual visits in American health care settings, where antibiotic prescription is historically higher [6], possibly due to a more market-controlled health care system with incentives for high patient satisfaction [7]. Consequently, there have been mixed results regarding antibiotic prescribing following virtual visits in various contexts [4,8-18], with most studies focusing on urinary tract infections (UTIs) and upper respiratory infections. For example, depending on the health care provider, virtual visits for sinusitis have been associated with both higher [14] and lower [10,13] prescriptions rates compared with office visits. Comparisons to urgent care settings often demonstrate lower prescription rates for virtual visits [8,9].

In Sweden, primary care accounts for 61% of medical antibiotic consumption [19], with 30% of consultations concerning infections [20], most commonly upper respiratory tract infections, tonsillitis, and UTIs [20,21]. Guideline adherence in management of these conditions is poor [22-24]. A study on virtual visits reported that 50% to 60% of cases diagnosed with viral pharyngitis had rapid streptococcal antigen testing (RST) performed or no antibiotics prescribed, while 90% of those diagnosed with streptococcal pharyngitis had RST performed or antibiotics prescribed [25]. However, no comparison was made with office visits. There is thus a paucity of literature concerning eVisit investigations, particularly in terms of head-to-head comparisons to office visits, as highlighted by systematic reviews [26,27].

The aim of this study was to investigate if management of sore throat using a specific eVisit platform led to significantly higher rates of antibiotic prescription compared with usual management using office visits. Secondary outcomes include prescription rate following dysuria and other respiratory symptoms, type of antibiotics prescribed, documentation of Centor criteria (used to identify the likelihood of a bacterial infection in adult patients complaining of a sore throat), and set diagnoses.

Methods

eVisit Platform

This retrospective cohort study specifically evaluates an eVisit platform (referred to as "the platform" in this paper) used by a major private health care provider. The platform combines automated patient interviewing software with an asynchronous 2-way text-based chat between patient and health care provider. Patients access the platform using their smartphone, tablet, or computer device and choose their chief complaint from a prespecified symptom list. A digital patient history is then taken, allowing the patient to formulate ideas, concerns, and expectations [28] in free-text with the addition of symptom-specific multiple-choice questions based on algorithms. Questions may address UTI symptoms and patient-assessed Centor criteria [29], such as “Do you have any of the following symptoms together with your sore throat?” with choices of “fever,” “swollen lymph nodes on the neck,” “severe pain when swallowing,” “cough,” “white exudates on your tonsils or in the back of your throat” (image not mandatory but recommended). If a patient reports fever, the question “Have you measured your body temperature?” may be asked with choices “no” or “yes” with an option to specify the highest value in degrees Celsius. Photos can be attached when relevant; this is recommended for the management of sore throat. Answers are summarized and presented to a physician for review, and further doctor-patient communication occurs through a text-based conversation, similar to text messaging, with patients and providers messaging each other at their convenience. Physicians can prescribe medications, order laboratory samples, provide patient information, or stay available for up to 72 hours for conservative management. If deemed necessary, the physician can schedule an office visit at a primary health care center of the same health care provider. At the time of the study, the platform used no machine learning technology.

Setting and Population

As the private health care provider offers both office visits and eVisits using the platform since July 31, 2017, data could be acquired for both visit types. A total of 16 primary health care centers in the county provided office visit data, while national eVisit data was acquired from the online platform. Inclusion criteria were physician visits with a chief complaint of sore throat, cough, cold/flu symptoms, or dysuria as specified by free-form text in the electronic medical record (EMR) as identified by data extraction software (Multimedia Appendix 1). We also included visits with a recorded diagnosis code J030 (streptococcal tonsillitis), J069 (acute upper respiratory infection), or N300 (cystitis). Visits were included if they occurred between March 30, 2016, and March 29, 2017 (office visits only) or March 30, 2018, and March 29, 2019 (eVisits and office visits). Exclusion criteria were patients aged younger than 18 years, male patients with dysuria, and identifiable visits for similar chief complaints in the past 21 days.

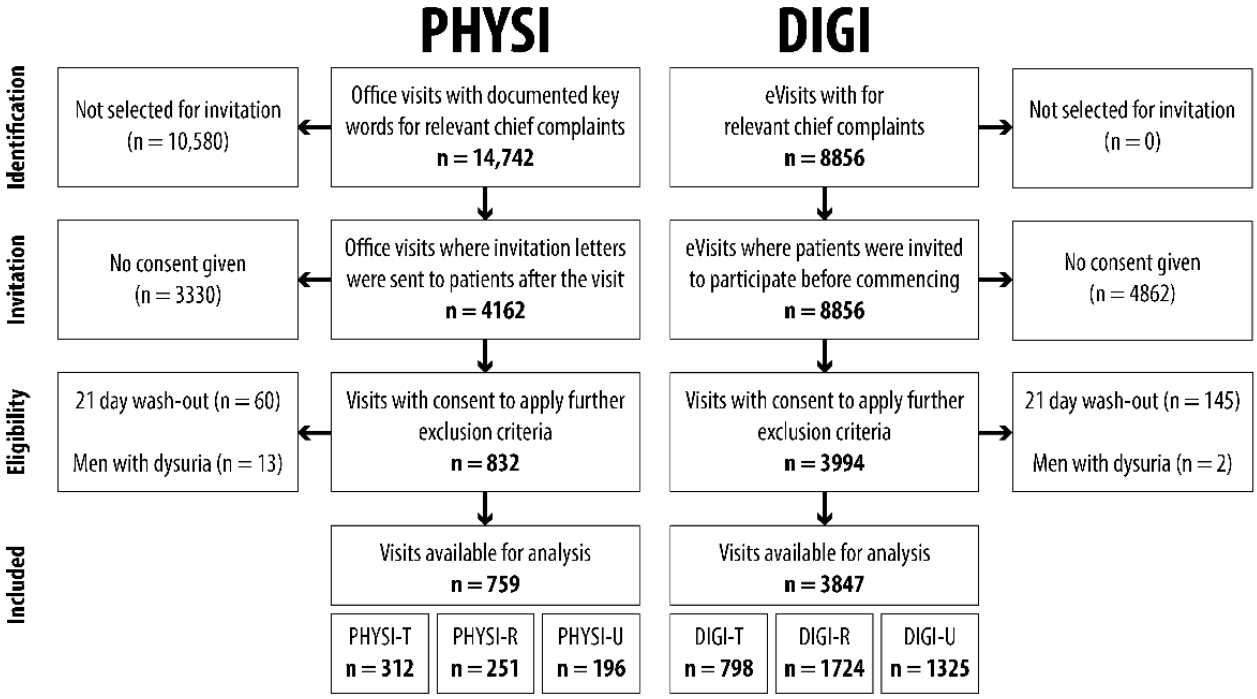
Power Calculation and Recruitment

Previous data from Sweden suggested an antibiotic prescription rate of 59% for patients with sore throat–related diagnoses [20]. Using a binary outcome power calculation with a noninferiority limit of 10%, an alpha level of .05, for 80% power, we estimated needing 300 sore throat visits per group.

Digital consent was acquired from eVisit patients at the beginning of the visits and recorded in the EMR. Written consent was acquired from office visit patients, with sore throat patients receiving letters including 2 reminders if no reply was received. Recruitment was completed after consent was acquired from at least 300 sore throat patients in each group. After recruitment, remaining exclusion criteria were applied before analysis commenced (Figure 1).

The health care provider identified 14,742 potential office visits eligible for participation. Letters were then sent to a random selection of 2000 patients with suspected sore throat, 1000 patients with suspected dysuria, and 1000 patients with suspected symptoms of cough, common cold, and influenza, comprising 4162 visits. For office visits with a chief complaint of sore throat (PHYSI-T), 87 patients were recruited after 1 month. An additional 117 patients were recruited after a second letter was sent 2 months later, and an additional 96 patients were recruited 1 month after the third recruitment letter was sent out. A total of 8856 relevant eVisits were identified, from which patients were also invited to participate. In total, we recruited patients from 832 office visits and 3994 eVisits. After exclusion of dysuria visits with male patients and visits within the 21-day washout period, 759 office visits and 3847 eVisits remained for analysis (Figure 1). Office visits were in 99.1% of cases identified via keywords in the free-form text the EMR, while 0.1% (2 sore throat visits, 22 respiratory visits, and 18 dysuria visits) were identified through set diagnoses.

Figure 1. Flowchart of patient recruitment. PHYSI: primary care office visits; DIGI: eVisits; PHYSI-T: office visits with a chief complaint of sore throat; PHYSI-R: office visits with a chief complaint of common cold/influenza or cough; PHYSI-U: office visits with a chief complaint of dysuria; DIGI-T: eVisits with a chief complaint of sore throat; DIGI-R: eVisits with a chief complaint of common cold/influenza or cough; DIGI-U: eVisits with a chief complaint of dysuria.



Diagnostic Criteria and Guideline Adherence

Swedish national guidelines recommend identifying at least 3 Centor criteria (tonsillar exudates, swollen tender anterior cervical nodes, lack of cough, and presence of fever over 38.5° Celsius) prior to ordering an RST [29]. Guidelines recommend that RST should only be performed if the advantages of antibiotic treatment are deemed to outweigh the disadvantages for the individual patient and subsequently recommend penicillin V as first-line treatment [30]. All cases of ordered RST in the presence of Centor criteria were assumed to be due to primary health care physicians deeming the advantages of antibiotic therapy outweighing the disadvantages. In the office visit group, Centor criteria are documented after a physical examination by a physician. For the eVisit group, patients self-assess and report Centor criteria in the automated patient interviewing software [25]. Answers are evaluated by a physician who then chooses which criteria to document in a specified template by, for example, being required to check a box specifying that temperature was above 38.5° Celsius. The physician may choose to document Centor criteria differently from how patients report the criteria depending on what information is acquired during the 2-way patient-provider chat.

Data Collection

Baseline variables included chief complaint, visiting date, age, and gender. The primary outcome was antibiotic prescription within 3 days following sore throat as the chief complaint, which is similar to previous studies [11,31,32]. Secondary outcomes included antibiotic prescription within 3 days of visits for

dysuria and cough/common cold/influenza, type of antibiotic prescribed, documentation of Centor criteria, laboratory tests ordered within 3 days (c-reactive protein [CRP] and RST). Guideline adherence for sore throat patients was also assessed in terms of following indications for antibiotic prescription.

Data extraction software was used to automatically extract data [33,34] with subsets manually validated by reading all free-form text in the EMR and evaluating deviations from automatically extracted data. Variables that were manually evaluated included chief complaint (n=783), Centor criteria (n=285), CRP ordered (n=294), RST ordered (n=284), antibiotic prescription (n=782), and antibiotic type (n=183).

As automatic extraction of free-form text was not possible, Centor criteria for PHYSI-T were manually extracted from a randomly selected subset of the cohort (n=125) while automatically extracted Centor criteria were manually validated for a subset of DIGI-T visits (n=160), resulting in a total of 285 visits with manually validated Centor criteria. Protocols were used for all interpretation of free-form text (Multimedia Appendix 1). For example, free-text documentation stating "fever" was deemed a Centor criterion since only a minority of cases specified temperature in this context.

Statistical Analyses

Analysis was conducted using SPSS Statistics version 26 (IBM Corporation). A 21-day washout period was applied, excluding past eVisits or office visits for similar chief complaints, similar to previous methods [4]. For this washout, sore throat, cough, and common cold or influenza were all deemed similar chief complaints as they are all respiratory symptoms.

Visits for cough and common cold or influenza, each a separate chief complaint for eVisits, were grouped together for analysis as these chief complaints often result in similar diagnoses, resulting in a total of 6 groups for analysis: sore throat office visit (PHYSI-T) and eVisit (DIGI-T), cough/common cold/influenza office visit (PHYSI-R) and eVisit (DIGI-R), and dysuria office visit (PHYSI-U) and eVisit (DIGI-U). Variables on type of antibiotics prescribed were recategorized to separate antibiotics not commonly recommended by guidelines (Multimedia Appendix 2). For analyses of guideline adherence, manually collected Centor criterion data were dichotomized so that undocumented symptoms were assumed to be absent.

The first diagnosis recorded at each visit was recategorized as UTI, viral upper and lower respiratory tract infection, tonsillitis, and 3 common diagnoses seen as more severe conditions following each of our chosen chief complaints: pneumonia, peritonsillar abscess, and pyelonephritis. Symptom-based codes and nondiagnostic codes were grouped as nonspecific or symptom-based diagnosis and remaining diagnoses were grouped as other (Multimedia Appendix 3). Continuous data were presented with mean and standard deviation and analyzed with Student *t* test, while categorical data were presented with percentage and analyzed with chi-square test.

We hypothesized that there would be no clinically relevant difference in antibiotic prescribing. Hypothesis testing was conducted by comparing office visits to eVisits for each chief complaint. As age and set diagnoses are potential confounding factors for the tendency to prescribe antibiotics, multiple binary logistic regressions were conducted for each chief complaint with antibiotic prescription as the dependent variable and visit type as the independent variable in an enter regression model. The models were then adjusted for age and diagnoses of tonsillitis, viral upper and lower respiratory tract infection, pneumonia, and other diagnoses. eVisits were used as the reference group.

No data were missing for the primary outcome analyses. For secondary outcomes, visits with missing data were compared with visits with valid data for patient age, prescription of antibiotics. and antibiotic choice to test whether data was missing at random. Visits with data missing at random were excluded from the analyses.

Exploratory analyses were conducted for sore throat patients from one county (n=289 for DIGI-T and n=312 for PHYSI-T) where data on Centor criteria and related variables were available for random subsets of the data. Two measures of guideline adherence for sore throat management were explored:

- Proportion of RST performed on properly documented indications (ie, 3 or more documented Centor criteria)
- Proportion of visits diagnosed with tonsillitis that were prescribed antibiotics with a positive RST performed on properly documented indications

Ethics and Registration

The study was approved by the Swedish Ethical Review Authority (reference number: 2019-00463). Permission to use regional medical record data was also granted (reference number: 062-18). The study was registered at ClinicalTrials.gov [NCT03474887] and reported using a Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist.

Data Sharing Statement

Data are available to the Department of Clinical Sciences in Malmö at Lund University and can be accessed for a prespecified purpose after approval upon reasonable request.

Results

Manual Validation of Data

Manual validation showed high accuracy of extracted data, with 98.7% (773/783) accuracy for antibiotic prescription within 3 days and chief complaint for office visits correctly classified in 98.5% (133/135) for PHYSI-T but less often so for PHYSI-R (212/234, 90.6%) and PHYSI-U (95/103, 92.2%). For PHYSI-U patients, many cases of misclassified patients had lower abdominal pain rather than dysuria.

Baseline Demographics

For all chief complaints, baseline demographics revealed a significantly higher patient age among office visits compared with eVisits. For both sore throat and respiratory symptoms, around one-third (343/1110, 30.9%, and 721/1975, 36.5%, for sore throat and respiratory symptoms, respectively) of the visits involved male patients, with slightly more men in DIGI-T compared with PHYSI-T (Table 1).

Table 1. Baseline demographics.

Chief complaint	Age in years, mean (SD)	*P* value for difference	Sex, male, n (%)	*P* value for difference
Sore throat (n=1110)	—[a]	<.001	—	.03
DIGI-T[b] (n=798)	35.1 (11.5)	—	262 (32.8)	—
PHYSI-T[c] (n=312)	44.5 (17.5)	—	81 (26.0)	—
Respiratory (n=1975)	—	<.001	—	.28
DIGI-R[d] (n=1724)	42.8 (14.5)	—	637 (36.9)	—
PHYSI-R[e] (n=251)	60.0 (16.2)	—	84 (33.5)	—
Dysuria (n=1521)	—	<.001	—	—
DIGI-U[f] (n=1325)	42.1 (15.4)	—	0 (0.0)	—
PHYSI-U[g] (n=196)	60.0 (18.9)	—	0 (0.0)	—

[a]Not applicable.

[b]DIGI-T: eVisits with a chief complaint of sore throat.

[c]PHYSI-T: Office visits with a chief complaint of sore throat.

[d]DIGI-R: eVisits with a chief complaint of common cold/influenza or cough.

[e]PHYSI-R: Office visits with a chief complaint of common cold/influenza or cough.

[f]DIGI-U: eVisits with a chief complaint of dysuria.

[g]PHYSI-U: Office visits with a chief complaint of dysuria.

Diagnoses

Based on the first diagnosis recorded by the physician, a total of 185 different diagnosis codes were recorded across the entire cohort, with 107 different diagnosis codes for office visits and 98 different diagnosis codes for eVisits.

Nonspecific or symptom-based diagnoses were recorded among 25.3% (973/3847) of eVisits compared with 14.2% (108/759) of office visits, while other diagnoses were recorded for 1.8% (70/3847) of eVisits compared with 19.1% (145/759) of office visits.

Tonsillitis was recorded among 25.8% (206/798) of DIGI-T compared with 33.3% (104/312) of PHYSI-T. Viral upper and lower respiratory diagnoses were recorded among 61.3% (1057/1724) of DIGI-R compared with 48.6% (122/251) of PHYSI-R.

A total of 0.7% (19/2522) recorded diagnoses were for pneumonia across DIGI-T and DIGI-R compared with 2.3% (13/563) across PHYSI-T and PHYSI-R. Peritonsillar abscess was recorded in 0.8% (6/798) of DIGI-T compared with 0.6% (2/312) of PHYSI-T. There was one recorded diagnosis of pyelonephritis among PHYSI-U and none among DIGI-U.

Antibiotic Prescription

Compared with eVisits, antibiotic prescription within 3 days of the visit was significantly higher for office visits for sore throat and respiratory symptoms. No significant difference in prescription rate was observed for dysuria visits (Table 2).

For respiratory symptoms and dysuria, office visits more often led to the prescription of antibiotics outside of guideline recommendations for tonsillitis and pneumonia, respectively (Table 2).

Odds ratio of antibiotic prescription as the dependent variable following a PHYSI-T visit compared with DIGI-T was 2.46 (95% CI 1.86-3.26; *P*<.001). Adjustment for age and differences in recorded diagnoses had a marginal impact on odds ratios (Table 3).

Table 2. Antibiotic-related outcomes. No data were missing among presented variables. See Multimedia Appendix 2 for guideline-recommended antibiotics.

Chief complaint	Antibiotic prescription within 3 days of visit, n (%)	*P* value for difference	Guideline-recommended antibiotics, n (%)	*P* value for difference
Sore throat (n=1110)	__[a]	<.001	—	.39
DIGI-T[b] (n=798)	169 (21.2)	—	163 (96.4)	—
PHYSI-T[c] (n=312)	124 (39.7)	—	117 (94.4)	—
Respiratory (n=1975)	—	<.001	—	.02
DIGI-R[d] (n=1724)	27 (1.6)	—	26 (96.3)	—
PHYSI-R[e] (n=251)	50 (19.9)	—	37 (74.0)	—
Dysuria (n=1521)	—	.25	—	<.001
DIGI-U[f] (n=1325)	1016 (76.7)	—	1009 (99.3)	—
PHYSI-U[g] (n=196)	143 (73.0)	—	135 (94.4)	—

[a]Not applicable.

[b]DIGI-T: eVisits with a chief complaint of sore throat.

[c]PHYSI-T: Office visits with a chief complaint of sore throat.

[d]DIGI-R: eVisits with a chief complaint of common cold/influenza or cough.

[e]PHYSI-R: Office visits with a chief complaint of common cold/influenza or cough.

[f]DIGI-U: eVisits with a chief complaint of dysuria.

[g]PHYSI-U: Office visits with a chief complaint of dysuria.

Table 3. Regression models for antibiotic prescription for office visits compared with eVisits.

Chief complaint	Antibiotic prescription within 3 days of office visits vs eVisits, UOR[a] (95% CI)	*P* value	Antibiotic prescription within 3 days of office visits vs eVisits, AOR[b,c] (95% CI)	*P* value
Sore throat (n=1110)	2.46 (1.85-3.26)	<.001	2.94 (1.99-4.33)	<.001
Respiratory (n=1975)	15.63 (9.58-25.53)	<.001	11.57 (5.50-24.32)	<.001
Dysuria (n=1521)	0.82 (0.58-1.15)	.25	1.01 (0.66-1.53)	.98

[a]UOR: unadjusted odds ratio.

[b]AOR: adjusted odds ratio.

[c]Each regression model was adjusted for age and diagnoses, tonsillitis, viral upper and lower respiratory tract infection, pneumonia, and other.

Antibiotic Choice

Antibiotic choice was similar for DIGI-T versus PHYSI-T as well as DIGI-U versus PHYSI-U (Figures 2 and 3, respectively). Antibiotic prescriptions following DIGI-R most often led to prescriptions of penicillin V, while PHYSI-R most often led to prescriptions of doxycycline (Figure 4). Penicillin V accounted for 89.3% (151/169) of all prescribed antibiotics among DIGI-T and 77.4% (96/124) of all prescribed antibiotics in PHYSI-T.

Among the 13 sore throat visits included in “Other” (6 DIGI-T, 7 PHYSI-T visits), there was one DIGI-T and one PHYSI-T visit each with UTI diagnoses receiving pivmecillinam, and one PHYSI-T visit with a diagnosis of acute bronchitis receiving doxycycline. Remaining visits had only sore throat–related diagnoses and were followed by prescriptions of doxycycline, erythromycin, and amoxicillin with and without clavulanic acid.

Figure 2. Prescription rates for various antibiotics following chief complaint of sore throat.

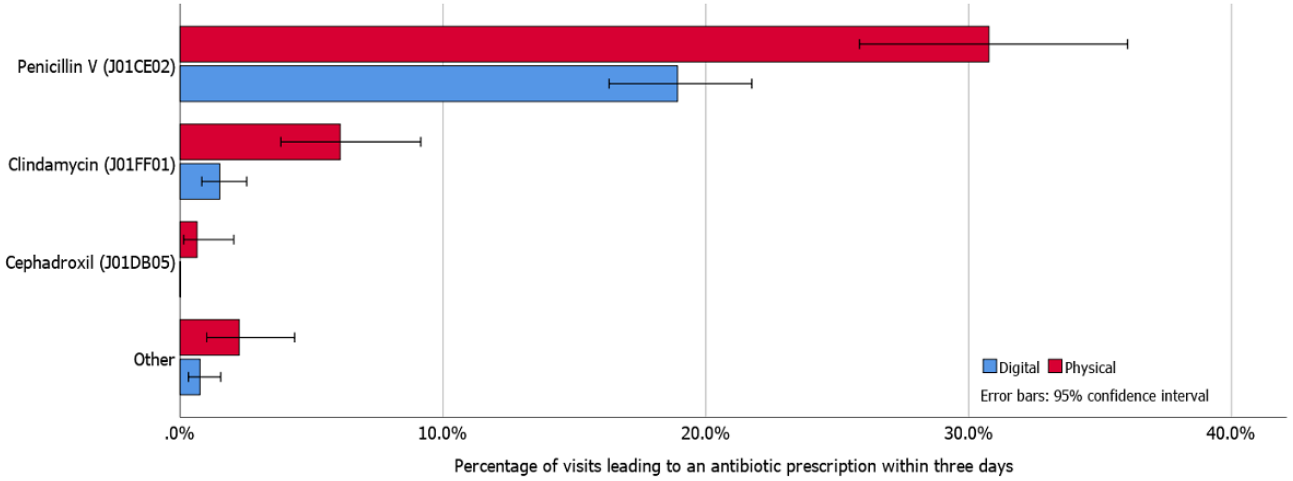


Figure 3. Prescription rates for various antibiotics following chief complaint of dysuria.

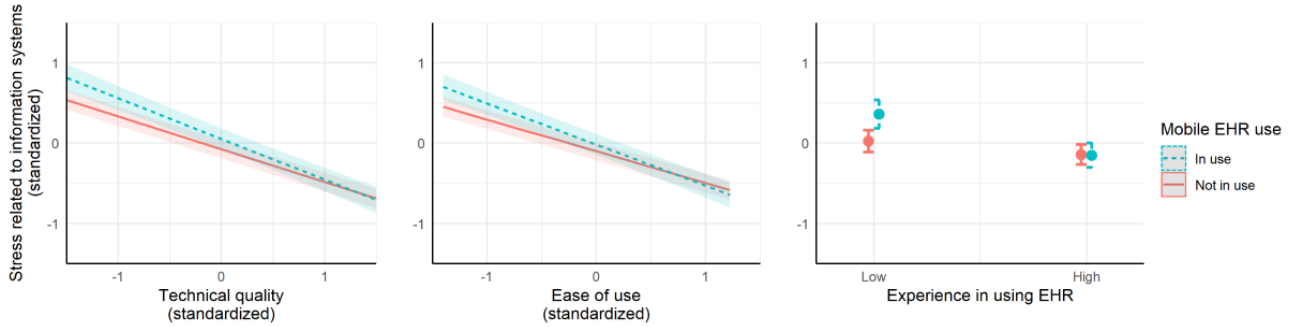
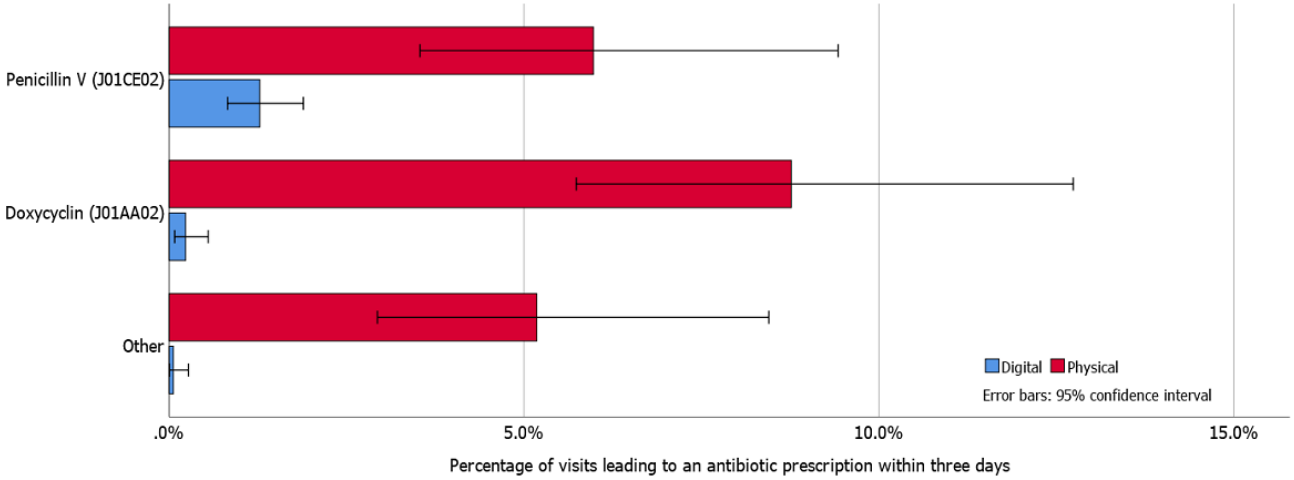


Figure 4. Prescription rates for various antibiotics following chief complaint of respiratory symptoms.



Among the 15 dysuria visits included in "Other" (7 DIGI-U, 8 PHYSI-U visits), 7 PHYSI-U visits led to prescriptions of trimethoprim sulfamethoxazole, methenamine, or ciprofloxacin without a relevant diagnosis to support the prescription given current guidelines, while one PHYSI-U visit led to a diagnosis with pyelonephritis and prescription of ciprofloxacin accordingly. A total of 5 DIGI-U visits had non-specified UTI diagnoses; 3 of these patients were prescribed ciprofloxacin, 1 trimethoprim sulfamethoxazole, and 1 lymecycline. The remaining 2 DIGI-U patients were diagnosed with acute cystitis and prescribed ciprofloxacin.

Among the 14 respiratory visits included in "Other" (1 DIGI-R, 13 PHYSI-R visits), 4 PHYSI-R patients were prescribed amoxicillin, erythromycin, or cefadroxil without a diagnosis supported by guidelines, and 2 PHYSI-R visit patients were prescribed amoxicillin with the diagnosis of pneumonia. A total of 3 PHYSI-R visit patients were diagnosed with concurrent UTIs, 2 of whom were prescribed pivmecillinam and 1 trimethoprim sulfamethoxazole. The remaining PHYSI-R visit

patients were diagnosed with chronic obstructive pulmonary disease or acute exacerbation and were prescribed amoxicillin.

Documentation of Centor Criteria

All 4 Centor criteria were documented for 100% (798/798) of DIGI-T visits and 28% (35/125) of PHYSI-T visits. Documentation did not differ among PHYSI-T visits prescribed antibiotics versus cases not prescribed antibiotics (13/45, 28.9%, versus 22/80, 27.5%, complete documentation, respectively). Specifically, presence or absence of tonsillar exudates, fever, lymphadenopathy, and cough were not documented in 4.8% (6/125), 21.6% (27/125), 26.4% (33/125), and 57.6% (72/125) of PHYSI-T visits, respectively.

Among the subset of sore throat patients from a specific county, there was no significant difference in documented fever between DIGI-T and PHYSI-*t* (116/289, 40.1%, vs 46/125, 36.8%; *P*=.52). PHYSI-T more often had absence of cough (96/125, 76.8%, vs 151/289, 52.2%; *P*<.001). DIGI-T had significantly more documented swollen tender anterior cervical nodes (182/289, 63.0%, vs 39/125, 31.2%; *P*<.001) and tonsillar exudates (136/289, 47.1%, vs 37/125, 29.6%; *P*=.001; Multimedia Appendix 4).

Among manually reviewed cases with documented tonsillar exudates among DIGI-T, 86.6% (116/134) had a photo attached of varying quality in terms of visualizing tonsillar exudates.

Guideline Adherence for Sore Throat

Exploratory analyses of sore throat visits with Centor criteria data (Multimedia Appendix 5) showed that RST testing was more often performed on properly documented indications in terms of Centor criteria among DIGI-T compared with PHYSI-*t* (105/132, 79.5%, vs 23/70, 32.9%; *P*<.001).

Among visits that were diagnosed with tonsillitis and prescribed antibiotics, there were more cases of positive RSTs performed on properly documented indications among DIGI-T compared with PHYSI-*t* (42/43, 97.7%, vs 8/20, 40.0%; *P*<.001).

Discussion

Principal Findings

Rates of antibiotic prescription following eVisits for sore throat, cough, common cold, and influenza were significantly lower than for office visits, while no differences in prescription rates were noted for dysuria. This difference persisted after adjusting for age and set diagnoses.

Limitations

Results should be interpreted with consideration for several limitations. First, as the groups were not randomized, we were unable to establish causality between visit type and antibiotic prescription rate. However, randomization in this context was not feasible as risk of spillover was high with patients free to seek other forms of care.

We cannot exclude that the lower prescription rate among eVisits reflects a self-selected group with different symptom severity, comorbidity frequency, patient expectations, and time constraints compared with those seeking office care. Differences between physicians working in the digital platform versus in the office setting may be another factor influencing differences in prescription rates.

Differences in recruitment strategy may have impacted the results. During eVisits, patients self-selected their chief complaint, which was then documented and used for recruitment, while office visit physicians chose which symptom to document as the chief complaint. eVisit physicians were not blinded for participation to the study, which may have influenced the outcome.

Regarding sore throat, the results of this study may not apply to countries preferentially using other scoring systems such as the McIsaac score to determine whether to perform an RST [35].

Finally, while we used a 21-day washout period, we cannot exclude that some visits may have been preceded by a visit from another health care provider within the washout period. Across the entire cohort, there were 12 patients who had both an eVisit and an office visit, with the eVisit preceding the office visit in 8 cases. However, visits were always separated by at least 21 days, making conversions clinically unlikely and warranting a novel assessment regarding indications for antibiotic prescription. Our sample size was relatively small but adequate to address the research question.

Strengths

Despite the above, this study has several strengths. As far as the authors know, this is the first study specifically comparing antibiotic prescription following asynchronous eVisits to office visits outside of the American health care setting. The dataset comes from one of the few health care providers of both eVisits and office visits, thus making the groups more comparable. Using chief complaint as opposed to diagnosis as inclusion criteria means prescription rates may better reflect clinical practice as many clinicians tend to choose diagnoses based on their choice to prescribe antibiotics, regardless of guideline adherence. Using data extraction software ensured reliability of data, and manually reviewing subsets of the data added validity regarding physician assessment and documentation. Findings were robust through logistic regression and several subgroup analyses.

Interpretation

Beyond potential unidentified confounding factors, the lower antibiotic prescription rate in DIGI-T may reflect the health care providers' use of a structured documentation platform requiring physicians to actively mark each Centor criterion prior to ordering an RST. It has previously been hypothesized that availability of guidelines may be the driving factor behind improved guideline adherence in virtual visits [8], and decision support systems have previously been shown to improve guideline adherence [16,36].

One must also consider the risk of misdiagnosis with eVisits. There is a risk that the system would lead the physician into a logical conclusion and apparently guideline-coherent decision, increasing the risk of cognitive biases such as confirmation bias, which may not have occurred face-to-face in an office setting.

eVisits may also facilitate physicians to better manage emotionally demanding patients [37], possibly reducing the risk of prescribing antibiotics without proper indications. In addition, eVisits provide a convenient way for physicians to use watchful waiting prior to antibiotic prescription as patients easily can access the chat within 72 hours of a consultation.

DIGI-T patients are required to visit their nearest primary health care center to take the RST prior to receiving antibiotics, which may create an additional barrier to antibiotic prescription not present in PHYSI-T. These barriers are absent for antibiotic prescription following UTIs, which may explain the similar rates between DIGI-U and PHYSI-U.

As previously mentioned, eVisits involve physician interpretation of patient reported Centor criteria prior to documentation, while office visits involve interpretation of Centor criteria through physical examination prior to documentation. For example, cough may be more correctly reported following eVisits as it is reported much more categorically than when asked in an office setting and interpreted by the physician with a working diagnosis. Conversely, lymphadenopathy may be overreported among eVisits due to self-palpating of cervical myalgia because of a sore throat. The use of patient-reported Centor criteria remains to be validated, prompting some organizations to dissuade management of sore throat patients using eVisits [38]. As future studies are required to validate specific criteria for eVisit diagnosis of streptococcal tonsillitis, this study's objective was to evaluate adherence to local health care provider protocols.

The seemingly higher proportion of nonspecific or symptom-based diagnoses recorded after eVisits may represent physicians' reluctancy or inability to make diagnoses through the platform.

A majority of DIGI-T but a minority of PHYSI-T visits with ordered RST had sufficiently documented Centor criteria. Furthermore, a larger proportion of prescribed antibiotics in DIGI-T had a positive RST ordered on correctly documented indications. These findings should be interpreted with caution and warrant replication given their basis in a small random sample of PHYSI-T visits. EMR notes after office visits are often short, and all symptoms may not have been documented in PHYSI-T visits. Thus, PHYSI-T physicians may still adhere to guidelines similarly to DIGI-T, even though this adherence is not documented. It is, however, worth considering that more complete documentation may be a strength of eVisits compared with office visits, regardless of guideline adherence. Antibiotic prescriptions without positive RST following office visits may also be a consequence of general practitioners relying on clinical gaze over laboratory test results [24].

Comparison With Other Studies

As most studies investigating antibiotic prescribing for visits were selected based on recorded diagnoses such as streptococcal tonsillitis, our findings are not directly comparable as each group in this study contains a range of set diagnoses. However, certain patterns can be noted when the current findings are placed in context.

The finding that antibiotic prescriptions are lower following eVisits for sore throat contrasts with most previous research finding higher prescription rates for virtual visits compared with office visits following diagnosis of pharyngitis [4,15,32], with the exception of one study finding lower prescription rates following diagnosis of nasopharyngitis [32]. Differences in antibiotic prescription in this study persisted after adjusting for age and differences in set diagnoses. However, a retrospective cohort study with a large, matched sample noted no differences in prescription rates for pharyngitis [15]. Given this disparity, the findings in this study warrant replication in a different population.

The finding that DIGI-T more often were prescribed antibiotics per guideline recommendations contrasts with previous studies suggesting overprescription of broad-spectrum antibiotics after virtual visits compared with office visits [15,32]. This may demonstrate that the platform specifically improves guideline adherence through a framework encouraging physicians to reflect on guidelines prior to prescription. This is partially reflected by 100% documentation of Centor criteria, higher than reported from other eVisit platforms [25]. Indeed, previous interventions involving the use of symptom templates demonstrate improved documentation [39].

Regarding respiratory symptoms, the lower prescription rate noted in this study is in line with most research on virtual visits finding similar or lower prescription rates for bronchitis and acute respiratory infections compared with office visits [4,15,18,32,40], although some studies found higher broad-spectrum prescription rates for bronchitis [18,32].

For dysuria, previous research noted higher prescription rates following virtual visits [4] as well as eVisits [11] compared with office visits. However, a recent study on management of UTIs using asynchronous eVisits found no differences in antibiotic prescription rates. Our findings support this latter finding and the use of telemedicine for the management of uncomplicated UTIs [12]. This also suggests that eVisits and virtual visits may differently impact antibiotic prescribing.

Conclusions

The use of asynchronous eVisits for the management of sore throat, dysuria, or respiratory symptoms does not appear to lead to an inherent overprescription of antibiotics compared with office visits, even after considering differences in age and recorded diagnoses. Antibiotic prescriptions do not seem to deviate from guidelines more often than usual management using office visits. Findings support the use of structured eVisits in the context of a platform with an infrastructure encouraging guideline adherence. Future research is needed to confirm the findings of this study and validate the use of Centor criteria or another set of criteria to use for differential diagnosis and treatment of conditions related to sore throat in the eVisit setting.

Acknowledgments

The authors would like to thank Capio Sweden for assisting with sending patient consent forms, Doctrin AB for collaborating on scientifically evaluating their platform Doctrin FLOW, Olof Larsson at Medrave Software AB for assisting in programming of the data extraction software, and medical students Hannes Rönnelid, Vivi Tang, and Maria Amundstad at Lund University for assisting in manual data collection. This study was partly funded by Capio Sweden, Region Skåne, and Västra Götalandsregionen to AE, and Avtal om Läkarutbildning och Forskning funding from Region Skåne to SC.

Authors' Contributions

AE, SC, VMN, LV, AL, UJ, and PM were responsible for study concept and design. AE and TB were responsible for acquisition of data. AE, UJ, and SC performed the analysis. All authors interpreted the data. AE drafted the manuscript. All authors were responsible for critical revision of the manuscript for important intellectual content and final approval.

Conflicts of Interest

AL is a cofounder of Doctrin AB. AE is currently employed by Capio AB. LV has previously been employed by Capio AB. Other authors declared no conflicts of interest.

Multimedia Appendix 1
Protocol for interpretation of free-form text for validation of data. In uncertain cases, dialogue occurred with a family medicine specialist in order to determine if symptoms should be deemed present or absent. As not all visits were manually validated, all visits were included in the analysis.
[DOCX File , 15 KB - medinform_v9i3e25473_app1.docx]

Multimedia Appendix 2
Anatomic therapeutic chemical classification codes for recategorization of prescriptions according to current Swedish guideline recommendations.
[DOCX File , 15 KB - medinform_v9i3e25473_app2.docx]

Multimedia Appendix 3
Recategorization of diagnoses.
[DOCX File , 13 KB - medinform_v9i3e25473_app3.docx]

Multimedia Appendix 4
Centor criteria for a subset of sore throat patients from a specific county. Denominators based on available data (unavailable data is missing at random).
[DOCX File , 14 KB - medinform_v9i3e25473_app4.docx]

Multimedia Appendix 5
Secondary outcomes related to guideline adherence for a subset of sore throat visits from a specific county. Denominators vary due to missing data (unavailable data is missing at random).
[DOCX File , 14 KB - medinform_v9i3e25473_app5.docx]

References

1. Ekman B. Cost analysis of a digital health care model in Sweden. Pharmacoecon Open 2018 Sep;2(3):347-354 [FREE Full text] [doi: 10.1007/s41669-017-0059-7] [Medline: 29623633]
2. Mehrotra A. The convenience revolution for treatment of low-acuity conditions. JAMA 2013 Jul 03;310(1):35-36 [FREE Full text] [doi: 10.1001/jama.2013.6825] [Medline: 23821082]
3. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 01;20(1):1193 [FREE Full text] [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]
4. Gordon AS, Adamson WC, DeVries AR. Virtual visits for acute, nonurgent care: a claims analysis of episode-level utilization. J Med Internet Res 2017 Feb 17;19(2):e35 [FREE Full text] [doi: 10.2196/jmir.6783] [Medline: 28213342]
5. O'Neill J. Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations. London: Wellcome Trust; 2014.
6. Lee GC, Reveles KR, Attridge RT, Lawson KA, Mansi IA, Lewis JS, et al. Outpatient antibiotic prescribing in the United States: 2000 to 2010. BMC Med 2014 Jun 11;12:96 [FREE Full text] [doi: 10.1186/1741-7015-12-96] [Medline: 24916809]

7. Foster CB, Martinez KA, Sabella C, Weaver GP, Rothberg MB. Patient satisfaction and antibiotic prescribing for respiratory infections by telemedicine. Pediatrics 2019 Sep;144(3):e20190844 [FREE Full text] [doi: 10.1542/peds.2019-0844] [Medline: 31371464]
8. Davis CB, Marzec LN, Blea Z, Godfrey D, Bickley D, Michael SS, et al. Antibiotic prescribing patterns for sinusitis within a direct-to-consumer virtual urgent care. Telemed J E Health 2019 Jun;25(6):519-522. [doi: 10.1089/tmj.2018.0100] [Medline: 30020851]
9. Halpren-Ruder D, Chang AM, Hollander JE, Shah A. Quality assurance in telehealth: adherence to evidence-based indicators. Telemed J E Health 2019 Jul;25(7):599-603 [FREE Full text] [doi: 10.1089/tmj.2018.0149] [Medline: 30070966]
10. Johnson K, Dumkow L, Burns K, Yee M, Egwuatu N. Comparison of diagnosis and prescribing practices between virtual visits and office visits for adults diagnosed with sinusitis within a primary care network. Open Forum Infect Dis 2019 Sep;6(9):ofz393 [FREE Full text] [doi: 10.1093/ofid/ofz393] [Medline: 31660415]
11. Mehrotra A, Paone S, Martich GD, Albert SM, Shevchik GJ. A comparison of care at e-visits and physician office visits for sinusitis and urinary tract infection. JAMA Intern Med 2013 Jan 14;173(1):72-74 [FREE Full text] [doi: 10.1001/2013.jamainternmed.305] [Medline: 23403816]
12. Murray MA, Penza KS, Myers JF, Furst JW, Pecina JL. Comparison of eVisit management of urinary symptoms and urinary tract infections with standard care. Telemed J E Health 2020 May;26(5):639-644. [doi: 10.1089/tmj.2019.0044] [Medline: 31313978]
13. Penza KS, Murray MA, Myers JF, Furst JW, Pecina JL. Management of acute sinusitis via e-Visit. Telemed J E Health 2020 Jun 10. [doi: 10.1089/tmj.2020.0047] [Medline: 32522103]
14. Ray KN, Shi Z, Gidengil CA, Poon SJ, Uscher-Pines L, Mehrotra A. Antibiotic prescribing during pediatric direct-to-consumer telemedicine visits. Pediatrics 2019 Apr 08;143(5):e20182491. [doi: 10.1542/peds.2018-2491]
15. Shi Z, Mehrotra A, Gidengil CA, Poon SJ, Uscher-Pines L, Ray KN. Quality Of care for acute respiratory infections during direct-to-consumer telemedicine visits for adults. Health Aff (Millwood) 2018 Dec;37(12):2014-2023 [FREE Full text] [doi: 10.1377/hlthaff.2018.05091] [Medline: 30633682]
16. Smith K, Tran D, Westra B. Sinusitis treatment guideline adherence in the e-visit setting: a performance improvement project. Appl Clin Inform 2016;7(2):299-307 [FREE Full text] [doi: 10.4338/ACI-2015-10-CR-0143] [Medline: 27437042]
17. Tan LF, Mason N, Gonzaga WJ. Virtual visits for upper respiratory tract infections in adults associated with positive outcome in a Cox model. Telemed J E Health 2017 Mar;23(3):200-204. [doi: 10.1089/tmj.2016.0018] [Medline: 27351543]
18. Uscher-Pines L, Mulcahy A, Cowling D, Hunter G, Burns R, Mehrotra A. Access and quality of care in direct-to-consumer telemedicine. Telemed J E Health 2016 Apr;22(4):282-287 [FREE Full text] [doi: 10.1089/tmj.2015.0079] [Medline: 26488151]
19. Swedes-Svarm: consumption of antibiotics and occurrence of antibiotic resistance in Sweden—2016. Public Health Agency of Sweden. URL: https://www.sva.se/media/y50fy2sl/rapport_swedres-svarm_2016.pdf [accessed 2021-02-26]
20. Tyrstrup M, Beckman A, Mölstad S, Engström S, Lannering C, Melander E, et al. Reduction in antibiotic prescribing for respiratory tract infections in Swedish primary care: a retrospective study of electronic patient records. BMC Infect Dis 2016 Nov 25;16(1):709 [FREE Full text] [doi: 10.1186/s12879-016-2018-9] [Medline: 27887585]
21. Wändell P, Carlsson AC, Wettermark B, Lord G, Cars T, Ljunggren G. Most common diseases diagnosed in primary care in Stockholm, Sweden, in 2011. Fam Pract 2013 Oct;30(5):506-513. [doi: 10.1093/fampra/cmt033] [Medline: 23825186]
22. Tell D, Engström S, Mölstad S. Adherence to guidelines on antibiotic treatment for respiratory tract infections in various categories of physicians: a retrospective cross-sectional study of data from electronic patient records. BMJ Open 2015 Jul 15;5(7):e008096 [FREE Full text] [doi: 10.1136/bmjopen-2015-008096] [Medline: 26179648]
23. Dekker ARJ, Verheij TJM, van der Velden AW. Antibiotic management of children with infectious diseases in Dutch Primary Care. Fam Pract 2017 Apr 01;34(2):169-174. [doi: 10.1093/fampra/cmw125] [Medline: 28122841]
24. Gröndal H, Hedin K, Strandberg EL, André M, Brorsson A. Near-patient tests and the clinical gaze in decision-making of Swedish GPs not following current guidelines for sore throat: a qualitative interview study. BMC Fam Pract 2015 Jul 04;16:81 [FREE Full text] [doi: 10.1186/s12875-015-0285-y] [Medline: 26141740]
25. Schoenfeld AJ, Davies JM, Marafino BJ, Dean M, DeJong C, Bardach NS, et al. Variation in quality of urgent health care provided during commercial virtual visits. JAMA Intern Med 2016 Apr 4;176(5):635-642. [doi: 10.1001/jamainternmed.2015.8248] [Medline: 27042813]
26. Hickson R, Talbert J, Thornbury WC, Perin NR, Goodin AJ. Online medical care: the current state of “eVisits” in acute primary care delivery. Telemed J E Health 2015 Feb;21(2):90-96. [doi: 10.1089/tmj.2014.0022] [Medline: 25474083]
27. Liddy C, Drosinis P, Keely E. Electronic consultation systems: worldwide prevalence and their impact on patient care: a systematic review. Fam Pract 2016 Jun;33(3):274-285. [doi: 10.1093/fampra/cmw024] [Medline: 27075028]
28. Larsen J, Neighbour R. Five cards: a simple guide to beginning the consultation. Br J Gen Pract 2014 Mar;64(620):150-151 [FREE Full text] [doi: 10.3399/bjgp14X677662] [Medline: 24567647]
29. Centor RM, Witherspoon JM, Dalton HP, Brody CE, Link K. The diagnosis of strep throat in adults in the emergency room. Med Decis Making 1981;1(3):239-246. [doi: 10.1177/0272989X8100100304] [Medline: 6763125]
30. Handläggning av faryngotonsilliter i öppenvård: ny rekommendation [Management of pharyngotonsillitis in ambulatory care—new recommendation]. Swedish Medical Products Agency. 2012 Dec 14. URL: https://www.lakemedelsverket.se/

48ff63/globalassets/dokument/behandling-och-forskrivning/behandlingsrekommendationer/behandlingsrekommendation/behandlingsrekommendation-antibiotika-vid-faryngotonsilliter-i-oppenvard.pdf [accessed 2021-03-01]
31. Ashwood JS, Mehrotra A, Cowling D, Uscher-Pines L. Direct-To-consumer telehealth may increase access to care but does not decrease spending. Health Aff (Millwood) 2017 Dec 01;36(3):485-491. [doi: 10.1377/hlthaff.2016.1130] [Medline: 28264950]
32. Uscher-Pines L, Mulcahy A, Cowling D, Hunter G, Burns R, Mehrotra A. Antibiotic prescribing for acute respiratory infections in direct-to-consumer telemedicine visits. JAMA Intern Med 2015 Jul;175(7):1234-1235. [doi: 10.1001/jamainternmed.2015.2024] [Medline: 26011763]
33. Skånér Y, Arrelöv B, Backlund LG, Fresk M, Aström AW, Nilsson GH. Quality of sickness certification in primary health care: a retrospective database study. BMC Fam Pract 2013 Apr 12;14:48 [FREE Full text] [doi: 10.1186/1471-2296-14-48] [Medline: 23586694]
34. Holte M, Holmen J. Program for data extraction in primary health records: a valid tool for knowledge production in general practice? BMC Res Notes 2020 Jan 10;13(1):23 [FREE Full text] [doi: 10.1186/s13104-020-4887-7] [Medline: 31924277]
35. Fine AM, Nizet V, Mandl KD. Large-scale validation of the Centor and McIsaac scores to predict group A streptococcal pharyngitis. Arch Intern Med 2012 Jun 11;172(11):847-852 [FREE Full text] [doi: 10.1001/archinternmed.2012.950] [Medline: 22566485]
36. Rubin MA, Bateman K, Donnelly S, Stoddard GJ, Stevenson K, Gardner RM, et al. Use of a personal digital assistant for managing antibiotic prescribing for outpatient respiratory tract infections in rural communities. J Am Med Inform Assoc 2006;13(6):627-634 [FREE Full text] [doi: 10.1197/jamia.M2029] [Medline: 16929045]
37. Entezarjou A, Bolmsjö BB, Calling S, Midlöv P, Milos Nymberg V. Experiences of digital communication with automated patient interviews and asynchronous chat in Swedish primary care: a qualitative study. BMJ Open 2020 Jul 23;10(7):e036585 [FREE Full text] [doi: 10.1136/bmjopen-2019-036585] [Medline: 32709650]
38. Rekommendationer för kvalitetsindikatorer vid digitala vårdmöten [Recommendations for quality indicators in digital care visits]. Swedish Strategic Programme Against Antibiotic Resistance. 2019. URL: https://strama.se/wp-content/uploads/2019/10/Kvalitetsindikatorer-f%C3%B6r-digitala-v%C3%A5rdm%C3%B6ten-191031.pdf [accessed 2021-03-01]
39. Razai M, Hussain K. Improving antimicrobial prescribing practice for sore throat symptoms in a general practice setting. BMJ Qual Improv Rep 2017;6(1) [FREE Full text] [doi: 10.1136/bmjquality.u211706.w4738] [Medline: 28469911]
40. Hersh AL, Stenehjem E, Daines W. RE: antibiotic prescribing during pediatric direct-to-consumer telemedicine visits. Pediatrics 2019 Aug;144(2):e20191786B [FREE Full text] [doi: 10.1542/peds.2019-1786B] [Medline: 31366684]

Abbreviations

CRP: c-reactive protein
EMR: electronic medical record
eVisit: asynchronous chat-based visit
PHYSI-T: office visit with a chief complaint of sore throat
RST: rapid streptococcal antigen testing
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
UTI: urinary tract infection

Edited by C Lovis; submitted 03.11.20; peer-reviewed by J Pecina, T Jamieson, A Hidki,Dr; comments to author 16.01.21; revised version received 27.01.21; accepted 31.01.21; published 15.03.21.

Please cite as:
Entezarjou A, Calling S, Bhattacharyya T, Milos Nymberg V, Vigren L, Labaf A, Jakobsson U, Midlöv P
Antibiotic Prescription Rates After eVisits Versus Office Visits in Primary Care: Observational Study
JMIR Med Inform 2021;9(3):e25473
URL: https://medinform.jmir.org/2021/3/e25473
doi:10.2196/25473
PMID:33720032

Original Paper

Users' Willingness to Share Health Information in a Social Question-and-Answer Community: Cross-sectional Survey in China

PengFei Li[1,2], MS; Lin Xu[1,2], MS; TingTing Tang[1,3], MS; Xiaoqian Wu[1,2], MS; Cheng Huang[1,2], MD

[1]College of Medical Informatics, Chongqing Medical University, Chongqing, China
[2]Medical Data Science Academy, Chongqing Medical University, Chongqing, China
[3]The Children's Hospital of Chongqing Medical University, Chongqing, China

Corresponding Author:
Cheng Huang, MD
College of Medical Informatics
Chongqing Medical University
No.1 Yixueyuan Road
Yuzhong District
Chongqing,
China
Phone: 86 023 6848 0060
Email: huangcheng@cqmu.edu.cn

Abstract

Background: Social question-and-answer communities play an increasingly important role in the dissemination of health information. It is important to identify influencing factors of user willingness to share health information to improve public health literacy.

Objective: This study explored influencing factors of social question-and-answer community users who share health information to provide reference for the construction of a high-quality health information sharing community.

Methods: A cross-sectional study was conducted through snowball sampling of 185 participants who are Zhihu users in China. A structural equation analysis was used to verify the interaction and influence of the strength between variables in the model. Hierarchical regression was also used to test the mediating effect in the model.

Results: Altruism (β=.264, P<.001), intrinsic reward (β=.260, P=.03), self-efficacy (β=.468, P<.001), and community influence (β=.277, P=.003) had a positive effect on users' willingness to share health information (WSHI). By contrast, extrinsic reward (β=−0.351, P<.001) had a negative effect. Self-efficacy also had a mediating effect (β=.147, 29.15%, 0.147/0.505) between community influence and WSHI.

Conclusions: The findings suggest that users' WSHI is influenced by many factors including altruism, self-efficacy, community influence, and intrinsic reward. Improving the social atmosphere of the platform is an effective method of encouraging users to share health information.

(JMIR Med Inform 2021;9(3):e26265) doi:10.2196/26265

KEYWORDS

health information; willingness to share information; ; structural equation model; Zhihu

Introduction

Background

Social question-and-answer (Q&A) communities collect a large amount of high-quality health information based on the informal and collaborative method of information generation. Therefore, they have become an important means for the public to obtain health information. They also play an increasingly important role in promoting public health literacy. Zhihu is one of the most representative Q&A communities in China. In 2019, Zhihu had over 220 million registered users, over 28 million questions, and 130 million answers. On this platform, 750,000 questions were health-related, and nearly 21 million people followed the topic of health. In the Healthy China 2030 plan, the Chinese government requires news media to strengthen the publicity of

health science knowledge. Moreover, news media is required to actively use social networks for health education. Therefore, exploring the influencing factors of users' willingness to share health information (WSHI) based on the Q&A community is necessary and meaningful.

Users in a social Q&A community gather considerable amounts of high-quality health information through the sharing mechanism of question-answer-feedback. This topic has become one of the hot spots in medical informatics research to encourage more users to participate in producing health information. Empirically, Zhao et al [1] found that the interaction of intrinsic and extrinsic motivations has a considerable effect on users' knowledge sharing willingness in a social Q&A community. He et al [2] updated the Open Access and Collaborative Consumer Health Vocabulary by mining user-generated health texts in such a social Q&A community to bridge the vocabulary gap between lay consumers and health care professionals. Exploration of the WSHI in a social Q&A community is the key to provision of appropriate services to users and an important guarantee for promotion of public health knowledge.

As public platforms, social Q&A communities were established on social networking sites (SNSs) for internet users to seek and share knowledge, experiences, and other information [3]. In a social Q&A community, users can ask or answer questions, comment on relevant content, agree or disagree with related views, and follow other users [4]. These features allow user information retrieval behavior not only by using keywords but also through the most direct form of asking questions about users' complex information needs [5,6]. In terms of structure and function, the earliest social Q&A community is Quora. In this platform, users can ask their own questions and invite other users in corresponding fields to answer [5,7]. Zhihu is one of the most popular social Q&A platforms in China. It is often called the Chinese Quora. There are many similarities in the two platforms, such as user information exchange, content recommendation, and UI design. However, Zhihu and Quora have different development directions and operational concepts because they originate from different countries and social cultures. At present, Zhihu is China's mainstream social Q&A community.

Thus far, the concept of WSHI has no unified definition. According to the self-determination theory proposed by Ryan et al [8], willingness is a psychological activity generated by an individual desire to perform a certain behavior based on various motivations. Health information sharing is one of the most important aspects in the research area of information sharing. Zhu et al [9] established an influencing factor model of patients' willingness to share health information. The model includes variables of privacy concerns, online information support, information sensitivity, and disease severity. Abdelhamid et al [10] found that privacy concerns have the most influence on individuals' intentions to share personal health information. Hah [11] analyzed health consumers' health information–sharing behaviors from the perspective of the habit of using internet banking. On these bases, we define the WSHI as an individual psychological behavior driven by internal or external motivation. In the social Q&A community, such psychological behavior is often manifested as the willingness to ask health questions based on consumer experience or knowledge and provide health knowledge answers and express their views based on the content of the responses.

Study Goal

The aim of this study is to establish a user WSHI model based on the social Q&A community environment. The study also seeks to explore factors that influence the sharing of health information among such users. There are many classical models in the area of research on health information sharing such as the social cognition theory [12], the theory of reasoned action [13], and the theory of planned behavior [14]. However, these classical theoretical models can only analyze users' information-sharing behaviors from the perspective of psychological or social relations. The social Q&A community is an emerging online social platform, and the more complex information flow in such an environment warrants our more comprehensive consideration of this area. Furthermore, in consideration of the influence of community characteristics on users' WSHI, it is necessary for one to establish a model suitable for the social Q&A community environment. This may be done by integrating various classical models. However, only a few studies put community influence into their models. Based on a structural equation, we attempted to bring the influences of community characteristics into this study. Meanwhile, we established a model of users' WSHI in the social Q&A community environment and analyzed the influencing factors of the WSHI by verifying the proposed hypotheses in the model.

Research Hypotheses

Altruism

Altruism is usually understood as an individual's behavior of offering help to others at the expense of their own interests [15]. According to social exchange theory, we believe that altruism is a very complex psychological activity, and there are few behaviors that only consider others. From the perspective of social norms, altruism is a self-moral requirement based on individual ability and social influence [16]. Typically, the pleasant psychological feelings such as self-value perception and self-satisfaction are the pursuits of altruists [16]. Health information sharing is one of the behaviors that could help others solve their health problems and promote their health literacy. Therefore, altruists are more likely to identify with health information–sharing behaviors. Andrews et al [17] found that altruism is an effective factor for parents with children with genetic conditions so that these parents would share their child's electronic health record. Obrenovic et al [14] separated tacit knowledge sharing from the scope of information sharing and found that altruism has a direct impact on tacit knowledge sharing. Lin et al [18] also suggested that altruism positively affects doctors' attitudes toward knowledge sharing. On these bases, we propose hypothesis 1:

- H1: Altruism positively affects WSHI.

Intrinsic and Extrinsic Rewards

Intrinsic and extrinsic rewards are two of the most important concepts in social exchange theory [19,20]. In the study of WSHI, health information with social exchange value and the

time and labor paid by individuals in these activities can be understood as a kind of commodity. Intrinsic and extrinsic rewards are the benefits that the users can expect to obtain after completing the commodity exchange. Health information can be regarded as a bargaining chip in a social exchange. Individuals can estimate how much they will be paid based on the health information they have shared. Social recognition such as respect and reputation are intrinsic rewards, whereas economic reward is an extrinsic reward. Researchers propose that intrinsic [21,22] and extrinsic rewards [23] are two of the main influencing factors in knowledge sharing. On these bases, we propose hypotheses 2 and 3:

- H2: Intrinsic reward positively affects WSHI.
- H3: Extrinsic reward positively affects WSHI.

Self-Efficacy

Self-efficacy refers to the subjective judgment of whether an individual can successfully implement a certain behavior and achieve the expected results in a specific environment and state [24]. This concept is derived from social cognition theory, which emphasizes the interaction among individual, behavior, and environment [12]. Self-efficacy is one of the most important individual factors in social cognition theory [12]. Kankanhalli et al [25] regarded self-efficacy as a factor for individuals to gain intrinsic benefits and believed that self-efficacy has a significant positive impact on users' knowledge-sharing behavior. Kye et al [26] proposed in their empirical research that internet-related self-efficacy is positively related to information sharing. On this basis, we propose hypothesis 4:

- H4: Self-efficacy positively affects WSHI.

Community Influence

The characteristics of the community platform, such as design (Q&A format, agree/disagree mechanism, the *like* form, and comments), user stereotypes about the platform, impact of platform in this field, and protection of information sharers and their information can influence users' WSHI in the social Q&A community. In this study, the features of the platform mentioned earlier are summarized as the variable community influence belonging to the category of objective variables of the influencing factors of user health information–sharing behavior. The improvement of community influence can reduce users' perceptions of the difficulty of sharing health information, improve users' self-efficacy, and promote actual sharing. On this basis, we propose the following hypotheses:

- H5a: Community influence positively affects user self-efficacy.
- H5b: Community influence positively affects WSHI.

Methods

Participant Selection

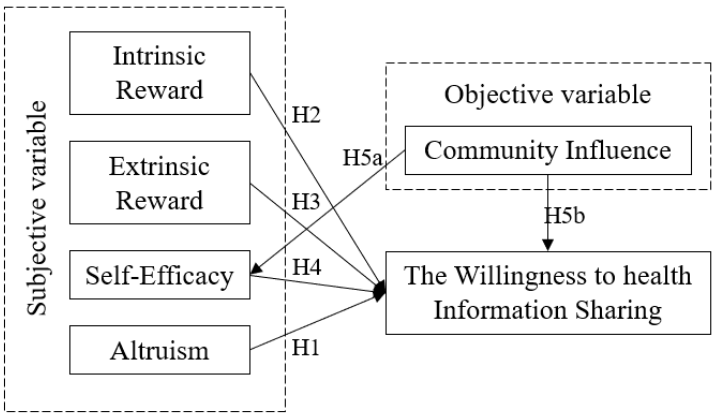
In this study, Zhihu users with a history of health information sharing were selected as the research objects. The sampling started with students at a medical university and extended by snowball sampling of an online questionnaire using WeChat and other message tools. This was done to maximize the population representation.

Zhihu users are mainly concentrated in the high knowledge-level population or people with a higher level educational background. Additionally, medical university students have high health literacy and will be the main health information disseminators in the future. Therefore, selecting medical university students as the starting point of snowball sampling is meaningful and necessary.

Modeling

This study explores the influencing factors of WSHI within the environment of a social Q&A community based on the interpretation of the above variables and related assumptions. Figure 1 illustrates the path model established from subjective and objective dimensions.

Figure 1. Willingness to share health information path model of the social question and answer community users.



Questionnaire

The observation indexes of the 6 variables in the model were screened according to the literature. A small-scale presurvey was conducted. Based on the results, the observation indexes of variables increased or decreased. Experts were asked to determine the questionnaire content in the form of a 5-point Likert scale. Table 1 lists the indexes and relevant explanations. The questionnaire is shown in Multimedia Appendix 1.

Table 1. Variables, indexes, and index descriptions in the model.

Variable, reference, and index description
Altruism—Lin [27]; Kankanhalli et al [25]
AL1: I like to share my health information with other users on Zhihu.
AL2: I think sharing health information on Zhihu can help others.
AL3: I enjoy the process of helping others by sharing health knowledge on Zhihu.
AL4: In my opinion, sharing health information on Zhihu is a manifestation of one's social value.
Intrinsic reward—Cho et al [28]; self-design
IR1: I think by sharing health information on Zhihu, we can gain others' respect.
IR2: I think by sharing health information on Zhihu, I can gain praise and recognition from others.
IR3: In my opinion, sharing health information on Zhihu can help me gain a more positive and confident attitude toward life.
Extrinsic reward—Kankanhalli et al [25]; Cho et al [28]
ER1: I think sharing health information can result in more followers.
ER2: I think sharing health information on Zhihu can bring money or other material benefits.
Community influence—self-design
CI1: I think the Zhihu platform has high credibility in solving health problems.
CI2: I think Zhihu is an important platform for me to obtain health information.
CI3: I think the public image of Zhihu can promote users to share health information.
CI4: I think Zhihu has certain security measures for sharers and the information they share.
CI5: In my opinion, the platform design of Zhihu (question-and-answer format, agree/disagree mechanism, the *like* form, and comments) can promote one's willingness to share health information.
Self-efficacy—Lin [27]; Hsu et al [29]; Chen et al [30]
SE1: I believe that the health information I have released on Zhihu is scientific and accurate.
SE2: I can express my opinions on a topic on Zhihu with confidence.
SE3: I can share new ideas and concepts about health information with others on Zhihu.
SE4: I can provide rich content in other aspects for a certain health problem on Zhihu.
SE5: I can accurately address the relevant issues and discuss them on Zhihu.
Willingness to share health information—Schwarzer et al [31]; Bock et al [32]
WSHI1: I am willing to share the health information I know on Zhihu.
WSHI2: I would like to continue the practice of sharing health information.
WSHI3: I will find more effective ways to share health information on Zhihu.
WSHI4: I would like to participate in the discussion of health information content and express my views.
WSHI5: I am willing to spend time to improve my knowledge system to provide others with better health information content.

Data Collection and Exclusion

Zhihu users with a history of health information sharing were selected as the research objects. In this study, a history of health information–sharing behavior (screening criteria) was defined as follows:

- Publishing health-related information (including asking questions, answering questions, posting articles or ideas)
- Commenting on health-related information (20 words or more)
- Sharing or forwarding health-related information (eg, to WeChat, microblog, and other platforms)

The online questionnaire was issued from June 5 to June 19, 2020 (14 days). At the end of the period, 921 Zhihu user responses were collected. Among them, 210 users had previously shared health information. After eliminating responses with missing values, 185 valid responses were obtained. This number accounts for an effective rate of 88.10% (185/210).

Statistical Analysis

Preprocessing, such as data filtering, was completed using Excel (Microsoft Corp) before importing information to the database. SPSS Statistics version 24.0 (IBM Corp) with AMOS version 24.0 and PROCESS [33] macro version 3.3 were used for data analysis. The continuous variables of demographic

characteristics were classified. Subsequently, the frequency and percentage of each indicator were calculated. A structural equation analysis was used to verify the hypotheses and calculate the coefficients of each path in the model. PROCESS was used to verify whether the mediating effect between variables is significant. A P value not more than the test level set at .05 was considered to be statistically significant.

Quality Control

The Cronbach alpha coefficient of the questionnaire was .961. It was well above .60 for each variable [34]. The composite reliability value was greater than 0.7 [35]. Table 2 presents details of the variables. All dimensions and the questionnaire as a whole have good internal consistency and reliability.

Table 2. Factor load, Cronbach alpha, average variance extracted, and composite reliability values of each variable.

Variable and index	Factor load	Cronbach alpha	AVE[a]	CR[b]
Altruism	—[c]	.875	.650	.881
AL1	.754	—	—	—
AL2	.792	—	—	—
AL3	.906	—	—	—
AL4	.764	—	—	—
Community influence	—	.894	.622	.892
CI1	.806	—	—	—
CI2	.776	—	—	—
CI3	.797	—	—	—
CI4	.772	—	—	—
CI5	.792	—	—	—
Extrinsic reward	—	.688	.638	.727
ER1	.842	—	—	—
ER2	.672	—	—	—
Intrinsic reward	—	.785	.755	.902
IR1	.918	—	—	—
IR2	.922	—	—	—
IR3	.756	—	—	—
Self-efficacy	—	.892	.628	.893
SE1	.648	—	—	—
SE2	.848	—	—	—
SE3	.853	—	—	—
SE4	.776	—	—	—
SE5	.819	—	—	—
Willingness to share health information	—	.928	.724	.929
WS1	.787	—	—	—
WS2	.877	—	—	—
WS3	.850	—	—	—
WS4	.910	—	—	—
WS5	.824	—	—	—

[a]AVE: Average variance extracted.

[b]CR: Critical ratio.

[c]Not applicable.

Content validity reflects the degree to which the description of measurement items affects the survey results. The measurement items in the questionnaire were mainly taken from the published literature. The self-designed indexes are obtained through expert discussion. They were then combined with the characteristics of the research object. Therefore, we believe that the scale has a good content validity. Structural validity includes both convergent and discriminant validities. The main measurement

indexes are the factor load and average variance extracted (AVE) [35]. Table 2 presents the specific analysis results and index values. The factor loads and AVE values of all variables are greater than 0.5, indicating that the model has good convergent validity [35]. Discriminant validity requires the lowest possible correlation among all variables. Moreover, the standard is that such value should be less than the square root of the AVE value of the variable itself [36]. Table 3 indicates that the data on the diagonal are the square roots of the AVEs of each variable, indicating that the model has acceptable discriminant validity.

Table 3. Discriminant validity matrix.

Variable	AL[a]	CI[b]	ER[c]	IR[d]	SE[e]	WSHI[f]
AL	0.650	—[g]	—	—	—	—
CI	0.420	0.623	—	—	—	—
ER	0.619	0.628	0.638	—	—	—
IR	0.637	0.454	0.655	0.754	—	—
SE	0.562	0.261	0.505	0.721	0.628	—
WSHI	0.688	0.324	0.654	0.735	0.628	0.724
AVE[h]	0.806	0.789	0.799	0.868	0.793	0.851

[a]AL: Altruism.

[b]CI: Community influence.

[c]ER: Extrinsic reward.

[d]IR: Intrinsic reward.

[e]SE: Self-efficacy.

[f]WSHI: Willingness to share health information.

[g]Not applicable.

[h]AVE: Average variance extracted.

Results

Demographic Characteristics

Table 4 lists demographic characteristics of the participants. The data indicate that, among the participants, 70.8% (131/185) were female and 90.8% (168/185) were aged between 19 and 38 years. Additionally, 96.8% (179/185) had an undergraduate degree or higher education level. This survey considers medical students as the starting point of the snowball sampling considering the good educational background of Zhihu users and the fact that medical students will be the main producers and disseminators of health information in the future.

Table 4. Demographic characteristics of participants.

Variable	Value, n (%)
Gender	
Male	54 (29.2)
Female	131 (70.8)
Age in years	
≤18	9 (4.9)
19-38	168 (90.8)
39-58	8 (4.3)
Education	
Senior high school and below	3 (1.6)
Junior college	3 (1.6)
Undergraduate	143 (77.3)
Master and above	36 (19.5)
Background of majors	
Medical science or related	139 (75.1)
Nonmedical-related	46 (24.9)
Profession	
Student	150 (81.1)
Government personnel	11 (5.9)
Professional technical personnel	11 (5.9)
Business and service personnel	4 (2.2)
Other	9 (4.9)

Model Test

Table 5 presents the path and model fitting using the SPSS Statistics 24.0 and AMOS 24.0 software. We selected the chi-square/degree of freedom (χ^2/df), root mean square error of approximation (RMSEA), incremental fit index (IFI), and cumulative fit index (CFI) as reference indexes of the model fitting. When χ^2/df < 3, RMSEA < 0.08 and IFI/TLI/CFI > 0.9, the model is considered to have a good fit [37-39]. Table 5 indicates that χ^2/df = 1.95 < 3 and RMSEA = 0.072 < 0.08. Other fitting indexes are all above 0.9. Therefore, the model fit is acceptable.

Table 5. Model fitting test index values.

Index	Value	Standard	Fitting
χ^2/df[a]	1.959	<5	acceptable
		<3	ideal
RMSEA[b]	0.072	<0.08	acceptable
		<0.05	ideal
IFI[c]	0.934	＞0.9	ideal
CFI[d]	0.933	＞0.9	ideal

[a]χ^2/df: Chi-square/degree of freedom.

[b]RMSEA: Root mean square error of approximation.

[c]IFI: Incremental fit index.

[d]CFI: Cumulative fit index.

Figure 2 is the path diagram of the structural equation model (SEM). By contrast, Table 6 lists the path coefficients. The influence path of each variable on health information–sharing intention is significant. Altruism, community influence, intrinsic reward, and self-efficacy all positively affect WSHI. According to the absolute value of the influence of independent variables

on dependent variables, the ranking is as follows: self-efficacy (β=.468; P<.001), extrinsic reward (β=−.351; P<.001), community influence (β=.277; P=.003), altruism (β=.264; P<.001), and intrinsic reward (β=.260; P=.03). All of the hypotheses (H1, H2, H4, H5a, and H5b) are true except for H3.

Figure 2. Structural equation model path diagram. AL: altruism; CI: community influence; IR intrinsic reward; ER: extrinsic reward; SE: self-efficacy; WS: willingness to share.

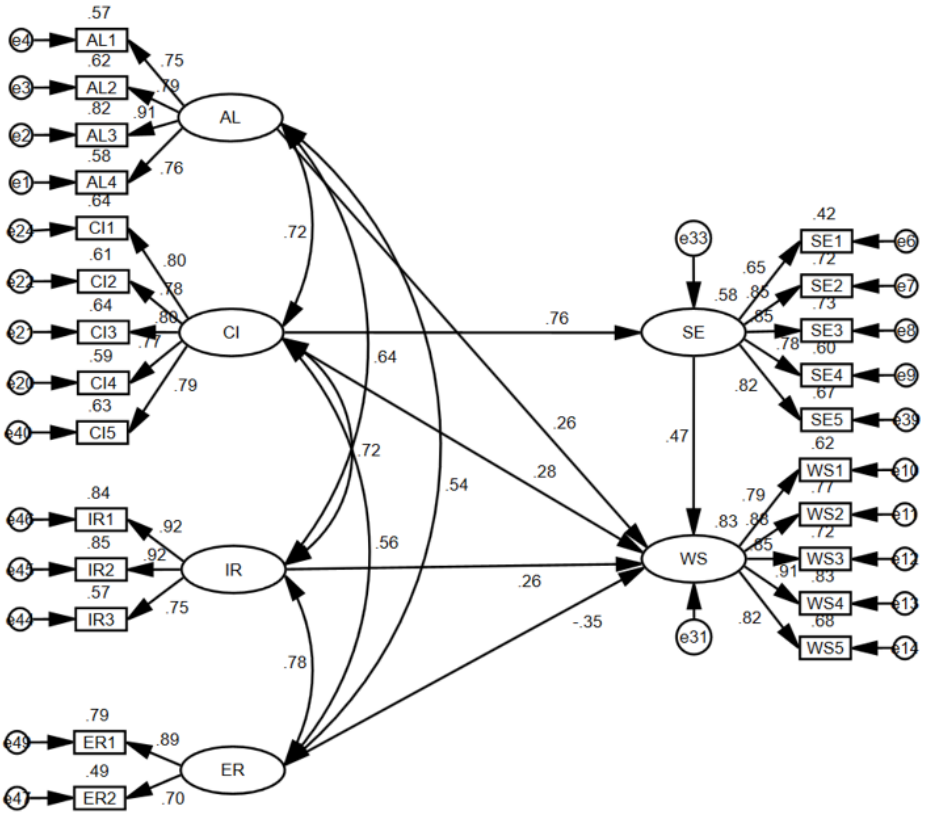


Table 6. Path testing.

Path	Unstandardized coefficient	Standardized coefficient	Standard error	CR[a]	P value
CI[b]→SE[c]	0.814	.764	0.092	8.818	<.001
SE→WSHI[d]	0.520	.468	0.095	5.313	<.001
AL[e]→WSHI	0.285	.264	0.087	3.446	<.001
CI→WSHI	0.328	.277	0.125	2.963	.003
IR[f]→WSHI	0.291	.260	0.153	2.157	.03
ER[g]→WSHI	–0.347	–.351	0.101	–3.433	<.001

[a]CR: Critical ratio.

[b]CI: Community influence.

[c]SE: Self-efficacy.

[d]WSHI: Willingness to share health information.

[e]AL: Altruism.

[f]IR: Intrinsic reward.

[g]ER: Extrinsic reward.

Mediating Effect Test

Based on the PROCESS [33] macro, we tested the mediating effect of self-efficacy through hierarchical regression. The study selected a bootstrap to sample 2000 times and set altruism, extrinsic reward, and intrinsic reward as control variables to test the mediating effect of self-efficacy in the influence of community influence on WSHI. Table 7 indicates that the SEM is significant (Δr^2=.063, ΔF=5.305, P<.001). The first line in Table 7 corresponds to the dependent variables of model 1, model 2, and model 3, respectively. The difference between model 2 and model 3 is whether the mediating effect of

self-efficacy is introduced. The r^2 and F score values for these three models were as follows: model 1 (r^2=.521, F=49.019), model 2 (r^2=.651, F=83.874), and model 3 (r^2=.714, F=89.179). As shown in Table 8, in the path community influence → self-efficacy → willingness to share, the mediating effect value of self-efficacy was 0.147. Moreover, the effect accounted for 29.15%. The bootstrap test indicated that the 95% confidence interval did not contain 0. Therefore, the mediating effect was significant.

Table 7. Model for testing the mediating effect of self-efficacy.

Variable	SE[a]		WSHI1[b, c]		WSHI2[d]	
	t score	P value	*t* score	P value	*t* score	P value
AL[e]	2.387	.02	5.169	<.001	4.506	<.001
ER[f]	0.609	.54	–2.298	.02	–2.811	.006
IR[g]	3.102	.002	3.883	<.001	2.754	.007
CI[h]	4.680	<.001	6.804	<.001	5.012	<.001
SE	__[i]	—	—	—	6.261	<.001

[a]SE: Self-efficacy.

[b]WSHI: Willingness to share health information.

[c]No Moderating variables.

[d]Self-efficacy was introduced as a moderating variable.

[e]AL: Altruism.

[f]ER: Extrinsic reward.

[g]IR: Intrinsic reward.

[h]CI: Community influence.

[i]Not applicable.

Table 8. Proportion of the mediating effect.

Mediating effect	β	Boot SE	Boot upper	Boot lower	%
Total	.505	0.088	0.327	0.670	—
Direction	.358	0.082	0.198	0.517	70.83
Mediation	.147	0.047	0.063	0.245	29.15

Discussion

Principal Findings

Influence of Altruism on Willingness to Share Health Information

Altruism has a positive effect on the user's WSHI. From the perspective of social norms, altruism is a moral requirement and standard for individual social values based on one's ability and social influence. In other words, this is the self-perception of "with great power comes great responsibility." In a social Q&A community, users tend to exert certain moral requirements and restrictions on themselves based on their own cognitive ability and knowledge. These include the sharing of the health information they know and grasping to help other community users. Raj et al [40] suggested that altruism is one of the important factors that promote the moral obligation of individuals to share health information for research. In a social Q&A community, a good social atmosphere can help users achieve more in-depth communication with others and maximize user exposure to health information needs. This finding, that optimizing the social atmosphere is an effective measure, is of considerable significance for generating altruistic psychology among users. To a certain extent, the discussion atmosphere in the community can be optimized through filtering of users and strict auditing of users' content publishing.

Influence of Intrinsic and Extrinsic Rewards on Willingness to Share Health Information

In the context of a social Q&A community, intrinsic reward has a positive effect on WSHI, but extrinsic reward has a negative impact. Users typically respect, praise, and thank the information sharers when users improve their health with the help of information shared by others. Intrinsic reward, such as respect and reputation, can promote a sense of satisfaction, pleasure, and fulfillment among health information sharers. In turn, this state of mind can continue to generate their WSHI. Thus, intrinsic reward (ie, reputation) can positively affect users' willingness in health information sharing [18,23]. To further improve users' perceptions of intrinsic reward, the community should, on the premise of ensuring that spam information is effectively filtered, magnify the exposure of other users to the effective feedback content of relevant health information. This may be done by means of group chat and push. Similarly, increasing intrinsic reward entails ensuring that the

magnification of this exposure is known to the health information sharer in the social Q&A community. Another necessity is an effective 2-way mutual evaluation function in which health information recipients can rate or express opinions on the sharers and their shared information. Moreover, the health information sharers should also be able to rate the opinions of recipients. At the same time, rewards (ie, membership points, experience value, and level promotion) are given based on the mutual recognition of both parties.

The conclusion that extrinsic reward negatively affects users' willingness in health information sharing differs from that of existing research. This finding may be caused by the demographic distribution characteristics of the study's current sample. The participants are mainly college students with their family or parents as their main sources of income. After the lower economic pressure is mapped to these users and their WSHI, it was found that intrinsic reward (ie, reputation and respect) can have a greater influence than extrinsic reward. Conversely, inappropriate extrinsic reward may cause user aversion or resistance. The users may feel that their sharing behavior is controlled by the organization if extrinsic reward is the intention of sharing health information. Just like the imposition of punishment from the organization, material reward is another mechanism of controlling individual behavior. This can cause an individual to lose interest and enthusiasm in sharing health information [41]. Tamir et al [42] found that individuals are willing to forgo money to share their experiences and knowledge. However, given the demographic distribution characteristics of the current samples, we do not deny that some professional web writers, who spend more time online and have more followers, will benefit from the flow economy by sharing health information. Therefore, it is necessary to further subdivide health information–sharing users to obtain more realistic and objective results. Meanwhile, the specific mechanism through which extrinsic reward influences WSHI also needs further study.

Effect of Community Influence and Self-Efficacy on Willingness in Health Information Sharing

Community influence and self-efficacy have a positive effect on users' WSHI. Moreover, self-efficacy has a mediating effect similar to the way that community influence, an objective variable in the environment, affects WSHI. This objective fact includes various aspects such as community development philosophy, platform design, information protection, and user group influence. These factors interact with each other, and they not only have a direct impact on the WSHI but also exert further influence by positively affecting users' sense of self-efficacy. The perception and evaluation of factors such as self-ability and environmental conditions are necessary steps before users share health information. Users tend to be satisfied with their information-sharing behavior in perceiving that their own knowledge can help other users [43]. Improving the design of the community platform and strengthening the information protection mechanism and publicity efforts of the community can attract more high-quality users to participate in sharing health information. This will in turn improve the overall influence of the community. At the same time, we should also pay attention to the effect of community influence on user self-efficacy. After realizing the accurate identification and tagging of social Q&A community users, it can push the needs of health information demanders to health information providers more efficiently through a reasonable information push mechanism. This shall stimulate the generation of user self-efficacy. Thus, the virtuous cycle of health information diffusion is effectively promoted.

Strengths and Limitations

Many scholars have conducted in-depth research in the field of knowledge sharing. They proposed different knowledge sharing models for different types of information or communication environments [44,45]. However, most of the relevant studies did not consider the impact of the characteristics of these information dissemination environments on users' willingness to share knowledge. This study attempted to bring influences of community characteristics into the model of users' WSHI. As a result, we found that the variable community influence has a positive impact on users' WSHI. Meanwhile, the variable self-efficacy has a mediating effect between community influence and users' WSHI. This study provides new ideas and directions in the research area of users' WSHI and proposes suggestions on promoting users to share health information. These suggestions can be used as a reference for health information service providers to formulate health intervention strategies.

This study also has certain limitations. First, only Zhihu users were taken as the objects of this research. Thus, not all social Q&A community users are covered. Follow-up research should further improve the coverage of social Q&A community users. Second, the samples mainly comprised ordinary Zhihu users, and key users with a large number of followers are underrepresented. Third, although the sample population, mainly comprising medical college students, can better represent the information-sharing behavior of the general population, the snowball sampling method used in this study still has systematic errors. Fourth, the method of using an online questionnaire may introduce bias of demographic characteristics. Finally, the data obtained were formed by subjective reports provided by participants. Overly conservative or exaggerated choices can lead to a certain degree of bias in the statistical results. More scientific experimental designs can be adopted to avoid the biases caused by these deficiencies in follow-up studies. Despite these shortcomings, this study presents novel ideas, and the results provide new insights into the promotion of WSHI.

Conclusions

Promoting the dissemination of high-quality health information is important for guiding users of a social Q&A community to actively participate in health information sharing. Compared with relevant research, this study introduces the variable community influence into the model based on the characteristics of a social Q&A community. Additionally, combining the variables intrinsic reward, extrinsic reward, altruism, and self-efficacy, WSHI's SEM in a social Q&A community is constructed. The results indicate that intrinsic reward, altruism, and self-efficacy have a positive effect on WSHI. By contrast, extrinsic reward has a negative effect. Self-efficacy has a mediating effect on the relationship between community

influence and WSHI. The generation of WSHI may be promoted by paying more attention to the social atmosphere of the community, optimizing the gratitude feedback mechanism, and striving to build good social relations among users. The results can provide a theoretical and practical reference for social Q&A community operators, health education and promotion, and other aspects.

Acknowledgments

This study was supported by the Western China Social Sciences Program entitled Research on Health Information Transmission Path and Diffusion Model Based on Mobile Internet (Item No. 16XTQ012). We would like to thank all the participants who took part in the questionnaire survey. Furthermore, we are also very grateful to all the reviewers for their careful examination of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire.
[DOCX File , 27 KB - medinform_v9i3e26265_app1.docx]

References

1. Zhao L, Detlor B, Connelly CE. J Manag Inform Syst 2016 Jun 17;33(1):70-100. [doi: 10.1080/07421222.2016.1172459]
2. He Z, Chen Z, Oh S, Hou J, Bian J. J Biomed Inform 2017 May;69:75-85 [FREE Full text] [doi: 10.1016/j.jbi.2017.03.016] [Medline: 28359728]
3. Zhao W, Lu P, Yu S, Lu L. BMC Med Inform Decis Mak 2020 Jul 09;20(Suppl 3):130 [FREE Full text] [doi: 10.1186/s12911-020-1124-1] [Medline: 32646418]
4. Charlie AM, Gao Y, Heller SL. What do patients want to know? Questions and concerns regarding mammography expressed through social media. J Am Coll Radiol 2018 Oct;15(10):1478-1486. [doi: 10.1016/j.jacr.2017.09.020] [Medline: 29221997]
5. Alasmari A, Zhou L. Int J Med Inform 2019 Nov;131:103958. [doi: 10.1016/j.ijmedinf.2019.103958] [Medline: 31521012]
6. Young I, Bhulabhai M, Papadopoulos A. Safe food handling advice provided on question-and-answer web sites is inconsistent. J Nutr Educ Behav 2020 Jul;52(7):688-696. [doi: 10.1016/j.jneb.2019.12.011] [Medline: 31948743]
7. Chen Z, Zhang C, Zhao Z, Yao C, Cai D. Question retrieval for community-based question answering via heterogeneous social influential network. Neurocomputing 2018 Apr;285:117-124. [doi: 10.1016/j.neucom.2018.01.034]
8. Ryan RM, Deci EL. Self-regulation and the problem of human autonomy: does psychology need choice, self-determination, and will? J Pers 2006 Dec;74(6):1557-1585. [doi: 10.1111/j.1467-6494.2006.00420.x] [Medline: 17083658]
9. Zhu P, Shen J, Xu M. Patients' willingness to share information in online patient communities: questionnaire study. J Med Internet Res 2020 Apr 01;22(4):e16546 [FREE Full text] [doi: 10.2196/16546] [Medline: 32234698]
10. Abdelhamid M, Gaia J, Sanders GL. Putting the focus back on the patient: how privacy concerns affect personal health information sharing intentions. J Med Internet Res 2017 Sep 13;19(9):e169 [FREE Full text] [doi: 10.2196/jmir.6877] [Medline: 28903895]
11. Hah H. Health consumers' daily habit of internet banking use as a proxy for understanding health information sharing behavior: quasi-experimental approach. J Med Internet Res 2020 Jan 08;22(1):e15585 [FREE Full text] [doi: 10.2196/15585] [Medline: 31913129]
12. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
13. Esmaeilzadeh P. The impacts of the perceived transparency of privacy policies and trust in providers for building trust in health information exchange: empirical study. JMIR Med Inform 2019 Nov 26;7(4):e14050 [FREE Full text] [doi: 10.2196/14050] [Medline: 31769757]
14. Obrenovic B, Jianguo D, Tsoy D, Obrenovic S, Khan MAS, Anwar F. The enjoyment of knowledge sharing: impact of altruism on tacit knowledge-sharing behavior. Front Psychol 2020;11:1496 [FREE Full text] [doi: 10.3389/fpsyg.2020.01496] [Medline: 32765348]
15. Batson CD, Shaw LL. Evidence for altruism: toward a pluralism of prosocial motives. Psychol Inq 1991 Apr;2(2):107-122. [doi: 10.1207/s15327965pli0202_1]
16. Sharp C, Randhawa G. Altruism, gift giving and reciprocity in organ donation: a review of cultural perspectives and challenges of the concepts. Transplant Rev (Orlando) 2014 Oct;28(4):163-168. [doi: 10.1016/j.trre.2014.05.001] [Medline: 24973193]

17. Andrews S, Raspa M, Edwards A, Moultrie R, Turner-Brown L, Wagner L, et al. "Just tell me what's going on": the views of parents of children with genetic conditions regarding the research use of their child's electronic health record. J Am Med Inform Assoc 2020 Mar 01;27(3):429-436 [FREE Full text] [doi: 10.1093/jamia/ocz208] [Medline: 31913479]
18. Lin TC, Lai MC, Yang SW. Factors influencing physicians' knowledge sharing on web medical forums. Health Informatics J 2016 Sep;22(3):594-607 [FREE Full text] [doi: 10.1177/1460458215576229] [Medline: 25888432]
19. Homans GC. Social behavior as exchange. Am J Sociol 1958 May;63(6):597-606. [doi: 10.1086/222355]
20. Emerson R. Exchange theory, part I: a psychological basis for social exchange. Sociol Theor Progr 1972;2:38-57. [doi: 10.4324/9781315573946-4]
21. Bock, Zmud, Kim, Lee. Behavioral intention formation in knowledge sharing: examining the roles of extrinsic motivators, social-psychological forces, and organizational climate. MIS Quarterly 2005;29(1):87. [doi: 10.2307/25148669]
22. Chen X, Sun M, Wu D, Song XY. Information-sharing behavior on WeChat moments: the role of anonymity, familiarity, and intrinsic motivation. Front Psychol 2019;10:2540 [FREE Full text] [doi: 10.3389/fpsyg.2019.02540] [Medline: 31798501]
23. Zhou J, Zuo M, Ye C. Understanding the factors influencing health professionals' online voluntary behaviors: evidence from YiXinLi, a Chinese online health community for mental health. Int J Med Inform 2019 Oct;130:103939. [doi: 10.1016/j.ijmedinf.2019.07.018] [Medline: 31434043]
24. Bandura A, Caprara GV, Barbaranelli C, Gerbino M, Pastorelli C. Role of affective self-regulatory efficacy in diverse spheres of psychosocial functioning. Child Dev 2003;74(3):769-782. [doi: 10.1111/1467-8624.00567] [Medline: 12795389]
25. Kankanhalli A, Tan B, Wei K. Contributing knowledge to electronic knowledge repositories: an empirical investigation. MIS Quarterly 2005;29(1):113. [doi: 10.2307/25148670]
26. Kye S, Shim M, Kim Y, Park K. Sharing health information online in South Korea: motives, topics, and antecedents. Health Promot Int 2019 Apr 01;34(2):182-192. [doi: 10.1093/heapro/dax074] [Medline: 29040499]
27. Lin H. Effects of extrinsic and intrinsic motivation on employee knowledge sharing intentions. J Inform Sci 2007 Feb 15;33(2):135-149. [doi: 10.1177/0165551506068174]
28. Cho H, Chen M, Chung S. Testing an integrative theoretical model of knowledge-sharing behavior in the context of Wikipedia. J Am Soc Inf Sci 2010;61(6):1198-1212. [doi: 10.1002/asi.21316]
29. Hsu M, Ju TL, Yen C, Chang C. Knowledge sharing behavior in virtual communities: the relationship between trust, self-efficacy, and outcome expectations. Int J Human Comput Studies 2007 Feb;65(2):153-169. [doi: 10.1016/j.ijhcs.2006.09.003]
30. Chen C, Hung S. To give or to receive? Factors influencing members' knowledge sharing and community promotion in professional virtual communities. Inform Manag 2010 May;47(4):226-236. [doi: 10.1016/j.im.2010.03.001]
31. Schwarzer R, Renner B. Social-cognitive predictors of health behavior: action self-efficacy and coping self-efficacy. Health Psychol 2000 Sep;19(5):487-495. [Medline: 11007157]
32. Bock GW, Kim Y. Breaking the myths of rewards: an exploratory study of attitudes about knowledge sharing. Inform Resource Manag J 2002 Apr;15(2):14-21. [doi: 10.4018/irmj.2002040102]
33. Hayes AF, Rockwood NJ. Regression-based statistical mediation and moderation analysis in clinical research: observations, recommendations, and implementation. Behav Res Ther 2017 Nov;98:39-57. [doi: 10.1016/j.brat.2016.11.001] [Medline: 27865431]
34. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull 1955 Jul;52(4):281-302. [doi: 10.1037/h0040957] [Medline: 13245896]
35. Bagozzi RP. Evaluating structural equation models with unobservable variables and measurement error: a comment. J Mark Res 2018 Nov 28;18(3):375-381. [doi: 10.1177/002224378101800312]
36. Chin WW, Gopal A, Salisbury WD. Advancing the theory of adaptive structuration: the development of a scale to measure faithfulness of appropriation. Inform Syst Res 1997 Dec;8(4):342-367. [doi: 10.1287/isre.8.4.342]
37. Harris PR, Sillence E, Briggs P. Perceived threat and corroboration: key factors that improve a predictive model of trust in internet-based health information and advice. J Med Internet Res 2011 Jul;13(3):e51 [FREE Full text] [doi: 10.2196/jmir.1821] [Medline: 21795237]
38. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
39. Marsh H, Balla J, Hau K. An evaluation of incremental fit indices: a clarification of mathematical and empirical properties. In: Advanced Structural Modeling: Issues and Techniques. London: Taylor and Francis; 1996.
40. Raj M, De Vries R, Nong P, Kardia SLR, Platt JE. Do people have an ethical obligation to share their health information? Comparing narratives of altruism and health information sharing in a nationally representative sample. PLoS One 2020;15(12):e0244767 [FREE Full text] [doi: 10.1371/journal.pone.0244767] [Medline: 33382835]
41. Hui J. Empirical study of impacts of intrinsic and extrinsic motivations on employee knowledge sharing: crowding-out and crowding-in effect. J Manag Sci (China) 2012;26:31-44. [doi: 10.3969/j.issn.1672-0334.2013.03.004]
42. Tamir DI, Mitchell JP. Disclosing information about the self is intrinsically rewarding. Proc Natl Acad Sci USA 2012 May 22;109(21):8038-8043 [FREE Full text] [doi: 10.1073/pnas.1202129109] [Medline: 22566617]
43. Chung N, Nam K, Koo C. Examining information sharing in social networking communities: applying theories of social capital and attachment. Telemat Inform 2016 Feb;33(1):77-91. [doi: 10.1016/j.tele.2015.05.005]

44. Bahrami M, Namnabati M, Mokarian F, Oujian P, Arbon P. Information-sharing challenges between adolescents with cancer, their parents and health care providers: a qualitative study. Support Care Cancer 2017 May;25(5):1587-1596. [doi: 10.1007/s00520-016-3561-z] [Medline: 28078477]
45. Kipkosgei F, Son SY, Kang S. Coworker trust and knowledge sharing among public sector employees in Kenya. Int J Environ Res Public Health 2020 Mar 18;17(6). [doi: 10.3390/ijerph17062009] [Medline: 32197432]

Abbreviations

AVE: average variance extracted
CFI: cumulative fit index
IFI: incremental fit index
Q&A: question-and-answer
RMSEA: root mean square error of approximation
SEM: structural equation model
SNS: social network site
WSHI: willingness to share health information
χ2/df: chi-square/degree of freedom

Edited by C Lovis; submitted 04.12.20; peer-reviewed by J Luan, S Rush, S Liu; comments to author 16.01.21; revised version received 10.02.21; accepted 07.03.21; published 30.03.21.

Please cite as:
Li P, Xu L, Tang T, Wu X, Huang C
Users' Willingness to Share Health Information in a Social Question-and-Answer Community: Cross-sectional Survey in China
JMIR Med Inform 2021;9(3):e26265
URL: https://medinform.jmir.org/2021/3/e26265
doi:10.2196/26265
PMID:33783364

Original Paper

Realistic High-Resolution Body Computed Tomography Image Synthesis by Using Progressive Growing Generative Adversarial Network: Visual Turing Test

Ho Young Park[1], MD; Hyun-Jin Bae[2], PhD; Gil-Sun Hong[1], MD; Minjee Kim[3], BSc; JiHye Yun[1], PhD; Sungwon Park[4], MD; Won Jung Chung[4], MD; NamKug Kim[1,5], PhD

[1]Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine & Asan Medical Center, Seoul, Republic of Korea

[2]Department of Medicine, University of Ulsan College of Medicine & Asan Medical Center, Seoul, Republic of Korea

[3]Department of Biomedical Engineering, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea, Seoul, Republic of Korea

[4]Department of Health Screening and Promotion Center, University of Ulsan College of Medicine & Asan Medical Center, Seoul, Republic of Korea

[5]Department of Convergence Medicine, University of Ulsan College of Medicine & Asan Medical Center, Seoul, Republic of Korea

Corresponding Author:
Gil-Sun Hong, MD
Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine & Asan Medical Center
88 Olympic-ro 43-gil, Songpa-gu
Seoul, 05505
Republic of Korea
Phone: 82 2 3010 1548
Email: hgs2013@gmail.com

Abstract

Background: Generative adversarial network (GAN)–based synthetic images can be viable solutions to current supervised deep learning challenges. However, generating highly realistic images is a prerequisite for these approaches.

Objective: The aim of this study was to investigate and validate the unsupervised synthesis of highly realistic body computed tomography (CT) images by using a progressive growing GAN (PGGAN) trained to learn the probability distribution of normal data.

Methods: We trained the PGGAN by using 11,755 body CT scans. Ten radiologists (4 radiologists with <5 years of experience [Group I], 4 radiologists with 5-10 years of experience [Group II], and 2 radiologists with >10 years of experience [Group III]) evaluated the results in a binary approach by using an independent validation set of 300 images (150 real and 150 synthetic) to judge the authenticity of each image.

Results: The mean accuracy of the 10 readers in the entire image set was higher than random guessing (1781/3000, 59.4% vs 1500/3000, 50.0%, respectively; $P<.001$). However, in terms of identifying synthetic images as fake, there was no significant difference in the specificity between the visual Turing test and random guessing (779/1500, 51.9% vs 750/1500, 50.0%, respectively; $P=.29$). The accuracy between the 3 reader groups with different experience levels was not significantly different (Group I, 696/1200, 58.0%; Group II, 726/1200, 60.5%; and Group III, 359/600, 59.8%; $P=.36$). Interreader agreements were poor ($\kappa=0.11$) for the entire image set. In subgroup analysis, the discrepancies between real and synthetic CT images occurred mainly in the thoracoabdominal junction and in the anatomical details.

Conclusions: The GAN can synthesize highly realistic high-resolution body CT images that are indistinguishable from real images; however, it has limitations in generating body images of the thoracoabdominal junction and lacks accuracy in the anatomical details.

(JMIR Med Inform 2021;9(3):e23328) doi:10.2196/23328

KEYWORDS

generative adversarial network; unsupervised deep learning; computed tomography; synthetic body images; visual Turing test

Introduction

Generative adversarial networks (GANs) is a recent innovative technology that generates artificial but realistic-looking images. Despite the negative views regarding the use of synthetic images in the medical field, GANs have been spotlighted in radiological research because of their undeniable advantages [1]. The use of diagnostic radiological images in the public domain always raises the problem of protecting patients' privacy [2-5]. This has been a great challenge to researchers in the field of deep learning. GANs may provide a solution to these privacy concerns. Moreover, GANs are powerful nonsupervised training methods. The traditional supervised learning methods have been challenged by a lack of high-quality training data labelled by experts. Building these data requires considerable time input from experts and leads to correspondingly high costs [6]. This problem has not yet been resolved despite several collaborative efforts to build large open access data sets [7]. Most radiological tasks using GANs include the generation of synthetic images for augmenting training images [8-11], translation between different radiological modalities [12-16], image reconstruction and denoising [17-20], and data segmentation [21-24].

The more recent noteworthy task using GANs is anomaly detection. Unlike other tasks using GANs, detecting abnormalities is based on learning the probability distribution of normal training data. Image data outside this distribution are considered as abnormal. Schlegl et al [25] demonstrated GAN-based anomaly detection in optical coherence tomography images. They trained GAN with normal data in an unsupervised approach and proposed an anomaly scoring scheme. Alex et al [26] showed that GAN can detect brain lesions on magnetic resonance images. This approach has attracted many radiologists for several reasons; the most critical is that this approach can achieve a broader clinical application than the current supervised deep learning–based diagnostic models. In daily clinical practice, diagnostic images are clinically acquired for patients with a variety of diseases. Therefore, before applying the supervised deep learning model, it is necessary to select suspected disease cases with disease categories similar to those of a training data set. For example, in the emergency department, a deep learning model trained by data from patients with acute appendicitis could hardly be applied to patients with different abdominal pathologies.

For this approach, we think that generating highly realistic images is a prerequisite. Previous studies [25,26] trained a GAN model with small patches (64×64 pixels), which are randomly extracted from original images. The trained model could only generate small patches and did not learn the semantics of the whole images. Hence, the GAN model may generate artificial features, which can lead to large errors in anomaly detection tasks. In addition, there are various kinds of small and subtle lesions in the actual clinical setting. Therefore, the previous low-resolution GAN approaches could not be used for this application. In this study, we trained GAN with whole-body computed tomography (CT) images (512×512 pixels); therefore, the model learned the semantics of the images. This may lead to robust performances in anomaly detection in CT images. Due to the aforementioned reasons, we have attempted to build large data sets of normal medical images to develop GAN-based diagnostic models for clinical application. As a preliminary study, we investigated and validated the unsupervised synthesis of highly realistic body CT images by using GAN by learning the probability distribution of normal training data.

Methods

Ethical Approval

This retrospective study was conducted according to the principles of the Declaration of Helsinki and was performed in accordance with current scientific guidelines. This study protocol was approved by the Institutional Review Board Committee of the Asan Medical Center (No. 2019-0486). The requirement for informed patient consent was waived.

Data Collection for Training

We retrospectively reviewed electronic medical records of patients who underwent chest CT or abdominopelvic CT (AP-CT) in the Health Screening and Promotion Center of Asan Medical Center between January 2013 and December 2017. We identified 139,390 patients. Their radiologic reports were then reviewed using the radiologic diagnostic codes "Code 0" or "Code B0," which indicated normal CT in our institution's disease classification system, and 17,854 patients with normal chest CT or normal AP-CT were identified. One board-certified radiologist (GSH) reviewed the radiological reports of the 17,854 patients and excluded 3650 cases with incidental benign lesions (eg, hepatic cysts, renal cysts, thyroid nodules) detected on body CT images. Benign lesions were defined as positive incidental findings on CT images, which did not require medical or surgical intervention. Our final study group included CT images showing anatomical variations (eg, right aortic arch, double inferior vena cava) and senile changes (eg, atherosclerotic calcification without clinical significance). Of the potentially suitable 14,204 cases, 2449 CT data sets were not available for automatic download using the inhouse system of our institution. Finally, this study included 11,755 body CT scans (473,833 axial slices) for training the GAN, comprising 5000 contrast-enhanced chest CT scans (172,249 axial slices) and 6755 AP-CT scans (301,584 axial slices, comprising 132,880 slices of contrast-enhanced AP-CT and 168,704 slices of contrast-enhanced low-dose AP-CT images).

Training PGGAN to Generate Body CT Images

A progressive growing GAN (PGGAN) was used to generate high-resolution (512×512 pixels) synthetic body CT images. Unlike PGGAN, previous GAN models such as deep convolutional GANs were able to generate relatively low-resolution (256×256 pixels) synthetic images [27]. However, PGGANs have demonstrated that high-resolution images (1024×1024 pixels) can be generated by applying progressive growing techniques [28]. Because CT images are acquired in high resolutions (512×512 pixels), PGGAN could be the GAN model that can train with whole CT images in full resolution. Consequently, the GAN model can preserve their semantics in the original resolution of CT images. While StyleGAN also demonstrates realistic synthetic images with the style feature [29], we chose the PGGAN model for training

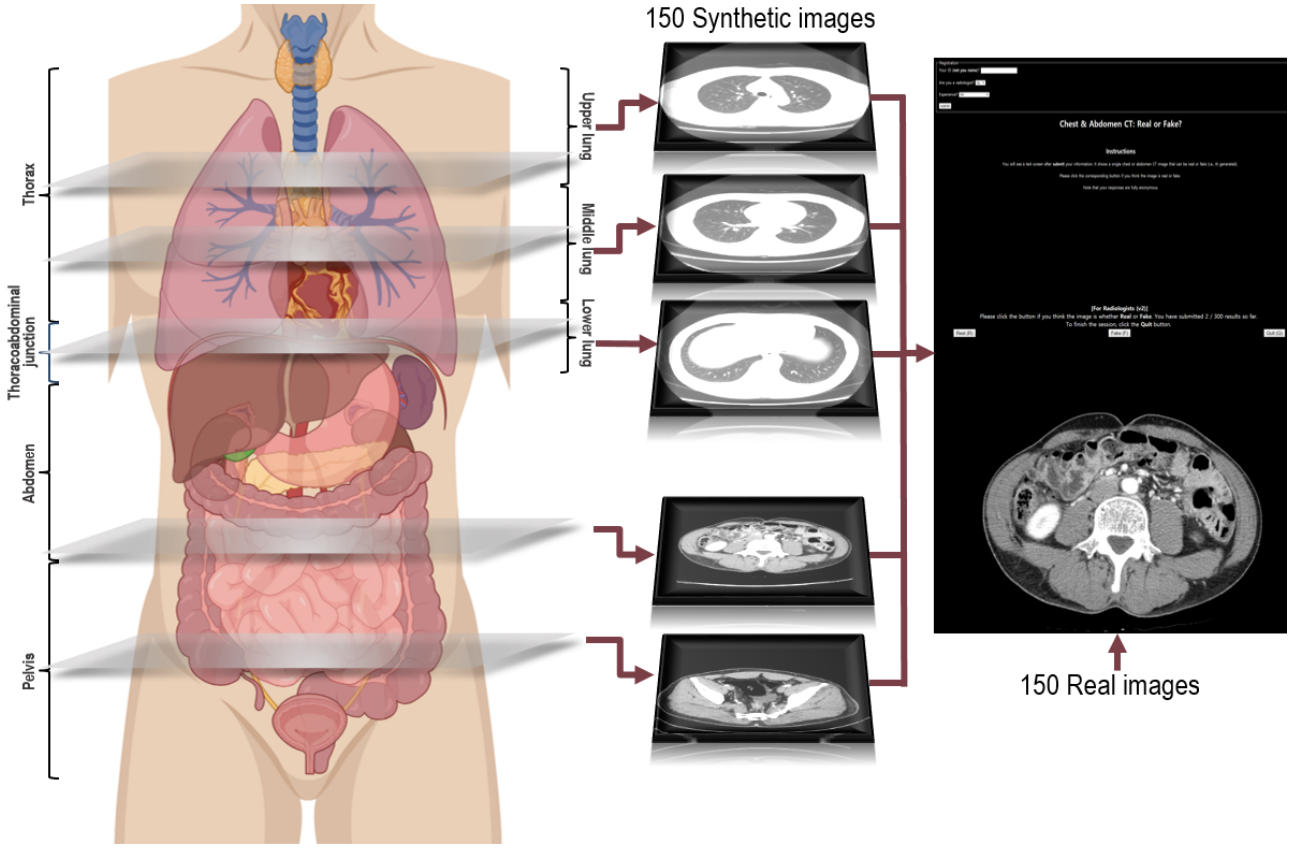
because of its simple yet powerful performance. In addition, we did not consider BigGAN because it is a conditional model [30]. To train the PGGAN with body CT images, the original 12-bit grayscale CT images were converted into 8-bit grayscale potable network graphics images with 3 different windowing settings: (1) a lung setting (window width 1500, window level 600), (2) a mediastinal setting (window width 450, window level 50) for chest CT images, and (3) a multiorgan setting (window width 350, window level 40) for AP-CT images. Images from each group with different windowing settings were used to train a PGGAN separately.

A publicly available official implementation of PGGAN using Tensorflow in Python was used [31]. While the sizes of the training images progressively grew from 4×4 to 512×512 (ie, $2^n \times 2^n$, where the integer n increases from 2 to 8), the batch sizes decreased from 512 to 16, respectively. The learning rate was fixed at 0.001 while training. We carefully monitored the training process (ie, training losses and generated images) with TensorBoard and intermediated image generation to determine whether the PGGAN was properly trained. The PGGAN training was completed after the network had evaluated around 20 million body CT images. The training took ~12.5 days with 2 NVIDIA Titan RTX graphic processing units for each group with different windowing settings (ie, total training for ~37.5 days).

Visual Turing Test to Assess the Realistic Nature of Synthetic CT Images

Figure 1 summarizes the study design for the visual assessment performed using an image Turing test. The validation set consisted of 300 axial body CT images (150 synthetic images and 150 real images). The 150 synthetic images comprised 50 chest CT-lung window (chest-L), 50 chest CT-mediastinal window (chest-M), and 50 AP-CT images. The validation set consisted of 7 subgroups based on the anatomical structure: 50 chest-L images were divided into upper lung, middle lung, and lower lung groups; and 50 chest-M and 50 AP-CT images were divided into thorax, thoracoabdominal junction, abdomen, and pelvis groups. To avoid any selection bias, all synthetic images in the validation set were automatically generated by the PGGAN model and were not individually selected by the researchers. For the real images, 50 CT images of each anatomical subgroup (ie, chest-L, chest-M, and AP-CT) were randomly selected from 50 normal whole-body CT scans (performed at the emergency department of Asan Medical Center) by 1 co-researcher (JHY) who did not otherwise participate in the realism assessment study. A website (validct.esy.es) was created to upload the validation set with 300 axial images posted and displayed in a random manner. Ten radiologists (4 radiologists with <5 years of experience [Group I], 4 radiologists with 5-10 years of experience [Group II], and 2 radiologists with >10 years of experience [Group III]) independently evaluated each of the 300 images slice-by-slice and decided whether each CT image was real or artificial by visual analysis with no time limit. To investigate the features of the images with obviously artificial appearance, we defined obviously artificial images as synthetic images that were identified as artificial by a majority of readers. Two radiologists (HYP and GSH) then visually reviewed these obviously artificial images. To determine whether the radiologists could learn to distinguish real from synthetic images, we performed an additional Turing test (postlearning visual Turing test). First, 2 board-certified radiologists (Group III) were educated in the obviously artificial findings in the synthetic images (not included in the test set). Then, 2 readers independently decided whether each of the 300 CT images were real or artificial by visual analysis. For accurate comparison of the results, the same test set as the index visual Turing test was used.

Figure 1. Graphical illustration of the method used to estimate the realism of the synthetic body computed tomography images. The validation set consisted of 150 synthetic and 150 real images. Synthetic images generated by the progressive growing generative adversarial network model and real images were randomly mixed and displayed on the website. Ten readers independently determined whether each image was real or artificial.



Statistical Analyses

The mean accuracy, sensitivity, and specificity of the 10 readers were calculated. The generalized estimating equations method was used to test whether the ratio of mean accuracy and random guessing was 1. The generalized estimating equations were used to compare the accuracy, sensitivity, and specificity across the reader groups with different experience levels (Group I, Group II, and Group III) and across the anatomical subgroups. To compare the diagnostic performance among subgroups, chest-L was classified into 3 image subgroups (upper, middle, and lower lung), and chest-M and AP-CT images were grouped into 4 image subgroups (thorax, thoracoabdominal junction, abdomen, and pelvis) on the basis of anatomical structures by visual inspection. The anatomical landmarks used in subgrouping of CT-L were as follows: (1) upper lung: apex to upper border of tracheal bifurcation; (2) middle lung: upper border of tracheal bifurcation to upper border of diaphragm; and (3) lower lung: upper border of diaphragm to lower border of diaphragm. The anatomical landmarks used in the subgroups of CT-M and AP-CT were as follows: (1) thorax: apex to upper border of diaphragm; (2) thoracoabdominal junction: upper border of diaphragm to lower border of diaphragm; (3) abdomen: lower border of diaphragm to upper border of iliac crest; and (4) pelvis: below the upper border of iliac crest. Chest-M and AP-CT images were combined for the subgroup classification because these images included the "soft tissue setting" used for the whole body. Figure 1 shows the subgroup classification according to the anatomical level. The significance level was corrected for multiple comparisons using the Bonferroni correction. Interreader agreement was evaluated using Fleiss kappa. To identify obviously artificial images, a histogram analysis was used to display the distribution of the number of correct answers from the 10 readers (ie, identification of synthetic images as artificial) and the number of artificial images. The cut-off values (ie, percentage of readers with correct answers) were set where dramatic changes in the histogram distribution was observed. When a cut-off ≥70% was used for chest-L and ≥80% for chest-M and AP-CT images, 1 subgroup (ie, upper lung for chest-L and thoracoabdominal junction for chest-M and AP-CT images) had the highest number of readers with correct answers. In the postlearning visual Turing test, the mean accuracy, sensitivity, and specificity of the 2 readers were calculated. SPSS software (version 23, IBM Corp) and R version 3.5.3 (R Foundation for Statistical Computing) were used for the statistical analyses with the significance level set at $P<.05$.

Results

Results of the Visual Turing Test

Table 1 summarizes the results of the realism assessment of all images by the 10 readers. The mean accuracy of the 10 readers in the entire image set was higher than the random guessing (1781/3000, 59.4% vs 1500/3000, 50.0%, respectively; $P<.001$). However, in terms of identifying synthetic images as fake, there was no significant difference in the specificity between the visual Turing test and random guessing (779/1500, 51.9% vs 750/1500, 50.0%, respectively; $P=.29$). There was no significant difference in the accuracy between the 3 reader groups with

different experience levels (Group I, 696/1200, 58.0%; Group II, 726/1200, 60.5%; and Group III, 359/600, 59.8%; *P*=.36). In the detection of synthetic images, Group III showed a significantly lower specificity than Group II (*P*=.01) but did not show a significant difference from Group I (*P*=.30). Multimedia Appendix 4 summarizes the results of the subgroup analysis of the realism assessment according to the anatomical region. There were no significant differences in the accuracy between the 3 CT groups (chest-L, 595/1000, 59.5%; chest-M, 615/1000, 61.5%; and AP-CT, 571/1000, 57.1%; *P*=.33). In addition, there was no significant difference in the accuracy between the upper, middle, and lower lung groups of the chest-L images (upper lung, 227/370, 61.4%; middle lung, 190/290, 65.5%; and lower lung, 136/240, 56.7%, *P*=.36). The thoracoabdominal junction showed a significantly higher accuracy (208/280, 74.3% vs 194/370, 52.4% to 361/600, 60.2%; *P*=.004) and specificity (154/200, 77.0% vs 93/220, 42.3% to 149/250, 59.6%; *P*<.001) compared with the other subgroups. Examples of the multilevel random generation of synthetic chest CT and AP-CT images by the PGGAN are shown in Figure 2 and in Multimedia Appendix 1, Multimedia Appendix 2, and Multimedia Appendix 3.

Table 1. Assessment of the realism of all images by the 10 readers.

Groups, readers (R)	Accuracy (%)[a]	Sensitivity (%)[b]	Specificity (%)[c]
Group I[d]			
R01	56.7	67.3	46.0
R05	48.3	53.3	43.3
R09	61.0	70.7	51.3
R10	66.0	70.7	61.3
Group II[e]			
R02	43.7	50.0	37.3
R06	73.0	68.7	77.3
R07	61.3	65.3	57.3
R08	64.0	77.3	50.7
Group III[f]			
R03	65.3	86.0	44.7
R04	54.3	58.7	50.0

[a]Mean (95% CI) accuracy: 59.4 (56.9-61.8), *P*=.36. *P* value was determined by generalized estimating equations.

[b]Mean (95% CI) sensitivity: 66.8 (63.9-69.5), *P*=.04.

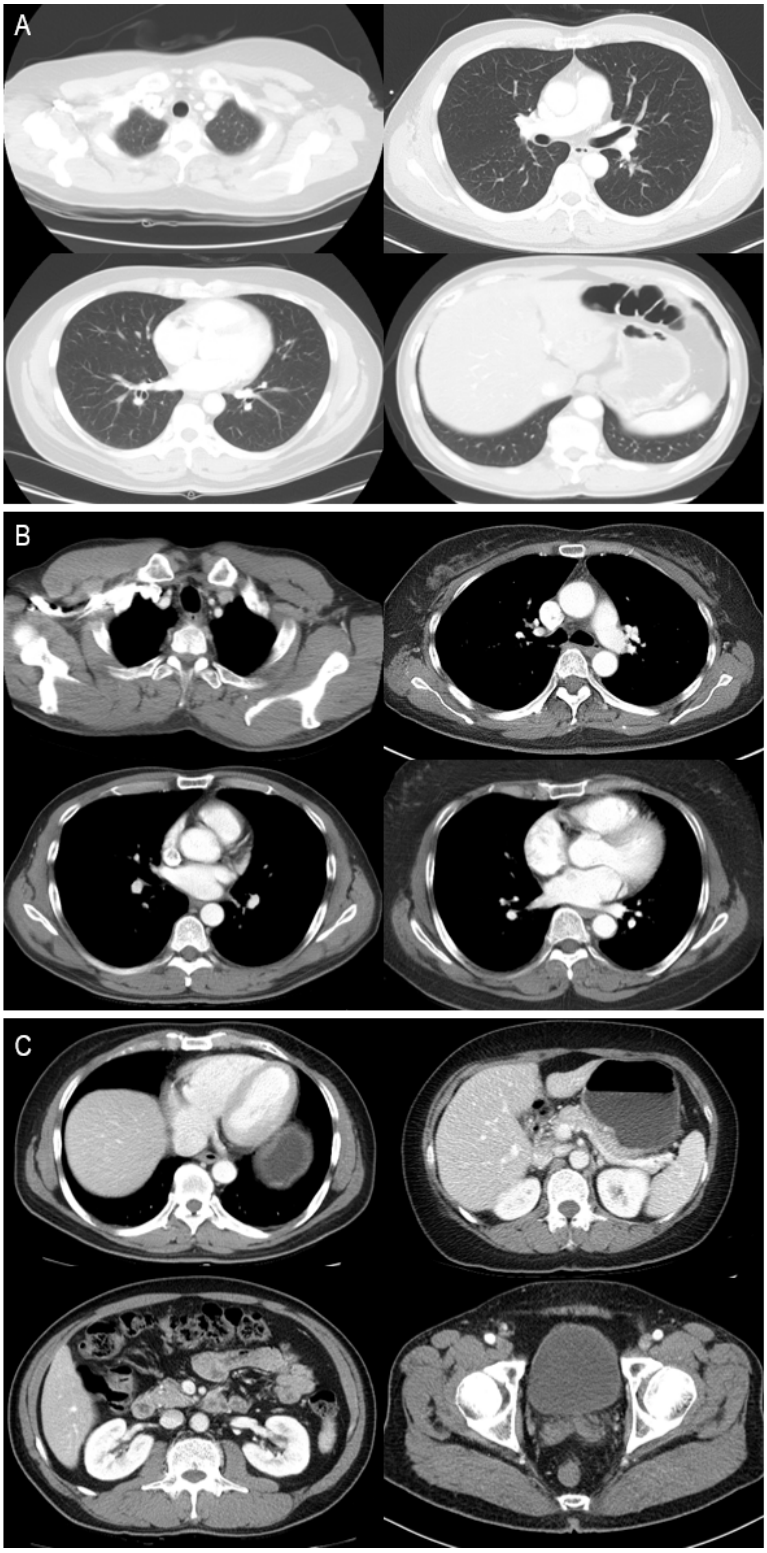
[c]Mean (95% CI) specificity: 51.9 (48.4-55.5), *P*=.02.

[d]Group I: radiologists with <5 years of experience. Mean (95% CI) accuracy 58.0 (55.0-61.0), sensitivity 65.5 (61.4-69.4), and specificity 50.5 (46.3-54.7).

[e]Group II: radiologists with 5-10 years of experience. Mean (95% CI) accuracy 60.5 (57.6-63.4), sensitivity 65.3 (61.4-69.0), and specificity 55.7 (51.4-59.9).

[f]Group III: radiologists with >10 years of experience. Mean (95% CI) accuracy 59.8 (55.5-64.1), sensitivity 72.3 (67.0-77.1), and specificity 47.3 (41.1-53.7).

Figure 2. Synthetic high-resolution body computed tomography images. A. Chest computed tomography images-lung window. B. Chest computed tomography images-mediastinal window. C. Abdominopelvic computed tomography images.



In the postlearning visual Turing test, the mean accuracy, sensitivity, and specificity of the 2 radiologists were 67.3%, 72.7%, and 62.0%, respectively. Compared with the results of the index visual Turing test, the accuracy was increased by 7.5% and the specificity was increased by 10.1% in the postlearning visual Turing test.

Interreader Agreement for Synthetic and Real Images

Interreader agreement was poor for the entire image set (κ=0.11) and for the 3 CT subsets (chest-L, chest-M, and AP-CT; κ=0.04-0.13). Interreader agreement was higher for the thoracoabdominal junction subset than for the other anatomical regions (κ=0.31 vs 0.03-0.14) (Table 2).

Table 2. Interreader agreement of the 10 readers with respect to the imaging subgroups.

Image type, subsets	Kappa values	95% CI
Entire image set	0.11	0.09 to 0.13
Image subsets		
Chest-L[a]	0.04	0.01 to 0.07
Chest-M[b]	0.13	0.10 to 0.15
AP-CT[c]	0.11	0.08 to 0.14
Chest-L		
Upper lung	0.04	–0.01 to 0.09
Middle lung	0.01	–0.04 to 0.07
Lower lung	0.06	0.00 to 0.12
Chest-M and AP-CT		
Thorax	0.03	–0.01 to 0.06
Thoracoabdominal junction	0.31	0.25 to 0.36
Abdomen	0.14	0.10 to 0.18
Pelvis	0.03	–0.02 to 0.08

[a]Chest-L: chest computed tomography images-lung window.

[b]Chest-M: chest computed tomography images-mediastinal window.

[c]AP-CT: abdominopelvic computed tomography images.

Analysis of the Features of Obviously Artificial Images

Figure 3 shows that the majority of readers characterized the synthetic images as artificial predominantly at the thoracoabdominal junction of the chest-M and AP-CT, followed by the upper lung of the chest-L. Using a histogram analysis, 24 of the 150 synthetic images (22 images of the chest-M and AP-CT groups and 2 images of the upper lung) were selected and reviewed by 2 radiologists to identify the features indicating that the images were artificial. Table 3 details the artificial features indicative of synthetic CT images. A total of 34 artificial features were found in the 24 synthetic images, the most common being vascular structures (24/34, 71%), followed by movable organs (ie, stomach, heart, small bowel, and mediastinal fat around the heart, 8/34, 24%). Among the vascular structures, intrahepatic vessels (ie, portal and hepatic veins) most frequently had abnormal configurations, directions, or diameters (Figure 4). In case of the movable organs, an abnormal organ contour was the main feature indicative of an artificially generated image (Figure 4C and Figure 4D).

Figure 3. Histogram analysis of the correct answers for the 150 synthetic images (accurate identification of the artificial images) by the 10 readers. A. When a cut-off for the percentage of readers with correct answers was set at ≥70% for the chest computed tomography-lung window group, only 1 subgroup (upper lung) remained (§). B. When a cut-off level for the percentage of readers with correct answers was set at ≥80% for the chest computed tomography-mediastinal window and abdominopelvic computed tomography groups, the thoracoabdominal (TA) junction group (*) showed dominance over the other subgroups.

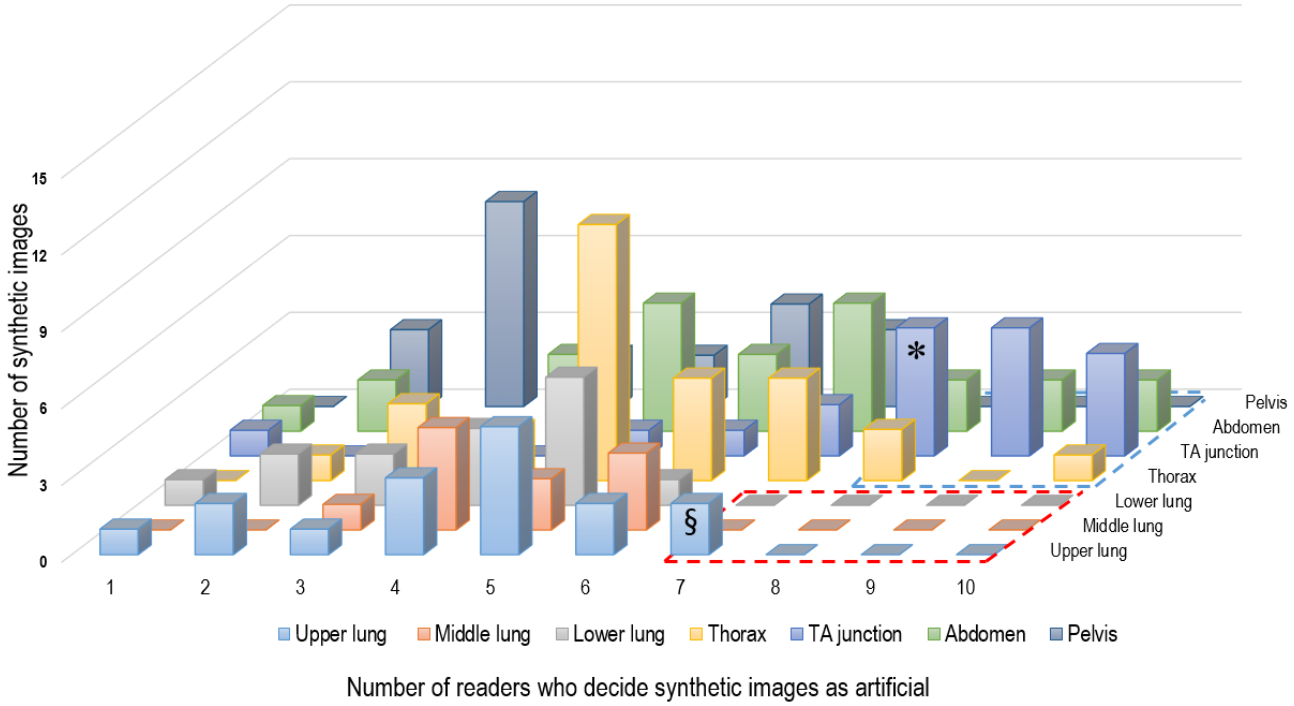


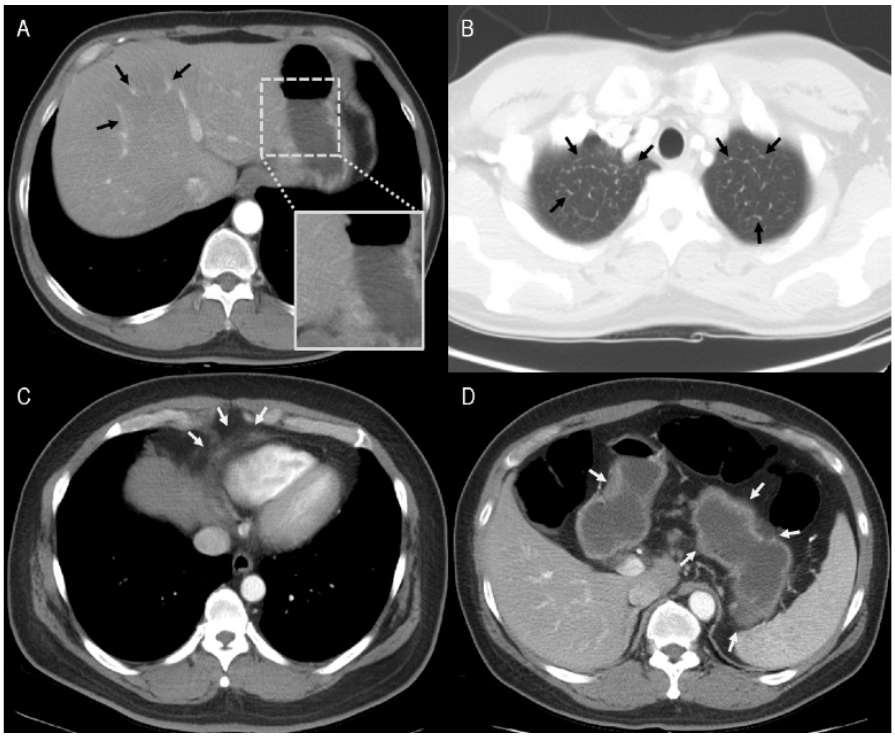
Table 3. Details of the obviously artificial body computed tomography images.

Configuration, artificial features	Images (n)
Abnormal vascular configuration[a]	
Hepatic vessel (portal vein and hepatic vein)	13
Gastric vessel	3
Mesenteric vessel	2
Pulmonary vessel	2
Others (peripancreatic, coronary, rectal, axillary vessel)	4
Abnormal contour or structure[b]	
Stomach	3
Pancreas	2
Heart	2
Mediastinal fat around the heart	2
Small bowel	1

[a]Ill-defined vascular margin, bizarre vascular course, or abnormal vascular diameter.

[b]Blurred margin of the organ, or bizarre structure of the soft tissue.

Figure 4. Obviously artificial body computed tomography images. A. Ill-defined margins and abnormal courses of intrahepatic vessels (arrows) in the liver. Note curvilinear structures (dotted rectangle) at the liver and stomach. B. Accentuated vascular markings in both upper lung apices (arrows). C. Abnormal infiltration in the pericardial fat (arrows). D. Irregular contours of the stomach body and antrum with blurred margins (arrows).



Discussion

Principal Findings

We showed that the GAN-based synthetic whole-body CT images have comparable image fidelity to real images. For this, our study validated the synthetic images by multiple radiology experts because the visual Turing test could be greatly influenced by the reader's level of expertise [10,32,33]. There was no significant difference in the accuracy between the reader groups. In addition, the interreader agreement was poor for the distinction between real and synthetic images. These results imply that a validation test was properly performed with mitigation of the impact of the reader's level of expertise. However, there was quite a significant disparity between sensitivity (66.8%) and specificity (51.9%). We presume that this is mainly due to factors affecting reader performance test. First, all readers had at least some exposure to real body CT images in clinical practice. In addition, the real images in the validation data set consisted of relatively uniform CT images because they were acquired using a similar CT machine with similar acquisition parameters. These factors affect the readers' confidence and decisions to identify real images, resulting in high sensitivity. This is supported by the fact that the sensitivity proposed here reached 72.3% in Group III (radiologists with long-term exposure to real CT images in our institution). In contrast, some obviously artificial features (eg, the ill-defined margin of the heart) in synthetic images are similar to the motion artifacts or noises in real images. This can cause reader confusion, resulting in lower specificity. In addition, the mean accuracy (59.4%) was higher than random guessing (50%); however, it is believed that the high sensitivity contributed significantly to this result. Therefore, in terms of identifying synthetic images as fake, the readers' performance was not much better than random guessing. For robust validation, using real CT images from other medical institutions (not experienced by the readers) in the validation set could be needed. Despite this limitation, our data suggest that the synthetic images are highly realistic and indistinguishable from real CT images.

One critical finding of this study was that the discrepancies between real and synthetic CT images occur mainly in the thoracoabdominal junction and in anatomical details. The thoracoabdominal junction is the most prone to motion artifacts due to respiratory movement. In addition, it has a complex anatomical structure due to multiple organs in small spaces [34]. These features of the thoracoabdominal junction might have contributed to the identification of unrealistic synthetic body images. This phenomenon in the areas with complex structures has been shown in other image syntheses using GANs [27,28]. It is worth noting that this study showed that GAN achieved highly realistic images for gross anatomy and not for detailed anatomical structures. The most common obviously artificial features in synthetic images were bizarre configurations and directions of small-to-medium vessels. This is probably due to the lack of the interslice shape continuity caused by the 2D CT image–training and the anatomical diversity of these vessels [10,35]. Therefore, to overcome these limitations, further work would require the generation of 3D CT images with larger and more diverse data sets. The second most obviously artificial feature was an abnormal contour of the movable organs. This could be another limitation in the GAN-based realistic image synthesis. Recently, more powerful GAN models have been introduced into the medical field. We believe that many problems raised here can serve as criteria to test the performance of the newly introduced GAN models.

As expected, learning artificial features in the synthetic images improved the performance of radiologists in identifying artificial images. However, it did not reach our expectations. This is because artificial features occurred mainly in some images of certain anatomical subgroups. In addition, as mentioned before, it is not easy for radiologists to distinguish these artificial features from motion artifacts or noise in real images. Furthermore, our visual Turing tests were based on reviewing 2D synthetic CT slices. However, although 3D data (eg, CT) are presented as 2D images, human perception of an anomaly is based on the imagination of space from 2D images. These factors could make it difficult to determine whether each CT image is real or artificial.

Comparison With Prior Work

Bermudez et al [36] reported that GAN can successfully generate realistic brain MR images. However, unlike this study, the previous GAN-based unconditional synthesis of advanced radiological images (CT or magnetic resonance images) has been confined to some specific pathologic lesions (eg, lung and liver lesions) and specific organs (eg, heart and brain) for a variety of purposes [8,36-40]. In contrast, this study shows that realistic high-resolution (512×512 pixels) whole-body CT images can be synthesized by GAN. GAN was trained with whole-body CT images (512×512 pixels) in this study; therefore, the model learned the semantics of the images. It is worth noting that the generated images cover a wide range of 2-dimensional (2D) slice CT images along the z-axis from the thorax to the pelvis and contain multiple organs. To the best of our knowledge, there has been no study that has investigated and validated the unsupervised synthesis of highly realistic body CT images by using a PGGAN.

Limitations

Our study had some limitations. First, technical novelty is lacking in this study. However, while state-of-the-art GAN models such as PGGAN and StyleGAN were introduced recently, there are still limited studies in the medical domain and a lack of published studies on anomaly detection tasks. As far as we know, this is the first attempt to generate high-quality medical images (whole-body CT) and to validate the generated medical images by expert radiologists. This study will provide readers a way to follow our approach and to achieve advances in anomaly detection tasks in medical imaging. Second, our training data are not enough to cover the probability distribution of normal data. This preliminary study used normal CT images from our institution. The training data consisted of relatively homogeneous CT images with similar acquisition parameters and CT machines. Therefore, further studies should focus on the collection of multi-center and multi-country diverse CT data to achieve better results. Third, due to limited graphics processing unit memory, our study only validated the realistic nature of separate 2D high-resolution body CT slices that were randomly generated by the GAN. This study did not handle 3D synthetic CT images, although real body CT images are volumetric data. Therefore, interslice continuity of pathologic lesions and organs may be a crucial factor for improving the performance of deep learning–based models. Further studies are needed to generate and validate 2.5D or 3D synthetic CT images in terms of detailed anatomical structures. Fourth, the number of synthetic images in the validation set varied between each anatomical region; thus, the statistical power may have been insufficient. However, we tried to avoid any researcher-associated selection bias in this process. Finally, we did not evaluate the correlation between the number of CT images in the training set and the generation of realistic images in the validation set. Our study showed that the PGGAN can successfully produce realistic body CT images by using a much smaller amount of training data in contrast to previous studies on the generation of celebrity face images with 1K pixels by 1K pixels [28,29]. However, we did not provide a cut-off value for the number of CT images required to generate realistic images. Therefore, further studies are needed to clarify the approximate data set size required for the generation of highly realistic normal or disease-state CT images.

Conclusions

GAN can synthesize highly realistic high-resolution body CT images indistinguishable from real images; however, it has limitations in generating body images in the thoracoabdominal junction and lacks accuracy in anatomical details.

Acknowledgments

The authors are grateful to Ju Hee Lee, MD, Hyun Jung Koo, MD, Jung Hee Son, MD, Ji Hun Kang, MD, Jooae Choe, MD, Mi Yeon Park, MD, Se Jin Choi, MD, and Yura Ahn, MD, for participating as readers. This work was supported by the National Research Foundation of Korea (NRF-2018R1C1B6006371 to GS Hong). The data sets are not publicly available due to restrictions in the data-sharing agreements with the data sources. Ethical approval for the use of the deidentified slides in this study was granted by the Institutional Review Board of the Asan Medical Center.

Authors' Contributions

HYP wrote the original draft, analyzed the data, and performed formal analysis. HJB wrote the original draft and provided technical guidance for the project. GSH and NKK conceptualized the project and provided the methodology for the project. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example video of the multi-level random generation of synthetic chest computed tomography-lung window by the progressive growing generative adversarial network.
[MP4 File (MP4 Video), 34971 KB - medinform_v9i3e23328_app1.mp4]

Multimedia Appendix 2
Example video of the multi-level random generation of synthetic chest computed tomography-mediastinal window by the progressive growing generative adversarial network.
[MP4 File (MP4 Video), 35117 KB - medinform_v9i3e23328_app2.mp4]

Multimedia Appendix 3
Example video of the multi-level random generation of synthetic abdominopelvic computed tomography images by the progressive growing generative adversarial network.
[MP4 File (MP4 Video), 34476 KB - medinform_v9i3e23328_app3.mp4]

Multimedia Appendix 4
Subgroup analysis of diagnostic performance with respect to the anatomical subgroups. A. Accuracy, B. Sensitivity, C. Specificity. There was a significant difference in accuracy (*) and specificity (†) between the thoracoabdominal junction (TA) and other image subgroups. Chest-L: chest computed tomography-lung window; chest-M: chest computed tomography-mediastinal window; AP-CT: abdominopelvic computed tomography.
[PNG File , 152 KB - medinform_v9i3e23328_app4.png]

References

1. Sorin V, Barash Y, Konen E, Klang E. Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) - A Systematic Review. Acad Radiol 2020 Aug;27(8):1175-1185. [doi: 10.1016/j.acra.2019.12.024] [Medline: 32035758]
2. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep Learning: A Primer for Radiologists. Radiographics 2017;37(7):2113-2131. [doi: 10.1148/rg.2017170077] [Medline: 29131760]
3. Ker J, Wang L, Rao J, Lim T. Deep Learning Applications in Medical Image Analysis. IEEE Access 2018;6:9375-9389 [FREE Full text] [doi: 10.1109/ACCESS.2017.2788044]
4. Lee J, Jun S, Cho Y, Lee H, Kim GB, Seo JB, et al. Deep Learning in Medical Imaging: General Overview. Korean J Radiol 2017;18(4):570-584 [FREE Full text] [doi: 10.3348/kjr.2017.18.4.570] [Medline: 28670152]
5. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018 Nov 27;19(6):1236-1246 [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
6. Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, et al. GANs for medical image analysis. Artificial Intelligence in Medicine 2020 Sep;109:101938. [doi: 10.1016/j.artmed.2020.101938]
7. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. Med Image Anal 2019 Dec;58:101552. [doi: 10.1016/j.media.2019.101552] [Medline: 31521965]
8. Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing 2018 Dec;321:321-331. [doi: 10.1016/j.neucom.2018.09.013]
9. Gadermayr M, Li K, Müller M, Truhn D, Krämer N, Merhof D, et al. Domain-specific data augmentation for segmenting MR images of fatty infiltrated human thighs with neural networks. J Magn Reson Imaging 2019 Jun;49(6):1676-1683. [doi: 10.1002/jmri.26544] [Medline: 30623506]
10. Kazuhiro K, Werner R, Toriumi F, Javadi M, Pomper M, Solnes L, et al. Generative Adversarial Networks for the Creation of Realistic Artificial Brain Magnetic Resonance Images. Tomography 2018 Dec;4(4):159-163 [FREE Full text] [doi: 10.18383/j.tom.2018.00042] [Medline: 30588501]
11. Russ T, Goerttler S, Schnurr A, Bauer DF, Hatamikia S, Schad LR, et al. Synthesis of CT images from digital body phantoms using CycleGAN. Int J Comput Assist Radiol Surg 2019 Oct;14(10):1741-1750. [doi: 10.1007/s11548-019-02042-9] [Medline: 31378841]
12. Ben-Cohen A, Klang E, Raskin SP, Soffer S, Ben-Haim S, Konen E, et al. Cross-modality synthesis from CT to PET using FCN and GAN networks for improved automated lesion detection. Engineering Applications of Artificial Intelligence 2019 Feb;78:186-194. [doi: 10.1016/j.engappai.2018.11.013]
13. Dar SU, Yurt M, Karacan L, Erdem A, Erdem E, Cukur T. Image Synthesis in Multi-Contrast MRI With Conditional Generative Adversarial Networks. IEEE Trans. Med. Imaging 2019 Oct;38(10):2375-2388. [doi: 10.1109/tmi.2019.2901750]
14. Jiang J, Hu Y, Tyagi N, Zhang P, Rimner A, Deasy JO, et al. Cross-modality (CT-MRI) prior augmented deep learning for robust lung tumor segmentation from small MR datasets. Med Phys 2019 Oct;46(10):4392-4404 [FREE Full text] [doi: 10.1002/mp.13695] [Medline: 31274206]

15. Lei Y, Harms J, Wang T, Liu Y, Shu H, Jani AB, et al. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. Med Phys 2019 Aug;46(8):3565-3581 [FREE Full text] [doi: 10.1002/mp.13617] [Medline: 31112304]
16. Vitale S, Orlando JI, Iarussi E, Larrabide I. Improving realism in patient-specific abdominal ultrasound simulation using CycleGANs. Int J Comput Assist Radiol Surg 2020 Feb;15(2):183-192. [doi: 10.1007/s11548-019-02046-5] [Medline: 31392671]
17. Kang E, Koo HJ, Yang DH, Seo JB, Ye JC. Cycle-consistent adversarial denoising network for multiphase coronary CT angiography. Med Phys 2019 Feb;46(2):550-562. [doi: 10.1002/mp.13284] [Medline: 30449055]
18. Kim KH, Do W, Park S. Improving resolution of MR images with an adversarial network incorporating images with different contrast. Med Phys 2018 Jul;45(7):3120-3131. [doi: 10.1002/mp.12945] [Medline: 29729006]
19. Liang X, Chen L, Nguyen D, Zhou Z, Gu X, Yang M, et al. Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy. Phys Med Biol 2019 Jun 10;64(12):125002. [doi: 10.1088/1361-6560/ab22f9] [Medline: 31108465]
20. You C, Cong W, Wang G, Yang Q, Shan H, Gjesteby L, et al. Structurally-Sensitive Multi-Scale Deep Neural Network for Low-Dose CT Denoising. IEEE Access 2018;6:41839-41855. [doi: 10.1109/access.2018.2858196]
21. Dong X, Lei Y, Wang T, Thomas M, Tang L, Curran WJ, et al. Automatic multiorgan segmentation in thorax CT images using U-net-GAN. Med Phys 2019 May;46(5):2157-2168 [FREE Full text] [doi: 10.1002/mp.13458] [Medline: 30810231]
22. Liu X, Guo S, Zhang H, He K, Mu S, Guo Y, et al. Accurate colorectal tumor segmentation for CT scans based on the label assignment generative adversarial network. Med Phys 2019 Aug;46(8):3532-3542. [doi: 10.1002/mp.13584] [Medline: 31087327]
23. Seah JCY, Tang JSN, Kitchen A, Gaillard F, Dixon AF. Chest Radiographs in Congestive Heart Failure: Visualizing Neural Network Learning. Radiology 2019 Feb;290(2):514-522. [doi: 10.1148/radiol.2018180887] [Medline: 30398431]
24. Xue Y, Xu T, Zhang H, Long LR, Huang X. SegAN: Adversarial Network with Multi-scale L Loss for Medical Image Segmentation. Neuroinformatics 2018 Oct;16(3-4):383-392. [doi: 10.1007/s12021-018-9377-x] [Medline: 29725916]
25. Schlegl T, Seeböck P, Waldstein S, Schmidt-Erfurth U, Langs G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. 2017 Presented at: International conference on information processing in medical imaging; 25-30 June 2017; Boone, United States. [doi: 10.1007/978-3-319-59050-9_12]
26. Alex V, Chennamsetty S, Krishnamurthi G. Generative adversarial networks for brain lesion detection. 2017 Presented at: The international society for optics and photonics; 26 February-2 March 2017; San Jose, California, United States. [doi: 10.1117/12.2254487]
27. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv. 2015. URL: https://arxiv.org/abs/1511.06434 [accessed 2021-01-02]
28. Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANS for improved quality, stability, and variation. arXiv. 2017. URL: https://arxiv.org/abs/1710.10196 [accessed 2021-01-02]
29. Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. IEEE Trans Pattern Anal Mach Intell 2020 Feb 02;PP. [doi: 10.1109/TPAMI.2020.2970919] [Medline: 32012000]
30. Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. arXiv. 2019. URL: https://arxiv.org/abs/1809.11096 [accessed 2021-01-02]
31. Mirsky Y, Mahler T, Shelef I, Elovici Y. CT-GAN: malicious tampering of 3D medical imagery using deep learning. arXiv. 2019. URL: https://arxiv.org/abs/1901.03597 [accessed 2021-01-02]
32. Monnier-Cholley L, Carrat F, Cholley BP, Tubiana J, Arrivé L. Detection of lung cancer on radiographs: receiver operating characteristic analyses of radiologists', pulmonologists', and anesthesiologists' performance. Radiology 2004 Dec;233(3):799-805. [doi: 10.1148/radiol.2333031478] [Medline: 15486213]
33. Quekel LG, Kessels AG, Goei R, van Engelshoven JM. Detection of lung cancer on the chest radiograph: a study on observer performance. European Journal of Radiology 2001 Aug;39(2):111-116. [doi: 10.1016/s0720-048x(01)00301-1]
34. Killoran JH, Gerbaudo VH, Mamede M, Ionascu D, Park S, Berbeco R. Motion artifacts occurring at the lung/diaphragm interface using 4D CT attenuation correction of 4D PET scans. J Appl Clin Med Phys 2011 Nov 15;12(4):3502 [FREE Full text] [doi: 10.1120/jacmp.v12i4.3502] [Medline: 22089005]
35. Cai J, Lu L, Xing F. Pancreas segmentation in CT and MRI images via domain specific network designing and recurrent neural contextual learning. arXiv 2018.
36. Bermudez C, Plassard A, Davis L, Newton A, Resnick S, Landman B. Learning Implicit Brain MRI Manifolds with Deep Learning. Proc SPIE Int Soc Opt Eng 2018 Mar;10574 [FREE Full text] [doi: 10.1117/12.2293515] [Medline: 29887659]
37. Bowles C, Chen L, Guerrero R, Bentley P, Gunn R, Hammers A. Gan augmentation: Augmenting training data using generative adversarial networks. arXiv 2018.
38. Bowles C, Gunn R, Hammers A, Rueckert D. Modelling the progression of Alzheimer's disease in MRI using generative adversarial networks. Medical Imaging 2018: Image Processing; 2018 Presented at: The international society for optics and photonics; 25 February-1 March 2018; San Jose, California, United States. [doi: 10.1117/12.2293256]

39. Chuquicusma M, Hussein S, Burt J, Bagci U. How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. 2018 Presented at: 2018 IEEE 15th international symposium on biomedical imaging; 4-7 April 2018; Washington, D.C. United States. [doi: 10.1109/isbi.2018.8363564]
40. Zhang L, Gooya A, Frangi A. Semi-supervised assessment of incomplete LV coverage in cardiac MRI using generative adversarial nets. 2017 Presented at: 2017 International Workshop on Simulation and Synthesis in Medical Imaging; 10 September 2017; Québec City, Canada. [doi: 10.1007/978-3-319-68127-6_7]

Abbreviations

AP-CT: abdominopelvic computed tomography
Chest-L: chest computed tomography-lung window
Chest-M: chest computed tomography-mediastinal window
CT: computed tomography
GAN: generative adversarial network
PGGAN: progressive growing generative adversarial network

Edited by C Lovis; submitted 10.08.20; peer-reviewed by HC Lee, H Arabnia; comments to author 21.09.20; revised version received 15.11.20; accepted 20.02.21; published 17.03.21.

Please cite as:
Park HY, Bae HJ, Hong GS, Kim M, Yun J, Park S, Chung WJ, Kim N
Realistic High-Resolution Body Computed Tomography Image Synthesis by Using Progressive Growing Generative Adversarial Network: Visual Turing Test
JMIR Med Inform 2021;9(3):e23328
URL: https://medinform.jmir.org/2021/3/e23328
doi:10.2196/23328
PMID:33609339

Original Paper

Machine Learning Approach to Predict the Probability of Recurrence of Renal Cell Carcinoma After Surgery: Prediction Model Development Study

HyungMin Kim[1,2], MSc; Sun Jung Lee[1,2], BSc; So Jin Park[1,2], MSc; In Young Choi[1,2*], PhD; Sung-Hoo Hong[3*], MD, PhD

[1]Department of Medical Informatics, College of Medicine, The Catholic University, Seoul, Republic of Korea

[2]Department of Biomedicine & Health Sciences, College of Medicine, The Catholic University, Seoul, Republic of Korea

[3]Department of Urology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Sung-Hoo Hong, MD, PhD
Department of Urology
Seoul St. Mary's Hospital
College of Medicine, The Catholic University
222, Banpo-daero, Seocho-gu
Seoul
Republic of Korea
Phone: 82 2 2258 6228
Email: toomey@catholic.ac.kr

Abstract

Background: Renal cell carcinoma (RCC) has a high recurrence rate of 20% to 30% after nephrectomy for clinically localized disease, and more than 40% of patients eventually die of the disease, making regular monitoring and constant management of utmost importance.

Objective: The objective of this study was to develop an algorithm that predicts the probability of recurrence of RCC within 5 and 10 years of surgery.

Methods: Data from 6849 Korean patients with RCC were collected from eight tertiary care hospitals listed in the KOrean Renal Cell Carcinoma (KORCC) web-based database. To predict RCC recurrence, analytical data from 2814 patients were extracted from the database. Eight machine learning algorithms were used to predict the probability of RCC recurrence, and the results were compared.

Results: Within 5 years of surgery, the highest area under the receiver operating characteristic curve (AUROC) was obtained from the naïve Bayes (NB) model, with a value of 0.836. Within 10 years of surgery, the highest AUROC was obtained from the NB model, with a value of 0.784.

Conclusions: An algorithm was developed that predicts the probability of RCC recurrence within 5 and 10 years using the KORCC database, a large-scale RCC cohort in Korea. It is expected that the developed algorithm will help clinicians manage prognosis and establish customized treatment strategies for patients with RCC after surgery.

(JMIR Med Inform 2021;9(3):e25635) doi:10.2196/25635

KEYWORDS

renal cell carcinoma; recurrence; machine learning; naïve Bayes; algorithm; cancer; surgery; web-based; database; prediction; probability; carcinoma; kidney; model; development

Introduction

Renal cell carcinoma (RCC) accounts for 90% of malignant tumors in the kidney and is twice as common in men as in women [1]. Kidney cancer, therefore, generally refers to RCC. It is the sixth most frequently diagnosed cancer in men and the 10th most frequently diagnosed cancer in women worldwide [2]. According to the cancer statistics from the National Cancer

Center, the number of new kidney cancer cases in Korea in 2017 was 5299, accounting for approximately 2.3% of the total of 232,255 cancer cases. Further, the incidence of kidney cancer per 100,000 people has been increasing since 1999 [3]. RCC is one of the most lethal types of malignant tumors in urology, with approximately 20% to 30% of patients with RCC suffering from metastatic diseases, and more than 40% of patients eventually die of the disease [4-6]. The main treatment for RCC is radical nephrectomy; for small tumors, partial nephrectomy is performed to preserve kidney function [7].

RCC can be completely cured through full surgical resection if there is no evidence of preoperative metastatic disease. However, it has a high recurrence rate of 20% to 30% [8,9], and approximately 50% of recurrences occur within 2 years [8,10]. RCC recurrence is generally classified as early recurrence or late recurrence based on the 5-year threshold [11]. Most recurrences occur during the early recurrence period (within 5 years) [11,12], whereas approximately 10% occur during the late recurrence period (after 5 years) [11,13].

RCC is generally resistant to radiation and chemotherapy, making treatment of its recurrence difficult [4]. Therefore, it is necessary to predict the probability of RCC recurrence so that risk factors can be managed in advance. The Memorial Sloan Kettering Cancer Center (MSKCC) in the United States developed a nomogram that predicts the probability of recurrence within 5 years using the symptoms and histology of 601 patients with kidney cancer who received surgical treatment in 2001 [14]. Additionally, in 2005, a nomogram was developed to predict the recurrence probability within 5 years using the pathological stage, Fuhrman nuclear grade, tumor size, necrosis, vascular invasion, and clinical presentation variables of 701 patients with kidney cancer [15]. Previous studies have used small-scale RCC cohorts from single institutions, and the data have included censored data, where the values of the observations were only partially known. If censored data are included, they can be applied in the Cox proportional hazards model, a standard statistical technique for modeling censored data, but they are difficult to apply to other machine learning (ML) techniques [16].

In this study, we used a multicenter, large-scale RCC cohort collected from eight tertiary care hospitals in Korea; we removed censored data and used only the fully observed data. ML focuses on building new predictive models by performing extensive searches on multiple models and parameters and then performing validation [17]. The objective of this study was to develop an algorithm that could predict the recurrence probability of RCC after surgery within 5 and 10 years by applying eight representative ML algorithms to a large-scale Korean RCC cohort. Using the developed algorithm, clinicians can manage postoperative patient outcomes and establish personalized treatment strategies.

Methods

Study Population

The data used in this study were obtained from a large-scale cohort of Korean patients with RCC assembled from the KOrean Renal Cell Carcinoma (KORCC) web-based database. It consisted of 206 variables, including demographic information such as age, height, and weight, as well as pathological information, including clinical stage, pathological stage, Fuhrman nuclear grade, and survival period [18]. The study protocol was approved by the institutional review board of the Catholic University of Korea (IRB No. KC20ZIDI0966). The data of 6849 patients who participated in the KORCC study group as of July 1, 2015, were collected from eight tertiary hospitals.

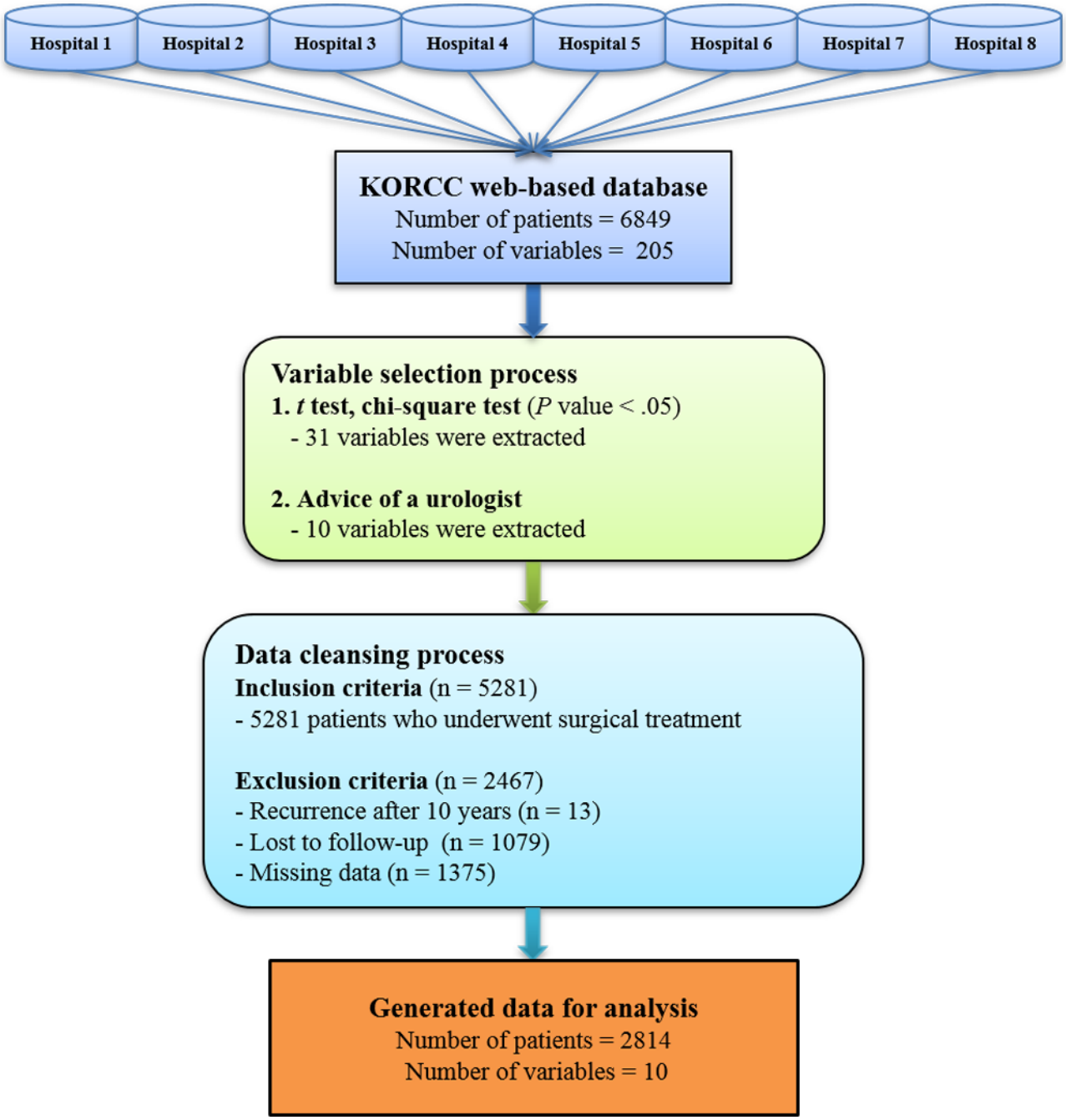
Variable Selection and Data Cleansing

The *t* test for continuous variables and the chi-square test for categorical variables were used to explore variables that significantly affect recurrence. In both tests, variables with missing values were removed to ensure that the data used were complete and without missing values. At a significance level of P=.05, we first extracted 31 variables showing significant differences between the recurring and nonrecurring groups. Of the 31 variables extracted, 10 variables that had significant effects on recurrence in actual clinical trials were finally extracted based on the expert advice of a urologist. The final 10 selected variables were gender, age, BMI, smoking, pathological tumor stage, histological type, necrosis, lymphovascular invasion, capsular invasion, and Fuhrman nuclear grade.

Several studies reported that age ≥60 years, Fuhrman nuclear grade ≥3, and pathological stage ≥pT2 were statistically associated with RCC recurrence [19]. In addition, women had better prognoses after surgery than men [20], and individuals with higher BMIs showed better prognoses than those with normal or lower BMIs [21]. Furthermore, the prognoses of smokers were worse than those of nonsmokers [22], and pathological variables such as histological type [23], necrosis [24], lymphovascular invasion [11], and capsular invasion [25] were all related to the recurrence of RCC.

Next, we cleansed the data to present them in a form suitable for analysis. Of the 6849 patients, only 5281 patients who received surgical treatment were included in the analysis. Of those 5281 patients, 13 patients with recurrence after 10 years, 1079 lost to follow-up, and 1375 with missing values in 10 variables were excluded from the analysis. Finally, a subset of 2814 patients with values for 10 variables was available for analysis (Figure 1).

Figure 1. Data generation process for analysis. KORCC: Korean Renal Cell Carcinoma.



Dealing with the Imbalanced Data Set

One of the most frequent problems in applying ML classification algorithms is data imbalance [26,27]. In the medical field, data asymmetry occurs between normal and abnormal classes because most patients are concentrated in the "normal" class, whereas relatively few—such as patients with cancer—are in the "abnormal" class. In this case, the ML algorithm attempts to improve the performance by predicting normal classes, in which most patients are concentrated, resulting in lower predictability of abnormal classes with small numbers of patients [27]. However, from a research perspective, it is more important to predict abnormal classes; hence, it is necessary to deal with the imbalanced data.

In this study, the synthetic minority oversampling technique (SMOTE) was applied to the training data set to solve the imbalance problem. SMOTE is an oversampling method that is widely used when ML is applied to data with high imbalance [28,29]. Before applying SMOTE, the ratio of patients in the recurrence group to patients in the nonrecurrence group in the training set was significantly asymmetrical—approximately 1:10; ML was applied after making the ratio of the two groups equal to 1:1 using SMOTE (Table 1). Because the volume of the data set was sufficiently large after SMOTE application, we verified the prediction model using the 20% hold-out validation method with the data partitioning of the training set and test set at 80:20 [30].

Table 1. Distribution of data sets before and after synthetic minority oversampling technique application.

	Training set (n=2251)		Test set (n=563)	
	Recurrence group, n (%)	Nonrecurrence group, n (%)	Recurrence group, n (%)	Nonrecurrence group, n (%)
Before	226 (10.04)	2025 (89.96)	52 (9.24)	511 (90.76)
After	2025 (50.00)	2025 (50.00)	52 (9.24)	511 (90.76)

Statistical Analysis and ML Model Development

In this study, we compared the performance of the following representative ML classification algorithms: kernel support vector machine (SVM) [31], logistic regression [32], decision tree [33], k-nearest neighbor (KNN) [34], naïve Bayes (NB) [35], random forest [36], AdaBoost [36], and gradient boost [37]. For each algorithm, we calculated four values: sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUROC). The algorithm with the highest performance was finally selected based on the AUROC value, which is one of the most important indicators for confirming the performance of a classification model [38]. We used Python (version 3.7.6) for statistical analysis and algorithm development.

Results

Characteristics and Distribution of Patients

We compared the patient characteristics and distribution of each variable between the recurrence and nonrecurrence groups (Table 2).

The mean age of patients in the recurrence group was higher than that of patients in the nonrecurrence group (58.4 years versus 55.4 years, respectively). The average BMIs of patients in the recurrence and nonrecurrence groups were 23.6 kg/m^2 and 24.7 kg/m^2, respectively. The results show the same characteristics as those found in studies that have revealed better prognoses for obese patients [21]. The proportion of smokers in the recurrence and nonrecurrence groups was 25.5% and 20.1%, respectively. The pathology stage—an important variable in predicting recurrence—showed that the proportion of patients with a pathological stage ≥pT2 was approximately 60.4% (168/278) in the recurrence group and 15.2% (386/2536) in the nonrecurrence group. Approximately 77.7% (216/278) of the patients in the recurrence group and 44.8% (1135/2536) of those in the nonrecurrence group had Fuhrman nuclear grades ≥3; thus, the recurrence group had higher Fuhrman nuclear grades. The distribution of each category of pathological variables is shown in Table 2.

Table 2. Baseline characteristics of patients (N=2814).

Variable	Recurrence group (n=278)	Nonrecurrence group (n=2536)
Age (years), mean (SD)	58.4 (11.9)	55.4 (12.7)
BMI (kg/m^2), mean (SD)	23.6 (3.2)	24.7 (3.3)
Gender, n (%)		
Male	212 (76.3)	1811 (71.4)
Female	66 (23.7)	725 (28.6)
Smoking, n (%)		
Nonsmoker	207 (74.5)	2026 (79.9)
Current smoker	71 (25.5)	510 (20.1)
Pathological tumor stage, n (%)		
1a	50 (18.0)	1663 (65.6)
1b	60 (21.6)	487 (19.2)
2a	30 (10.8)	106 (4.2)
2b	12 (4.3)	29 (1.1)
3a	82 (29.5)	201 (7.9)
3b	34 (12.2)	36 (1.4)
3c	1 (0.4)	3 (0.1)
4	9 (3.2)	11 (0.4)
Histologic type, n (%)		
Clear cell	242 (87.1)	2243 (88.4)
Papillary	14 (5.0)	44 (1.7)
Chromophobe	4 (1.4)	180 (7.1)
Collecting duct	5 (1.8)	4 (0.2)
Unclassified	5 (1.8)	15 (0.6)
Multilocular cystic	0 (0.0)	19 (0.7)
Mixed	6 (2.2)	24 (0.9)
Xp11.2 translocation	1 (0.4)	3 (0.1)
Clear cell papillary	1 (0.4)	4 (0.2)
Necrosis, n (%)		
No	143 (51.4)	2272 (89.6)
Microscopic	30 (10.8)	126 (5.0)
Macroscopic	105 (37.8)	138 (5.4)
Lymphovascular invasion, n (%)		
No	200 (71.9)	2436 (96.1)
Yes	78 (28.1)	100 (3.9)
Capsular invasion, n (%)		
No	148 (53.2)	2114 (83.4)
Yes	130 (46.8)	422 (16.6)
Fuhrman nuclear grade, n (%)		
1	5 (1.8)	108 (4.3)
2	57 (20.5)	1293 (51.0)
3	141 (50.7)	1008 (39.7)
4	75 (27.0)	127 (5.0)

Prediction Model Performance

We trained eight ML algorithms on the training data set and calculated the sensitivity, specificity, accuracy, and AUROC values using the test data set (Table 3). The NB algorithm showed higher performance than the other algorithms, with an AUROC of 0.836 within 5 years and 0.784 within 10 years. The NB approach calculates the conditional probability, which is the likelihood that a conclusion will be observed based on the evidence given [35]. The NB algorithm is simple and fast [39] and has proven effective in text classification and medical diagnosis [40,41]. However, the NB approach has a limitation in that its prediction probability becomes zero when a new value that is not in the training data set is entered; Laplace smoothing is a means of solving this problem [42]. The predictive model we developed also had a problem in that the probability value became zero when a new type of data that was not in the training data set was entered; hence, the algorithm was optimized by adjusting the α value—a parameter in Laplace smoothing (Table 4).

Table 3. Diagnostic performance of machine learning algorithms for the prediction of renal cell carcinoma recurrence.

Algorithm (parameter name) and parameter value (in 5 years, in 10 years)	Sensitivity		Specificity		Accuracy		AUROC[a]	
	5-year	10-year	5-year	10-year	5-year	10-year	5-year	10-year
Kernel SVM[b,c]	0.733	0.673	0.805	0.853	0.800	0.837	0.769	0.763
Logistic regression[c]	0.644	0.692	0.839	0.816	0.823	0.805	0.741	0.754
Decision tree[c]	0.533	0.442	0.866	0.869	0.839	0.829	0.700	0.656
KNN[d] (n-neighbors)								
(100, 100)[c]	0.556	0.519	0.905	0.898	0.877	0.863	0.730	0.709
(10, 10)	0.467	0.426	0.947	0.928	0.909	0.881	0.707	0.675
(50, 50)	0.511	0.461	0.931	0.922	0.898	0.879	0.722	0.692
(200, 200)	0.556	0.481	0.899	0.902	0.871	0.863	0.727	0.691
NB[e] (alpha)								
(10, 100)[c]	0.822	0.731	0.850	0.828	0.848	0.819	0.836	0.784
Random forest (number of trees)								
(5, 5)[c]	0.578	0.500	0.858	0.853	0.835	0.821	0.718	0.677
(10, 10)	0.511	0.423	0.866	0.861	0.837	0.821	0.688	0.642
(50, 50)	0.511	0.442	0.875	0.861	0.846	0.822	0.693	0.652
(100, 100)	0.511	0.462	0.864	0.861	0.835	0.824	0.687	0.661
AdaBoost (number of trees)								
(50, 200)[c]	0.733	0.692	0.815	0.810	0.809	0.800	0.774	0.751
(10, 10)	0.600	0.577	0.895	0.845	0.871	0.821	0.747	0.711
(50, 50)	0.733	0.673	0.815	0.824	0.809	0.810	0.774	0.748
(100, 100)	0.711	0.692	0.835	0.802	0.825	0.792	0.773	0.747
(200, 200)	0.711	0.692	0.837	0.810	0.826	0.800	0.774	0.751
Gradient boost (number of trees)								
(50, 100)[c]	0.688	0.635	0.819	0.826	0.809	0.808	0.754	0.730
(10, 10)	0.756	0.596	0.667	0.849	0.674	0.825	0.711	0.723
(50, 50)	0.688	0.615	0.819	0.826	0.809	0.806	0.754	0.721
(100, 100)	0.555	0.635	0.823	0.826	0.805	0.808	0.711	0.730
(200, 200)	0.533	0.558	0.848	0.832	0.823	0.806	0.691	0.695

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

[b]SVM: support vector machine.

[c]Final algorithms selected by adjusting parameters.

[d]KNN: k-nearest neighbor.

[e]NB: naïve Bayes.

Table 4. Performance according to the α value in the naïve Bayes model.

α value	Sensitivity		Specificity		Accuracy		AUROC[a]	
	5-year	10-year	5-year	10-year	5-year	10-year	5-year	10-year
0 (no smoothing)	0.800	0.731	0.848	0.828	0.844	0.819	0.824	0.779
1	0.822	0.731	0.848	0.828	0.846	0.819	0.835	0.779
10	0.822	0.731	0.850	0.834	0.848	0.824	0.836	0.782
20	0.800	0.731	0.850	0.834	0.846	0.824	0.825	0.782
30	0.800	0.731	0.852	0.834	0.848	0.824	0.826	0.782
100	0.800	0.731	0.854	0.840	0.850	0.828	0.827	0.784
200	0.756	0.692	0.860	0.845	0.852	0.831	0.807	0.769

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

For predictions within 5 years, the AUROC was found to be 0.836 when α=10, which was the highest performance compared with that before smoothing was applied (α= 0, AUROC 0.824). For predictions within 10 years, the AUROC was 0.784 when α=100, which was the highest performance compared with that before smoothing was applied (α=0, AUROC 0.779). When comparing the area by drawing the ROC curve of the prediction algorithm within 5 and 10 years, the NB curve line was close to the upper left corner, which means that the area for that algorithm was the widest (Figures 2 and 3).

Figure 2. Receiver operating characteristic (ROC) curves of recurrence prediction algorithms within 5 years. KNN: k-nearest neighbor; SVM: support vector machine.

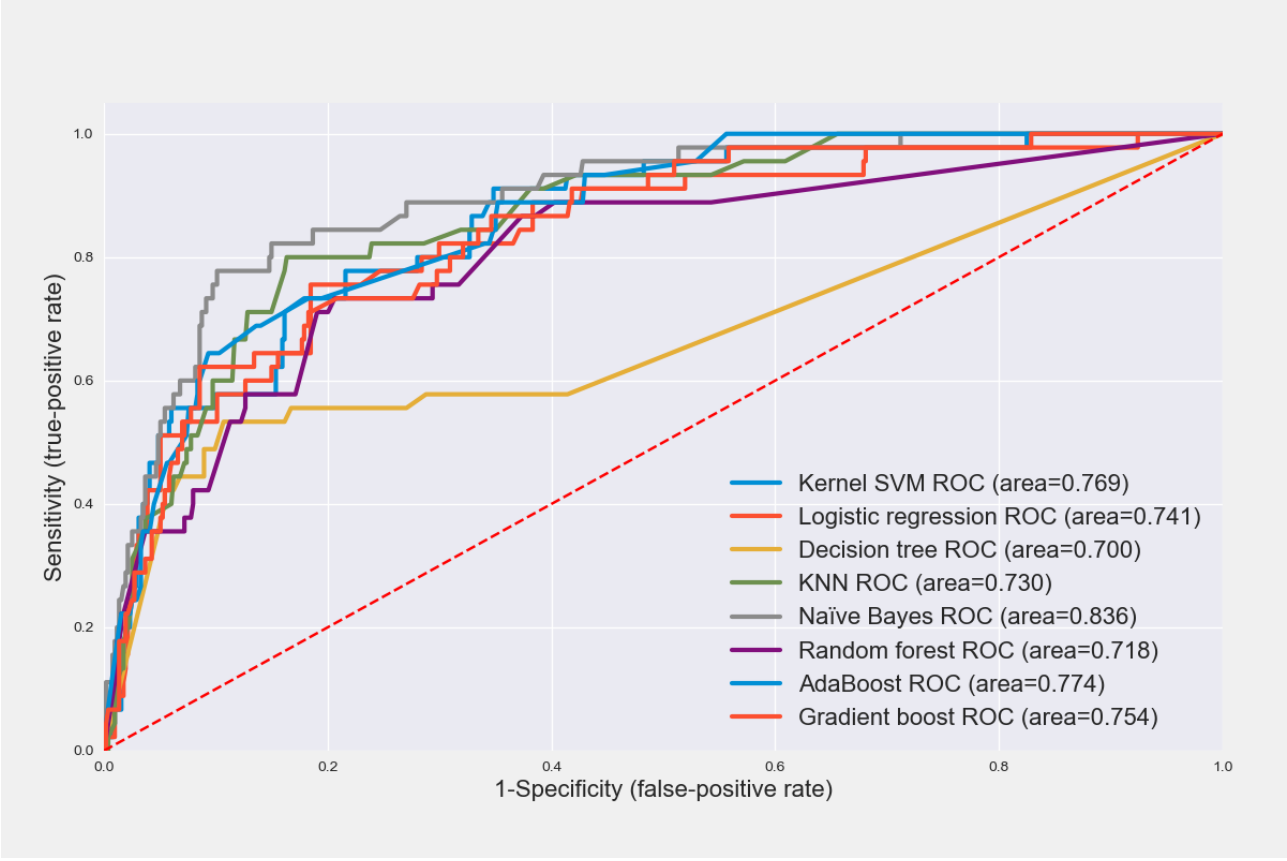
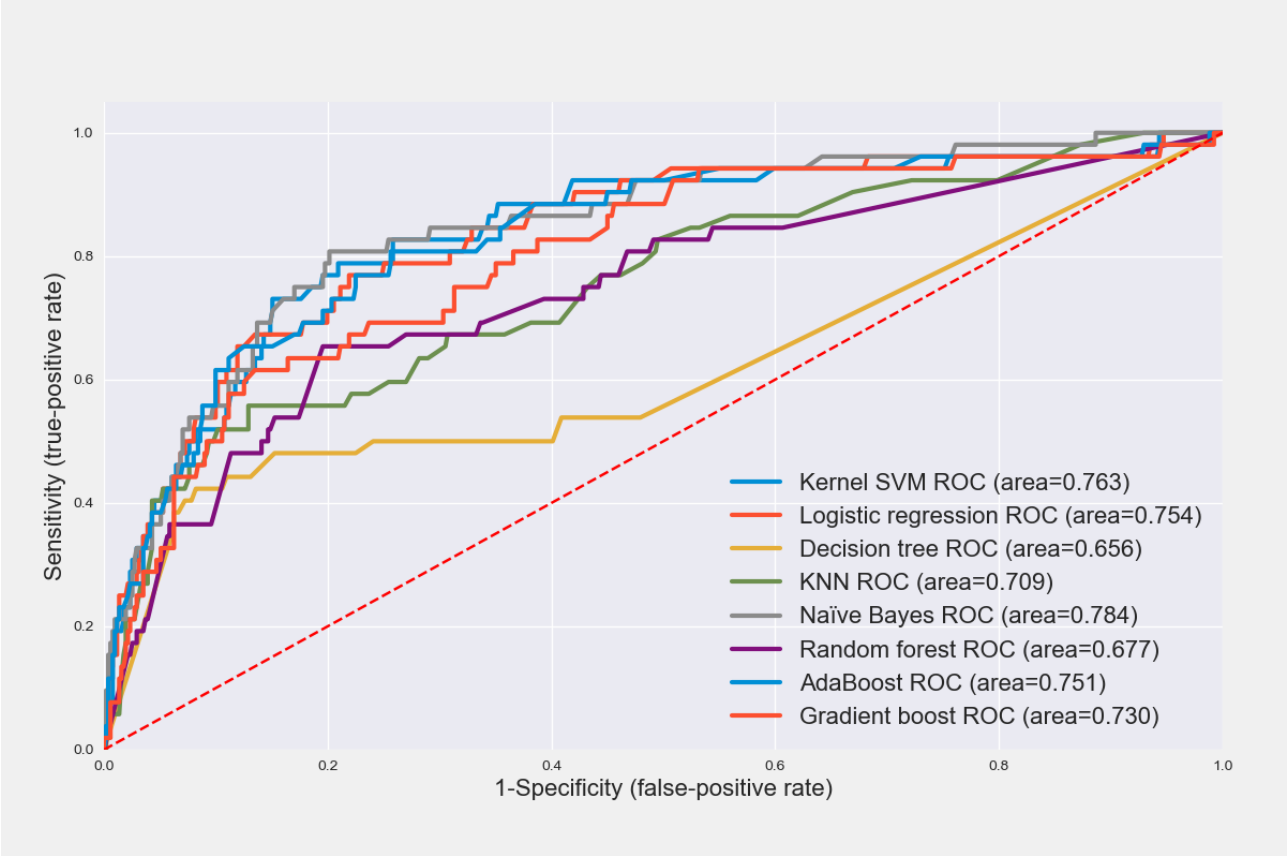


Figure 3. Receiver operating characteristic (ROC) curves of recurrence prediction algorithms within 10 years. KNN: k-nearest neighbor; SVM: support vector machine.



Discussion

Principal Findings

In this study, we developed an algorithm to predict the probability of RCC recurrence within 10 years by selecting 10 variables that significantly affect recurrence. The AUROC of the algorithm was 0.84 for models of recurrence within 5 years and 0.79 for models of recurrence within 10 years. Our proposed algorithm achieved better prediction performance than the previously developed 5-year prediction algorithm by MSKCC, which yielded AUROCs of 0.74 [14] and 0.82 [15].

In the previous studies, 66 recurrences in 601 patients [14] and 72 recurrences in 701 patients [15] were used to form the data set for analysis. Because the data were collected from a single institution, the scale was small, and the data included censored data. The methods that can be applied to analyze censored data are limited. Therefore, in previous studies, an algorithm was developed using the Cox proportional hazards model—the most representative survival analysis method—and its performance was presented.

Because the results of previous studies were based on a single institutional analysis, the characteristics of patients in various regions were likely not reflected, meaning biased results may have been obtained. Thus, a data set composed of data from eight institutions in various regions of Korea was used in this study. In our data, 278 out of 2814 patients experienced RCC recurrence, and censored data were not included. We attempted to improve the prediction performance using more diverse and significant variables than those used by the prediction algorithms in previous studies. Finally, we developed a prediction algorithm by applying ML techniques that are typically used in classification tasks. Because we used large-scale data that sufficiently reflect the characteristics of patients with RCC in Korea, the proposed algorithm achieved stable results with high accuracy and low bias.

To the best of our knowledge, this is the first study to predict the recurrence of RCC within 10 years after surgery using ML techniques. The recurrence of most cancers is typically within 5 years. Because RCC has a late recurrence [12], it is vital to predict the late recurrence in advance and establish a personalized treatment strategy for managing the prognosis of patients with RCC. Thus, our study makes an important contribution by accurately predicting the likelihood of late recurrence of RCC.

Limitations

We utilized the data of patients with RCC recurrence after 1 to 10 years in the recurrence prediction model within 10 years. However, in several studies, a difference between variables that affect early recurrence and late recurrence was observed [12,43]. Therefore, the prediction models for 1 to 5 years and 5 to 10 years should be distinct from each other and should be constructed using different combinations of variables. However, despite being a large cohort representing the whole of Korea, it was difficult to create a single model, as only 23 cases occurred after 5 to 10 years. Therefore, in this study, we developed a predictive model by integrating both groups within

10 years. Hence, the algorithm for within 10 years seems to have lower performance than the model for within 5 years because of the heterogeneity between the 1- to 5-year recurrence group and the 5- to 10-year recurrence group. We plan to develop additional stable and accurate models to predict late recurrence when data are collected after 5 to 10 years.

Furthermore, we used large-scale cohort data showing the characteristics of patients with RCC in Korea. Therefore, the algorithm we developed exhibits stable performance when applied to Korean patients with RCC. However, patients with RCC have different demographic and clinical characteristics; hence, the performance may be reduced when applied to different ethnicities [44,45].

Conclusions

Using the KORCC database, a large-scale cohort of RCC in Korea, we developed an algorithm to predict the probability of RCC recurrence after surgery using a representative ML technique. Among the eight ML algorithms, the NB algorithm showed the best diagnostic performance in both the 5-year model and the 10-year model in terms of the AUROC. The developed algorithm can help clinicians establish postoperative prognosis management and personalized treatment strategies for patients with RCC.

Acknowledgments

This study was supported by the R&D Performance Creation Promotion Project 2019 of Seoul St Mary's Hospital. We thank the KOrean Renal Cell Carcinoma (KORCC) group for assisting us in analyzing the data.

Authors' Contributions

HMK contributed to the work as the first author. SJL and SJP contributed to data preparation and discussion. IYC and S-HH equally supervised the entire process as corresponding authors.

Conflicts of Interest

None declared.

References

1. Choueiri TK, Motzer RJ. Systemic Therapy for Metastatic Renal-Cell Carcinoma. N Engl J Med 2017 Jan 26;376(4):354-366. [doi: 10.1056/nejmra1601333]
2. Capitanio U, Bensalah K, Bex A, Boorjian SA, Bray F, Coleman J, et al. Epidemiology of Renal Cell Carcinoma. European Urology 2019 Jan;75(1):74-84. [doi: 10.1016/j.eururo.2018.08.036]
3. Hong S, Won Y, Park YR, Jung K, Kong H, Lee ES. Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2017. Cancer Res Treat 2020 Apr;52(2):335-350. [doi: 10.4143/crt.2020.206]
4. Chin AI, Lam JS, Figlin RA, Belldegrun AS. Surveillance strategies for renal cell carcinoma patients following nephrectomy. Rev Urol 2006;8(1):1-7 [FREE Full text] [Medline: 16985554]
5. Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, et al. Cancer statistics, 2005. CA Cancer J Clin 2005;55(1):10-30 [FREE Full text] [doi: 10.3322/canjclin.55.1.10] [Medline: 15661684]
6. Janzen NK, Kim HL, Figlin RA, Belldegrun AS. Surveillance after radical or partial nephrectomy for localized renal cell carcinoma and management of recurrent disease. Urol Clin North Am 2003 Nov;30(4):843-852. [doi: 10.1016/s0094-0143(03)00056-9] [Medline: 14680319]
7. Jang HA, Kim JW, Byun SS, Hong SH, Kim YJ, Park YH, et al. Oncologic and Functional Outcomes after Partial Nephrectomy Versus Radical Nephrectomy in T1b Renal Cell Carcinoma: A Multicenter, Matched Case-Control Study in Korean Patients. Cancer Res Treat 2016 Apr;48(2):612-620 [FREE Full text] [doi: 10.4143/crt.2014.122] [Medline: 26044158]
8. Tyson MD, Chang SS. Optimal Surveillance Strategies After Surgery for Renal Cell Carcinoma. J Natl Compr Canc Netw 2017 Jun;15(6):835-840. [doi: 10.6004/jnccn.2017.0102] [Medline: 28596262]
9. van der Mijn JC, Al Hussein Al Awamlh B, Islam Khan A, Posada-Calderon L, Oromendia C, Fainberg J, et al. Validation of risk factors for recurrence of renal cell carcinoma: Results from a large single-institution series. PLoS One 2019;14(12):e0226285 [FREE Full text] [doi: 10.1371/journal.pone.0226285] [Medline: 31815952]
10. Quinlan M, Wei G, Davis N, Poyet C, Perera M, Bolton D, et al. Renal Cell Carcinoma Follow-Up - Is it Time to Abandon Ultrasound? Curr Urol 2019 Sep;13(1):19-24 [FREE Full text] [doi: 10.1159/000499299] [Medline: 31579200]
11. Acar Ö, Şanlı Ö. Surgical Management of Local Recurrences of Renal Cell Carcinoma. Surg Res Pract 2016;2016:2394942 [FREE Full text] [doi: 10.1155/2016/2394942] [Medline: 26925458]
12. Park Y, Baik K, Lee Y, Ku J, Kim H, Kwak C. Late recurrence of renal cell carcinoma >5 years after surgery: clinicopathological characteristics and prognosis. BJU Int 2012 Dec;110(11 Pt B):E553-E558. [doi: 10.1111/j.1464-410X.2012.11246.x] [Medline: 22578274]
13. Kirkali Z, Van Poppel H. A critical analysis of surgery for kidney cancer with vena cava invasion. Eur Urol 2007 Sep;52(3):658-662. [doi: 10.1016/j.eururo.2007.05.009] [Medline: 17548146]

14. Kattan MW, Reuter V, Motzer RJ, Katz J, Russo P. A postoperative prognostic nomogram for renal cell carcinoma. J Urol 2001 Jul;166(1):63-67. [Medline: 11435824]
15. Sorbellini M, Kattan MW, Snyder ME, Reuter V, Motzer R, Goetzl M, et al. A postoperative prognostic nomogram predicting recurrence for patients with conventional clear cell renal cell carcinoma. J Urol 2005 Jan;173(1):48-51. [doi: 10.1097/01.ju.0000148261.19532.2c] [Medline: 15592023]
16. Zupan B, Demsar J, Kattan MW, Beck J, Bratko I. Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artif Intell Med 2000 Aug;20(1):59-75. [doi: 10.1016/s0933-3657(00)00053-1] [Medline: 11185421]
17. Mani S, Ozdas A, Aliferis C, Varol HA, Chen Q, Carnevale R, et al. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. J Am Med Inform Assoc 2014;21(2):326-336 [FREE Full text] [doi: 10.1136/amiajnl-2013-001854] [Medline: 24043317]
18. Byun S, Hong SK, Lee S, Kook HR, Lee E, Kim HH, et al. The establishment of KORCC (KOrean Renal Cell Carcinoma) database. Investig Clin Urol 2016 Jan;57(1):50-57 [FREE Full text] [doi: 10.4111/icu.2016.57.1.50] [Medline: 26966726]
19. Lee SH, Son HS, Cho S, Kim SJ, Yoo DS, Kang SH, et al. Which Patients Should We Follow up beyond 5 Years after Definitive Therapy for Localized Renal Cell Carcinoma? Cancer Res Treat 2015 Jul;47(3):489-494 [FREE Full text] [doi: 10.4143/crt.2014.013] [Medline: 25622589]
20. Fukushima H, Saito K, Yasuda Y, Tanaka H, Patil D, Cotta BH, et al. Female Gender Predicts Favorable Prognosis in Patients With Non-metastatic Clear Cell Renal Cell Carcinoma Undergoing Curative Surgery: Results From the International Marker Consortium for Renal Cancer (INMARC). Clin Genitourin Cancer 2020 Apr;18(2):111-116.e1. [doi: 10.1016/j.clgc.2019.10.027] [Medline: 32001181]
21. Choi Y, Park B, Jeong BC, Seo SI, Jeon SS, Choi HY, et al. Body mass index and survival in patients with renal cell carcinoma: a clinical-based cohort and meta-analysis. Int J Cancer 2013 Feb 01;132(3):625-634 [FREE Full text] [doi: 10.1002/ijc.27639] [Medline: 22610826]
22. Xu Y, Qi Y, Zhang J, Lu Y, Song J, Dong B, et al. The impact of smoking on survival in renal cell carcinoma: a systematic review and meta-analysis. Tumour Biol 2014 Jul;35(7):6633-6640. [doi: 10.1007/s13277-014-1862-8] [Medline: 24699995]
23. Yoo S, You D, Jeong IG, Song C, Hong B, Hong JH, et al. Histologic subtype needs to be considered after partial nephrectomy in patients with pathologic T1a renal cell carcinoma: papillary vs. clear cell renal cell carcinoma. J Cancer Res Clin Oncol 2017 Sep;143(9):1845-1851. [doi: 10.1007/s00432-017-2430-6] [Medline: 28451753]
24. Abel EJ, Raman JD, Shapiro DD, Chan W, Allen GO, Patil D, et al. Defining individual recurrence risk following surgery for high risk non-metastatic renal cell carcinoma. J Clin Oncol 2018 Feb 20;36(6_suppl):664-664. [doi: 10.1200/jco.2018.36.6_suppl.664]
25. Ha U, Lee KW, Jung J, Byun S, Kwak C, Chung J, et al. Renal capsular invasion is a prognostic biomarker in localized clear cell renal cell carcinoma. Sci Rep 2018 Jan 09;8(1):202 [FREE Full text] [doi: 10.1038/s41598-017-18466-9] [Medline: 29317731]
26. Ali A, Shamsuddin SM, Ralescu AL. Classification with class imbalance problem: A review. Int J Adv Soft Comput Appl 2015;7(3):176-204 [FREE Full text]
27. Li D, Liu C, Hu SC. A learning method for the class imbalance problem with medical data sets. Comput Biol Med 2010 May;40(5):509-518. [doi: 10.1016/j.compbiomed.2010.03.005] [Medline: 20347072]
28. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PLoS One 2017;12(7):e0179805 [FREE Full text] [doi: 10.1371/journal.pone.0179805] [Medline: 28738059]
29. Blagus R, Lusa L. Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. BMC Bioinformatics 2015 Nov 04;16:363 [FREE Full text] [doi: 10.1186/s12859-015-0784-9] [Medline: 26537827]
30. Foster KR, Koprowski R, Skufca JD. Machine learning, medical diagnosis, and biomedical engineering research - commentary. Biomed Eng Online 2014 Jul 05;13:94 [FREE Full text] [doi: 10.1186/1475-925X-13-94] [Medline: 24998888]
31. Huang M, Chen C, Lin W, Ke S, Tsai C. SVM and SVM Ensembles in Breast Cancer Prediction. PLoS One 2017;12(1):e0161501 [FREE Full text] [doi: 10.1371/journal.pone.0161501] [Medline: 28060807]
32. Liao J, Chin K. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. Bioinformatics 2007 Aug 01;23(15):1945-1951. [doi: 10.1093/bioinformatics/btm287] [Medline: 17540680]
33. Song Y, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry 2015 Apr 25;27(2):130-135 [FREE Full text] [doi: 10.11919/j.issn.1002-0829.215044] [Medline: 26120265]
34. Deng Z, Zhu X, Cheng D, Zong M, Zhang S. Efficient kNN classification algorithm for big data. Neurocomputing 2016 Jun;195:143-148. [doi: 10.1016/j.neucom.2015.08.112]
35. Subbalakshmi G, Ramesh K, Chinna Rao M. Decision Support in Heart Disease Prediction System using Naive Bayes. Indian J Comput Sci Eng 2011;2(2):170-176 [FREE Full text]
36. Chan JC, Paelinckx D. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sensing of Environment 2008 Jun;112(6):2999-3011. [doi: 10.1016/j.rse.2008.02.011]

37. Chang Y, Chang K, Wu G. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. Applied Soft Computing 2018 Dec;73:914-920. [doi: 10.1016/j.asoc.2018.09.029]
38. Jin Huang, Ling C. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005 Mar;17(3):299-310. [doi: 10.1109/tkde.2005.50]
39. Rennie J, Shih L, Teevan J, Karger D. Tackling the Poor Assumptions of Naive Bayes Text Classifiers Jason. 2003 Presented at: Proc 20th Int Conf Mach Learn. Published online; 2003; Washington DC p. 616-623.
40. Rish I. An empirical study of the naive Bayes classifier. 2001 Presented at: IJCAI 2001 Workshop on empirical methods in artificial intelligence; 2001; Seattle, USA p. 4863-4869.
41. Hellerstein JL, Jayram TS, Rish I. Recognizing end-user transactions in performance management. 2000 Presented at: Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence; 2000; Austin, Texas, USA p. 596-602.
42. Cherian V, Bindu M. Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique. Int J Comput Sci Trends Technol 2017;5(2):68-73 [FREE Full text]
43. Adamy A, Chong KT, Chade D, Costaras J, Russo G, Kaag MG, et al. Clinical characteristics and outcomes of patients with recurrence 5 years after nephrectomy for localized renal cell carcinoma. J Urol 2011 Feb;185(2):433-438. [doi: 10.1016/j.juro.2010.09.100] [Medline: 21167521]
44. Chow W, Shuch B, Linehan WM, Devesa SS. Racial disparity in renal cell carcinoma patient survival according to demographic and clinical characteristics. Cancer 2013 Jan 15;119(2):388-394 [FREE Full text] [doi: 10.1002/cncr.27690] [Medline: 23147245]
45. Olshan AF, Kuo T, Meyer A, Nielsen ME, Purdue MP, Rathmell WK. Racial difference in histologic subtype of renal cell carcinoma. Cancer Med 2013 Oct;2(5):744-749 [FREE Full text] [doi: 10.1002/cam4.110] [Medline: 24403240]

Abbreviations

AUROC: area under the receiver operating characteristic curve
KNN: k-nearest neighbor
KORCC: KOrean Renal Cell Carcinoma
ML: machine learning
MSKCC: Memorial Sloan Kettering Cancer Center
NB: naïve Bayes
RCC: renal cell carcinoma
SMOTE: synthetic minority oversampling technique
SVM: support vector machine

Edited by G Eysenbach; submitted 11.12.20; peer-reviewed by X Zhang; comments to author 13.01.21; revised version received 23.01.21; accepted 29.01.21; published 01.03.21.

Please cite as:
Kim H, Lee SJ, Park SJ, Choi IY, Hong SH
Machine Learning Approach to Predict the Probability of Recurrence of Renal Cell Carcinoma After Surgery: Prediction Model Development Study
JMIR Med Inform 2021;9(3):e25635
URL: https://medinform.jmir.org/2021/3/e25635
doi:10.2196/25635
PMID:33646127

Original Paper

Predictive Modeling of 30-Day Emergency Hospital Transport of German Patients Using a Personal Emergency Response: Retrospective Study and Comparison with the United States

Jorn op den Buijs[1], MSc, PhD; Marten Pijl[1], PhD; Andreas Landgraf[2], PhD

[1]Philips Research, Eindhoven, Netherlands

[2]Philips DACH, Hamburg, Germany

Corresponding Author:
Jorn op den Buijs, MSc, PhD
Philips Research
High Tech Campus 34
Eindhoven, 5656 AE
Netherlands
Phone: 31 631926890
Email: jorn.op.den.buijs@philips.com

Abstract

Background: Predictive analytics based on data from remote monitoring of elderly via a personal emergency response system (PERS) in the United States can identify subscribers at high risk for emergency hospital transport. These risk predictions can subsequently be used to proactively target interventions and prevent avoidable, costly health care use. It is, however, unknown if PERS-based risk prediction with targeted interventions could also be applied in the German health care setting.

Objective: The objectives were to develop and validate a predictive model of 30-day emergency hospital transport based on data from a German PERS provider and compare the model with our previously published predictive model developed on data from a US PERS provider.

Methods: Retrospective data of 5805 subscribers to a German PERS service were used to develop and validate an extreme gradient boosting predictive model of 30-day hospital transport, including predictors derived from subscriber demographics, self-reported medical conditions, and a 2-year history of case data. Models were trained on 80% (4644/5805) of the data, and performance was evaluated on an independent test set of 20% (1161/5805). Results were compared with our previously published prediction model developed on a data set of PERS users in the United States.

Results: German PERS subscribers were on average aged 83.6 years, with 64.0% (743/1161) females, with 65.4% (759/1161) reported 3 or more chronic conditions. A total of 1.4% (350/24,847) of subscribers had one or more emergency transports in 30 days in the test set, which was significantly lower compared with the US data set (2455/109,966, 2.2%). Performance of the predictive model of emergency hospital transport, as evaluated by area under the receiver operator characteristic curve (AUC), was 0.749 (95% CI 0.721-0.777), which was similar to the US prediction model (AUC=0.778 [95% CI 0.769-0.788]). The top 1% (12/1161) of predicted high-risk patients were 10.7 times more likely to experience an emergency hospital transport in 30 days than the overall German PERS population. This lift was comparable to a model lift of 11.9 obtained by the US predictive model.

Conclusions: Despite differences in emergency care use, PERS-based collected subscriber data can be used to predict use outcomes in different international settings. These predictive analytic tools can be used by health care organizations to extend population health management into the home by identifying and delivering timelier targeted interventions to high-risk patients. This could lead to overall improved patient experience, higher quality of care, and more efficient resource use.

(JMIR Med Inform 2021;9(3):e25121) doi:10.2196/25121

KEYWORDS

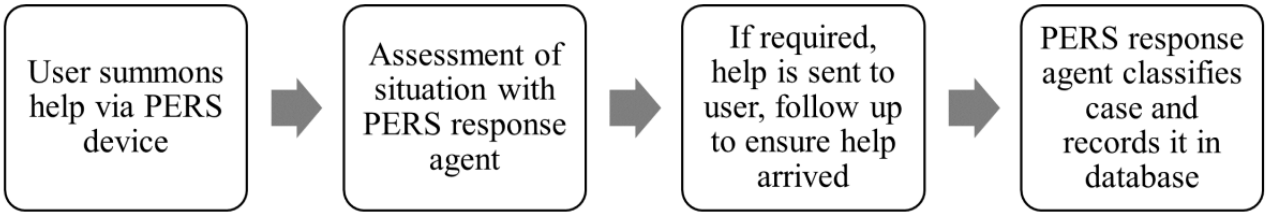
emergency hospital transport; predictive modeling; personal emergency response system; population health management; emergency transport; emergency response system; emergency response; health management

Introduction

The German population is one of five super-aged societies, and its population aged 65 years and older is projected to grow to about 24 million in 2050—roughly one-third of the total German population [1]. As a result, increasing demands are placed on the health care system due to chronic diseases that are more common in the elderly [2]. Emergency care in Germany is chronically overloaded: the number of patients seen in the emergency department (ED) doubled between 2005 and 2015 to around 25 million per year [3]. Older multimorbid patients make up an average of 30% of the patients treated in German EDs [4]. The number of hospital admissions by individuals aged 65 years and older is more than 49,000 per 100,000 annually [5]. The country's health care sector has begun to leverage digital technology and eHealth solutions as part of a broader effort to accommodate a healthier and more engaged older population. Smartification of a person's home through connected technologies has the potential to alleviate the shortage of nursing staff, support the desire of many elderly people to stay at home longer, and reduce costs for municipalities and health care services [6,7].

Such connected technologies include a personal emergency response system (PERS), which can help older adults get immediate assistance when a home-based incident occurs and where delayed response may result in preventable health care use such as ED visits [8]. The work described here builds on a previous study in which we used data obtained from a US PERS service to predict patients at high risk for imminent ambulance transport to the hospital [9]. PERS is a widely used wearable technology with a help button that is worn either as a bracelet or pendant. Subscribers may press the help button at any time to activate an in-home communication system that connects to a 24/7 response center. The response center associate may contact an informal responder (eg, a neighbor or family member) or emergency medical services (EMS) based on the subscriber's specific situation, and follows up with the subscriber to confirm that help has arrived. The response center associate records notes from conversations with the subscribers in an electronic record and classifies the type, situation, and outcome of the case (Figure 1). In combination with user enrollment data, such as demographics, caregiver network, and medical condition data, these case data provide valuable information about subscriber status.

Figure 1. Overview of the personal emergency response system process and case data collection. PERS: personal emergency response system.



PERS services collect information while the subscriber is at home, including details such as timestamp, type, situation, and outcome of calls, that have either medical (eg, falls, respiratory issues, chest pain, or general pain) or social (eg, check-in calls) nature [10]. Such events may be indicative of decline in patient status, which may be captured earlier with PERS-based prediction models than with models based on only clinical data [11-14]. Previous efforts to predict health care use include predictive modeling of hospital readmission [11], repeat ED visits [12,13], and the use of specialized discharge services [14]. The LACE index uses 4 variables (length of stay [L], acuity of the admission [A], comorbidity of the patient [C], and emergency department use in the duration of 6 months before admission (E)) and was designed for the prediction of death or unplanned readmissions after hospitalization [15], achieving a predictive performance of AUC=0.68. HOSPITAL, a risk score for predicting 30-day potentially avoidable readmission, achieved a performance of AUC=0.72 as evaluated in 9 hospitals in 4 different countries [16]. Yet another study used 1-year retrospective electronic medical record data to predict 30-day ED revisits achieving AUC=0.70 in a prospective validation cohort [12].

As a next step, we designed and executed a 2-arm randomized control trial, which demonstrated that PERS-based risk prediction with targeted interventions could reduce health care use and costs [17,18]. In this study, a study nurse contacted high-risk subscribers, conducted additional triaging and, if deemed necessary, provided them with interventions including educational support, nurse home visits, or primary care physician referral. Based on the positive findings in the United States, we are investigating if a PERS-based risk prediction system with tailored interventions could also be applied in the German health care setting [19]. Similar to the US study, the German study requires a predictive model of risk of hospital transport in PERS users. Therefore, the objectives of this paper are to (1) develop and validate a predictive model of 30-day emergency hospital transport based on German PERS provider data and (2) compare the German and US models. It should be noted that various structural differences between the German and US PERS data prevented us from applying the US predictive model to the German data directly or using a transfer learning approach. Therefore, we opted to train a new prediction model on the German data.

Methods

Retrospective Data Set

The first study aim was to develop a 30-day predictive model of emergency hospital transport for a German PERS subscriber population. The initial retrospective data set used to develop the predictive model was extracted from the German PERS service provider ServiceCall AG [20]. It contained data from

8374 former PERS subscribers covering the period March 2006 through November 2018. Subscribers used a variety of PERS devices commercially available in Germany. At the time of study data collection, subscribers in the data set were deceased for at least 1 year to minimize impact on data privacy. This retrospective data study was approved by the Internal Committee for Biomedical Experiments of Philips (ICBE-2-24827).

The extract contained historical data including subscriber demographics such as gender, subscriber age at enrollment, and number of responders the subscriber had listed who could be contacted by the response center. The latter served as an indication of the size of the subscriber's support network. In addition, the data set included self-reported medical conditions and medications provided by the subscriber at the time of enrollment.

Finally, the data set contained case data, which represent interactions with the response center such as incidents (where the subscriber requires assistance) or nonincidents such as test calls, false alarms, or technical issues. For interactions classified as incidents, a number of different situations were recorded (eg, subscriber has fallen), as well as a number of different follow-up actions, including contacting a friend or family member, having a conversation with the subscriber to jointly resolve the problem, or, in some cases, contact EMS.

Inclusion and Exclusion Criteria

Subscribers were included in the analysis if they were active on the service at any time between January 1, 2012, and January 1, 2018. Subscribers were included in the analysis if they had a listed age between 18 and 100 years at the time of enrollment on the PERS service. Furthermore, subscribers were excluded if their contract end date predated the start date, presumably due to administrative error. Subscribers who did not have a unique identifier in the data set (ie, that shared a pseudonym with another subscriber) were also excluded. After applying inclusion and exclusion criteria, data from 5805 subscribers remained for analysis.

Data Processing

The retrospective data set included a table consisting of subscriber data with a single row for each subscriber and a case data table with each row representing a single case. The tables were processed in the statistical programming software R (R Foundation for Statistical Computing).

The case data were characterized in terms of case types, case reasons, and case outcomes (Table 1). The case data for each subscriber were then aggregated by determining the frequency and recency of each of the case types, reasons, and outcomes. The frequency represents the number of a particular case that the subscriber has experienced, while the recency represents the time that has passed since the subscriber has experienced a particular case. Up to 2 years of historic case data were used to derive these features. The frequency and recency features of case data cases were then combined with subscriber demographics, support network, and self-reported medical conditions and medications from the subscriber data table. Tables were merged based on the pseudonymized subscriber IDs.

Table 1. Case types, reasons, and outcomes for which frequency and recency features are derived for input into the predictive model. Examples are given per category.

Classification example	Description
Case type	
Incident	Case where the subscriber is in need of help
Accidental	Subscriber accidentally pushed the help button
Test	Test call by subscriber
Case reason	
Fall	Subscriber fell
Breathing problems	Subscriber has breathing problems
Heart problems	Subscriber has heart problems
Case outcome	
Nurse	Nurse visit
Ambulance transport	Emergency medical services dispatched to bring subscriber to hospital
No assistance required	Subscriber did not require further assistance

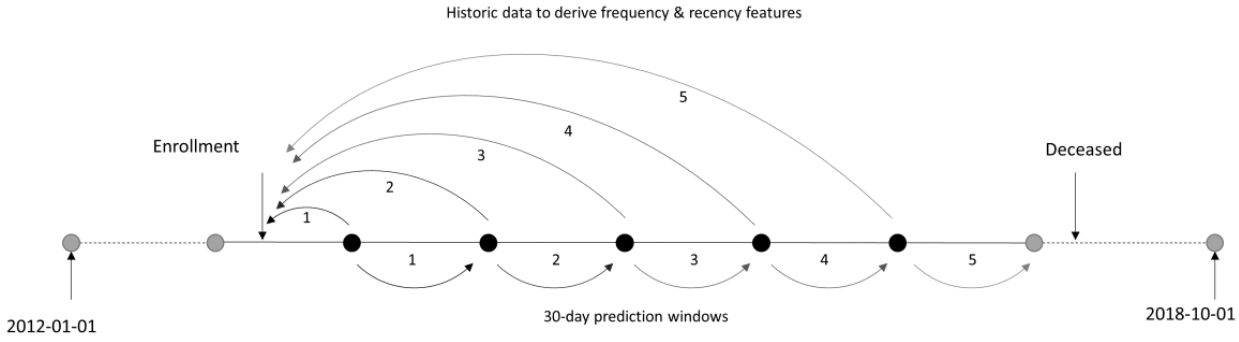
Predictive Model Development

The 5805 German PERS users were randomized into a training and test set in an 80:20 ratio (Figure 2). Originally, the US predictive model was trained using a 50:50 split of training and test set [9]. To eliminate the difference in training/test set ratio, the US predictive model was retrained on the US data set using an 80:20 split. Because the German data set was much smaller than the US data set, it was decided to create multiple prediction windows for each subscriber in order to use the data to the fullest extent possible. This was achieved by computing frequency and recency features and the dependent variable by splitting the data for each subscriber at multiple 30-day intervals (Figure 3). The dependent variable for the predictive model was determined as whether or not a subscriber had an event with the outcome "ambulance transport" in a 30-day prediction window (ie, the

prediction was treated as a binary classification problem). The frequency and recency features for the predictive model were derived from the entire case data history prior to the 30-day window. It was ensured that training and test sets were independent (ie, data from a single subscriber were either in the training or the test set but not in both).

Figure 2. Overview of the study design to develop and evaluate the predictive models of emergency hospital transport. PERS: personal emergency response system.

Figure 3. Schematic of using 30-day intervals to train and validate the prediction models.



Based on the processed features, a predictive model for hospital transport was created using extreme gradient boosting, a variation of the boosted regression trees approach. Extreme gradient boosting is an ensemble approach where new models are added over a number of iterations in order to improve upon and correct the errors of the previous set of models. The models themselves take the form of small regression trees. In this study, XGBoost, an extreme gradient boosting algorithm implemented in R, was used since it has proven to perform well on nonlinear problems, including many high-ranking finishes in Kaggle data science competitions [21].

Predictive Model Evaluation

Discriminatory accuracy of the predictive models was evaluated using area under the receiver operator characteristic curve (AUC), which indicates the probability of the predictive model ranking a randomly selected subscriber with 30-day emergency transport higher than a randomly selected subscriber without the event. Furthermore, the positive predictive value (PPV) indicates the percentage of subscribers having emergency transport in the group classified as positive (ie, having a 30-day emergency transport). The threshold for classifying subscribers as positive was varied using risk scores >90th, 95th and 99th percentile, such that 10%, 5%, and 1% of subscribers were classified as high risk, respectively. For these thresholds, the

PPV, sensitivity, specificity, and accuracy were computed. Confidence intervals for performance metrics were derived using a stratified bootstrapping method with 1000 bootstrap replicates. The agreement between the predictions made by the model and the observed outcome was evaluated by plotting the average of the predicted probabilities and the observed percentage of users having 30-day emergency transport in deciles of the prediction score.

Statistical Analysis

Differences between subscriber characteristics in the training and test sets of the German data and between both test sets of the German and US data were analyzed using Student *t* tests for age and chi-square tests for the categorical variables. Differences were considered to be statistically significant if $P<.05$.

Results

Subscriber Characteristics

Characteristics of subscribers in the training and test set of the Germany model are presented in Table 2 and compared with the test set used for the US predictive model. Data of 5805 unique PERS users were used in the Germany predictive model. A total of 4644 (80%) individuals were randomly selected for the training set and 1161 (20%) for the test set—training and test sets were mutually exclusive with regard to users. The US test set comprised 109,966 PERS users. Since the German data set was smaller, multiple prediction dates were considered by splitting the time range January 1, 2012, through January 1, 2018, into 73 equally spaced 30-day windows. This resulted in a training set containing 96,273 (79.5%) prediction dates and a test set with 24,847 (20.5%) prediction dates.

PERS users were on average aged 84.0 years in the German training set. Average age was slightly, but statistically significantly, younger in the test set (83.6 years). The average age in the US test set was statistically significantly lower at 81.2 years compared with the German test set. About two-thirds (2997/4644, 64.5%) of German PERS users were female, with no statistically significant difference between training and test sets. However, the US test set showed a significantly higher proportion of females (88,433/109,966, 80.4%).

In the German training set, more than half of users (2598/4644, 55.9%) were on the service 2 years or less, 23.2% (1079/4644) of users were 2 to 4 years on the service, and 20.8% (967/4644) of users were more than 4 years on the service. These percentages were not statistically significantly different in the test set. In the US test set, 44.4% (48,922/109,966) of users were less than 2 years on the service, which was significantly lower than in the Germany test set (630/1161, 54.3%). A similar percentage of US PERS users were 2 to 4 years on the service (26,193/109,966, 23.8%, vs 264/1161, 22.7%), while more users were 4 or more years on the service (34,851/109,966, 31.7%, vs 267/1161, 23.0%).

In the training set for the Germany predictive model, 94.3% (4397/4644) of users had at least one self-reported medical condition, with 35.7% (1657/4644) reporting 5 or more conditions. There were no statistically significant differences between the number of self-reported conditions in the training and test set for the Germany predictive model. In contrast, 77.3% (85,056/109,966) of users in the test set for the US predictive model self-reported one or more medical conditions, which was statistically significantly lower than in the test set for the Germany predictive model.

The prevalence of the dependent variable “emergency hospital transport in the next 30 days” was 1.6% (1506/96,273) in the training set and 1.4% (350/24,847) in the test set for the Germany predictive model. The latter was statistically significantly lower than the prevalence of the dependent variable in the test set of the US predictive model (2455/109,966, 2.2%).

Table 2. Subscriber characteristics and prevalence of the dependent variable in the training and test sets for the Germany predictive model compared with the previously published results of the US predictive model in the test set. *P* values are reported for differences between German test and training sets, and between US and German test sets.

Characteristics	Germany predictive model (this study)			US predictive model (from [9])	
	Training set	Test set	*P* value (test vs training Germany)	Test set	*P* value (US vs Germany test)
General					
Prediction dates	Jan 1, 2012, to Jan 1, 2018	Jan 1, 2012, to Jan 1, 2018	—	Feb 1, 2014	—
# of unique PERS[a] users, n (%)	4644 (80)	116 (20)	—	109,966 (20)	—
# of prediction windows, n (%)	96,273 (79.5)	24,847 (20.5)	—	109,966 (20)	—
Age in years, mean (SD)	84.0 (8.2)	83.6 (8.3)	.001	81.1 (11.4)	<.001
Female gender, n (%)	2997 (64.5)	743 (64.0)	.37	88,433 (80.4)	<.001
Years on PERS service, n (%)					
0-2	2598 (55.9)	630 (54.3)	.32	48,922 (44.4)	<.001
2-4	1079 (23.2)	264 (22.7)	.75	26,193 (23.8)	.41
4 or more	967 (20.8)	267 (23.0)	.11	34,851 (31.7)	<.001
Number of PERS self-reported medical conditions, n (%)					
None	265 (5.7)	73 (6.3)	.49	24,910 (22.6)	<.001
1-2	1325 (28.5)	329 (28.3)	.93	26,515 (24.1)	<.001
3-4	1397 (30.1)	370 (31.9)	.25	28,561 (26.0)	<.001
5 or more	1657 (35.7)	389 (33.5)	.18	29,980 (27.3)	<.001
30-day emergency hospital transport (% of prediction windows)	1506 (1.6)	350 (1.4)	.08	2455 (2.2)	<.001

[a]PERS: personal emergency response system.

Predictive Model Evaluation

The performance of the Germany predictive model on the test set is detailed in Table 3 for various prediction score thresholds. AUC was 0.749 (95% CI 0.721-0.777) for emergency hospital transport in 30 days. This was slightly but not statistically significantly lower than the AUC for the US predictive model (0.778 [95% CI 0.769-0.788]), as 95% CIs were overlapping.

Positive predictive values for the Germany predictive model were low due to the low prevalence of 30-day emergency hospital transport, which was 1.4% (350/24,847) in the test set (Table 1). By increasing the prediction score threshold, PPV increased but at the expense of decreased sensitivity. At a prediction score threshold corresponding to the 90th percentile, the Germany predictive model identified 40.3% (95% CI 35.1%-45.4%) of the subscribers who had emergency transport in the 30 days following the prediction date (sensitivity); however, only 5.7% (95% CI 4.9%-6.4%) of flagged subscribers had emergency transport in the following 30 days (PPV) at this threshold. At thresholds corresponding to the 95th and 99th percentiles, the sensitivity dropped to 26.9% (95% CI 22.3%-31.1%) and 10.6% (95% CI 7.4%-14.0%), respectively, while the PPV increased to 7.5% (95% CI 6.3%-8.8%) and 15.0% (95% CI 10.7%-19.3%), respectively. When the threshold was set at the 99th percentile, the PPV was 10.7 times higher than the prevalence of 1.4%. This lift of the prediction model was similar for the US prediction model, namely 11.9.

The US predictive model demonstrated similar sensitivity and specificity values for the different thresholds. However, PPV was significantly higher across all thresholds compared with the Germany predictive model due to the higher prevalence of the target variable in the US data set.

Table 3. Performance of the Germany and US predictive models on the corresponding test sets, evaluated by positive predictive value, sensitivity, and specificity using the 90th, 95th, and 99th percentiles as a threshold and area under receiver operator characteristic curve.

Performance metric and threshold (percentile)	Germany predictive model (this study), % (95% CI)	US predictive model (adapted from [9]), % (95% CI)
PPV[a]		
90%	5.7 (4.9-6.4)[b]	9.4 (8.9-9.8)
95%	7.5 (6.3-8.8)[b]	13.6 (12.9-14.3)
99%	15.0 (10.7-19.3)[b]	26.2 (23.7-28.5)
Sensitivity		
90%	40.3 (35.1-45.4)	41.9 (40.0-43.9)
95%	26.9 (22.3-31.1)	30.3 (28.6-32.0)
99%	10.6 (7.4-14.0)	11.7 (10.5-13.0)
Specificity		
90%	90.4 (90.1-90.8)	90.8 (90.6-91.0)
95%	95.3 (95.1-95.6)	95.6 (95.6-95.7)
99%	99.1 (99.0-99.2)	99.3 (99.2-99.3)
AUC[c]		
—	0.749 (0.721-0.777)	0.778 (0.769-0.788)

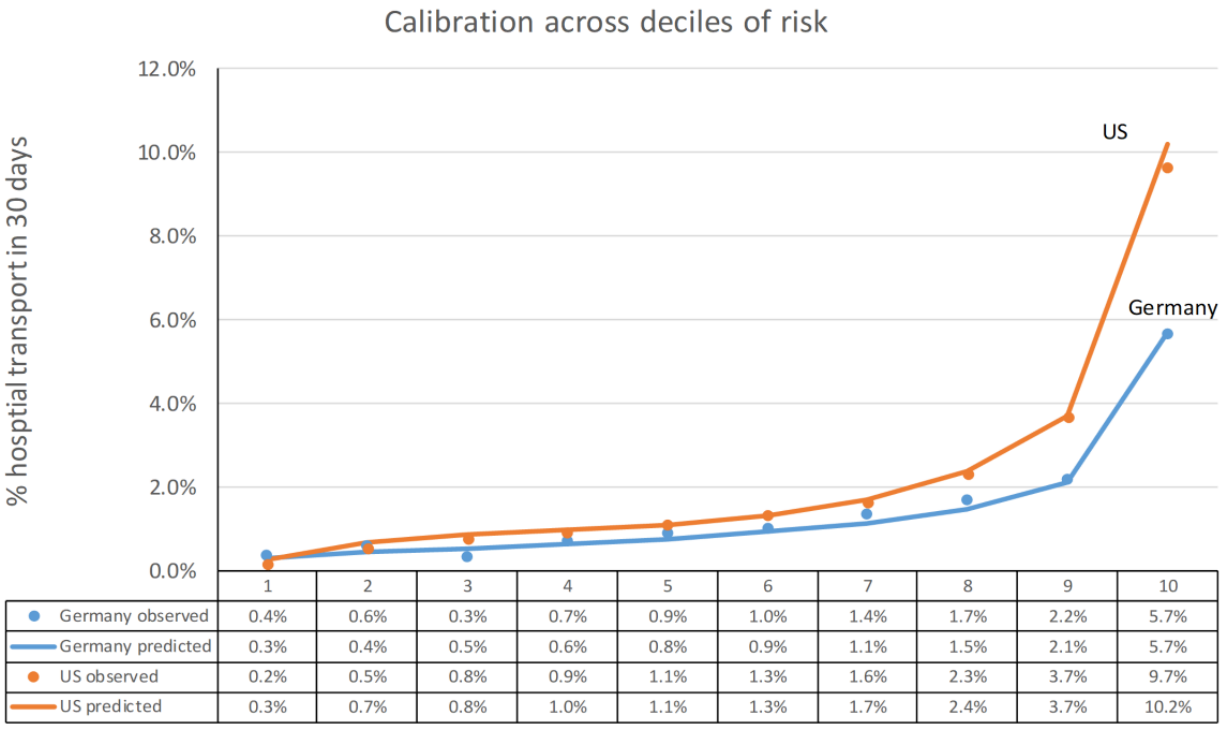
[a]PPV: positive predictive value.

[b]Nonoverlapping 95% CI between Germany and US predictive models.

[c]AUC: area under the receiver operator characteristic curve.

The predictive model produced for each subscriber a probability from 0% to 100% indicating the risk of having 30-day emergency hospital transport. The actually observed percentage of subscribers with 30-day emergency hospital transport and average predicted probabilities are presented in Figure 4 to indicate calibration across deciles of risk for both models. Each decile consists of 10% of the test set sorted by predicted probability. Probabilities increased from 0.4% in the lowest risk decile to 5.7% in the highest risk decile observed in the Germany test set and from 0.2% to 9.7% for the US test set. Both models were well calibrated with R^2=.9935 and R^2=.9992 for the Germany and US predictive models, respectively.

Figure 4. Observed percentage of subscribers needing 30-day emergency hospital transport versus model predicted probability across deciles of risk.



	1	2	3	4	5	6	7	8	9	10
Germany observed	0.4%	0.6%	0.3%	0.7%	0.9%	1.0%	1.4%	1.7%	2.2%	5.7%
Germany predicted	0.3%	0.4%	0.5%	0.6%	0.8%	0.9%	1.1%	1.5%	2.1%	5.7%
US observed	0.2%	0.5%	0.8%	0.9%	1.1%	1.3%	1.6%	2.3%	3.7%	9.7%
US predicted	0.3%	0.7%	0.8%	1.0%	1.1%	1.3%	1.7%	2.4%	3.7%	10.2%

Predictor Importance

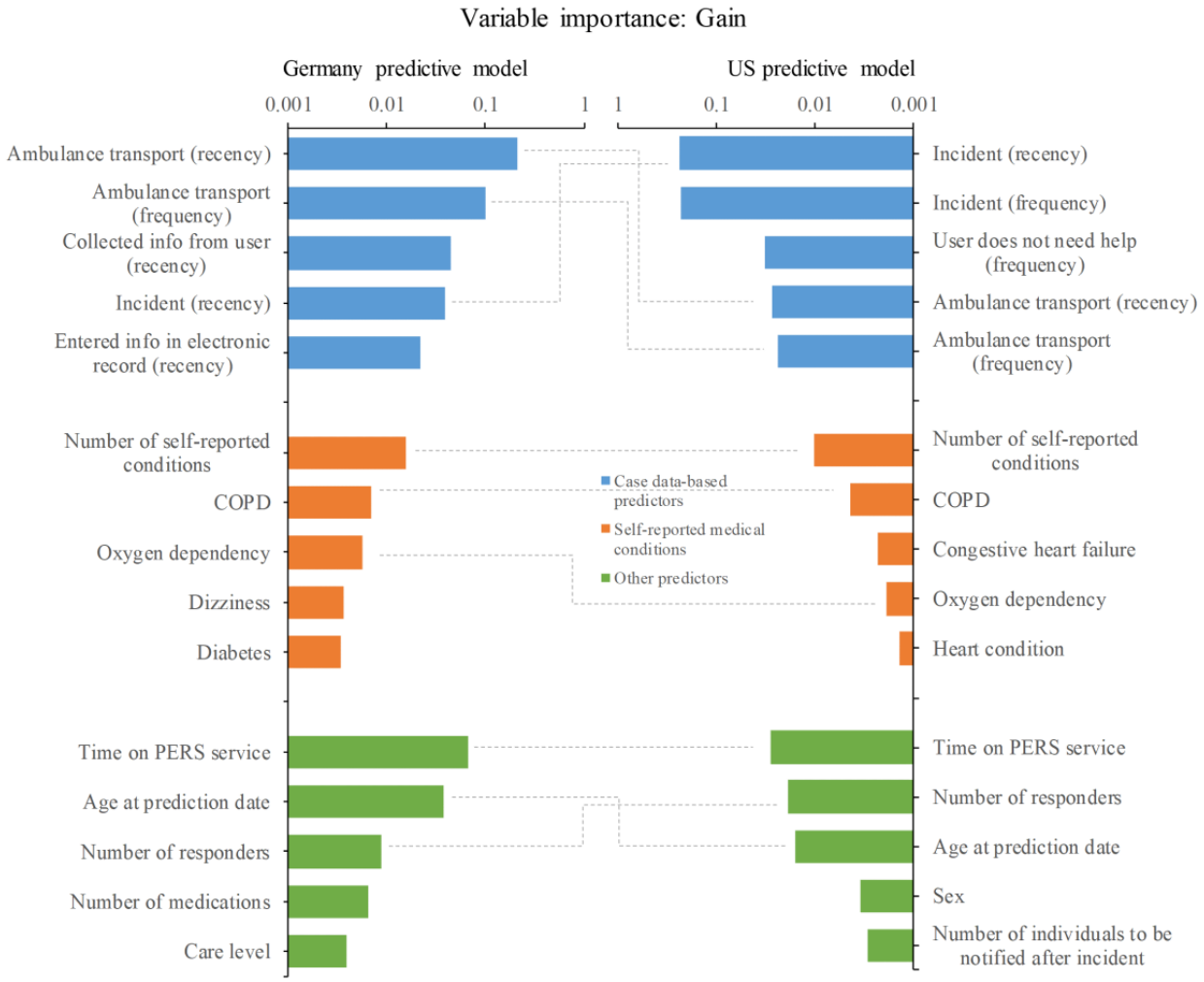
The Germany predictive model of 30-day emergency hospital transport included 98 variables with nonzero values for the gain compared with 121 for the US predictive model. For each broad category of predictors, Table 4 provides the number of predictors and the gain. Here, gain is calculated by the XGBoost algorithm and represents a combined statistic of the information gain over all trees for a particular predictor. As such, gain represents a measure of the relative importance of individual predictors. Predictors from the case data form the most important predictor category for both predictive models, although percentage-wise, their contribution to the Germany predictive model was lower compared with the US predictive model (72.9% vs 87.7%, respectively). On the other hand, the relative importance of self-reported medical conditions (9.6% vs 3.7%, respectively) and other predictors (17.5% vs 8.7%, respectively) was higher in the Germany predictive model.

Table 4. Number of predictors and total gain per predictor category.

Predictor category	Number of predictors		Total gain, %	
	Germany	US	Germany	US
Case data-based predictors	48	62	72.9	87.7
Self-reported medical conditions	43	44	9.6	3.7
Other predictors	7	15	17.5	8.7
Total	98	121	100	100

The 5 most important predictors for each category are shown in Figure 5 for both models. In each category, 3 out of 5 predictors were overlapping between the Germany and US predictive models. For case data–based predictors, these included recency and frequency of ambulance transport and recency of incidents. In the Germany predictive model, features based on the collection of further case information from the user and entering additional information into the electronic record were found among the important predictors. We expect that extracting features from these free text notes in the electronic record could lead to further model improvement; however, text notes were left out of the analysis due the risk of including privacy-sensitive information.

Figure 5. Predictor importance as measured by the gain of the 5 most important variables in 3 categories of predictors for both Germany and US predictive models. Predictor categories: case data, self-reported medical conditions, and other predictors. Dashed lines indicate features present in both predictive models. COPD: chronic obstructive pulmonary disease; PERS: personal emergency response system.



Furthermore, the number of self-reported medical conditions, chronic obstructive pulmonary disease, and oxygen dependency were among the 5 most important predictors in both models in the category self-reported medical conditions. Other important predictors that were shared by both models included time on the PERS service, age, and number of responders. It should be noted that in the Germany predictive model, number of medications and care level (Pflegestufe), a Germany-specific categorization of the level of (financial) support individuals need with activities of daily living, were among the most important other predictors, and this information was not available in the US PERS data.

Discussion

Principal Findings

In previous work, we have shown that PERS service data from US subscribers can be used to predict risk for emergency hospital transport [9]. This study is an extension of that work to determine the feasibility to develop such a prediction model on data from German PERS subscribers. Comparison of data from the German and US PERS providers shows that PERS users in both countries have, on a high level, similar characteristics—average age over 80 years, predominantly female, and more than half reporting on 3 or more medical conditions. On the other hand, Table 2 shows various subtle differences between the characteristics of the two populations, justifying the effort to develop a Germany-specific predictive model.

In the US and German PERS populations, the prevalence of hospital transport in 30 days among PERS users was significantly lower in the German PERS population (1.4% vs 2.2%, respectively). Health insurance is obligatory in Germany, and the health care systems covers all costs of both inpatient and outpatient treatment [22]. Compared with the United States, where patients often self-refer to the ED out of financial considerations [23], this is therefore less likely to play a role in Germany, which might explain the difference in prevalence of hospital transport. Nevertheless, the increasing number of ED visits that could have been prevented via treatment in the primary care setting is a growing issue in Germany.

Evaluation of the German prediction model of 30-day emergency hospital transport on a test set of data from different

PERS users demonstrated that at-risk subscribers could be identified with discriminatory accuracy similar to the US prediction model (AUC=0.749 vs AUC=0.778). Furthermore, calibration across deciles indicated that the predicted probabilities for both the Germany and US prediction models closely matched with observed outcomes. Calibration refers to the agreement between observed outcomes and predictions (ie, if we predict a 10% risk of 30-day hospital transport, the observed frequency of hospital transport should be approximately 10 out of 100 subscribers with such a prediction [24]). Finally, analysis of variable importance indicated that predictors derived from the medical alert pattern data, including the frequency and recency of prior ambulance transports, were most predictive of future hospital transport in both the German and US prediction models. Similarly, Poole et al [25] found that the timing and frequency of prior ED use are the strongest predictors of future ED visits using a random forest model.

Our previous study on health care use in US PERS users indicates that 21% of hospital admissions are considered potentially avoidable [10]. A recent study on hospitalizations by German nursing home patients classified 27% as potentially avoidable [26]. Therefore, we believe that prediction of emergency transport risk in combination with appropriate interventions could potentially reduce health care use. Case managers and health professionals should integrate risk prediction of patients into their clinical workflows to obtain the clinical and financial benefits from predictive models, which requires a detailed guideline that clarifies how the algorithm will inform care [27]. In a recently completed randomized clinical trial, we developed workflows that integrate daily PERS-based risk of 30-day emergency hospital transport with care pathways [17], resulting in 49% fewer EMS encounters in the intervention group [18]. In a currently running prospective study in Germany [19], the predictive model described herein is used to predict subscribers risk for 30-day emergency transport, followed by a case manager assessment, and tailored interventions for high-risk subscribers. The number of patients who will ultimately benefit from a combination of prediction and intervention will depend on various factors including the population size and prevalence of emergency health care use, performance of the predictive model and risk threshold above which patients are considered to be high risk, and efficacy of the interventions provided to high-risk patients.

In our prospective study in Germany [19], the predicted risk scores drive proactive outreach—if the risk is above a certain threshold, the patient may be contacted by the case manager. Due to the low prevalence of hospital transport in the German PERS population, setting the value of the risk threshold is a trade-off between finding many true positive cases (ie, a high sensitivity) and reducing the number of false positives (ie, a high PPV), as shown in Table 2 and also reported by other predictive modeling studies of emergency health care use [12,28]. Despite this, our recent study in a US PERS population has demonstrated that health care use and cost can be reduced by combining risk prediction with preventive interventions [18].

Limitations

This study had a few limitations. The PERS population is mostly older and primarily female, and the service is to a certain extent privately paid for by subscribers (ie, not fully covered by their health insurance). This may limit the generalizability of the study to older women who can afford the service. Furthermore, the predictive model may have been influenced by confounding of unobserved variables, including when and where users wear the PERS device [29].

Subscribers may have initiated emergency hospital transport outside of the PERS service, in which case there are no records in the PERS data, which may have affected predictive model development. As a mitigation measure, participants of our prospective study are instructed to use their emergency pendant for all incidents where they require help.

Conclusions

This study showed that remotely collected subscriber data from a German PERS service can be used to predict 30-day hospital transport with similar discriminatory accuracy and calibration as our previously published prediction model developed on data from a US PERS population. Health care providers could potentially benefit from our validated predictive model by estimating the risk of 30-day emergency hospital transport for individual subscribers and target timely preventive interventions to high-risk subscribers. Due to a lower prevalence of emergency hospital transport in Germany compared with the United States, it needs further investigation if combining risk prediction with interventions will effectively reduce health care use. We are currently testing this hypothesis in a prospective study where risk predictions are combined with a stepped intervention pathway. This approach could lead to overall improved patient experience, higher quality of care, and more efficient resource use.

Acknowledgments

The authors thank Mr Sandrock and Mr Runge from ServiceCall AG for provision of the data. JB, MP, and AL designed the research; JB and MP conducted the analyses. AL provided feedback on analyses and interpretation of results; JB, MP, and AL wrote the paper. JB had primary responsibility for the final content. All authors read and approved the final manuscript.

Conflicts of Interest

Philips funded the study. JB, MP, and AL are employed by Philips.

References

1. The Aging Readiness & Competitiveness Report—Germany. URL: https://arc.aarpinternational.org/File%20Library/Full%20Reports/ARC-Report---Germany.pdf [accessed 2021-02-19]
2. Nowossadeck E. Demografische Alterung und stationäre Versorgung chronischer Krankheiten. Dtsch Arztebl Int Deutscher Arzte-Verlag GmbH 2012;109:157. [doi: 10.1007/978-3-8349-6787-9_6]
3. Bundesärztekammer. (Politische) Rahmenbedingungen einer sektorenübergreifenden Versorgung in Notfallpraxen und Notaufnahmen. 2017. URL: https://www.bundesaerztekammer.de/politik/programme-positionen/notfallversorgung/ [accessed 2021-02-19]
4. Singler K, Heppner HJ. [Acute and emergency care of geriatric patients: old ways—new paths]. Z Gerontol Geriatr 2017 Dec;50(8):669-671. [doi: 10.1007/s00391-017-1305-4] [Medline: 28900726]
5. Eckdaten der Krankenhauspatientinnen und -patienten. Statistisches Bundesamt. URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Krankenhaeuser/Tabellen/entlassene-patienten-eckdaten.html [accessed 2021-01-08]
6. Christiansen S, Klötzer JP. [Ambient assisted living: an overview]. Versicherungsmedizin 2015 Sep 01;67(3):130-132. [Medline: 26548006]
7. Uhlig M, Bahrmann A. [Ambient assisted living: a sleeping giant]. MMW Fortschr Med 2014 Nov 06;156(19):45-48. [doi: 10.1007/s15006-014-3639-9] [Medline: 25510023]
8. Bakk L. Does PERS and falls prevention information reduce stress, anxiety, falls risk, hospital admission and emergency room use for older adults on the wait list for home-and community- based services?. 2011. URL: http://www.aaa1b.com/wp-content/uploads/2010/07/PERS-Report-Final-May-2011.pdf [accessed 2021-02-20]
9. Op den Buijs J, Simons M, Golas S, Fischer N, Felsted J, Schertzer L, et al. Predictive modeling of 30-day emergency hospital transport of patients using a personal emergency response system: prognostic retrospective study. JMIR Med Inform 2018 Nov 27;6(4):e49 [FREE Full text] [doi: 10.2196/medinform.9907] [Medline: 30482741]
10. Golas S, Fischer N, Agboola S, Jethwani K, Simons-Nikolova M, Op Den Buijs J. Reasons patients are using Personal Emergency Response Services...and it's not just falls. 2018. URL: http://incenter.medical.philips.com/doclib/getDoc.aspx?func=ll&objId=16015470&objAction=Open [accessed 2021-02-19]
11. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M. Risk prediction models for hospital readmission: a systematic review. VA Evidence-based Synth Progr Rep 2011 Oct. [Medline: 22206113]
12. Hao S, Jin B, Shin AY, Zhao Y, Zhu C, Li Z, et al. Risk prediction of emergency department revisit 30 days post discharge: a prospective study. PLoS One 2014;9(11):e112944 [FREE Full text] [doi: 10.1371/journal.pone.0112944] [Medline: 25393305]
13. Meldon SW, Mion LC, Palmer RM, Drew BL, Connor JT, Lewicki LJ, et al. A brief risk-stratification tool to predict repeat emergency department visits and hospitalizations in older patients discharged from the emergency department. Acad Emerg Med 2003 Mar;10(3):224-232 [FREE Full text] [doi: 10.1111/j.1553-2712.2003.tb01996.x] [Medline: 12615588]
14. Holland DE, Harris MR, Leibson CL, Pankratz VS, Krichbaum KE. Development and validation of a screen for specialized discharge planning services. Nurs Res 2006;55(1):62-71. [doi: 10.1097/00006199-200601000-00008] [Medline: 16439930]
15. van Walraven C, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. Can Med Assoc J 2010 Apr 06;182(6):551-557 [FREE Full text] [doi: 10.1503/cmaj.091117] [Medline: 20194559]
16. Donzé JD, Williams MV, Robinson EJ, Zimlichman E, Aujesky D, Vasilevskis EE, et al. International validity of the HOSPITAL score to predict 30-day potentially avoidable hospital readmissions. JAMA Intern Med 2016 Apr;176(4):496-502 [FREE Full text] [doi: 10.1001/jamainternmed.2015.8462] [Medline: 26954698]
17. Palacholla RS, Fischer NC, Agboola S, Nikolova-Simons M, Odametey S, Golas SB, et al. Evaluating the impact of a web-based risk assessment system (CareSage) and tailored interventions on health care utilization: protocol for a randomized controlled trial. JMIR Res Protoc 2018 May 09;7(5):e10045 [FREE Full text] [doi: 10.2196/10045] [Medline: 29743156]
18. Golas SB, Nikolova-Simons M, Palacholla R, Op Den Buijs J, Garberg G, Orenstein A. Predictive analytics and tailored interventions improve clinical outcomes in older adults: a randomized controlled trial [in press]. NPJ Dig Med 2021 Mar 15 (forthcoming).
19. Pijl M, Op den Buijs J, Landgraf A. Evaluating the impact of a risk assessment system with tailored interventions in Germany: protocol for a prospective study with matched controls. JMIR Res Protoc 2020 Oct 01;9(10):e17584 [FREE Full text] [doi: 10.2196/17584] [Medline: 33001038]
20. ServiceCall. URL: http://www.servicecall.de/startseite/index.php [accessed 2021-01-08]
21. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. ArXiv. Preprint posted online on March 9, 2016 2016. [doi: 10.1145/2939672.2939785]
22. Schmiedhofer M, Möckel M, Slagman A, Frick J, Ruhla S, Searle J. Patient motives behind low-acuity visits to the emergency department in Germany: a qualitative study comparing urban and rural sites. BMJ Open 2016 Nov 16;6(11):e013323 [FREE Full text] [doi: 10.1136/bmjopen-2016-013323] [Medline: 27852722]
23. Kraaijvanger N, van Leeuwen H, Rijpsma D, Edwards M. Motives for self-referral to the emergency department: a systematic review of the literature. BMC Health Serv Res 2016 Dec 09;16(1):685 [FREE Full text] [doi: 10.1186/s12913-016-1935-z] [Medline: 27938366]

24. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010 Jan;21(1):128-138 [FREE Full text] [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
25. Poole S, Grannis S, Shah NH. Predicting Emergency Department Visits. AMIA Jt Summits Transl Sci Proc 2016;2016:438-445 [FREE Full text] [Medline: 27570684]
26. Leutgeb R, Berger SJ, Szecsenyi J, Laux G. Potentially avoidable hospitalisations of German nursing home patients? A cross-sectional study on utilisation patterns and potential consequences for healthcare. BMJ Open 2019 Jan 21;9(1):e025269 [FREE Full text] [doi: 10.1136/bmjopen-2018-025269] [Medline: 30670526]
27. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 2014 Jul;33(7):1123-1131. [doi: 10.1377/hlthaff.2014.0041] [Medline: 25006137]
28. Billings J, Dixon J, Mijanovich T, Wennberg D. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. BMJ 2006 Aug 12;333(7563):327 [FREE Full text] [doi: 10.1136/bmj.38870.657917.AE] [Medline: 16815882]
29. Stokke R. The personal emergency response system as a technology innovation in primary health care services: an integrative review. J Med Internet Res 2016 Jul 14;18(7):e187 [FREE Full text] [doi: 10.2196/jmir.5727] [Medline: 27417422]

Abbreviations

AUC: area under the receiver operator characteristic curve
ED: emergency department
EMS: emergency medical services
PERS: personal emergency response system
PPV: positive predictive value

Edited by C Lovis; submitted 19.10.20; peer-reviewed by D Li, H Yasuda; comments to author 05.12.20; revised version received 08.01.21; accepted 07.02.21; published 08.03.21.

Please cite as:
op den Buijs J, Pijl M, Landgraf A
Predictive Modeling of 30-Day Emergency Hospital Transport of German Patients Using a Personal Emergency Response: Retrospective Study and Comparison with the United States
JMIR Med Inform 2021;9(3):e25121
URL: https://medinform.jmir.org/2021/3/e25121
doi:10.2196/25121
PMID:33682679

Original Paper

A Novel Convolutional Neural Network for the Diagnosis and Classification of Rosacea: Usability Study

Zhixiang Zhao[1,2*], MD, PhD; Che-Ming Wu[3*], MEng; Shuping Zhang[1,2], MD, PhD; Fanping He[1,2], MD, PhD; Fangfen Liu[1,2], MD, PhD; Ben Wang[1,2], MD, PhD; Yingxue Huang[1,2], MD, PhD; Wei Shi[1,2], MD, PhD; Dan Jian[1,2], MD, PhD; Hongfu Xie[1,2], MD, PhD; Chao-Yuan Yeh[3*], MD; Ji Li[1,2,4,5*], MD, PhD

[1]Department of Dermatology, Xiangya Hospital of Central South University, Changsha, China

[2]Hunan Key Laboratory of Aging Biology, Xiangya Hospital of Central South University, Changsha, China

[3]aetherAI, Co Ltd, Taipei, Taiwan, China

[4]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital of Central South University, Changsha, China

[5]Key Laboratory of Organ Injury, Aging and Regenerative Medicine of Hunan Province, Changsha, China

[*]these authors contributed equally

Corresponding Author:
Ji Li, MD, PhD
Department of Dermatology
Xiangya Hospital of Central South University
87 Xiangya Rd.
Changsha, 410008
China
Phone: 86 073189753406
Email: liji_xy@csu.edu.cn

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2024/1/e57654

Abstract

Background: Rosacea is a chronic inflammatory disease with variable clinical presentations, including transient flushing, fixed erythema, papules, pustules, and phymatous changes on the central face. Owing to the diversity in the clinical manifestations of rosacea, the lack of objective biochemical examinations, and nonspecificity in histopathological findings, accurate identification of rosacea is a big challenge. Artificial intelligence has emerged as a potential tool in the identification and evaluation of some skin diseases such as melanoma, basal cell carcinoma, and psoriasis.

Objective: The objective of our study was to utilize a convolutional neural network (CNN) to differentiate the clinical photos of patients with rosacea (taken from 3 different angles) from those of patients with other skin diseases such as acne, seborrheic dermatitis, and eczema that could be easily confused with rosacea.

Methods: In this study, 24,736 photos comprising of 18,647 photos of patients with rosacea and 6089 photos of patients with other skin diseases such as acne, facial seborrheic dermatitis, and eczema were included and analyzed by our CNN model based on ResNet-50.

Results: The CNN in our study achieved an overall accuracy and precision of 0.914 and 0.898, with an area under the receiver operating characteristic curve of 0.972 for the detection of rosacea. The accuracy of classifying 3 subtypes of rosacea, that is, erythematotelangiectatic rosacea, papulopustular rosacea, and phymatous rosacea was 83.9%, 74.3%, and 80.0%, respectively. Moreover, the accuracy and precision of our CNN to distinguish rosacea from acne reached 0.931 and 0.893, respectively. For the differentiation between rosacea, seborrheic dermatitis, and eczema, the overall accuracy of our CNN was 0.757 and the precision was 0.667. Finally, by comparing the CNN diagnosis with the diagnoses by dermatologists of different expertise levels, we found that our CNN system is capable of identifying rosacea with a performance superior to that of resident doctors or attending physicians and comparable to that of experienced dermatologists.

Conclusions: The findings of our study showed that by assessing clinical images, the CNN system in our study could identify rosacea with accuracy and precision comparable to that of an experienced dermatologist.

(JMIR Med Inform 2021;9(3):e23415) doi:10.2196/23415

KEYWORDS

rosacea; artificial intelligence; convolutional neural networks

Introduction

Rosacea is a common chronic inflammatory disease, which mainly affects the convex facial areas such as nose, cheek, chin, and glabella, with estimated prevalence ranging from 2% to 22% worldwide [1,2] and leading to impaired physical appearance, self-abasement, frustration, and poor quality of life in millions of patients with rosacea [3]. The clinical manifestations of rosacea are quite diversified, including flushing, erythema, angiotelectasis, papules, pustules, and phymatous changes [4], which vary largely from patient to patient, and some of these manifestations usually overlap [5]. Besides, the clinical features of rosacea resemble those of a series of facial inflammatory diseases such as acne, seborrheic dermatitis/eczema, and lupus, thereby making the correct recognition of rosacea even more difficult [6]. In addition, the existing clinical diagnostic criteria for rosacea are still debatable and cause confusion in clinical practice [7,8]. Thus, the correct diagnosis of rosacea remains a big challenge for the medical community, and there is a desperate need for a universal reliable diagnostic system for rosacea.

In recent years, with the rapid development of computer science, artificial intelligence has emerged as a promising tool for face recognition, image analysis, and deciphering genomics [9-13]. Among them, the utility of deep convolutional neural networks (CNNs) in medical practice has caught great attention, especially in the field of dermatology [14,15]. Much efforts have been made to apply machine learning in the detection of malignant skin tumors such as melanoma and basal cell carcinoma [16-21]. Early screening and accurate detection of these skin cancers are the premises for timely treatment and would be of great benefit for patients. Furthermore, machine learning can serve as a potential method for identifying other common skin diseases such as psoriasis, atopic dermatitis, and onychomycosis [14,15]. By objectively analyzing and summarizing dermatological images, artificial intelligence can offer clinicians unbiased suggestions for clinical assessment and outcome prediction [14,22], which would effectively narrow the gap between physicians with different educational backgrounds or clinical experience.

In this study, we trained a deep CNN to analyze clinical images (from 3 different angles) of thousands of patients with rosacea versus those of patients with other common diseases, which could be easily confused with rosacea in clinic (eg, acne, facial seborrheic dermatitis, eczema). We aimed to evaluate the ability of our CNN to identify and classify rosacea. We also compared the accuracy and specificity of our CNN in distinguishing rosacea from other skin diseases with those of clinicians with different levels of clinical experiences.

Methods

The concept of CNN was proposed by Lecun et al [23]. CNN uses various filters to capture features from local regions of an image and shows state-of-the-art performances in many image-based machine learning tasks such as image classification [24], object detection [25], and object segmentation [26]. The common architecture of CNN can be divided into 2 parts: feature extractor and classifier. The feature extractor is composed of stacked convolutional layers and pooling layers. Each convolutional layer contains many filters, which scan the image and do a Hadamard product operation.

After scanning, a filter will generate a 2D matrix called as the feature map (Multimedia Appendix 1). This feature map will progress to an activation function. The most common function is the rectified linear unit, as shown below [23].

$$y=x, \text{ if } x\geq 0$$

$$y=0, \text{ if } x<0$$

Another common layer in the feature extractor part is the pooling layer. The pooling layer will subsample the feature maps in the height and width domain. It is applied to execute a denoising process. It will sample a value from every 2×2 or 3×3 subregions of the feature maps (Multimedia Appendix 2).

In this way, the pooling layer can reduce redundant information. Reducing the feature map size also decreases the calculation in the following convolutional layers. The sampling strategy in the pooling layers can be done in many ways. In recent years, max pooling is considered as the most efficient strategy in image classification and is used in many CNN architectures. It samples the maximum value from a subregion. The below equation shows how max pooling works.

$$y = \max(X)$$

The second part of the CNN is the classifier. It is composed of one or many fully connected layers. The feature maps from the feature extractor module will be flattened or downsampled into a 1D vector and fed to the fully connected layers. Each fully connected layer executes a matrix calculation as shown below.

$$y = h(WX + b)$$

W means weights, which is a 2D matrix; b means bias, which is a 1D vector; and h (·) is an activation function. The whole CNN will be optimized by backpropagation [27] and gradient descent algorithm [28]. All learnable parameters, including filters, weights, and bias, will be updated during the optimizing procedure.

In our model, we used ResNet-50, which is a variant of CNN, to distinguish rosacea from other facial diseases [29]. ResNet-50 is known as a CNN model with a very “deep” architecture. ResNet-50 overcomes the gradient vanishing problem in case a model becomes deeper and has better generalization than other

architectures. The architecture of ResNet-50 is shown below (Multimedia Appendix 3).

The major structure of ResNet-50 is the residual block. ResNet-50 contains 16 residual blocks. Each residual block is composed of 3 convolutional layers. The following figure displays the structure of a residual block (Multimedia Appendix 4).

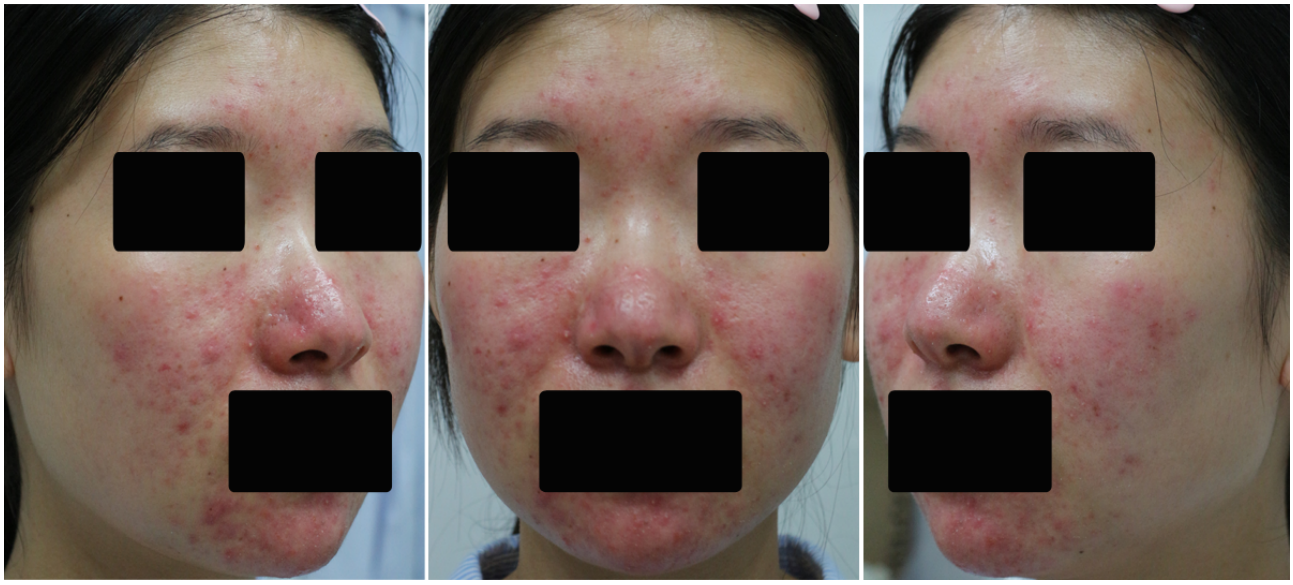
The first 2 layers in a residual block have the same number of filters. The last layer always has 4 times the number of filters present in the previous layers. The output of a residual block is computed by adding the original input and the output from 3 convolutional layers. After passing 16 blocks, a global average pooling layer samples a value from each feature map by averaging them. Finally, a fully connected layer generates a vector with the output of the global average pooling layer. This vector is the prediction of the model, which contains 2 values and means the scores of rosacea and other facial diseases that might be easily confused with rosacea (such as acne, facial seborrheic dermatitis, and eczema). We applied transfer learning to our model because parameters in the model after pretraining can be considered as a better initialization than initializing parameters randomly at the beginning of the training [30]. Therefore, our model was pretrained with an ImageNet data set, which contains 1 million images, and fine-tuned on our data set [31]. Since our raw images have different resolutions, we should unify them before feeding them into the model. We resized each image to 256 pixels at their shorter side and kept their aspect ratio. After that, we only reserved the central 256×256 region. In this way, we could obtain many 256 pixel×256 pixel images without aspect ratio distortion.

We used facial cropping, rotation, and flipping to augment our data set. For each image, we randomly sampled a 224×224 crop. Then, we rotated the image by 0, 90, 180, or 270 degrees; 25 of the images in a batch were flipped vertically. Further, 25% of the images in the same batch were flipped horizontally. Images may be chosen to flip vertically or horizontally at the same time. We tried to use more affine transformation to augment our data such as more rotation angles, scaling, and shifting. However, we found that too much affine transformation would cause overfitting more easily. Moreover, the performances were also worse than using only a few augmentation methods. We did not consider color augmentation because we believed that the color of patients' facial skin is one of the keys to determine their disease. Changing the contrast, saturation, and hue would confuse the model.

We optimized our model with mini-batch gradient descent with a momentum of 0.9 and a batch size of 32. Our model was trained for 100 epochs. The initial learning rate was set to 0.0001. If validation loss did not decrease in continuous 10 epochs, the learning rate was divided by 5. The minimum learning rate was not lower than 0.000001. Before training, we randomly split 20% of the data from the training set as the validation set. Further, we used the performance of the validation set to select the best model.

A total of 24,736 photos comprising of 18,647 photos of patients with rosacea and 6089 photos of patients with different skin diseases such as acne, facial seborrheic dermatitis, and eczema were included in our study. The patients in this study gave written informed consent to publish their case details. In order to cover the whole face, the photos of each patient were taken using smartphones (iPhone X and Huawei P20) or digital camera (Canon Rebel 550) from 3 different angles (Figure 1): left face (45 degrees from the left), middle face, and right face (45 degrees from the right).

Figure 1. Examples of 3 photos taken from 3 different angles for each patient.



To build a test set without class imbalance, we did not split the data randomly. Instead, we made sure that each class has the same number of test examples. For each binary classification task, we chose 768 photos from 256 patients (128 patients for rosacea, 128 patients for other skin diseases) as the test set. For rosacea subtype prediction, we chose 576 images from 192 patients (64 patients for each subtype) as the test set. For data analysis, the area under the receiver operating characteristic curve (AUROC) was calculated for each of these curves to quantify the CNN's performance. A confusion matrix was

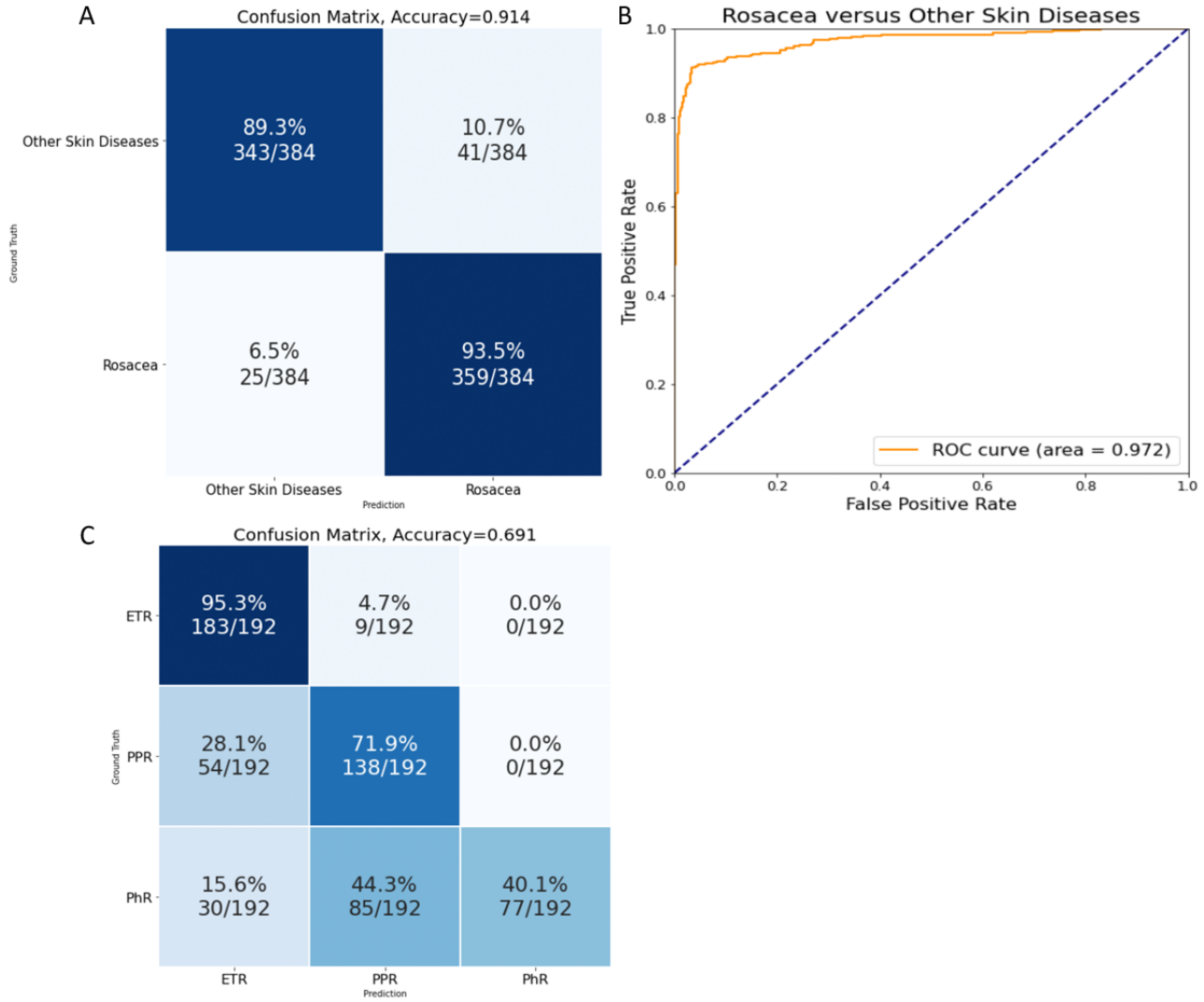
constructed from the results of the testing images to evaluate the performance.

Results

Using Deep CNN to Identify and Classify Rosacea

First, we tested the ability of CNN to identify rosacea (18,647 images) and other skin diseases, which could be easily confused with rosacea in clinic (6089 images). The latter included acne, facial eczema and seborrheic dermatitis, lupus erythematosus, chronic solar dermatitis, corticosteroid-dependent dermatitis, and lupus miliaris disseminatus faciei. Among them, 23,768 images were used for training and the rest were used for testing. The accuracy and precision pf the CNN for the classification of rosacea against other skin diseases were 0.914 and 0.898, respectively, with an AUROC of 0.972 (Figure 2A and Figure 2B), thereby indicating that CNN was able to identify rosacea effectively and accurately from other skin diseases on the face that might be easily confused with rosacea. Next, we tried to utilize the CNN to further classify the 3 major subtypes of rosacea: erythematotelangiectatic rosacea (ETR), papulopustular rosacea (PPR), and phymatous rosacea (PhR). The accuracy of the CNN to classify one subtype against the others was 83.9%, 74.3%, and 80.0% for ETR, PPR, and PhR, respectively (Figure 2C). To be more specific, 28.1% (54/192) of the patients with PPR were mistakenly recognized as having ETR, while 15.6% (30/192) and 44.3% (85/192) of the patients with PhR were misinterpreted as having ETR and PPR, respectively (Figure 2C).

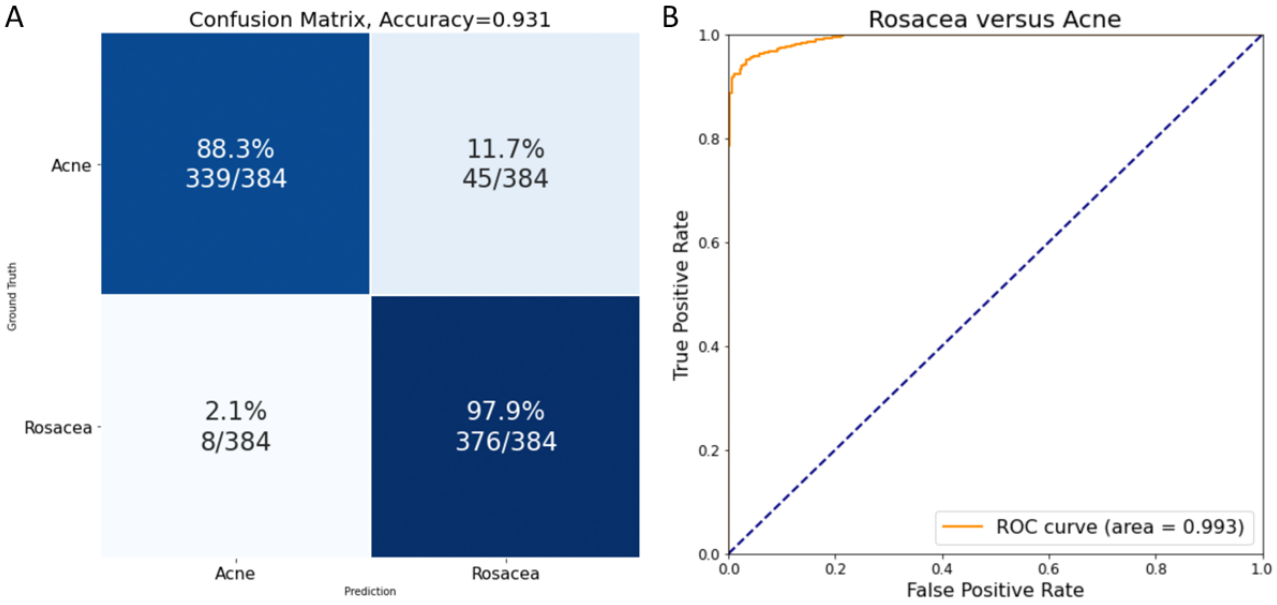
Figure 2. Performance of the convolutional neural network in the identification and classification of rosacea and other skin diseases. A. Confusion matrix showing the accuracy and precision of 0.914 and 0.898, respectively; B. Receiver operating characteristic curve showing that the area under the receiver operating characteristic curve reached 0.972; C. Performance of the convolutional neural network in the classification of subtypes of rosacea. ETR: erythematotelangiectatic rosacea; PhR: phymatous rosacea; PPR: papulopustular rosacea.



Using Deep CNN to Distinguish Rosacea From Acne

Acne is one of the most important disorders considered in the differential diagnosis of rosacea; therefore, we further proceeded to apply our CNN to distinguish rosacea from acne. The total number of images incorporated into this study was 18,647 for rosacea and 3552 for acne. Among them, 21,431 images were used for training and 768 for testing. The accuracy of this test was 0.931 with a precision of 0.893 (Figure 3A). The AUROC was 0.993 (Figure 3B) and the recall was 0.982. These results demonstrated that our CNN was capable of accurately distinguishing rosacea from acne.

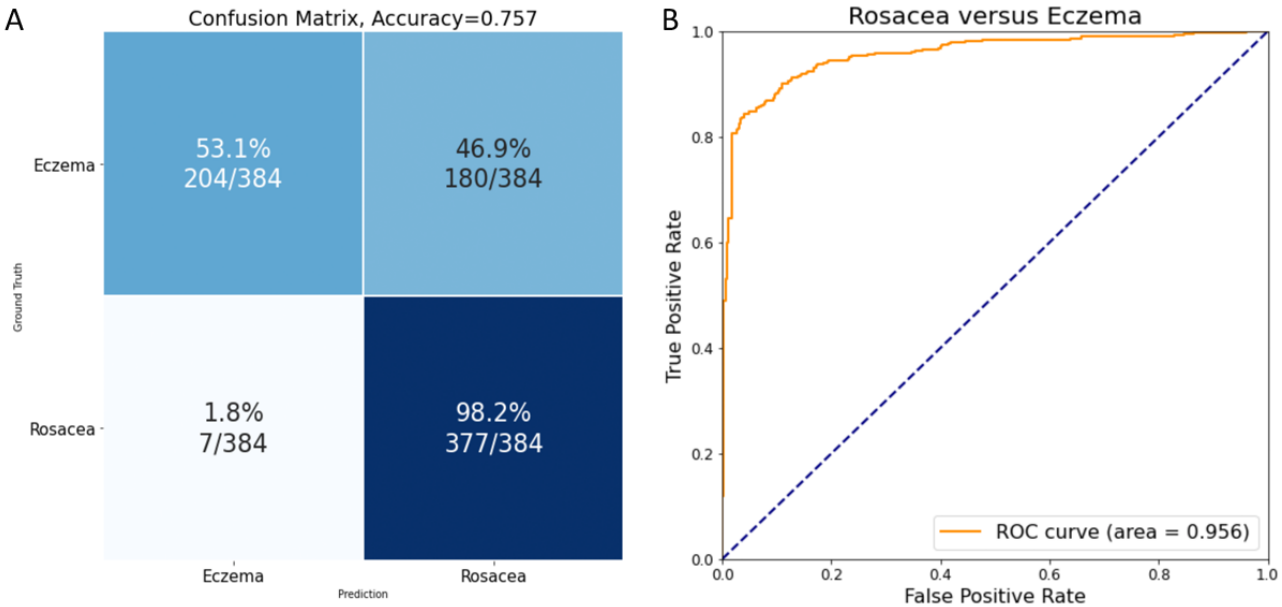
Figure 3. Performance of the convolutional neural network in the identification of rosacea and acne. A. Confusion matrix showing that the accuracy and precision were 0.931 and 0.893, respectively; B. Receiver operating characteristic curve showing that the area under the receiver operating characteristic curve reaches 0.993.



Using Deep CNN to Distinguish Rosacea From Facial Seborrheic Dermatitis/Eczema

Facial seborrheic dermatitis and eczema are other types of facial dermatitis that can be easily misdiagnosed as rosacea in clinical practice. We collected 18,647 images of rosacea and 1896 facial seborrheic dermatitis/eczema images for CNN assessment and identification. After being trained with 19,775 images, the CNN achieved 0.757 for accuracy and 0.677 for precision in the differentiation of rosacea from facial seborrheic dermatitis/eczema on the test set of 768 images (Figure 4A). The overall AUROC of this test was 0.956 (Figure 4B).

Figure 4. Performance of the convolutional neural network in the identification of rosacea and facial seborrheic dermatitis/eczema. A. Confusion matrix showing that the accuracy and precision were 0.757 and 0.677, respectively; B. Receiver operating characteristic curve showing that the area under the receiver operating characteristic curve reaches 0.956.

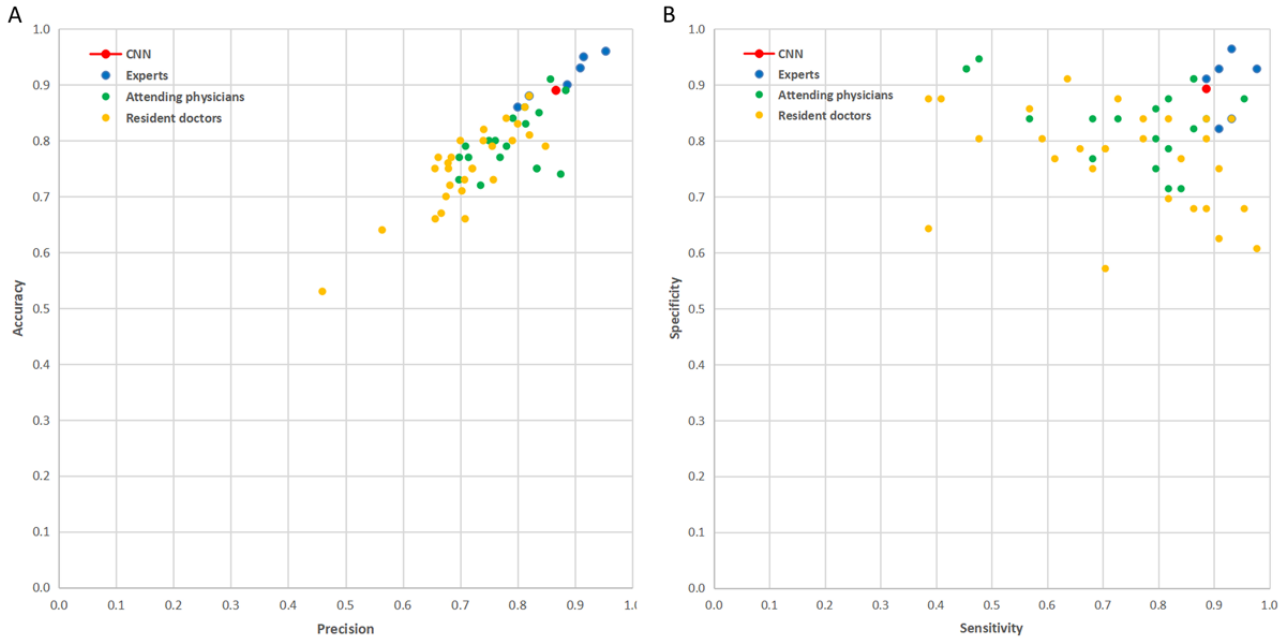


Comparing the Performance of Deep CNN With That of Dermatologists of Different Expertise Levels

We compared the performance of our CNN with that of dermatologists of different expertise levels in the identification of rosacea and other skin diseases. The latter consisted of 6 experts dedicated in the clinical research of rosacea, 19 attending physicians, and 28 resident doctors of dermatology; 44 images of patients with rosacea and 56 images of patients with different skin diseases were used for the test. Compared with our CNN, which achieved an accuracy of 0.890 and precision of 0.867, the overall mean accuracy and precision of the experts were 0.913 (SD 0.040) and 0.881 (SD 0.059), respectively. By contrast, the overall mean accuracy and precision of attending

physicians were 0.803 (SD 0.058) and 0.791 (SD 0.063), while those of resident doctors were 0.75 (SD 0.075) and 0.714 (SD 0.081), respectively (Figure 5A and Figure 5B). In summary, these results indicated that the performance of our CNN was significantly superior to that of resident doctors and attending physicians and was comparable to that of experienced dermatologists in the identification of rosacea.

Figure 5. Performance of the convolutional neural network and dermatologists of different expertise levels in the identification of rosacea and other skin diseases. A. Precision and accuracy of the convolutional neural network compared to those of resident doctors, attending physicians, and experts; B. Sensitivity and specificity of the convolutional neural network compared to those of resident doctors, attending physicians, and experts.



Discussion

Our study offers a novel CNN that can correctly identify and classify the subtypes of rosacea, and the performance of our CNN is comparable to that of expert dermatologists specialized in the diagnosis and treatment of rosacea. Previous efforts have been made to apply CNN to identify rosacea. However, previous work focused mainly on the development of networks or analysis of images instead of practically applying CNN for the identification of rosacea and differentiating it from other skin diseases or for the classification of subtypes of rosacea [32]. Besides, the number of images for model development was quite limited (less than 100) in the previous studies and the sensitivity or specificity were barely satisfactory [33]. In our work, a vast number of images were incorporated for the training of CNN, and the precision and accuracy of our deep CNN system were 0.914 and 0.898, respectively, for the identification of rosacea among other skin diseases. In addition, in the test for detecting rosacea, our CNN system significantly outperformed the resident doctors and the attending physicians and the performance was comparable to that of experienced dermatologists. Thus, our CNN can serve as a unified detection tool and as a promising adjunct for grassroot health care workers (such as family doctors) to improve their capability in recognizing rosacea and narrow the gap between doctors with different clinical experiences. This, in turn, would be also of great benefit for patients with rosacea since it is not easy for the general public to obtain access to experts for dermatological consultation in daily life.

Traditionally, rosacea is categorized into the following different subtypes: ETR, PPR, PhR, and ocular rosacea [4]. However, the boundary between these subtypes (specially ETR and PPR) is quite obscure and these subtypes may overlap or transform from one subtype to another [7]. The correct classification of the different subtypes in clinical practice has always been a challenge for clinicians. In this study, we tried to utilize our CNN to classify ETR, PPR, and PhR, and the precision was 83.9%, 74.3%, and 80.0%, respectively. To be more specific, 28.1% (54/192) of the patients with PPR were mistakenly recognized as having ETR. One possible explanation for this difference in performance could be the overlapping presentations of ETR and PPR, which is commonly seen in clinic. Moreover, 15.6% (30/192) and 44.3% (85/192) of the patients with PhR were misinterpreted as having ETR and PPR, respectively. Possible reasons for these mistakes could be that the erythema of some patients with ETR and the papules of some patients with PPR were confined to the nasal part, making it difficult to distinguish ETR and PPR from PhR, especially when phymas were not prominent. Nowadays, with the growing understanding of rosacea, this traditional subtype classification has been abrogated by the experts committee for rosacea due to the impractical sorting criteria [34]. The confusions in the classification of the subtypes of rosacea by CNN in our study would in turn support the current consensus of abolishing the impractical classification method.

Generally, the data sources for the interpretation of machine learning varies from digital medical records [35,36], histopathological pictures [37], and clinical photos [38-41] to dermoscopic images [42-50]. The advantage of machine learning

over the human eyes is that the CNN is able to objectively record, process, and summarize all the subtle features included in the images of patients with rosacea—even those details that would have been neglected by clinicians otherwise. Each type of data provides unique clinical information for disease diagnosis and at the same time has its own benefits and drawbacks. For example, dermoscopic images are standardized high-resolution pictures, which can clearly present all the microscopic details of the lesion and eliminate the potential interference of image quality and lighting angles with ordinary photos, which makes them especially useful for identifying diseases with specific microscopic patterns (eg, atypical network and irregular dots/globules for melanoma, large blue-grey ovoid nests and spoke-wheel areas for basal cell carcinoma). However, dermoscopic images have a rather limited field of view. For skin diseases such as rosacea, one single dermoscopic image covers only a small proportion of the whole lesion, which hardly represents all the clinical characteristics of the disease comprehensively. In this scenario, clinical photographs taken using the digital camera from 3 different angles covering the whole face (as shown in Figure 1 of our study) would be the preferred image source. Further, clinical photos from different angles also allowed us to analyze areas that were not commonly implicated in rosacea, such as the zygomatic process of the maxilla and the lower mandible. The implication of these areas could be a key point for the differentiation between rosacea and other skin diseases. Additionally, given that each image type has its own limitations, it would be interesting if different types of images (clinical photos, dermoscopic images, and histopathological pictures) were integrated for artificial intelligence assessment at the same time. Future work is encouraged to integrate different types of images providing both microcosmic and macroscopic features of diseases for machine learning to achieve greater performance.

To date, little is known about how exactly deep CNNs analyze, process, and summarize these images and distinguish one disease from another. It remains to be explored what factors (image number, light, background, definition, dimensions) can affect this process. Further efforts are required to investigate the possible ways for improving the accuracy and specificity for the detection of diseases by using artificial intelligence.

In conclusion, our study demonstrated the capability of our deep CNN to identify rosacea and differentiate it from other skin diseases. The performance of our CNN was superior to that of resident doctors and attending physicians and was comparable to that of experienced dermatologists. Our results offer a new potential way for the proper diagnosis of rosacea, thereby indicating that artificial intelligence would be of great help for physicians in the diagnosis of rosacea in clinical practice in the future.

Acknowledgments

This work was supported by The Educational Science and Planning Project of Hunan Province (XTK20BGD008).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Generation of feature map.
[PNG File , 6 KB - medinform_v9i3e23415_app1.png]

Multimedia Appendix 2
Schema of the denoising process.
[PNG File , 3 KB - medinform_v9i3e23415_app2.png]

Multimedia Appendix 3
The architecture of ResNet-50.
[PNG File , 17 KB - medinform_v9i3e23415_app3.png]

Multimedia Appendix 4
The structure of a residual block.
[PNG File , 14 KB - medinform_v9i3e23415_app4.png]

References

1. Rainer BM, Kang S, Chien AL. Rosacea: Epidemiology, pathogenesis, and treatment. Dermatoendocrinol 2017;9(1):e1361574 [FREE Full text] [doi: 10.1080/19381980.2017.1361574] [Medline: 29484096]
2. Li J, Wang B, Deng Y, Shi W, Jian D, Liu F, et al. Epidemiological features of rosacea in Changsha, China: A population-based, cross-sectional study. J Dermatol 2020 May;47(5):497-502. [doi: 10.1111/1346-8138.15301] [Medline: 32207167]

3. van der Linden MMD, van Rappard DC, Daams J, Sprangers M, Spuls P, de Korte J. Health-related quality of life in patients with cutaneous rosacea: a systematic review. Acta Derm Venereol 2015 Apr;95(4):395-400 [FREE Full text] [doi: 10.2340/00015555-1976] [Medline: 25270577]
4. Wilkin J, Dahl M, Detmar M, Drake L, Feinstein A, Odom R, et al. Standard classification of rosacea: Report of the National Rosacea Society Expert Committee on the Classification and Staging of Rosacea. J Am Acad Dermatol 2002 Apr;46(4):584-587. [doi: 10.1067/mjd.2002.120625] [Medline: 11907512]
5. Xie HF, Huang YX, He L, Yang S, Deng YX, Jian D, et al. An observational descriptive survey of rosacea in the Chinese population: clinical features based on the affected locations. PeerJ 2017;5:e3527 [FREE Full text] [doi: 10.7717/peerj.3527] [Medline: 28698821]
6. Zhou MS, Xie H, Cheng L, Li J. Clinical characteristics and epidermal barrier function of papulopustular rosacea: A comparison study with acne vulgaris. Pak J Med Sci 2016;32(6):1344-1348 [FREE Full text] [doi: 10.12669/pjms.326.11236] [Medline: 28083023]
7. Gallo RL, Granstein RD, Kang S, Mannis M, Steinhoff M, Tan J, et al. Standard classification and pathophysiology of rosacea: The 2017 update by the National Rosacea Society Expert Committee. J Am Acad Dermatol 2018 Jan;78(1):148-155. [doi: 10.1016/j.jaad.2017.08.037] [Medline: 29089180]
8. Wang YA, James WD. Update on rosacea classification and its controversies. Cutis 2019 Jul;104(1):70-73. [Medline: 31487337]
9. Nichols JA, Herbert Chan HW, Baker MAB. Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophys Rev 2019 Feb;11(1):111-118 [FREE Full text] [doi: 10.1007/s12551-018-0449-9] [Medline: 30182201]
10. Miller DD, Brown EW. Artificial Intelligence in Medical Practice: The Question to the Answer? Am J Med 2018 Feb;131(2):129-133. [doi: 10.1016/j.amjmed.2017.10.035] [Medline: 29126825]
11. Meng H, Jin W, Yan C, Yang H. The Application of Machine Learning Techniques in Clinical Drug Therapy. Curr Comput Aided Drug Des 2019;15(2):111-119. [doi: 10.2174/1573409914666180525124608] [Medline: 29804538]
12. Li C, Yao Z, Zhu M, Lu B, Xu H. Biopsy-Free Prediction of Pathologic Type of Primary Nephrotic Syndrome Using a Machine Learning Algorithm. Kidney Blood Press Res 2017;42(6):1045-1052 [FREE Full text] [doi: 10.1159/000485592] [Medline: 29197864]
13. Feng Z, Rong P, Cao P, Zhou Q, Zhu W, Yan Z, et al. Machine learning-based quantitative texture analysis of CT images of small renal masses: Differentiation of angiomyolipoma without visible fat from renal cell carcinoma. Eur Radiol 2018 Apr;28(4):1625-1633. [doi: 10.1007/s00330-017-5118-z] [Medline: 29134348]
14. Thomsen K, Iversen L, Titlestad TL, Winther O. Systematic review of machine learning for diagnosis and prognosis in dermatology. J Dermatolog Treat 2020 Aug;31(5):496-510. [doi: 10.1080/09546634.2019.1682500] [Medline: 31625775]
15. Hogarty DT, Su JC, Phan K, Attia M, Hossny M, Nahavandi S, et al. Artificial Intelligence in Dermatology-Where We Are and the Way to the Future: A Review. Am J Clin Dermatol 2020 Feb;21(1):41-47. [doi: 10.1007/s40257-019-00462-6] [Medline: 31278649]
16. Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, et al. Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review. J Med Internet Res 2018 Oct 17;20(10):e11936 [FREE Full text] [doi: 10.2196/11936] [Medline: 30333097]
17. Phillips M, Marsden H, Jaffe W, Matin RN, Wali GN, Greenhalgh J, et al. Assessment of Accuracy of an Artificial Intelligence Algorithm to Detect Melanoma in Images of Skin Lesions. JAMA Netw Open 2019 Oct 02;2(10):e1913436 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.13436] [Medline: 31617929]
18. Brinker TJ, Hekler A, Hauschild A, Berking C, Schilling B, Enk AH, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. Eur J Cancer 2019 Apr;111:30-37 [FREE Full text] [doi: 10.1016/j.ejca.2018.12.016] [Medline: 30802784]
19. Aractingi S, Pellacani G. Computational neural network in melanocytic lesions diagnosis: artificial intelligence to improve diagnosis in dermatology? Eur J Dermatol 2019 Apr 01;29(S1):4-7. [doi: 10.1684/ejd.2019.3538] [Medline: 31017580]
20. Fujisawa Y, Otomo Y, Ogata Y, Nakamura Y, Fujita R, Ishitsuka Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. Br J Dermatol 2019 Feb;180(2):373-381. [doi: 10.1111/bjd.16924] [Medline: 29953582]
21. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Jan 25;542(7639):115-118. [doi: 10.1038/nature21056]
22. Choudhury A, Asan O. Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review. JMIR Med Inform 2020 Jul 24;8(7):e18599 [FREE Full text] [doi: 10.2196/18599] [Medline: 32706688]
23. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. IEEE 1998 Nov;86(11):2278-2324. [doi: 10.1109/5.726791]
24. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun. ACM 2017 May 24;60(6):84-90. [doi: 10.1145/3065386]
25. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell 2017 Jun 1;39(6):1137-1149. [doi: 10.1109/tpami.2016.2577031]

26. Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Trans. Pattern Anal. Mach. Intell 2017 Apr 1;39(4):640-651. [doi: 10.1109/tpami.2016.2572683]
27. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986 Oct;323(6088):533-536. [doi: 10.1038/323533a0]
28. Ruder S. An overview of gradient descent optimization algorithms. arXiv.org. 2017 Jun 15. URL: https://arxiv.org/abs/1609.04747v2 [accessed 2020-01-15]
29. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 Jun 30 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; Las Vegas, NV, USA p. 770-778. [doi: 10.1109/CVPR.2016.90]
30. Caruana R. Multitask Learning. Machine Learning 1997:41-75. [doi: 10.1023/A:1007379606734]
31. Jia D, Wei D, Richard S, Li-Jia L, Kai L, Li FF. ImageNet: A large-scale hierarchical image database. 2009 Jun 25 Presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009; Miami, FL, USA p. 248-255. [doi: 10.1109/CVPR.2009.5206848]
32. Binol H, Plotner A, Sopkovich J, Kaffenberger B, Niazi MKK, Gurcan MN. Ros-NET: A deep convolutional neural network for automatic identification of rosacea lesions. Skin Res Technol 2020 May;26(3):413-421. [doi: 10.1111/srt.12817] [Medline: 31849118]
33. Aggarwal SLP. Data augmentation in dermatology image recognition using machine learning. Skin Res Technol 2019 Nov;25(6):815-820. [doi: 10.1111/srt.12726] [Medline: 31140653]
34. Schaller M, Almeida LMC, Bewley A, Cribier B, Del Rosso J, Dlova NC, et al. Recommendations for rosacea diagnosis, classification and management: update from the global ROSacea COnsensus 2019 panel. Br J Dermatol 2020 May;182(5):1269-1276 [FREE Full text] [doi: 10.1111/bjd.18420] [Medline: 31392722]
35. Gustafson E, Pacheco J, Wehbe F, Silverberg J, Thompson W. A Machine Learning Algorithm for Identifying Atopic Dermatitis in Adults from Electronic Health Records. IEEE Int Conf Healthc Inform 2017 Aug;2017:83-90 [FREE Full text] [doi: 10.1109/ICHI.2017.31] [Medline: 29104964]
36. Zhang H, Ni W, Li J, Zhang J. Artificial Intelligence-Based Traditional Chinese Medicine Assistive Diagnostic System: Validation Study. JMIR Med Inform 2020 Jun 15;8(6):e17608 [FREE Full text] [doi: 10.2196/17608] [Medline: 32538797]
37. Hekler A, Utikal JS, Enk AH, Berking C, Klode J, Schadendorf D, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. Eur J Cancer 2019 Jul;115:79-83 [FREE Full text] [doi: 10.1016/j.ejca.2019.04.021] [Medline: 31129383]
38. Min S, Kong H, Yoon C, Kim HC, Suh DH. Development and evaluation of an automatic acne lesion detection program using digital image processing. Skin Res Technol 2013 Feb;19(1):e423-e432. [doi: 10.1111/j.1600-0846.2012.00660.x] [Medline: 22891680]
39. Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. Computer-aided diagnosis of psoriasis skin images with HOS, texture and color features: A first comparative study of its kind. Comput Methods Programs Biomed 2016 Apr;126:98-109. [doi: 10.1016/j.cmpb.2015.11.013] [Medline: 26830378]
40. Lu J, Kazmierczak E, Manton JH, Sinclair R. Automatic Segmentation of Scaling in 2-D Psoriasis Skin Images. IEEE Trans. Med. Imaging 2013 Apr;32(4):719-730. [doi: 10.1109/tmi.2012.2236349]
41. Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. A novel and robust Bayesian approach for segmentation of psoriasis lesions and its risk stratification. Comput Methods Programs Biomed 2017 Oct;150:9-22. [doi: 10.1016/j.cmpb.2017.07.011] [Medline: 28859832]
42. Marchetti MA, Codella NC, Dusza SW, Gutman DA, Helba B, Kalloo A, International Skin Imaging Collaboration. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018 Feb;78(2):270-277.e1 [FREE Full text] [doi: 10.1016/j.jaad.2017.08.016] [Medline: 28969863]
43. Yu C, Yang S, Kim W, Jung J, Chung K, Lee SW, et al. Acral melanoma detection using a convolutional neural network for dermoscopy images. PLoS ONE 2018 Mar 7;13(3):e0193321. [doi: 10.1371/journal.pone.0193321]
44. Tschandl P, Kittler H, Argenziano G. A pretrained neural network shows similar diagnostic accuracy to medical students in categorizing dermatoscopic images after comparable training conditions. Br J Dermatol 2017 Sep;177(3):867-869. [doi: 10.1111/bjd.15695] [Medline: 28569993]
45. Xie F, Fan H, Li Y, Jiang Z, Meng R, Bovik A. Melanoma Classification on Dermoscopy Images Using a Neural Network Ensemble Model. IEEE Trans. Med. Imaging 2017 Mar;36(3):849-858. [doi: 10.1109/tmi.2016.2633551]
46. Shimizu K, Iyatomi H, Celebi ME, Norton K, Tanaka M. Four-Class Classification of Skin Lesions With Task Decomposition Strategy. IEEE Trans. Biomed. Eng 2015 Jan;62(1):274-283. [doi: 10.1109/tbme.2014.2348323]
47. Bi L, Kim J, Ahn E, Kumar A, Fulham M, Feng D. Dermoscopic Image Segmentation via Multistage Fully Convolutional Networks. IEEE Trans. Biomed. Eng 2017 Sep;64(9):2065-2074. [doi: 10.1109/tbme.2017.2712771]
48. Li Y, Shen L. Skin Lesion Analysis towards Melanoma Detection Using Deep Learning Network. Sensors (Basel) 2018 Feb 11;18(2) [FREE Full text] [doi: 10.3390/s18020556] [Medline: 29439500]

49. Lingala M, Stanley RJ, Rader RK, Hagerty J, Rabinovitz HS, Oliviero M, et al. Fuzzy logic color detection: Blue areas in melanoma dermoscopy images. Comput Med Imaging Graph 2014 Jul;38(5):403-410 [FREE Full text] [doi: 10.1016/j.compmedimag.2014.03.007] [Medline: 24786720]
50. Premaladha J, Ravichandran KS. Novel Approaches for Diagnosing Melanoma Skin Lesions Through Supervised and Deep Learning Algorithms. J Med Syst 2016 Apr;40(4):96. [doi: 10.1007/s10916-016-0460-2] [Medline: 26872778]

Abbreviations

AUROC: area under the receiver operating characteristic curve
CNN: convolutional neural network
ETR: erythematotelangiectatic rosacea
PhR: phymatous rosacea
PPR: papulopustular rosacea

Edited by G Eysenbach; submitted 12.08.20; peer-reviewed by I Gabashvili, S Gupta; comments to author 11.09.20; revised version received 30.11.20; accepted 12.12.20; published 15.03.21.

Please cite as:
Zhao Z, Wu CM, Zhang S, He F, Liu F, Wang B, Huang Y, Shi W, Jian D, Xie H, Yeh CY, Li J
A Novel Convolutional Neural Network for the Diagnosis and Classification of Rosacea: Usability Study
JMIR Med Inform 2021;9(3):e23415
URL: https://medinform.jmir.org/2021/3/e23415
doi:10.2196/23415
PMID:33720027

Original Paper

Regional Resource Assessment During the COVID-19 Pandemic in Italy: Modeling Study

Pietro H Guzzi[1*], PhD; Giuseppe Tradigo[2*], PhD; Pierangelo Veltri[1], PhD

[1]Department of Surgical and Medical Sciences, University of Catanzaro, CZ, Italy

[2]Ecampus University, Novedrate, Italy

[*]these authors contributed equally

Corresponding Author:
Pietro H Guzzi, PhD
Department of Surgical and Medical Sciences
University of Catanzaro
Catanzaro
CZ,
Italy
Phone: 39 09613694148
Email: hguzzi@unicz.it

Abstract

Background: COVID-19 has been declared a worldwide emergency and a pandemic by the World Health Organization. It started in China in December 2019, and it rapidly spread throughout Italy, which was the most affected country after China. The pandemic affected all countries with similarly negative effects on the population and health care structures.

Objective: The evolution of the COVID-19 infections and the way such a phenomenon can be characterized in terms of resources and planning has to be considered. One of the most critical resources has been intensive care units (ICUs) with respect to the infection trend and critical hospitalization.

Methods: We propose a model to estimate the needed number of places in ICUs during the most acute phase of the infection. We also define a scalable geographic model to plan emergency and future management of patients with COVID-19 by planning their reallocation in health structures of other regions.

Results: We applied and assessed the prediction method both at the national and regional levels. ICU bed prediction was tested with respect to real data provided by the Italian government. We showed that our model is able to predict, with a reliable error in terms of resource complexity, estimation parameters used in health care structures. In addition, the proposed method is scalable at different geographic levels. This is relevant for pandemics such as COVID-19, which has shown different case incidences even among northern and southern Italian regions.

Conclusions: Our contribution can be useful for decision makers to plan resources to guarantee patient management, but it can also be considered as a reference model for potential upcoming waves of COVID-19 and similar emergency situations.

(JMIR Med Inform 2021;9(3):e18933) doi:10.2196/18933

KEYWORDS

COVID-19; data analysis; ICU; management; intensive care unit; pandemic; outbreak; infectious disease; resource; planning

Introduction

COVID-19 is a disease that was reported in Wuhan, China in December 2019 [1]. It has been stressing health structures and governments worldwide due to the difficulties in containing its diffusion [2-4]. The virus appears as a flu, but it attacks the pulmonary apparatus, and on average, 6% of symptomatic patients require hospitalization in intensive care units (ICUs) due to severe respiratory syndromes. The rapid diffusion of the virus caused a high number of infections. Only a fraction of patients who are infected need hospitalization in ICUs, a relatively high number of which do not survive the virus. In Italy, in almost 2 months during the first peak, there have been more than 100,000 known infections and nearly 35,000 COVID-19–related deaths [5-7].

The virus also spread in other countries across the world with different modalities and aggressiveness. During its first wave, from March to May 2020, Italy had the second highest number

of patients who were infected. All countries reported difficulties in answering to the high number of requests for ICU beds and reliable detection of the real infection numbers. In countries where the virus spread later with respect to Italy, such as the United States, France, Spain, and other European countries, it diffused with similar trends but with different absolute numbers, as reported by the World Health Organization [3,8-13]. Nevertheless, the impact on health structure resources has been similar. In fact, the number of infections detected is strictly related to the number of swab tests performed in the population. Tradigo et al [14] assessed the real number of people who are infected with respect to the known ones. In contrast, the number of infections is much higher than the known ones and includes patients who are asymptomatic (ie, those who got the infection but who did not manifest any symptoms). The number of ICU beds is always related to the number of real infections.

Regions such as Lombardia in the north of Italy have been strongly affected by COVID-19, and it seems that this is due to a large number of patients who were asymptomatic already in the middle of January 2020 and late adoption of containment measures such as limitation of circulation and delay in applying rules such as smart working [15-17].

We focus on the problem of rapidly estimating resources during the exponential phase of the COVID-19 emergency, in particular, being aware of the differences among regions in terms of health structure resources such as ICU beds. We show how the proposed model scales at the regional level and how it can help decision makers plan expansions of resources near saturation or reroute patients to neighboring regions.

The prediction of COVID-19 diffusion is a relevant problem, and it has been discussed in other papers [18-20]. Some papers do not consider the regional level and local differences that are relevant in some countries such as Italy. For instance, Li et al [21] developed a model starting from Hubei Province data, and they used the model for prediction in other countries such as Italy and Korea. Other papers did not consider the prediction of ICU resources. Conversely, our study is scalable on a regional level, and it is able to predict ICU needs. In a previous study [22-24], we developed a preliminary model for resource planning; here, we present an extension of the model with a particular focus on the assessment of the predictions.

The model presented here has been assessed comparing the simulated and predicted resource values with the measured ones. The rapid diffusion trend in high-income countries' populations and in cities with high density stressed the health structure in many countries. Indeed, a small portion of patients with COVID-19 require ICU admission. The exponential diffusion in terms of an increased number of infections per day required a larger number of ICU beds than the ones available. We report our model as being scalable at both the regional and subregional levels. We claim that it can be used in different countries and in future contexts where virus diffusion will require well-planned health resource management [13,25].

The paper is structured as follows. The Methods section reports the proposed assessed model and the Italian infection data. The Results section reports the application of the model on three sample regions out of 20, Lombardia (north), Toscana (central), and Sicilia (south). The Discussion section reports on the limitations of this study and comparisons with other work. However, our model is general enough to be successfully applied to other pandemic situations in other contexts.

Methods

ICU Situation in Italy

Italy was affected by COVID-19 by the end of January 2020, starting from northern regions such as Lombardia and Veneto. By the end of February, the increasing trend of infection numbers per day obliged the governments at the regional level—and at the national level starting on March 10–to introduce containment measures.

For example, on March 26, 2020, in Italy, we had 24,747 total reported COVID-19 infection cases, of which 20,603 had the disease, 1809 had died, and 2335 had recovered from it. Regarding patients who were infected, 9268 were treated in their homes since they did not have severe illness, 9663 were hospitalized, and 1672 were admitted to ICUs. The trend continued increasing until April 19, 2020, which has been the peak of COVID-19 infections in Italy.

In reaction to the exponential growth of patients who are infected that require hospitalization, one possible measure adopted by many countries has been to build emergency hospitals dedicated to patients with COVID. In Italy, one strategy consisted in improving existing structures by extending the number of ICU resources and beds, and using dedicated health structures. For example, one study [26] focuses on accelerating the process of acquiring and furnishing hospitals with assisted breathing devices.

Italy has approximately 5200 ICU beds in total, which have been dimensioned by design to be equal to 80% of their average occupancy at any given time. In addition, they are allocated at a regional level proportional to the local population and are usually managed locally. Table 1 reports the ICU bed distribution among regions associated with the demography. The COVID-19 pandemic called these choices into question, thus introducing the necessity of emergency units in cities where the virus rapidly diffused and where existing resources were limited.

Table 1. Distribution of ICU beds in each Italian region ordered by regional population. The number of beds could increase in the future due to government investments for the emergency.

Region	ICU[a] beds, n	Population, n	ICU beds per citizen (%)
Lombardia	1067	10,060,574	0.0106
Lazio	590	5,879,082	0.0100
Campania	350	5,801,692	0.0060
Sicilia	346	4,999,891	0.0069
Veneto	498	4,905,854	0.0102
Emilia Romagna	539	4,459,477	0.0121
Piemonte	320	4,356,406	0.0073
Puglia	210	4,029,053	0.0052
Toscana	450	3,729,641	0.0121
Calabria	110	1,947,131	0.0056
Sardegna	150	1,639,591	0.0091
Liguria	70	1,550,640	0.0045
Marche	108	1,525,271	0.0071
Abruzzo	73	1,311,580	0.0056
Friuli Venezia Giulia	80	1,215,220	0.0066
Trentino Alto Adige	71	1,072,276	0.0066
Umbria	30	882,015	0.0034
Basilicata	49	562,869	0.0087
Molise	30	305,617	0.0098
Valle D'Aosta	15	125,666	0.0119
Italy (total)	5156	60,359,546	0.0085

[a]ICU: intensive care unit.

Because of ICU bed limitations, many patients have been moved from ICUs to subintensive units or to other regions to free up spaces. Indeed, ICU slots are often used for treating postsurgery patients and patients affected by pulmonary diseases. At the date of the peak (ie, April 19, 2020), almost 2635 ICU beds were occupied, 108,257 infections were confirmed by swab tests, and 25,033 patients had recovered without using ICU beds. Thus, most of the infections were asymptomatic, and patients quarantined at home. Even if the number of required ICU beds is less than the total number of available ICU beds in Italy (see Table 1), the infection distribution is not homogeneous among regional departments and does not follow a regular geographical distribution.

Thus, performing a flexible and reliable model that can predict and control resource requirements and distribution at a regional scale is required. The number of patients in the ICUs is also related to the requests of other clinical units such as emergency units for non–COVID-19, but still serious, diseases (eg, cardiovascular-affected patients).

Moreover, considering that the average survival time for patients with COVID-19 that die has been measured to be approximately 10 days after ICU admission, the need to plan resources is urgent. It may involve making new ICU beds and planning logistics to move patients among regions or to optimize the grouping of patients with COVID-19 in dedicated health structures. It is trivial that such a decision must be based on the correct estimation of ICU beds that are occupied by patients, but this estimation is still a matter of discussion [26].

Model Description and Assessment

We report here the description and assessment of the proposed model by using Italian cases.

We start by considering a time window of six consecutive infection values (one reading per day) from the official Italian COVID-19 data set. We then calculate an exponential fitting function for these values, since we know that the viral phase follows an exponential growth.

In Figure 1, the first four time windows (ie, 6, 7, 8, and 9 days) and the related fitting functions are reported. The exponential fitting function for the first window is  , where x is time (ie, days) and y is the number of infections. We use the calculated fitting equation to predict the number of patients infected with COVID-19 for the succeeding days; for the first time window, we predicted 1086 total infections on the seventh day. We compared the predicted value with the observed infections (ie, real number of infections) from the data set, and we calculated the difference and the percentage increase, which will be useful during the assessment phase. We then proceeded by extending

the time window by 1 day, and we redid all the steps (exponential fitting, prediction of the infections for the succeeding day, difference, and percentage increase).

To assess the precision of the first step, we considered the calculated percentage increase between the predicted values and the observed ones. As reported in Figure 2, the percentage increase was around 40%.

In the second step, we consider the number of occupied ICU beds as a function of the number of COVID-19 infections. In this case, we adopted the weighted average as a fitting function. Figure 3 depicts this correlation (in blue) between total infections (x-axis) and ICU beds occupied (y-axis) together with the weighted average fitting for the whole data set (in light blue).

Figure 1. Exponential fitting of infection levels for the first four time windows. Each is longer than the previous by 1 day. The shown time windows are W1 (6 days), W2 (7 days), W3 (8 days), and W4 (9 days).

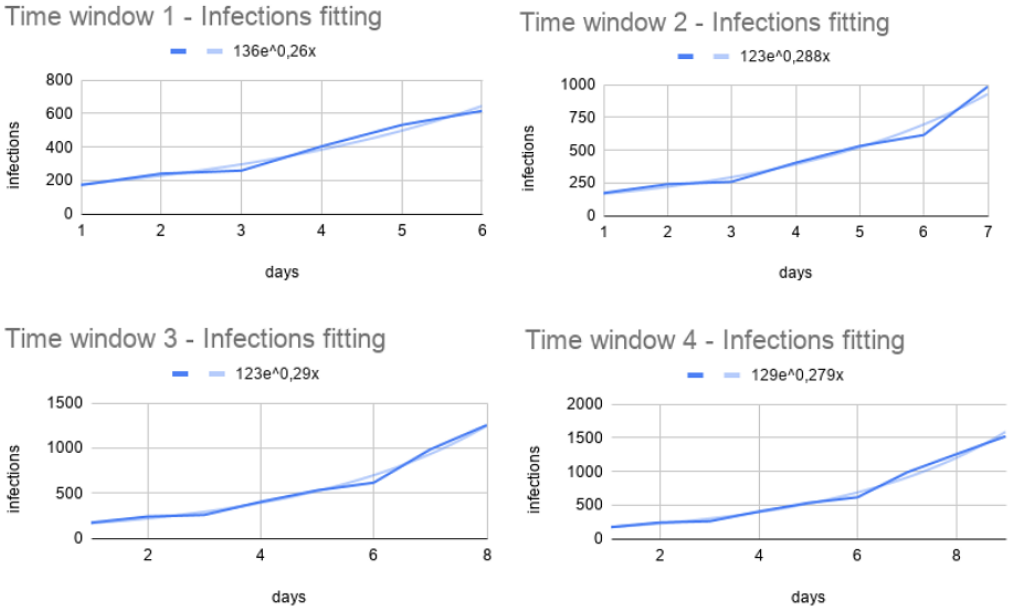


Figure 2. Percentage increase between the predicted and the observed number of infections. On the x-axis, we have time windows and, on the y-axis, the percentage increase of the predicted infection values with respect to the observed ones.

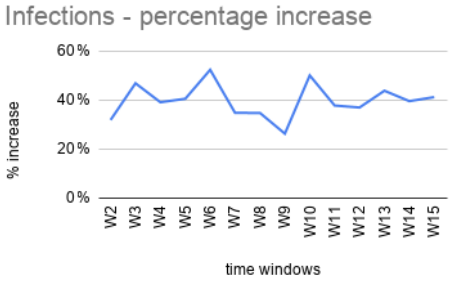
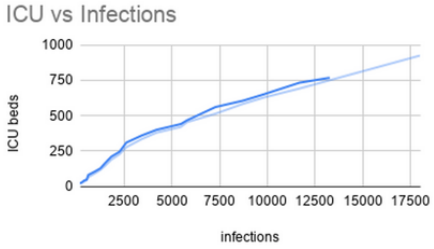


Figure 3. Correlation between occupied ICU beds and COVID-19 infections. ICU: intensive care unit.



For each time window (W1, W2, …, W15; see Figure 4), we consider three adjacent data point coordinates (infections for the x-axes and ICU for the y-axes) to calculate a linear equation as a weighted average fitting function for the values contained in it. We then used such a function to estimate the future ICU bed occupation for the following day by using the predicted infected value. We then calculated the difference between the predicted and the observed ICU values, and similarly to the first step, we reported the percentage increase between the two. Table 2 reports an example of percentage increase values of the predicted versus observed ICU resources for an Italian region's data set. The percentage increase is above 40% for only a few values, but the majority are near 20%, as represented in Figure 5.

Figure 4. Time windows W1, W2, …, W15 are defined incrementally starting from a period of 6 consecutive days, which has been considered the minimum number of data points to calculate the fitting function.

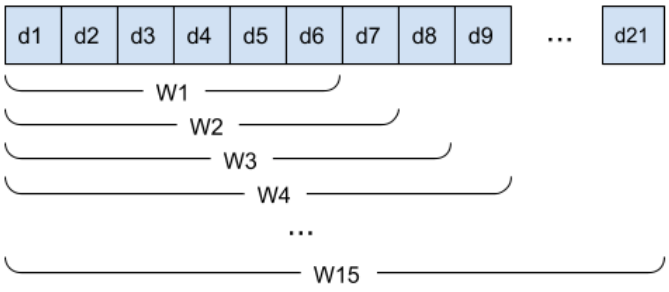


Table 2. Observed versus predicted ICU beds for the Lombardia Region for each time window (W1, ..., W15) considered in the time period from February 24 to March 15, 2020.

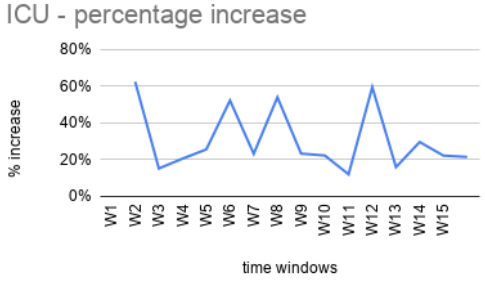
Observed ICU[a] beds[b], n	Predicted ICU beds[c], n	Percentage increase (%)
106	172	62
127	146	15
167	201	20
209	262	25
244	371	52
309	380	23
359	552	54
399	491	23
440	537	22
466	521	12
560	892	59
605	700	16
650	841	29
732	893	22
767	930	21

[a]ICU: intensive care unit.

[b]Observed (ie, real) ICU beds measured during the COVID-19 emergency.

[c]Calculated by our model.

Figure 5. Percentage increase between the predicted and the observed ICU beds occupied. ICU: intensive care unit.



We applied the described method both at the national and regional scale, and we report the results in the next section.

Results

In this section, we show the application of the described model to the Italian official COVID-19 data set [27], and we briefly discuss the limitations of the current model and its application to the Italian use case at a regional level. We show that, by using our method, it has been possible to predict future ICU bed occupancy with fair accuracy.

The proposed model works in the exponential phase of the infection spread, while for the nonexponential stage, other models can be used such as the Verhulst logistic model [9]. In

its present form, the model is tailored to the Italian COVID-19 data set. However, with minimal adaptation, it could work with other data sets of infectious diseases with different data schemas.

The presented model works in the exponential phase of the infection. Model sensitivity has not been considered in the case of this model because we focused only on the exponential growth of the infections, which is the crucial moment that ICU bed and resource availability are most stressed and inadequate.

Changing the assumption (ie, nonexponential growth) is considering a time window in which ICU resources are surely available with respect to the requests. For instance, we performed experiments that modeled the nonexponential infection phase with a Verhulst logistic model, but again, when there is no emergency, the ICU predictions are not useful since resources are largely available.

The final goal is to predict the number of future beds needed in the ICUs as a function of the level of infection in a given region. The ability to predict future resource occupation can be a powerful and useful tool for local decision makers with the responsibility of managing and optimizing clinical resources during the emergency.

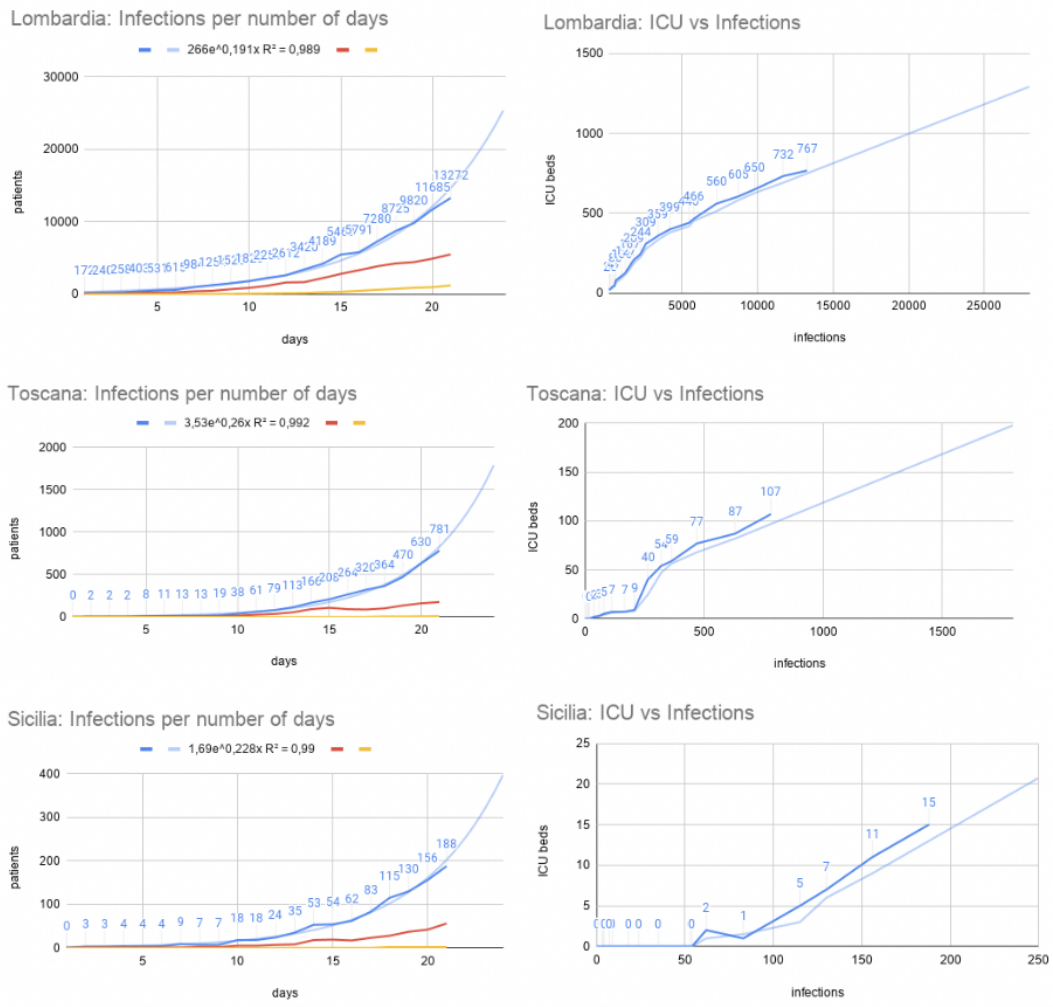
We report the application of the presented model for three Italian regions: Lombardia, Toscana, and Sicilia, which represent a balanced sample of northern, central, and southern Italian regions. We extracted and transformed the relevant data from the official Italian COVID-19 data set and considered the total number of patients who were infected and the ICU beds occupied by patients with COVID-19 reported by the various local health structures.

In the data set, we have 17 total features (eg, latitude, longitude, date, the total number of infections, number of patients who are hospitalized, number of deaths, and number of recoveries) and one reading per day for each region. We selected the features of interest and aggregated the tuples by region before estimating the prediction model parameters. For each region, we considered the number of infections, and we calculated an exponential fitting equation. By using such an equation, we were able to estimate the number of patients who will be infected in the succeeding days. We then considered the relation between ICUs and infections, with which we can use the predicted infection levels to estimate future values of ICU and resource occupation. These predictions can be a valuable tool for rapidly planning ICU resources in case of shortages during clinical emergencies in general, for instance, by reallocating patients in other regions with lower levels of ICU occupancies.

In Figure 6 (upper left), we report COVID-19 data about the Lombardia region between February 24 and March 15, 2020, a time period in which infection levels (in blue) were growing in an exponential fashion. We report the exponential fitting function for the infections (in light blue), the number of hospitalized with symptoms (in red), and the number of deaths related to COVID-19 (in yellow). Figure 6 (upper right) depicts the occupied ICU beds as a function of the number of people infected with COVID-19 in the same aforementioned time period. The fitting function (in light blue) is a weighted average with a modulus of 3, with which we predict future ICU bed occupation from a given infection value.

Similar to the Lombardia region, we report the application of the proposed model to COVID-19 infection data and resource necessity prediction for both the Toscana and Sicilia regions (central and lower part of Figure 6, respectively). We report three example regions (Figure 6) to represent the northern regions (ie, Lombardia; which are the more affected part of the country), the central regions (ie, Toscana), and the south (ie, Sicilia). They are needed to show how COVID-19 diffused differently from the north to the south of Italy.

Figure 6. The figure depicts Lombardia, Toscana, and Sicilia regions as representative of infection situations in the northern, central, and southern regions of Italy, respectively, in a time window between days (0 days to 21 days; ie, February 24 to March 15, 2020) of infection in the official data set (ie, 3 weeks). In the left column, we report the number of infections per day, while in the right column, we have the occupied ICU beds as a function of the number of COVID-19 infections. In light blue, we report the fitting functions for the considered data (left column: exponential function; right column: weighted average with a modulus of 3). The red lines in the left column show the hospitalized patients with symptoms, and the yellow lines show the number of deaths. ICU: intensive care unit.



Discussion

Limitations

The presented model only works in a scenario of an exponential growth of infections, since it is a crucial moment in which ICU bed and resource availability are most stressed and inadequate.

The nonexponential phase might be modeled, for instance, by using models such as the Verhulst logistic one. However, our focus is on time windows in which resources are scarcely available with respect to the rapidly increasing requests, such as ICU beds in the COVID-19 pandemic.

Another limitation of the proposed model regards the impossibility of considering predictions too far ahead in the future, limiting the applicability of the prediction to a few days. However, this is generally sufficient to help in planning ICU resources during an emergency.

Comparison With Prior Work

Modeling of the COVID-19 spread is currently a hot topic considering the pandemic. Consequently, many different works have been proposed. Some of them are based on a deterministic model that uses ordinary differential equations for predicting the number of infected people (eg, [11,12,28]). Some other approaches use Markov modeling and compartmental models (eg, [29-31]). To the best of our knowledge, only the work of Rossman et al [10] presents a scalable granularity (at a state level). With respect to those works, we also tried to predict ICU needs (including data about existing ICU occupancy and the trend of ICU use) with the aim to support health care managers.

Conclusion

The COVID-19 pandemic has been characterized by the rapid spread of an aggressive virus, which has stressed the health system. We think that patient management is strictly related to the ability of health structures to deal with this kind of disease, which requires nonstandard protocols such as the use of respiratory devices. We think that, by using a scalable predictive

model at regional and district levels, the granularities may support decision makers (eg, national governments) in better managing the emergency.

The COVID-19 pandemic has reached different regions in various countries worldwide. Furthermore, it is expected that the virus will cyclically reappear in the near future. To this end, the proposed model could be applied during these new outbreaks and as a decision support tool in other similar pandemics or situations where resource prediction is necessary.

Acknowledgments

We thank Italian Protezione Civile for freely providing online data, which allows studies on the COVID-19 pandemic.

Conflicts of Interest

None declared.

References

1. Malta M, Rimoin AW, Strathdee SA. The coronavirus 2019-nCoV epidemic: is hindsight 20/20? EClinicalMedicine 2020 Mar;20:100289 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100289] [Medline: 32154505]
2. Arabi YM, Fowler R, Hayden FG. Critical care management of adults with community-acquired severe respiratory viral infection. Intensive Care Med 2020 Feb;46(2):315-328 [FREE Full text] [doi: 10.1007/s00134-020-05943-5] [Medline: 32040667]
3. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science 2020 Apr 24;368(6489):395-400 [FREE Full text] [doi: 10.1126/science.aba9757] [Medline: 32144116]
4. Day M. Covid-19: Italy confirms 11 deaths as cases spread from north. BMJ 2020 Feb 26;368:m757. [doi: 10.1136/bmj.m757] [Medline: 32102793]
5. Gaeta G. Data analysis for the COVID-19 early dynamics in Northern Italy. arXiv. Preprint posted online March 4, 2020.
6. Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020 Apr 30;382(18):1708-1720. [doi: 10.1056/nejmoa2002032]
7. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020 Feb 15;395(10223):497-506 [FREE Full text] [doi: 10.1016/S0140-6736(20)30183-5] [Medline: 31986264]
8. WHO Coronavirus Disease (COVID-19) Dashboard. World Health Organization. 2020. URL: https://covid19.who.int/ [accessed 2020-07-05]
9. Bacaër N. Verhulst and the logistic equation (1838). In: A Short History of Mathematical Population Dynamics. London: Springer; 2011:35-39.
10. Rossman H, Keshet A, Shilo S, Gavrieli A, Bauman T, Cohen O, et al. A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys. Nat Med 2020 May;26(5):634-638 [FREE Full text] [doi: 10.1038/s41591-020-0857-9] [Medline: 32273611]
11. Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. J Travel Med 2020 Mar 13;27(2) [FREE Full text] [doi: 10.1093/jtm/taaa021] [Medline: 32052846]
12. Tang B, Bragazzi NL, Li Q, Tang S, Xiao Y, Wu J. An updated estimation of the risk of transmission of the novel coronavirus (2019-nCov). Infect Dis Model 2020;5:248-255 [FREE Full text] [doi: 10.1016/j.idm.2020.02.001] [Medline: 32099934]
13. Zhao S, Musa SS, Lin Q, Ran J, Yang G, Wang W, et al. Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of January 2020: a data-driven modelling analysis of the early outbreak. J Clin Med 2020 Feb 01;9(2) [FREE Full text] [doi: 10.3390/jcm9020388] [Medline: 32024089]
14. Tradigo G, Guzzi PH, Veltri P. On the assessment of more reliable COVID-19 infected number: the italian case. medRxiv. Preprint posted online March 27, 2020 2021. [doi: 10.1101/2020.03.25.20043562]
15. Keeling MJ, Rohani P. Modeling Infectious Diseases in Humans and Animals. Princeton, NJ: Princeton University Press; 2008.
16. Liew MF, Siow WT, MacLaren G, See KC. Preparing for COVID-19: early experience from an intensive care unit in Singapore. Crit Care 2020 Mar 09;24(1):83 [FREE Full text] [doi: 10.1186/s13054-020-2814-x] [Medline: 32151274]
17. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020 Feb 22;395(10224):565-574 [FREE Full text] [doi: 10.1016/S0140-6736(20)30251-8] [Medline: 32007145]
18. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]

19. Huang C, Xu X, Cai Y, Ge Q, Zeng G, Li X, et al. Mining the characteristics of COVID-19 patients in China: analysis of social media posts. J Med Internet Res 2020 May 17;22(5):e19087 [FREE Full text] [doi: 10.2196/19087] [Medline: 32401210]
20. Ambikapathy B, Krishnamurthy K. Mathematical modelling to assess the impact of lockdown on COVID-19 transmission in India: model development and validation. JMIR Public Health Surveill 2020 May 07;6(2):e19368 [FREE Full text] [doi: 10.2196/19368] [Medline: 32365045]
21. Li L, Yang Z, Dang Z, Meng C, Huang J, Meng H, et al. Propagation analysis and prediction of the COVID-19. Infect Dis Model 2020;5:282-292 [FREE Full text] [doi: 10.1016/j.idm.2020.03.002] [Medline: 32292868]
22. Rezaeetalab F, Mozdourian M, Amini M, Javidarabshahi Z, Akbari F. COVID-19: a new virus as a potential rapidly spreading in the worldwide. J Cardio-Thoracic Med 2020;8(1):563-564.
23. Yaesoubi R, Cohen T. Generalized Markov models of infectious disease spread: a novel framework for developing dynamic health policies. Eur J Oper Res 2011 Dec 16;215(3):679-687 [FREE Full text] [doi: 10.1016/j.ejor.2011.07.016] [Medline: 21966083]
24. Guzzi PH, Tradigo G, Veltri P. Spatio-temporal resource mapping for intensive care units at regional level for COVID-19 emergency in Italy. Int J Environ Res Public Health 2020 May 12;17(10) [FREE Full text] [doi: 10.3390/ijerph17103344] [Medline: 32408508]
25. Jung S, Akhmetzhanov AR, Hayashi K, Linton NM, Yang Y, Yuan B, et al. Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: inference using exported cases. J Clin Med 2020 Feb 14;9(2) [FREE Full text] [doi: 10.3390/jcm9020523] [Medline: 32075152]
26. Remuzzi A, Remuzzi G. COVID-19 and Italy: what next? Lancet 2020 Apr 11;395(10231):1225-1228 [FREE Full text] [doi: 10.1016/S0140-6736(20)30627-9] [Medline: 32178769]
27. pcm-dpc / COVID-19. GitHub. URL: https://github.com/pcm-dpc/COVID-19 [accessed 2020-07-05]
28. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 2020 Feb 29;395(10225):689-697 [FREE Full text] [doi: 10.1016/S0140-6736(20)30260-9] [Medline: 32014114]
29. Arango-Londoño D, Ortega-Lenis D, Muñoz E, Cuartas D, Caicedo D, Mena J, et al. Predicciones de un modelo SEIR para casos de COVID-19 en Cali, Colombia. Rev salud pública 2020 Mar 01;22(2):1-6. [doi: 10.15446/rsap.v22n2.86432]
30. Dye C, Gay N. Modeling the SARS epidemic. Science 2003 Jun 20;300(5627):1884-1885. [doi: 10.1126/science.1086925] [Medline: 12766208]
31. Gatto M, Bertuzzo E, Mari L, Miccoli S, Carraro L, Casagrandi R, et al. Spread and dynamics of the COVID-19 epidemic in Italy: effects of emergency containment measures. Proc Natl Acad Sci U S A 2020 May 12;117(19):10484-10491 [FREE Full text] [doi: 10.1073/pnas.2004978117] [Medline: 32327608]

Abbreviations

ICU: intensive care unit

Edited by G Eysenbach; submitted 27.03.20; peer-reviewed by B Smith, M Hamasha; comments to author 04.05.20; revised version received 05.06.20; accepted 17.01.21; published 09.03.21.

Please cite as:
Guzzi PH, Tradigo G, Veltri P
Regional Resource Assessment During the COVID-19 Pandemic in Italy: Modeling Study
JMIR Med Inform 2021;9(3):e18933
URL: https://medinform.jmir.org/2021/3/e18933
doi:10.2196/18933
PMID:33629957

Original Paper

Medical Morphology Training Using the Xuexi Tong Platform During the COVID-19 Pandemic: Development and Validation of a Web-Based Teaching Approach

Qinlai Liu[1], PhD; Wenping Sun[1], MD; Changqing Du[2], MD; Leiying Yang[1], MD; Na Yuan[1], MD; Haiqing Cui[2], MD; Wengang Song[3*], PhD; Li Ge[2*], PhD

[1]Department of Pathology, Shandong First Medical University & Shandong Academy of Medical Sciences, Tai'an, China

[2]Department of Histology and Embryology, Shandong First Medical University & Shandong Academy of Medical Sciences, Tai'an, China

[3]Department of Immunology, Shandong First Medical University & Shandong Academy of Medical Sciences, Tai'an, China

[*]these authors contributed equally

Corresponding Author:
Li Ge, PhD
Department of Histology and Embryology
Shandong First Medical University & Shandong Academy of Medical Sciences
2 Ying Sheng East Road
Tai'an
China
Phone: 86 0538622203
Email: juliagl@126.com

Abstract

Background: Histology and Embryology and Pathology are two important basic medical morphology courses for studying human histological structures under healthy and pathological conditions, respectively. There is a natural succession between the two courses. At the beginning of 2020, the COVID-19 pandemic suddenly swept the world. During this unusual period, to ensure that medical students would understand and master basic medical knowledge and to lay a solid foundation for future medical bridge courses and professional courses, a web-based medical morphology teaching team, mainly including teachers of courses in Histology and Embryology and Pathology, was established.

Objective: This study aimed to explore a new teaching mode of Histology and Embryology and Pathology courses during the COVID-19 pandemic and to illustrate its feasibility and acceptability.

Methods: From March to July 2020, our team selected clinical medicine undergraduate students who started their studies in 2018 and 2019 as recipients of web-based teaching. Meanwhile, nursing undergraduate students who started their studies in 2019 and 2020 were selected for traditional offline teaching as the control group. For the web-based teaching, our team used the Xuexi Tong platform as the major platform to realize a new "seven-in-one" teaching method (ie, videos, materials, chapter tests, interactions, homework, live broadcasts, and case analysis/discussion). This new teaching mode involved diverse web-based teaching methods and contents, including flipped classroom, screen-to-screen experimental teaching, a drawing competition, and a writing activity on the theme of "What I Know About COVID-19." When the teaching was about to end, a questionnaire was administered to obtain feedback regarding the teaching performance. In the meantime, the final written pathology examination results of the web-based teaching and traditional offline teaching groups were compared to examine the mastery of knowledge of the students.

Results: Using the Xuexi Tong platform as the major platform to conduct "seven-in-one" teaching is feasible and acceptable. With regard to the teaching performance of this new web-based teaching mode, students demonstrated a high degree of satisfaction, and the questionnaire showed that 71.3% or more of the students in different groups reported a greater degree of satisfaction or being very satisfied. In fact, more students achieved high scores (90-100) in the web-based learning group than in the offline learning control group (P=.02). Especially, the number of students with objective scores >60 in the web-based learning group was greater than that in the offline learning control group (P=.045).

Conclusions: This study showed that the web-based teaching mode was not inferior to the traditional offline teaching mode for medical morphology courses, proving the feasibility and acceptability of web-based teaching during the COVID-19 pandemic. Our findings lay a solid theoretical foundation for follow-up studies of medical students.

(JMIR Med Inform 2021;9(3):e24497) doi:10.2196/24497

KEYWORDS

COVID-19; histology and embryology; pathology; web-based teaching; Xuexi Tong platform

Introduction

The outbreak of COVID-19 rapidly became a worldwide pandemic [1] and necessitated rapid changes to higher education worldwide [2]. Many educators were dedicated to deliver knowledge via distance learning and web-based pedagogies, without stopping the teaching process [3]. This is an unprecedented form of teaching, and it has created substantial challenges to teaching as well as unprecedented opportunities. The Ministry of Education in China called for "suspension of classes without suspending the school" [4]. With active response to the call of the country, the teaching team, mainly including teachers of histology and embryology and pathology courses, immediately organized and strived to explore the web-based teaching mode for medical morphology. In this new era of information, web-based learning creates conditions for life-long learning and greater flexibility for people to learn on their own personal time; also, varied locations, good availability, and cost-effectiveness are associated with web-based teaching [5-9]. However, web-based teaching still has the disadvantage of learners feeling isolated within the virtual environment [10]. To ensure the effectiveness of teaching, diverse teaching practices that improve the communication between teachers and students during web-based education have been conducted, promoting the autonomy of students in learning.

Histology and Embryology is a medical morphological foundation course that studies the microstructure of the healthy body, organ functioning, and embryonic development using a microscope [11]. Additionally, Pathology is a course on the histological structural changes of disease conditions in the body; it explores the etiology as well as the pathogenesis of disease, functional changes, and basic outcomes [12]. Pathology is a bridge course that is based on the Histology and Embryology course, and it is included in basic medicine and clinical medicine courses.

Both these courses are theoretical and practical basic courses, and previous theoretical and experimental teaching and assessment models were no longer considered to be suitable during the global situation of the COVID-19 pandemic. Therefore, it is necessary to adjust the teaching and training strategies for medical students to ensure successful completion of the curriculum [13]. To enable students to smoothly transition from the study of histology and embryology to the study of pathology, teachers from both departments should cooperate closely and create new teaching modes.

Web-based teaching has placed greater demands on teachers' own qualities. To transform traditional classroom teaching to web-based teaching, teachers must master web-based teaching software and lead students to use virtual experiment platforms to grasp the essence of some morphological tissue sections. Therefore, teachers should consider students as the main body while the teachers themselves play the leading role, as in, "teach by learning and research by teaching." In this study, our team teaches Histology and Embryology and Pathology courses as points of entry to explore a new medical morphological web-based teaching mode and methodology using the Xuexi Tong platform as the major approach. This provides a reference sample for teaching at the Shandong First Medical University and other associated medical universities during the epidemic and postepidemic eras.

Methods

Participants

This study was conducted in the Shandong First Medical University from March to July 2020. Although all clinical medicine undergraduate students in who began their studies in 2018 and 2019 (Grades 2018 and 2019) have applied the Xuexi Tong platform throughout their studies, 512 students from 10 teaching classes were selected as the web-based teaching experiment group, which included 254 students learning histology and embryology in Grade 2019 (freshmen) and 258 students learning pathology in Grade 2018 (sophomores). Among these 512 students, 253 (49.4%) were from rural areas, 259 (50.6%) were from urban areas, 508 (99.2%) were Han Chinese, and 4 (0.8%) were members of ethnic minority groups (2 Hui, 1 Manchu, and 1 Mongol, respectively). At present, as the COVID-19 pandemic has been basically brought under control in China, most students have returned to school. For comparison with the performance of traditional offline teaching, 5 classes of Grade 2020 (freshmen, nursing majors) were used as the control group for the Histology and Embryology course, and 5 classes of Grade 2019 (sophomores, nursing majors) were used as the offline teaching control group of the Pathology course. In the meantime, to demonstrate differences in the written test scores, the final written examinations in Pathology of Grade 2018 (web-based teaching) and Grade 2017 (traditional offline teaching) students were compared among different score intervals.

Design of Teaching Methods

All the selected web-based teaching students did come into contact with other web-based teaching platforms, such as the Zhihui Shu platform, in other courses; however, we guaranteed that they had not come into contact with the Xuexi Tong platform before participating in this study.

Teaching standards were followed according to the normal teaching time, teaching content, and plans; also, attention was paid to attendance, homework, teaching content, and assessment.

Before Class

The electronic textbook and PowerPoint presentations were uploaded to the Xuexi Tong platform for students to preview and download. Content from Chinese University Massive Open Online Course (MOOC) and the school MOOC for the class

was selected, and students were instructed to proceed to the related platform to learn.

Classroom

Sign-in 10 minutes before class was recommended. The important and difficult points of the courses were recorded as videos by the teacher and uploaded in the task point section of the Xuexi Tong platform. Students were informed of the classroom content and time schedule in advance, and the classroom tests and quick response questions were distributed at regular intervals. Before the end of the class, a certain amount of time was set for problem-solving, and teachers were expected to provide timely answers in the web-based chat group. After each chapter, a case discussion question related to the teaching content was arranged in the discussion area of the Xuexi Tong platform, and the teachers provided scores and comments based on the students' analyses.

After Class

Once per week, a Tencent Meeting live broadcast question and answer link was sent to the students to check the deficiencies in web-based teaching knowledge and to address the problems encountered by the students. Teachers were asked to use EV screen recording software to record the important and difficult content of the chapter and upload the videos to the Xuexi Tong platform for students to review and summarize as well as to guide them in writing the chapter mind maps.

During the COVID-19 epidemic, to more objectively evaluate the learning effects of students under this new teaching mode, the final grade was divided into two parts. The final written grade accounted for 60% of the grade, and the formative evaluation accounted for 40%. The formative evaluation also had two parts; the writing of experimental reports accounted for 15% of the grade, and web-based learning accounted for 25%. The web-based learning included sign-in (10%), case analysis (10%), chapter tests (10%), task point viewing (30%), classroom tests (10%), interaction (5%), flipped classroom performance (10%), theme activity performance (5%), and mind map writing (10%).

Applying the Flipped Classroom Teaching Mode

In the "Respiratory System" chapter, the team used Tencent Live Conference to "flip the classroom." The teachers extracted 10 concepts from the "Respiratory System" chapter: (1) anatomical structure, microstructure, and embryogenesis of the respiratory system, (2) lobar pneumonia, (3) lobar pneumonia, (4) interstitial pneumonia, (5) chronic bronchitis, (6) chronic obstructive pulmonary emphysema, (7) silicosis, (8) pulmonary heart disease, (9) lung cancer, and (10) nasopharyngeal carcinoma.

Before the class, the students were asked to form study groups using WeChat and QQ to flexibly learn using various resources. The students learned the contents of the MOOC and task points, and they were divided into 10 random groups based on the concept of the group's presentation. Each group chose a representative to speak and give the presentation through Tencent Live Conference (Multimedia Appendix 1). At the end of the class, the teacher provided comments, and every student could go to the Xuexi Tong platform to rate the expression of the topic. The maximum total score of the flipped classroom presentation for each group was 10, and it was divided into three parts; the score within the group ratings accounted for 20%, the between-group ratings accounted for 40%, and the teacher's rating accounted for 40%.

Screen-to-Screen Experimental Teaching

Our team selected a medical morphological digital teaching platform [14] to guide students in observing virtual sections. The teacher carefully selected the most suitable tissue sections, and they used EV screen recording software to make videos to provide a dynamic view of the tissue sections and the PowerPoint presentations of the lectures. The videos were uploaded to the group chat section of the Xuexi Tong platform during the classes, and each student was required to post the picture they had drawn to the group chat; later, the teacher provided comments on the drawings.

Organizing a Drawing Competition

Students were encouraged to actively participate in the "Morphological Drawing Competition of College Students in Medical Colleges" organized by the Basic Medicine Group of the Joint Association of the National Experimental Teaching Demonstration Center of Colleges and Universities. The drawings that required evaluation were divided into two groups: the Histology and Embryology drawing group and the Pathology drawing group. The Pathology drawing group was then divided into two subgroups, namely Gross Pathology and Histopathology. Finally, 10 drawings were selected to proceed further and were then sent to the organizing committee as entries in the competition.

Conducting a Theme Activity

Before conducting a theme activity, the teachers sent many links on COVID-19 to students through WeChat in advance to enrich their knowledge. The topic of this theme activity was "What I Know About COVID-19." The answer format of the topic was not limited, and the submission could be an essay, poem, song, or even a video based on the medical knowledge the student had learned.

Evaluation and Statistical Analysis

Two weeks before the end of the teaching, a Questionnaire Star survey was used to investigate the long-term application performance of the new web-based teaching model and the traditional offline teaching model. The designed questions were imported into Questionnaire Star, and a WeChat link was generated and sent to the students. Next, the students answered the questions directly via the WeChat interface. Participation was voluntary, and complete anonymity was ensured. After conducting the survey, the final statistical results were automatically obtained and analyzed through Questionnaire Star.

In addition, as the Grade 2018 clinical undergraduates returned to the university at the end of August 2020, they underwent a written examination on pathology in their classroom. Therefore, SPSS version 19.0 (IBM Corporation) was used to compare the written examination results of these students with those of the

clinical undergraduates of the five classes of Grade 2017 who were taught offline the previous year. The corresponding classes of the two grades were taught by the same teacher, and the question type of the final written examination retained the same difficulty, and $P<.05$ was considered to be statistically significant. GraphPad Prism (GraphPad Software) was used to generate the corresponding histograms.

Results

The New Web-Based Teaching Model Achieves Good Teaching Performance

All the students completed the task points on time. Most of the students took the initiative to discuss case analysis, and they submitted their writings with regard to the mind map on time (Figure 1). Additionally, most of the students scored above 90 in the web-based learning evaluation.

All groups scored 9 or above on the flipped classroom assignment. This activity increased the students' interest in learning and enhanced their learning and collaboration abilities.

During screen-to-screen experimental teaching, based on the teachers' comments, most of the students mastered the content of the teaching task during the classroom, drew ideal pictures as required by the teacher, and posted their pictures in the group chat.

Figure 1. The mind map exercise performed by the students.

When participating in the drawing competition, all students drew good pictures: some of them used computer drawing software, while a few drew by hand. The subjects of the students' drawings included a general specimen of lung cancer, the normal structure of liver tissue (Figure 2), the normal structure of the small intestinal villi, squamous cell carcinoma (Figure 3), chronic pulmonary congestion, pneumonia caused by COVID-19, and fatty liver. After the teacher's instructions, all the participants drew better and more accurately.

Based on the theme activity, the students not only mastered the professional knowledge but also expressed their values. The

students responded enthusiastically to the activity, and each student uploaded an assignment through the Xuexi Tong platform; responses included long papers and references, good poems, adapted song lyrics, meaningful pictures (Figure 4), PowerPoint presentations, and recordings of uplifting videos.

Figure 2. The winning entry in the drawing competition of the Histology and Embryology course.

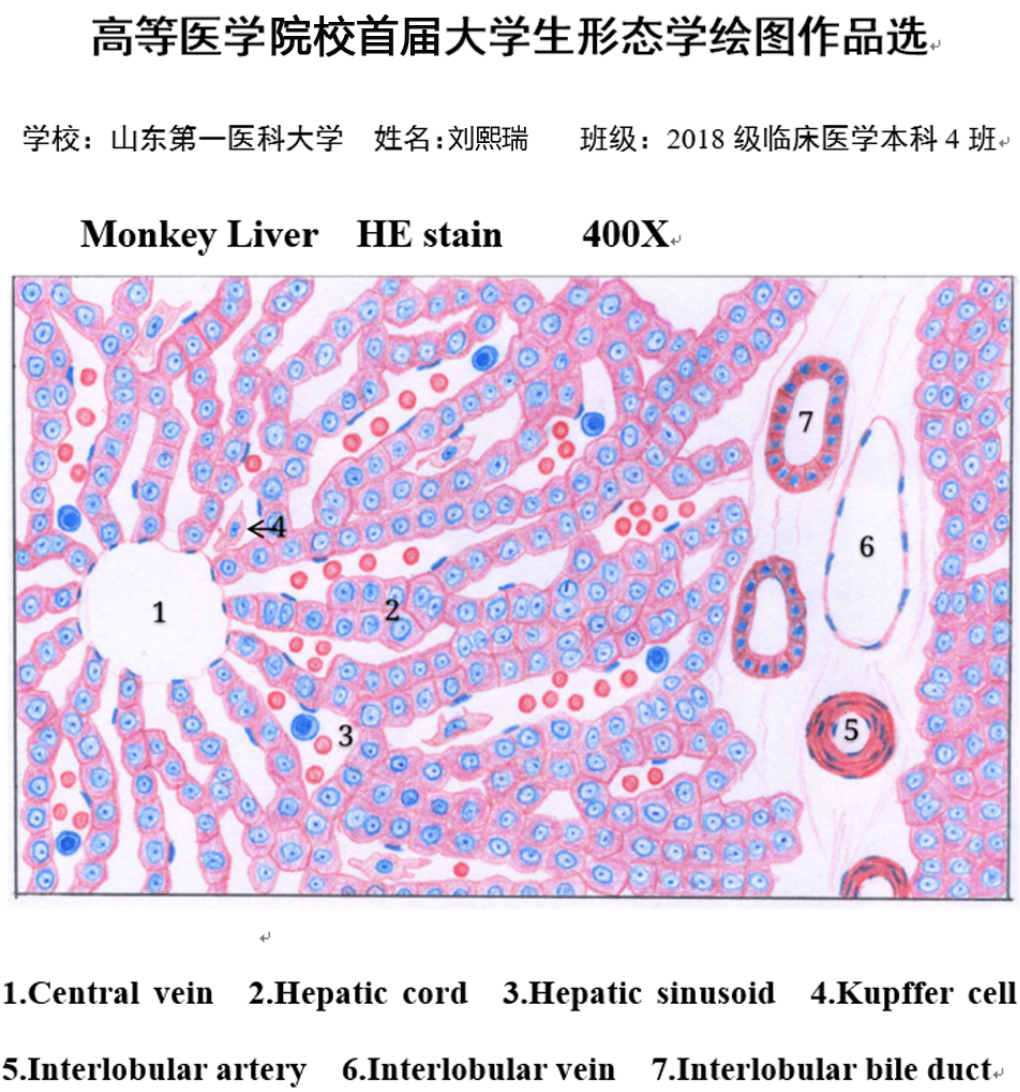


Figure 3. Representative image of pathological sections at low (1), medium (2) and high resolution power (3).

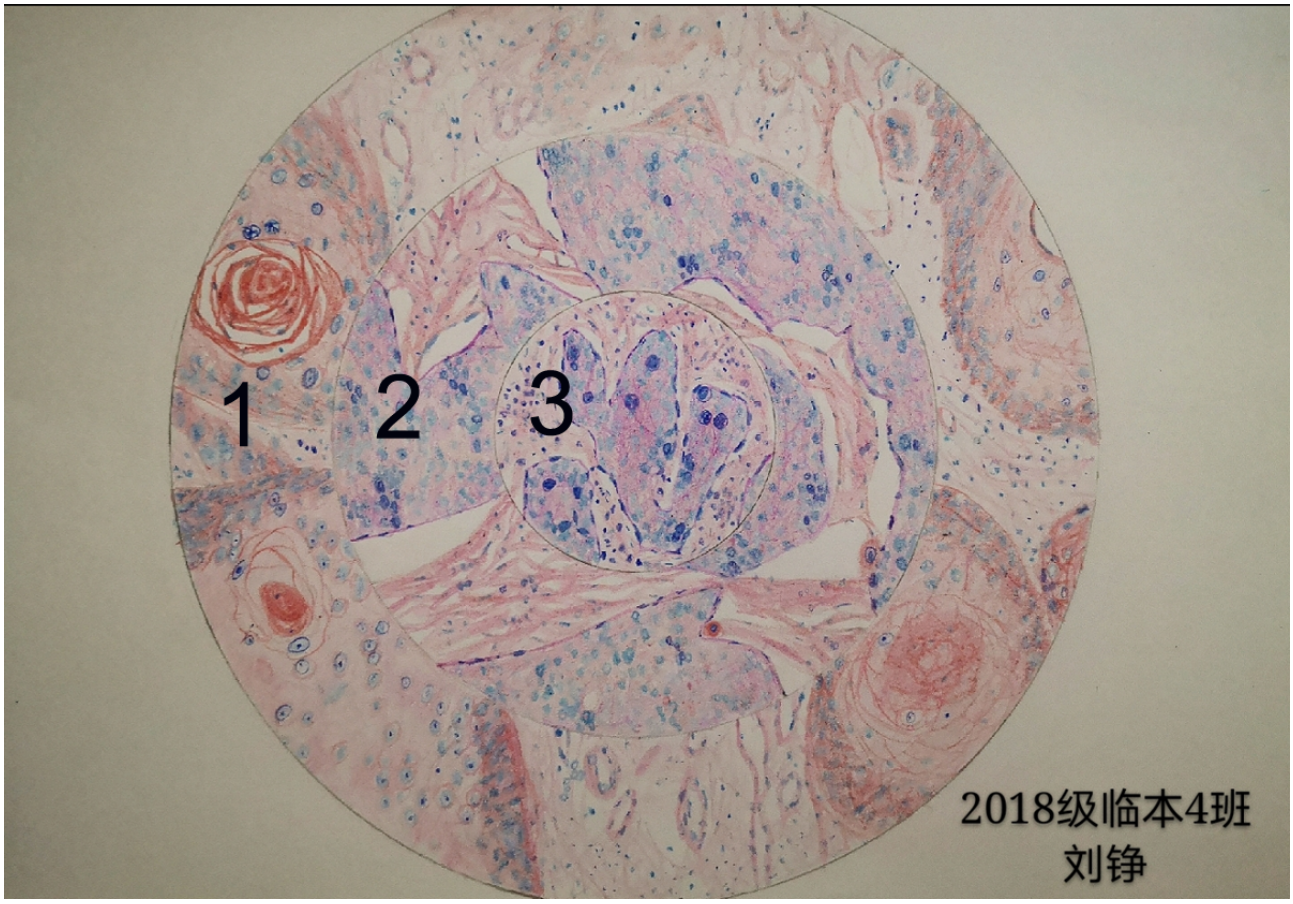
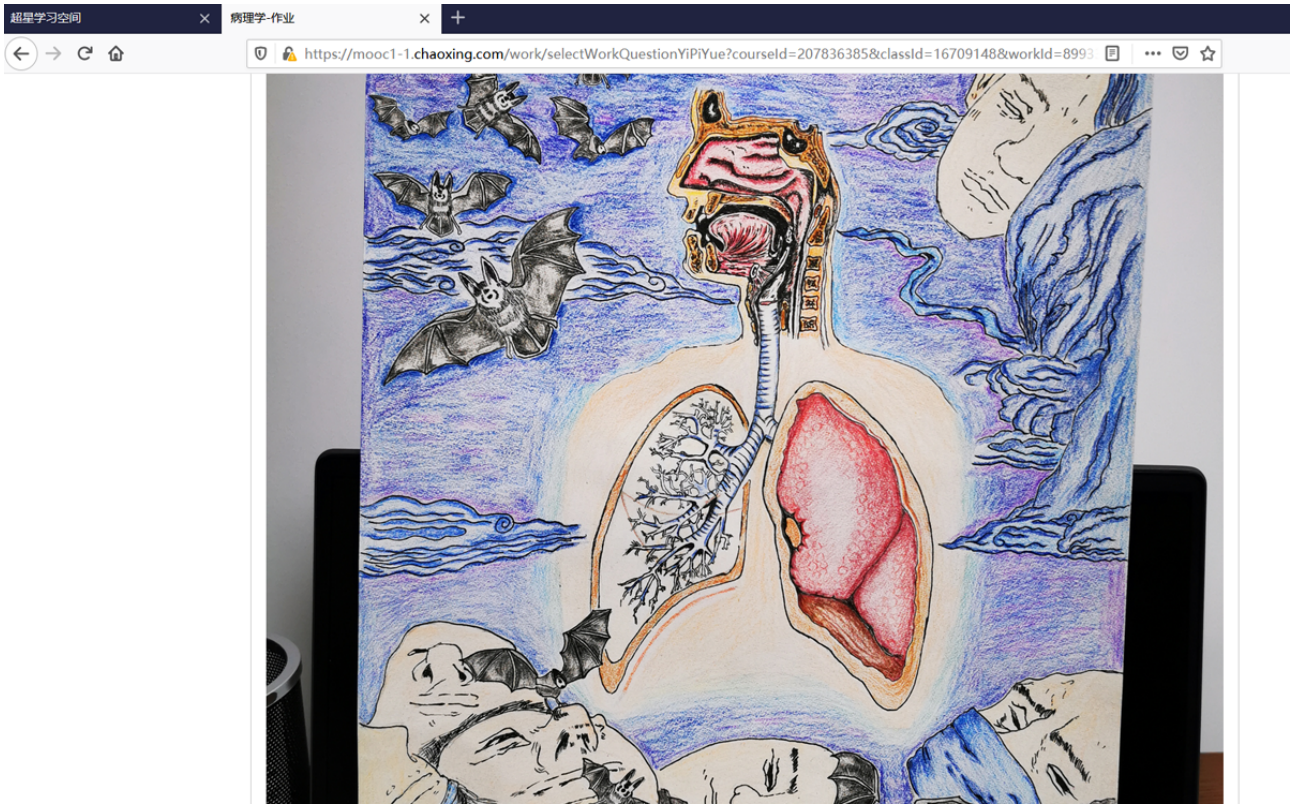


Figure 4. Representative drawing by a student for the theme activity.



Long-term Evaluation

Through the semester of web-based teaching practice, the results of the questionnaire survey and final written examination indicated that the new web-based teaching model was not inferior to the traditional teaching model. A total of 244 valid Histology and Embryology questionnaires and 246 valid Pathology questionnaires were answered by the students in the web-based teaching experiment group. The response rate of the students to the questionnaire was 95.7% (490/512). The rates of students who reported being very satisfied or having a greater degree of satisfaction with the theoretical teaching mode were 71.3% (174/244) in the Histology and Embryology course and 82.5% (203/246) in the Pathology course (Table 1). In terms of feelings about using the experimental platform, the rates of students who were very satisfied or had aa greater degree of satisfaction were 70.9% (173/244) in the Histology and Embryology course and 80.1% (197/246) in the Pathology course (Table 1), respectively. As shown in Table 2, considering the effectiveness of web-based teaching of Histology and Embryology and Pathology courses, more than 18.4% and 23.2% of the students in the experimental groups thought that the web-based courses were better than traditional teaching. Moreover, in the control group, a total of 253 valid Histology and Embryology questionnaires and 250 valid Pathology questionnaires were received through the Questionnaire Star tool. The effective ratio was 97.5% (503/516). As shown in Table 3, more than 7.1% (18/253) of students thought that the offline learning effect of the Histology and Embryology course was inferior to the web-based learning effect, while more than 11.6% (29/250) thought that the offline learning effect of the Pathology course was inferior to the web-based learning effect.

Table 1. Results of the first part of the questionnaire showing the students' satisfaction with the web-based teaching.

Survey question and course	Responses, n (%)[a]				
	Very satisfied	A greater degree of satisfaction	Generally satisfied	Less satisfied	Dissatisfied
1. What do you think of the current teaching mode? Are you satisfied?					
Histology and Embryology	64 (26.2)	110 (45.1)	57 (23.4)	7 (2.9)	6 (2.5)
Pathology	78 (31.7)	125 (50.8)	36 (14.6)	3 (1.2)	4 (1.6)
2. Is the web-based platform satisfactory?					
Histology and Embryology	63 (25.8)	110 (45.1)	58 (23.8)	8 (3.3)	5 (2.0)
Pathology	85 (34.6)	112 (45.5)	36 (14.6)	8 (3.3)	5 (2.0)
3. Do you think the current teaching goals are clear?					
Histology and Embryology	73 (29.9)	109 (44.7)	51 (20.9)	7 (2.9)	4 (1.6)
Pathology	92 (37.4)	113 (45.9)	36 (14.6)	3 (1.2)	2 (0.8)
4. Do you think the current teaching arrangements are reasonable?					
Histology and Embryology	71 (29.1)	114 (46.7)	50 (20.5)	7 (2.9)	2 (0.8)
Pathology	102 (41.5)	116 (47.2)	24 (9.8)	2 (0.8)	2 (0.8)
5. What is your attitude toward the quality of the web-based teaching microvideos?					
Histology and Embryology	53 (21.7)	114 (46.7)	67 (27.5)	7 (2.9)	3 (1.2)
Pathology	88 (35.8)	111 (45.1)	37 (15.0)	7 (2.9)	3 (1.2)

[a]Percentages calculated based on 244 valid Histology and Embryology questionnaires and 246 valid Pathology questionnaires.

Table 2. Results of the second part of the questionnaire showing the students' satisfaction with the web-based teaching performance.

Survey content and course	Responses, n (%)[a]			
	Very helpful, better than traditional classroom teaching	Helpful, similar to traditional classroom teaching	Helpful, but not as good as traditional classroom teaching	Not helpful
1. In terms of improving learning efficiency and quality and solving difficult problems				
Histology and Embryology	47 (19.3)	96 (39.3)	95 (38.9)	6 (2.5)
Pathology	57 (23.2)	97 (39.4)	88 (35.8)	4 (1.6)
2. In terms of mobilizing students' learning enthusiasm and initiative				
Histology and Embryology	48 (19.7)	93(38.1)/	94 (38.5)	9 (3.7)
Pathology	58 (23.6)	90(36.6)	89 (36.2)	9 (3.7)
3. In terms of knowledge summary, integration, application, and systematic understanding and memory				
Histology and Embryology	45 (18.4)	97 (39.8)	94 (38.5)	8 (3.3)
Pathology	57 (23.1)	100 (40.7)	84 (34.1)	5 (2.0)
4. For unity and cooperation, language and communication skills, and problem-solving				
Histology and Embryology	46 (18.9)	96 (39.3)	92 (37.7)	10 (4.1)
Pathology	61 (24.8)	92 (37.4)	84 (34.1)	9 (3.7)

[a]Percentages calculated based on 244 valid Histology and Embryology questionnaires and 246 valid Pathology questionnaires.

Table 3. Results of the survey questionnaire showing the students' satisfaction with the offline teaching performance in the control group.

Survey item and course	Responses, n (%)[a]			
	Very helpful, better than web-based classroom teaching	Helpful, similar to web-based classroom teaching	Helpful, but inferior to web-based classroom teaching	Not helpful
1. In terms of improving learning efficiency and quality and solving difficult problems				
Histology and Embryology	118 (46.6)	105 (41.5)	25 (9.9)	5 (2.0)
Pathology	115 (46)	95 (38.0)	39 (15.6)	1 (0.4)
2. In terms of mobilizing students' learning enthusiasm and initiative				
Histology and Embryology	129 (51.0)	99 (39.1)	20 (7.9)	5 (2.0)
Pathology	121 (48.4)	93 (37.2)	35 (14.0)	1 (0.4)
3. In terms of knowledge summary, integration, application, and systematic understanding and memory				
Histology and Embryology	123 (48.6)	95 (37.5)	27 (10.7)	8 (3.2)
Pathology	115 (46.0)	92 (36.8)	38 (15.2)	5 (2.0)
4. For unity and cooperation, language and communication skills, problem-solving				
Histology and Embryology	138 (54.5)	89 (35.2)	18 (7.1)	8 (3.2)
Pathology	131 (52.4)	89 (35.6)	29 (11.6)	1 (0.4)

[a]Percentages calculated based on 253 valid Histology and Embryology questionnaires and 250 valid Pathology questionnaires.

Through web-based learning, more than 82.4% (201/244) of students in the Histology and Embryology course (Figure 5) and more than 84.1% (207/246) of students in the Pathology course (Figure 6) achieved a score of more than 60% in the course by gaining theoretical knowledge. 17/244 students (7.0%) in the Histology and Embryology course (Figure 7) and 20/246 students (8.1%) in the Pathology course (Figure 8) with strong self-learning ability stated they would opt for web-based learning in the future. Of the 246 students in the Pathology course who answered the survey, 91 (37.0%) wished to return to offline teaching, and 124 (50.4%) (Figure 8) preferred web-based and offline blended learning. Of the 244 students in the Histology and Embryology course who answered the questionnaire, 127 (52.1%) (Figure 7) wished to return to offline classroom teaching after the epidemic; meanwhile, in the control group, 19/253 (7.5%) students in the Histology and Embryology course (Figure 9) and 30/250 (12%) students in the Pathology course (Figure 10) opted for web-based learning in the future, while 141/253 (55.7%) students in the Histology and Embryology course (Figure 9) and 141/250 (56.4%) students in the Pathology course (Figure 10) preferred web-based and offline blended learning.

A comparison of the final written examination scores of students in Grade 2018 with those in Grade 2017 showed that

significantly more students in Grade 2018 scored between 90 and 100 than those in Grade 2017 (P=.02), while a few others obtained scores of 80-90, 70-80, 60-70, and <60; the number of students in Grade 2018 showed no significant differences from that of students in Grade 2017 (Figure 11). In terms of objective scores, the number of students who scored >60 in Grade 2018 was significantly higher than that in Grade 2017 (P=.045); meanwhile, for other scores, including 50-60, 40-50, 30-40, and <30, no significant difference was observed between Grade 2018 and Grade 2017 (Figure 12). Moreover, for the subjective questions, no significant difference was observed in any grade (Figure 13).

Figure 5. Scores showing the extent of students' understanding and mastery of the theoretical knowledge of histology and embryology through web-based learning.

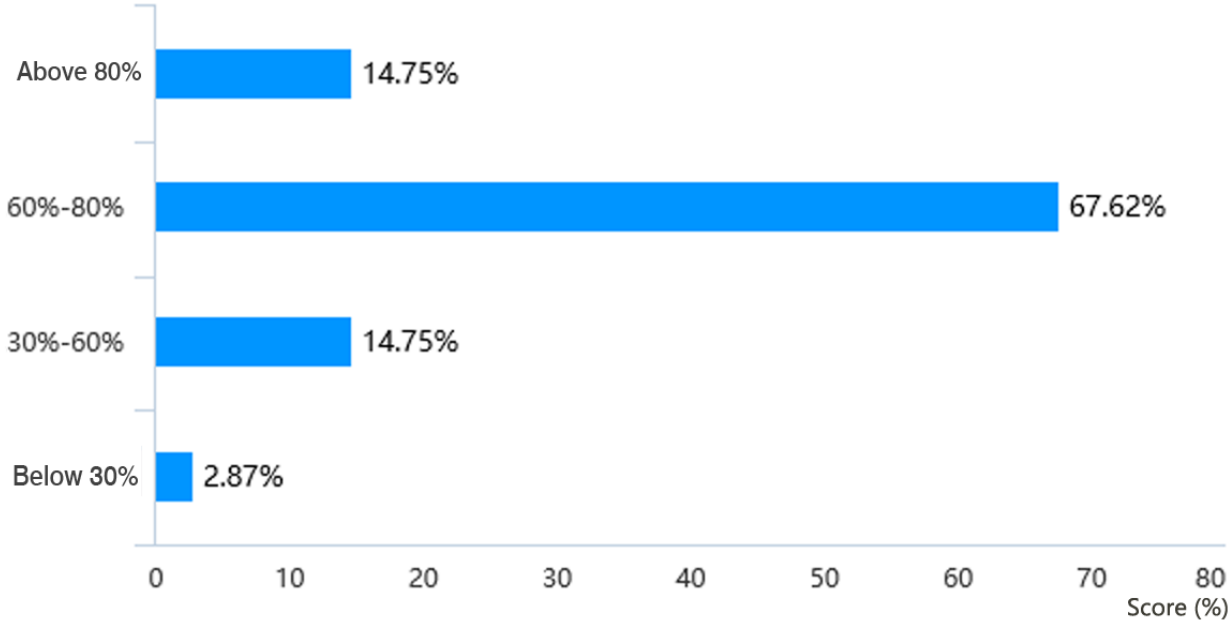


Figure 6. Scores showing the extent of students' understanding and mastery of the theoretical knowledge of pathology through web-based learning.

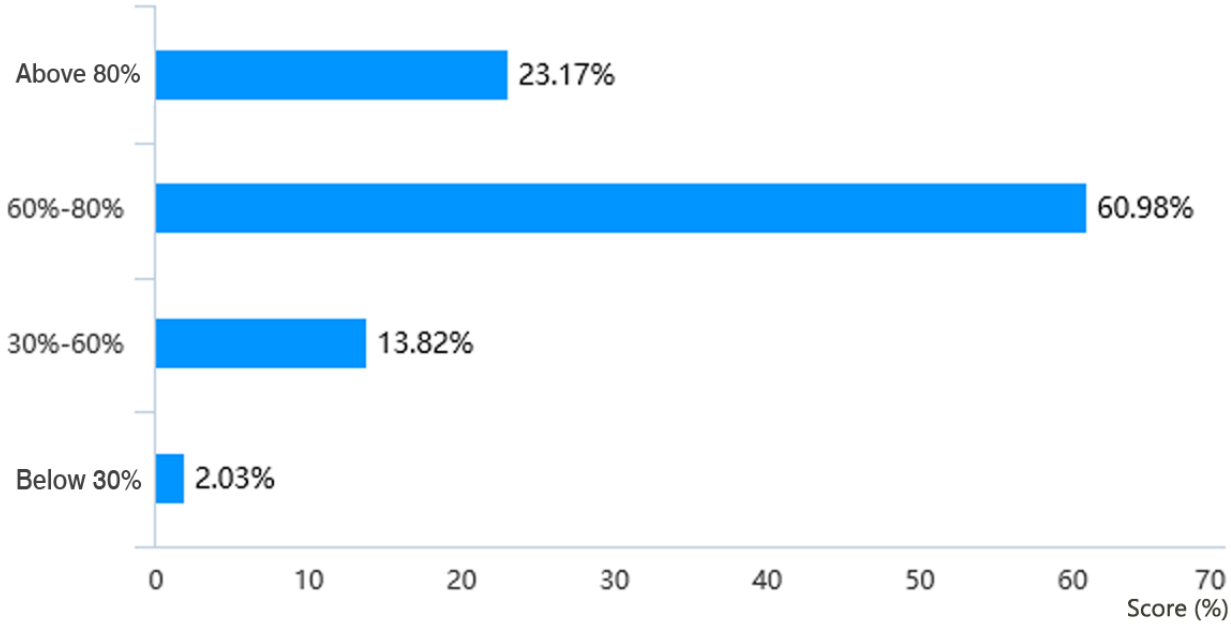


Figure 7. The proportions of students who chose different teaching methods of histology and embryology after the COVID-19 epidemic.

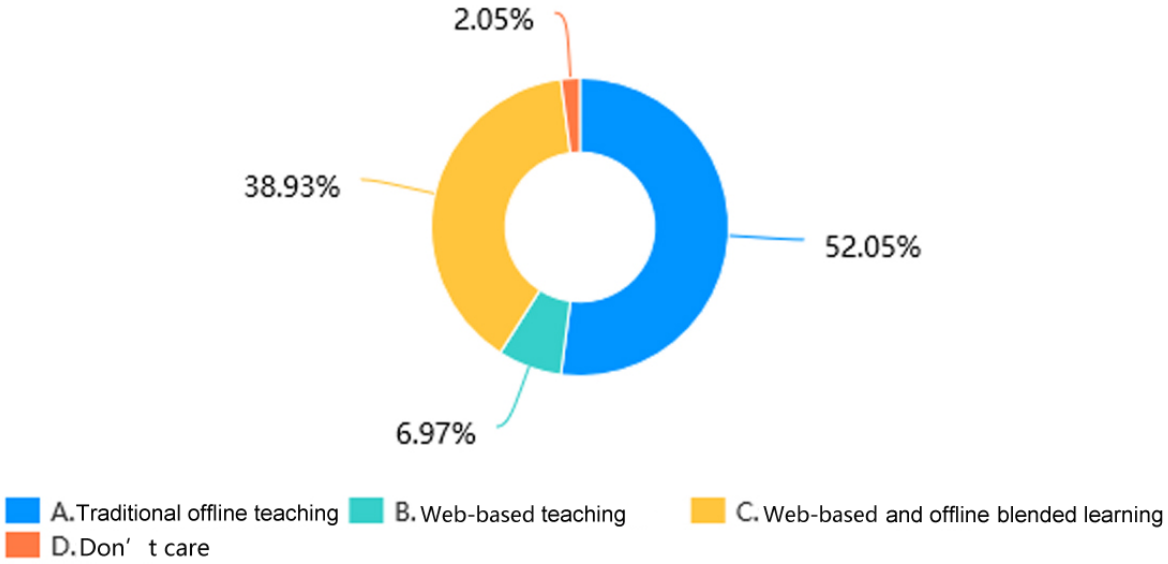


Figure 8. The proportions of students who chose different teaching methods of pathology after the COVID-19 epidemic.

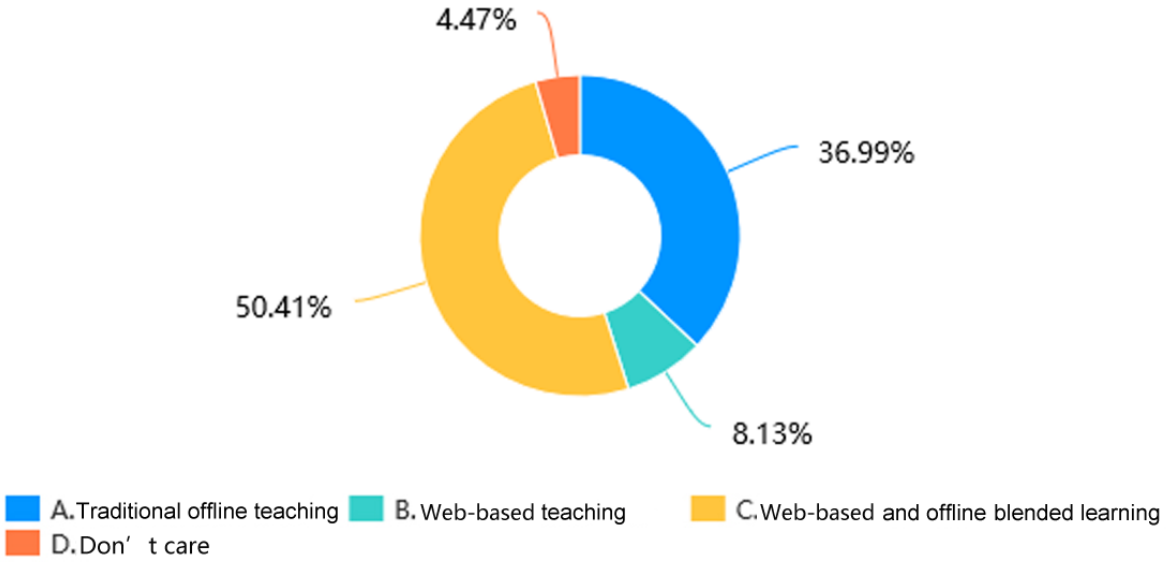


Figure 9. The proportions of students who would choose different teaching methods of histology and embryology in the future in the control group.

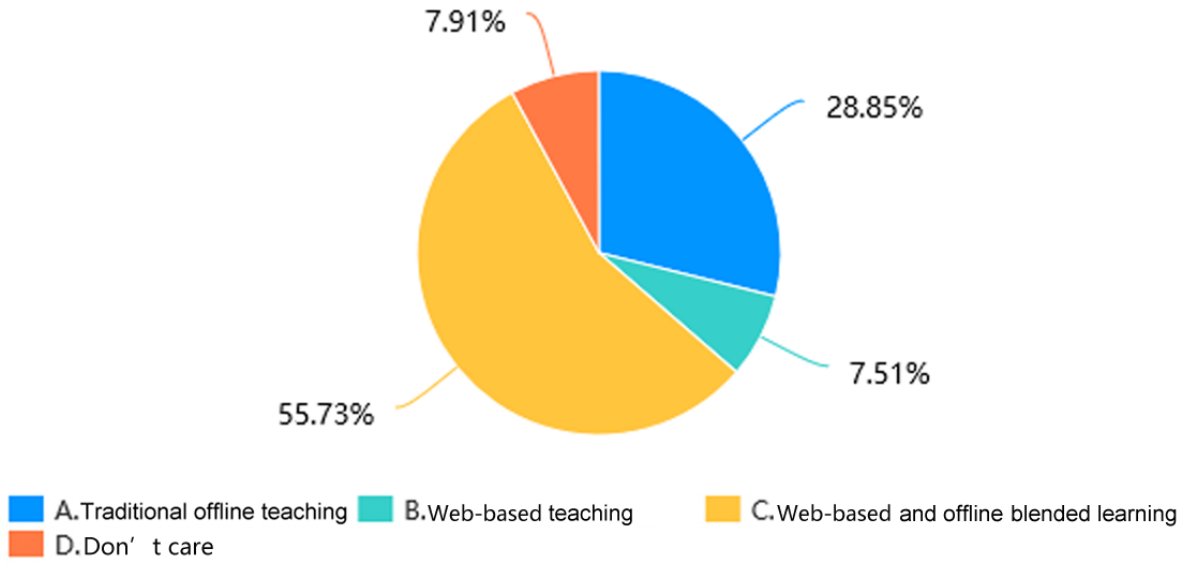


Figure 10. The proportions of students who would choose different teaching methods of pathology in the future in the control group.

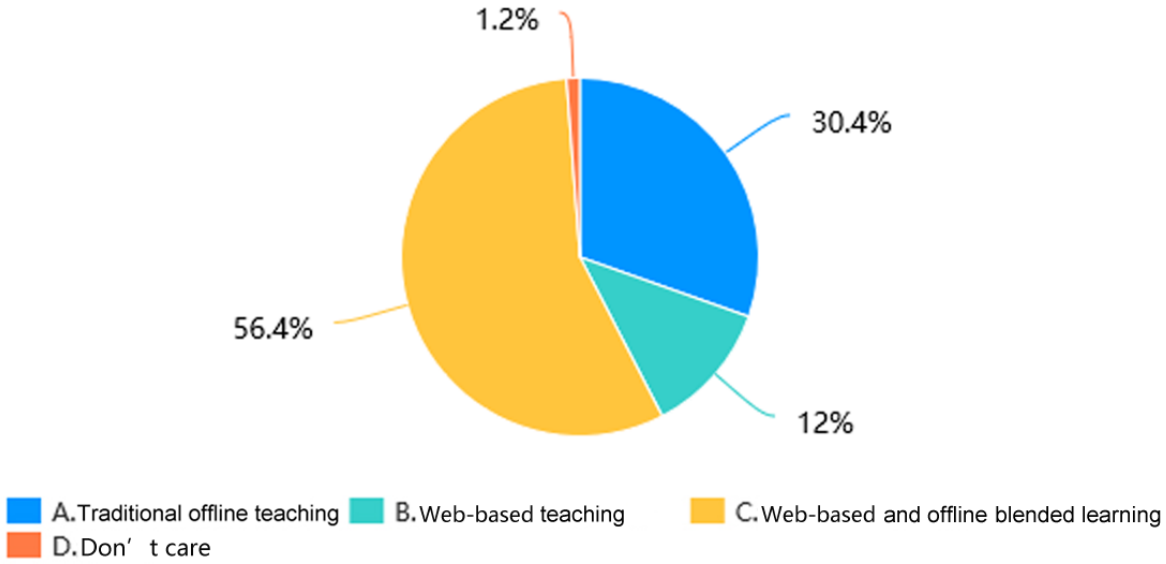


Figure 11. Statistical differences in the different score intervals in the final written pathology examination results. *: statistical significance; ns: no statistically significant difference.

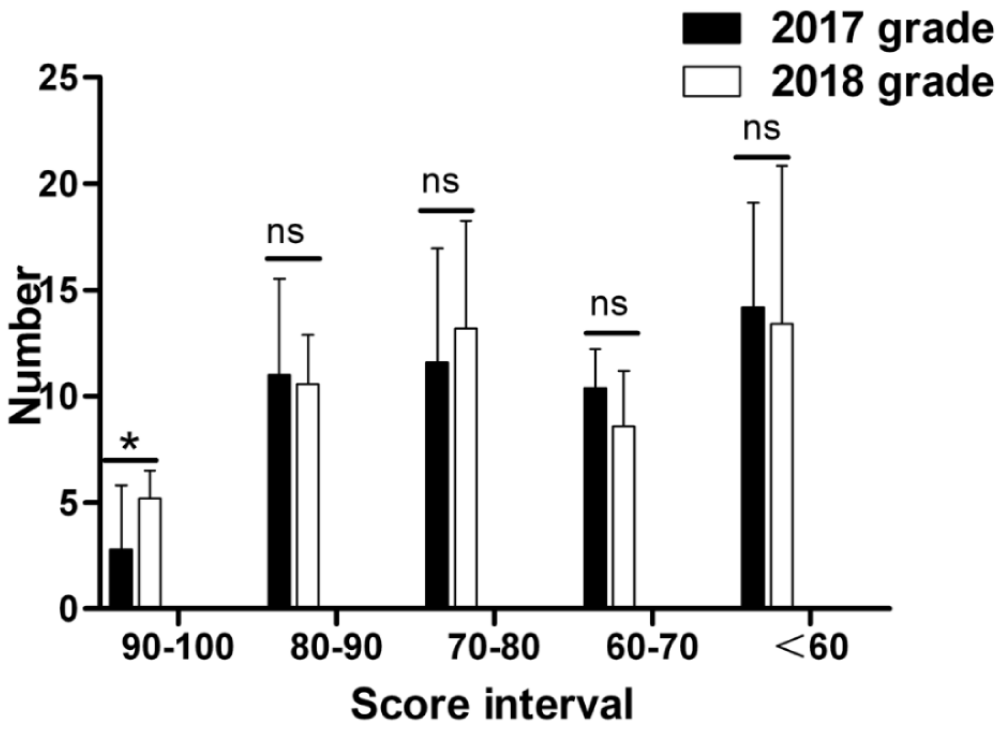


Figure 12. Statistical differences in the different score intervals for objective scores in the final written pathology examination results. *: statistical significance; ns: no statistically significant difference.

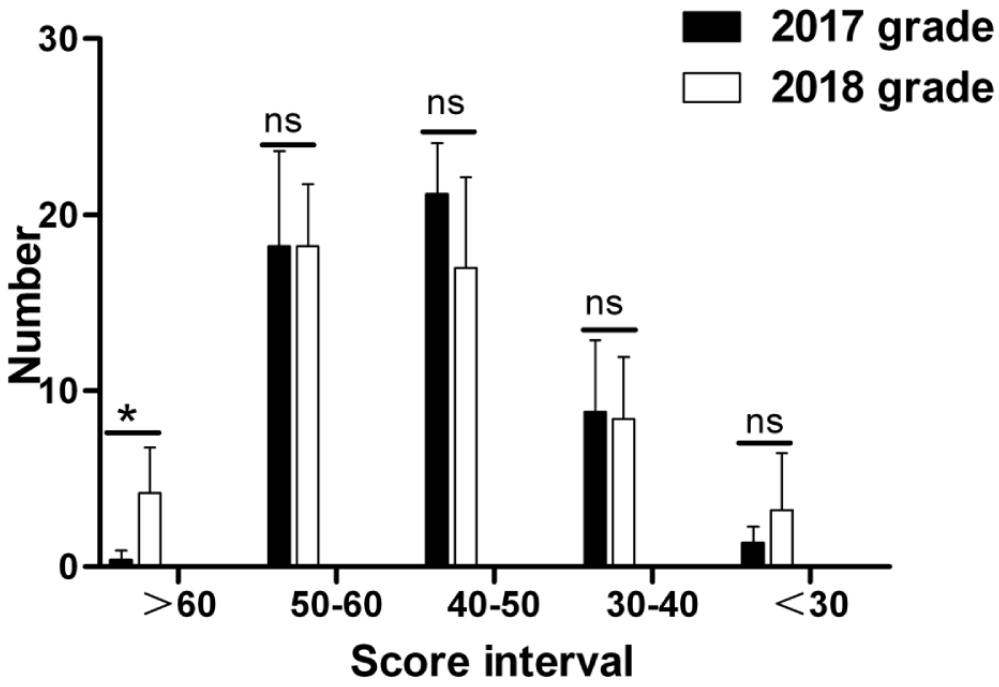
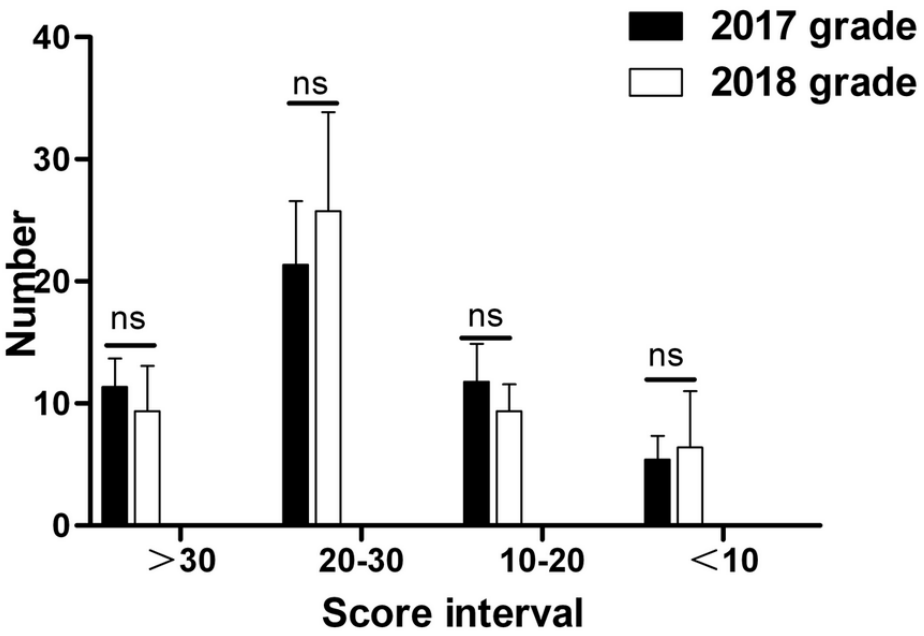


Figure 13. Statistical differences in the different score intervals for subjective scores in the final written pathology examination results. *: statistical significance; ns: no statistically significant difference.



Discussion

Principal Results

Our teaching team emphasized "internet + education" and adopted a "Seven-in-One" teaching mode of videos, materials, chapter tests, interaction, homework, live broadcasts, and case analysis/discussion. We adopted six web-based applications, namely the Xuexi Tong platform, Tencent Live Conference, Chinese University MOOC/school MOOC, a medical morphology digital teaching platform, WeChat/QQ, and Questionnaire Star, to guide our web-based teaching. The teaching of morphological courses with a series of education methods ensured that the web-based teaching was not inferior to traditional offline teaching.

The teaching team explored a new teaching mode wherein it could attract the students' attention and mobilize their enthusiasm for learning inside and outside the classroom. Teachers changed from knowledge imparters to instructors to help students learn, and they turned the "teaching-centered" notion into the "learning-centered" notion. Case analysis is an inquiry-based approach that can prompt students to actively engage in knowledge construction and develop competencies across multiple contexts [15,16]; it is required in the final written examination, and it is also a very good material to develop problem-based learning or Clinical Pathology Conference teaching. Our training mode of case analysis can cultivate clinical thinking ability in students. Considering the expression "A picture is worth a thousand words," the training of writing mind maps can strengthen students' understanding and memory of knowledge, improve their learning efficiency, cultivate divergent thinking, and encourage students to build a complete knowledge system and to integrate knowledge.

The teaching mode of unilateral instillation of knowledge through video has been changed, and the flipped classroom teaching approach was applied to strengthen the students' interactions and to improve their "hands-on" experience. Hew and Lo's study [17] indicated that more students favored the flipped classroom approach over traditional classroom teaching. Many of the in-class activities. such as small-group discussions, promoted students' interactions with their peers in flipped classes. Teachers also felt that they had a greater opportunity to provide more feedback during in-class sessions. In addition, there were greater opportunities for students to apply their knowledge [18].

The experimental courses of Histology and Embryology and Pathology belong to the category of morphology, and both require microscopic observation of tissue sections. Virtual microscopy has advantages in learning histology and pathology; it enables users to view slides almost anytime or anywhere, produce annotations that enhance student learning, and integrate slides into other digital resources [19]. Through screen-to-screen experiment teaching, the classroom teaching content is consolidated and the theoretical course understanding is deepened.

Organizing a drawing competition not only consolidates students' learning and understanding of human morphology but also improves their mastery and integration of knowledge of anatomy, histology and embryology, and pathology. This is more conducive for students to accept and master the surgical professional theories and their related skills.

Because COVID-19 is a "living teaching material," the teachers simply used the opportunity of the theme activity to carry out a character-education movement of "patriotism, love for mankind, and love for medicine." Many students who participated believed that in the face of the COVID-19 pandemic, it is necessary not only to solve problems using scientific knowledge but also to hone their minds and skills to become capable and responsible medical workers. All students said that they would take initiative to come forward when the country encounters difficulties in the future. This activity

improved the students' literary expression ability and thinking ability, and it also cultivated the students' medical and humanistic qualities.

The data from the Questionnaire Star survey showed that the overall performance of web-based teaching remained good; however, 127/244 (52.1%) students in the Histology and Embryology course preferred to return to offline classroom teaching after the epidemic. According to our understanding of students undergoing web-based learning of theoretical knowledge, the reason for this finding is that the students experienced a strong desire for practical operations, such as making tissue slices and other hands-on experiments, which can only be realized by offline teaching. Therefore, students opted to go back to the classroom or chose web-based and offline blended learning.

We found that the final written examination results of the web-based learning students were not inferior to those who experienced traditional offline learning. There were more high-scoring students (90-100) in the Grade 2018 web-based teaching classes than in the Grade 2017 offline learning classes (P=.02). These results suggest that for those students with good ability to self-study, the web has become more conducive to learn and self-study, to summarize and expand, and finally to improve the corresponding results.

Limitations

First, web-based teaching requires students to stare at the computer, mobile phone, or iPad for a long time; therefore, it is difficult to concentrate. There is a lack of school atmosphere; thus, students lacked interest in peer feedback [20]. Second, as web-based teaching was implemented throughout the country, the network jammed at times, and a few students had insufficient hardware at home and could only learn through the 4G network of mobile phones with insufficient traffic, restricting the implementation of web-based teaching to a certain extent. Third, the formative evaluations may have been affected by academic dishonesty and thus failed to objectively reflect the students' scores [21]. In addition, this study was conducted during the emergency situation of the epidemic, and we lacked a suitable randomized control group. Moreover, the Questionnaire Star survey only focused on the perception and satisfaction of the network; therefore, longer-term research is warranted to evaluate the medium-term and long-term impact. Finally, this research lacked a rigorous sampling process, and all students in the classes of the experimental group and control group were included in our research.

Conclusions

In short, the web-based teaching of histology and embryology and pathology in one semester demonstrated good performance and achieved the purpose of teaching with "suspension of classes without suspending the school." This new teaching mode not only successfully satisfied the urgent needs of students to acquire knowledge during the epidemic period but also provided a practical foundation for blended learning in the future.

Web-based teaching has the advantage of being more flexible in enabling students to choose when and where to study. Under the new teaching mode that was established, the course content of web-based teaching showed no shrinkage, and the requirements for students were not reduced. The web-based teaching was simply the Level I response to "a public health emergency of international concern over the global outbreak of novel coronavirus" [22], and most universities will continue to conduct face-to-face learning [23]. Therefore, for the teacher, web-based teaching is a challenge as well as an opportunity. Moreover, teachers can take advantage of this opportunity to master relevant skills and methods by constant exploration and practice. The COVID-19 epidemic will have a profound impact on education for the foreseeable future [24].

Our teachers are the education reformers and the main force of development. We are not only teachers but also learners, researchers, and innovators. Web-based teaching and distance education are growing parts of medical education [25]. However, one of the problems in evaluating web-based teaching is the lack of a proper evaluation tool [26]. Therefore, we need to further consider this evaluation in future studies. Additionally, the main task of teacher training and development is to organically combine teacher learning techniques by using timely techniques to explore new education models that combine web-based and offline learning [8] and to use technology as an important carrier for the development of education.

Therefore, after the COVID-19 epidemic, the combination of web-based and offline teaching methods, which is a new trend, would be advocated. Meanwhile, teachers should change their teaching concepts, establish "internet + education" thinking modes, learn to share and use big data, and better meet and serve the learning needs of their students.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (81401561, 81771711); General Program of Natural Science Foundation of Shandong Province (ZR2019MH123); and Shandong Provincial Key Laboratory of Animal Biotechnology and Disease Control and Prevention (ABDC-201901)

Authors' Contributions

QLL, LG, and WGS contributed to the study conception and design, the analysis and interpretation of data, and the drafting of the paper. WPS, CQD, LYY, NY, and HQC contributed to the analysis and interpretation of data as well as the drafting and revising of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
In the flipped classroom, a student used the tencent conference for presentation.
[MP4 File (MP4 Video), 5438 KB - medinform_v9i3e24497_app1.mp4]

References

1. Ahmed H, Allaf M, Elghazaly H. COVID-19 and medical education. Lancet Infect Dis 2020 Jul;20(7):777-778. [doi: 10.1016/s1473-3099(20)30226-7]
2. Longhurst GJ, Stone DM, Dulohery K, Scully D, Campbell T, Smith CF. Strength, weakness, opportunity, threat (SWOT) analysis of the adaptations to anatomical education in the United Kingdom and Republic of Ireland in response to the Covid-19 pandemic. Anat Sci Educ 2020 May;13(3):301-311 [FREE Full text] [doi: 10.1002/ase.1967] [Medline: 32306550]
3. Ortiz PA. Teaching in the time of COVID-19. Biochem Mol Biol Educ 2020 May 02;48(3):201-201 [FREE Full text] [doi: 10.1002/bmb.21348] [Medline: 32239800]
4. During the epidemic prevention and control period, universities should organize and manage online teaching. Webpage in Chinese. Ministry of Education of the People's Republic of China. URL: http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202002/t20200205_418131.html [accessed 2021-03-03]
5. Ruiz JG, Mintzer MJ, Leipzig RM. The impact of e-learning in medical education. Acad Med 2006 Mar;81(3):207-212. [doi: 10.1097/00001888-200603000-00002] [Medline: 16501260]
6. Ward R, Stevens C, Brentnall P, Briddon J. The attitudes of health care staff to information technology: a comprehensive review of the research literature. Health Info Libr J 2008 Jun;25(2):81-97 [FREE Full text] [doi: 10.1111/j.1471-1842.2008.00777.x] [Medline: 18494643]
7. Choules AP. The use of elearning in medical education: a review of the current situation. Postgrad Med J 2007 Apr;83(978):212-216 [FREE Full text] [doi: 10.1136/pgmj.2006.054189] [Medline: 17403945]
8. Pei L, Wu H. Does online learning work better than offline learning in undergraduate medical education? A systematic review and meta-analysis. Med Educ Online 2019 Dec;24(1):1666538 [FREE Full text] [doi: 10.1080/10872981.2019.1666538] [Medline: 31526248]
9. Keis O, Grab C, Schneider A, Öchsner W. Online or face-to-face instruction? A qualitative study on the electrocardiogram course at the University of Ulm to examine why students choose a particular format. BMC Med Educ 2017 Nov 09;17(1):194 [FREE Full text] [doi: 10.1186/s12909-017-1053-6] [Medline: 29121902]
10. Hara N. Student distress in a web-based distance education course. Inf Commun Soc 2010 Dec 02;3(4):557-579. [doi: 10.1080/13691180010002297]
11. Michael H, Wojciech P. Histology: A Text and Atlas: With Correlated Cell and Molecular Biology, 7th Edition. Philadelphia, PA: Lippincott Williams & Wilkins; 2015.
12. Vinay K, Abul A, Jon A. Robbins Basic Pathology, 10th Edition. London, UK: Elsevier; 2017.
13. Chang WJ, Jiang YD, Xu JM. Experience of teaching and training for medical students at gastrointestinal surgery department under COVID-19 epidemic situation. Article in Chinese. Zhonghua Wei Chang Wai Ke Za Zhi 2020 Jun 25;23(6):616-618. [doi: 10.3760/cma.j.cn.441530-20200603-00334] [Medline: 32521987]
14. Digital Human EMPT. URL: http://pt6.humanyun.com [accessed 2021-03-03]
15. Prosser M, Sze D. Problem-based learning: student learning experiences and outcomes. Clin Linguist Phon 2014;28(1-2):131-142. [doi: 10.3109/02699206.2013.820351] [Medline: 23944271]
16. Jin J, Bridges SM. Educational technologies in problem-based learning in health sciences education: a systematic review. J Med Internet Res 2014 Dec 10;16(12):e251 [FREE Full text] [doi: 10.2196/jmir.3240] [Medline: 25498126]
17. Hew KF, Lo CK. Flipped classroom improves student learning in health professions education: a meta-analysis. BMC Med Educ 2018 Mar 15;18(1):38 [FREE Full text] [doi: 10.1186/s12909-018-1144-z] [Medline: 29544495]
18. Galway LP, Corbett KK, Takaro TK, Tairyan K, Frank E. A novel integration of online and flipped classroom instructional models in public health higher education. BMC Med Educ 2014 Aug 29;14:181 [FREE Full text] [doi: 10.1186/1472-6920-14-181] [Medline: 25169853]
19. Lee LMJ, Goldman HM, Hortsch M. The virtual microscopy database-sharing digital microscope images for research and education. Anat Sci Educ 2018 Sep;11(5):510-515. [doi: 10.1002/ase.1774] [Medline: 29444388]
20. Latif MZ, Hussain I, Saeed R, Qureshi MA, Maqsood U. Use of Smart Phones and Social Media in Medical Education: Trends, Advantages, Challenges and Barriers. Acta Inform Med 2019 Jun;27(2):133-138 [FREE Full text] [doi: 10.5455/aim.2019.27.133-138] [Medline: 31452573]
21. Bell BS, Federman JE. E-learning in postsecondary education. Future Child 2013;23(1):165-185. [doi: 10.1353/foc.2013.0007] [Medline: 25522650]

22. WHO Director-General's statement on IHR Emergency Committee on Novel Coronavirus (2019-nCoV). World Health Organization. 2020 Jan 30. URL: https://www.who.int/director-general/speeches/detail/who-director-general-s-statement-on-ihr-emergency-committee-on-novel-coronavirus-(2019-ncov) [accessed 2021-03-03]
23. Currie G, Hewis J, Nelson T, Chandler A, Nabasenja C, Spuur K, et al. COVID-19 impact on undergraduate teaching: medical radiation science teaching team experience. J Med Imaging Radiat Sci 2020 Dec;51(4):518-527 [FREE Full text] [doi: 10.1016/j.jmir.2020.09.002] [Medline: 32981889]
24. Chick RC, Clifton GT, Peace KM, Propper BW, Hale DF, Alseidi AA, et al. Using technology to maintain the education of residents during the COVID-19 pandemic. J Surg Educ 2020;77(4):729-732 [FREE Full text] [doi: 10.1016/j.jsurg.2020.03.018] [Medline: 32253133]
25. de Leeuw RA, Westerman M, Walsh K, Scheele F. Development of an instructional design evaluation survey for postgraduate medical e-learning: content validation study. J Med Internet Res 2019 Aug 09;21(8):e13921 [FREE Full text] [doi: 10.2196/13921] [Medline: 31400102]
26. de Leeuw RA, Walsh K, Westerman M, Scheele F. Consensus on quality indicators of postgraduate medical e-learning: Delphi study. JMIR Med Educ 2018 Apr 26;4(1):e13 [FREE Full text] [doi: 10.2196/mededu.9365] [Medline: 29699970]

Abbreviations

MOOC: massive open online course

Edited by G Eysenbach; submitted 12.10.20; peer-reviewed by M Wu, Z Ren; comments to author 21.10.20; revised version received 04.12.20; accepted 16.01.21; published 15.03.21.

Please cite as:
Liu Q, Sun W, Du C, Yang L, Yuan N, Cui H, Song W, Ge L
Medical Morphology Training Using the Xuexi Tong Platform During the COVID-19 Pandemic: Development and Validation of a Web-Based Teaching Approach
JMIR Med Inform 2021;9(3):e24497
URL: https://medinform.jmir.org/2021/3/e24497
doi:10.2196/24497
PMID:33566792

Original Paper

Systematic Delineation of Media Polarity on COVID-19 Vaccines in Africa: Computational Linguistic Modeling Study

Sefater Gbashi[1*], PhD; Oluwafemi Ayodeji Adebo[1*], PhD; Wesley Doorsamy[2*], PhD; Patrick Berka Njobeh[1*], PhD

[1]Faculty of Science, University of Johannesburg, Johannesburg, South Africa

[2]Institute for Intelligent Systems, University of Johannesburg, Johannesburg, South Africa

[*]all authors contributed equally

Corresponding Author:
Sefater Gbashi, PhD
Faculty of Science
University of Johannesburg
55 Beit Street, Doornfontein
Johannesburg, 2028
South Africa
Phone: 27 620402580
Email: sefatergbashi@gmail.com

Abstract

Background: The global onset of COVID-19 has resulted in substantial public health and socioeconomic impacts. An immediate medical breakthrough is needed. However, parallel to the emergence of the COVID-19 pandemic is the proliferation of information regarding the pandemic, which, if uncontrolled, cannot only mislead the public but also hinder the concerted efforts of relevant stakeholders in mitigating the effect of this pandemic. It is known that media communications can affect public perception and attitude toward medical treatment, vaccination, or subject matter, particularly when the population has limited knowledge on the subject.

Objective: This study attempts to systematically scrutinize media communications (Google News headlines or snippets and Twitter posts) to understand the prevailing sentiments regarding COVID-19 vaccines in Africa.

Methods: A total of 637 Twitter posts and 569 Google News headlines or descriptions, retrieved between February 2 and May 5, 2020, were analyzed using three standard computational linguistics models (ie, TextBlob, Valence Aware Dictionary and Sentiment Reasoner, and Word2Vec combined with a bidirectional long short-term memory neural network).

Results: Our findings revealed that, contrary to general perceptions, Google News headlines or snippets and Twitter posts within the stated period were generally passive or positive toward COVID-19 vaccines in Africa. It was possible to understand these patterns in light of increasingly sustained efforts by various media and health actors in ensuring the availability of factual information about the pandemic.

Conclusions: This type of analysis could contribute to understanding predominant polarities and associated potential attitudinal inclinations. Such knowledge could be critical in informing relevant public health and media engagement policies.

(JMIR Med Inform 2021;9(3):e22916) doi:10.2196/22916

KEYWORDS

COVID-19; coronavirus; vaccine; infodemiology; infoveillance; infodemic; sentiment analysis; natural language processing; media; computation; linguistic; model; communication

Introduction

COVID-19

COVID-19, a communicable disease caused by SARS-CoV-2, has resulted in the ongoing COVID-19 pandemic [1], with significant health and socioeconomic effects in almost all countries of the world, mainly due to the shutdowns, social distancing, and the resultant business interruptions [2-5]. On March 11, 2020, the World Health Organization (WHO) declared a COVID-19 outbreak due to its global adverse (health and economic) impact [6,7], which is still unfortunately escalating as the world eagerly awaits a medical solution. According to projections by the International Monetary Fund, global economic growth could fall by 0.5% for the year 2020

due to the impact of COVID-19 [7]. A United Nations study revealed that, for the first time since 1900, global poverty could increase [8]. The world unemployment rate has been estimated to hit a mark of 9.4% by the end of 2020, in contrast to the 5.3% seen in 2019 [9]. A forecast by the World Bank projects that Africa is heading toward its first recession in the last 25 years, being driven by the impact of COVID-19 [2]. There is a compelling necessity to achieve an immediate medical breakthrough to curtail the pandemic. Although currently there is no known cure or vaccine for COVID-19, various agencies and governments are working round the clock to proffer much needed solutions. There are indications that some of the vaccines could be ready for clinical trials soon, while others are already undergoing trials.

Infodemic and COVID-19

Amid the spread of the COVID-19 pandemic and the efforts to provide a solution, there has equally been substantial media interest and widespread misinformation on the web about the scale, origin, diagnosis, prevention, and treatment of the disease [10]. The WHO has expressed particular concern about this *infodemic* and its impact on the fight against the COVID-19 pandemic [11-13]. This "infodemic" may have been facilitated by the global lockdown, isolation, and social distancing situation worldwide, which has led to increased internet and social media use [14,15]. For example, as of May 2020, global internet use was already up by 7% (since the same time in 2019), while the number of social media users increased by 8% (reaching 3.81 billion) since April 2019 [14]. Rozenblum and Bates [16] described the internet and social media as the "perfect storm" with regard to patient-oriented health. Singh [17] called the COVID-19 pandemic–induced media use a "compulsive social media use" and wonders whether it is an addictive behavior. Information diffusion and media coverage can have a positive effect on an epidemic by shortening the duration of the outbreak and reducing its burden [18]; however, if media communication is not properly managed, this can have a negative effect [19,20]. In a study by Brainard and Hunter [21], the authors noted that making 20% of a population unable to share fake advice or reducing the amount of damaging advice spreading online by just 10% can reduce the severity of a disease outbreak.

It is known that medical interventions involving infectious diseases and vaccines in particular are usually predisposed to infodemics [22-24]. Indeed, disease proliferation and associated potential effects take place in a dynamic social space, wherein public knowledge, sentiments, and individual health decisions are influenced by the media and cultural norms [19,25,26]. Studies have shown that an important factor in determining the success of a medical intervention, particularly a vaccination program, is the degree of public acceptance of the proposed intervention [27-29]. Lewandowsky et al [30] observed that the spread of myths regarding vaccinations has led to more parents being reluctant to adopt such measures. The withdrawal of the first US Food and Drug Administration licensed vaccine against the rotavirus in 1999 has been strongly correlated with negative media coverage [31]. Infodemics are known to foster the denial of scientific evidence [32]; incite apathy, cynicism, and extremism [33]; and cause people and institutions to take actions that may not necessarily be in their best interests or for the good of society in general [19,34]. Yu et al [19] observed that, even though medical interventions were able to reduce hepatitis B virus infection by up to 90% among children younger than 5 years using effective and safe hepatitis B (HepB) vaccines, negative media reports about infant death after HepB vaccination resulted in a loss of confidence in vaccines and an approximately 19% decrease in the use of vaccines within the monitored provinces in China. This further emphasizes the importance of individuals' actions in controlling the spread of the disease and the role of the media and psychosocial effects in this regard, as also observed in the literature [12,13,25,35].

In recognizing the importance of information and corresponding human actions in the management of diseases and other medical emergencies or outbreaks, medical researchers are increasingly exploring the application of computer and mathematical models to analyze communications relating to health problems and to derive value from such resources [18,25,35-37]. One such application is sentiment analysis [38,39]. It is important to understand the prevailing emotions in public media communications, as sentiments could provide a richer set of information about people's choices and reactions, and in many cases, even determine their decisions [40-42].

Sentiment Analysis

Sentiment analysis is a multidisciplinary field involved with machine learning (ML) and artificial intelligence (AI), and a subset of natural language processing (NLP) concerned with the systematic extraction, analysis, classification, quantification, and interpretation of affective tonality, opinions, and subjective information in human communications (written or spoken) using computational linguistic methods to derive value of the opinions people express. Sentiment analysis is important because opinions and emotions constitute a critical aspect of humans and play a key role in perceptions of reality, choices, actions, and behaviors [37]. Basic tasks in sentiment analysis include classifying emotional polarity in a communication as positive, negative, or neutral and often scaling such polarities to reflect the depth of the expressed emotions. By understanding prevailing sentiments in communications, health actors can then make more relevant and informed decisions of public relevance and act accordingly to improve the medical experience.

In performing sentiment analysis, language limitations remain a major challenge, as most of the sentiment analysis models and libraries are built in English, which limits their applicability for texts written in other languages. Another problem with traditional sentiment analysis models is the lack of context in labeling lexical items and extracting features, which results in ambiguity in polarity representations. A word or group of words can have different meanings and polarities in different contexts; hence, the global representation of words and sentences could influence the semantics of the words [43]. More contemporary developments in sentiment analysis are shifting toward increasingly context-based sentiment analysis and applicability in other languages [43-45]. Context-aware (ie, context-based) sentiment analysis attempts to tackle the challenge of ambiguity by taking into account all of the text around any given word or words, then processing the logical structure of the sentences, establishing the relations between semantic concepts and

assigning logical grammatical roles to the lexical elements to decipher the most relevant meaning of a word or group of words that have more than one definition.

Generally, sentiment analysis models can be largely grouped into two main categories: lexical rule-based techniques and learning-based approaches [43]. The former (ie, lexical rule-based methods) involve the use of predefined lexicons annotated with sentiment polarities, "positive," "negative," or "neutral," which are then used to determine the sentiment of the analyzed text [46]. Lexical rule-based sentiment analyzers such as TextBlob and Valence Aware Dictionary and Sentiment Reasoner (VADER) [47,48], although they use predefined dictionaries, take in to account punctuation such as exclamation marks, emoticons, capitalizations (which gives added intensities to assigned polarities), and negation words (which reverse the polarities). Lexical rule-based sentiment methods have the advantage of being easy to implement because there is no need for the exhausting and arduous task of tagging texts for training. In addition, the approach prevents overfitting because the lexicons are predefined independently of the data being analyzed, a feature that also enables it (lexical rule-based sentiment models) to be used on multiple data sets [46]. On the other hand, the main advantage of (machine) learning-based methods is that they enable the use of domain-specific data sets to train the models, which are then used to analyze a given text. Given relevant and adequate training data sets, this (domain-oriented training) substantially increases the accuracy and level of confidence in text classification and sentiment analysis using learning-based methods. Examples of successful implementation of ML-based sentiment models can be found in the literature [49].

Applications of Sentiment Analysis

Sentiment analysis has been applied in the fields of medical research, political science, business, mass communication, and education. Specific applications include brand monitoring and reputation management, customer feedback and support, product analysis, market research and competitive research, competitor analysis [50], attitudinal analysis [51], medical records analysis, patient experience analysis, and infodemiology [52-54]. More than ever before, the application of sentiment analysis in the medical field has seen a surge, as both patients and medical practitioners increasingly use web-based platforms such as websites, social media, and blogs to search for treatment-related information and to convey opinions on health matters [55,56]. Melzi et al [57] implemented a supervised learning model to extract sentiments from forums of Spine Health to retrieve patient knowledge. Yang et al [58] demonstrated a sentiment analysis–based framework to derive insights from user-generated content from health social media.

In the context of the ongoing COVID-19 pandemic, Barkur and Vibha [59] performed a sentiment analysis of Indians' Twitter communications about the nationwide lockdown due to the COVID-19 outbreak. An investigation of the sentiment and public discourse during the pandemic was performed using latent Dirichlet allocation for topic modeling on 1.9 million Tweets written in the English language related to COVID-19 collected from January 23 to March 7, 2020 [60]. Deep long short-term memory models were used to scrutinize the reaction of citizens, public sentiment, and emotions from different cultures about COVID-19 and the subsequent actions taken by different countries [61]. These studies reiterate the usefulness of sentiment analysis in the health domain. Nonetheless, there is no study on the sentiment analysis of media communications regarding COVID-19 vaccines in Africa that can guide health actors to understand the prevailing polarities and to make relevant polices, as the continent awaits the availability of viable vaccines to combat the pandemic.

Research Justification, Goal, and Questions

Indeed, opinion mining and sentiment analysis in public media is a daunting task because of the large amount of information shared on the internet and social media, particularly in recent times. To give some perspective, Google handles approximately 4.5 million queries every 60 seconds, and approximately 500 million tweets are made every 24 hours [62,63]. In the month of March 2020 alone, about 550 million tweets were made, which included the terms *COVID-19*, *COVID19*, *COVID_19*, *coronavirus*, *corona virus*, or *pandemic*, according to the Pan American Health Organization [64]. Considering the enormity of internet information, researchers are increasingly adopting computer-based algorithms to process, analyze, and interpret such data. This does not only substantially limit the need for human power but also eliminates the associated human biases and inefficiencies in many ways [65,66].

The goal of this paper is to retrieve timely and relevant information that is publicly available regarding COVID-19 vaccines in Africa and interrogate such resources in an attempt to extract the sentiment polarities using computational linguistics models. In this regard, we outlined the following research questions, which constitute the main contributions of this paper:

- What was the media activity patterns regarding COVID-19 vaccines in Africa within the study period of February 2 to May 5, 2020?
- What are the prevailing sentiments (ie, positive, negative, or neutral) in the communications?
- How does the sentiment polarities in Twitter posts compare to those in Google News communications?
- Using three different sentiment analysis approaches, how does the sentiment results from these models compare with each other?
- What were the specific activities or events that might have triggered the prevailing emotions in the communications?

Methods

This study adopts computational linguistic models (TextBlob, VADER, and Word2Vec–bidirectional long short-term memory [BiLSTM]) to scrutinize communications from popular web-based media sources regarding COVID-19 vaccines in Africa.

Data Collection and Cleaning

In the collection of data, the following criteria and procedure were used [67].

Data Source

Twitter and Google News were selected as the data sources. This is because both media outlets are important web-based information sources for many people. Von Nordheim et al [68] examined the use of different social media platforms (ie, Twitter and Facebook) as journalistic sources in newspapers of three different countries; the authors observed that Twitter is more commonly used as a news source than Facebook, and in comparison to Facebook, Twitter was primarily used as an elite channel [68]. In addition, previous studies have shown that most of the communications on the Twitter platform are truthful, although sometimes it may also be used to propagate false information and rumors [67,69,70]. Moreover, the microblog (ie, Twitter) facilitates the propagation of real-time information to a large group of people, making it an ideal tool for the dissemination of breaking news [69]. Twitter has become a favored communication platform for various social protest movements [71]. On the other hand, Google News is an online media service that aggregates and presents a continuous flow of news content from more than 20,000 publishers worldwide and is available on the web, iOS, and Android. For news in the English language, it covers about 4500 sources. This creates a unique information space, providing a major gateway for consumers to access news, reducing the time and effort needed to regularly check different media sources for updates, and increasing overall user engagement.

Query Terms

English language keywords were used to query the data sources instead of hashtags. Specific search terms used were “COVID-19 vaccine Africa,” “COVID19 vaccine Africa,” and “Coronavirus vaccine Africa.” This approach is more inclusive, as it includes communications that have hashtags of the keywords [67]. Moreover, hashtags are often short-lived on the web.

Query Period

The sources were queried within the period of February 2 to May 5, 2020, using the search terms. The study period was chosen on the basis of timelines, relevance, and particularly because this was the time frame in which opinions, perspectives, and narratives about the search terms were emerging and could set the tone for subsequent communications as well as make enduring impressions on people’s dispositions.

Search Tools

Google News data was retrieved using the Google News application processing interface (API) for Python by Hu [72]. Twitter data were obtained by manually scraping the Twitter web application for publicly available English language posts containing the query terms. This was essentially due to limited resources to subscribe for the Twitter API, which is a paid service. To eliminate any biases, a new Twitter account was used with no previous search history, liked pages, or topics or interests associated with the account.

A total of 569 news headlines and snippets published in English were obtained from Google News and saved as a comma-separated values (CSV) file (Multimedia Appendix 1). Likewise, 637 Twitter posts were extracted and saved in a CSV file (Multimedia Appendix 2) for further analysis. The obtained Twitter and Google News texts were supplied to an algorithm written in Python for data cleaning and “normalization” as described by Sahni et al [73] and Salas-Zárate et al [74] with minor modifications. Briefly, all numbers, special characters, punctuations, and symbols except for commas, periods, exclamation marks, question marks, and apostrophes were removed. Repeated white spaces were replaced with a single space followed by the removal of all stop words. In addition, we autocorrected wrongly spelled English words, removed non-English words, and performed lemmatization. The final cleaned corpus was saved into a Pandas dataframe object for further analysis.

Opinion Mining and Sentiment Analysis

To provide a more symmetrical and global perspective of the emotive patterns in the data sets, two rule- and lexicon–based ML techniques (ie, TextBlob and VADER) and an advance algorithm, which combines a ML model and a neural network (NN) for NLP, (ie, Word2Vec-BiLSTM) were used to process the cleaned data.

TextBlob Sentiment Analysis

The TextBlob Python library [48] is an easy-to-use publicly available library for NLP tasks such as sentiment analysis, part-of-speech tagging, classification, noun phrase extraction, translation, tokenization, and n-grams. In this study, we used the NaiveBayes analyzer implementation of TextBlob for sentiment analysis, which is a conditional probabilistic classifier implementation as explained by Hasan et al [75] and Singh et al [76]. The sentiment feature of TextBlob makes use of a predefined dictionary classifying positive and negative words and returns two values, polarity and subjectivity. Polarity is a float between −1 and +1, where −1 means a negative sentiment, +1 means a positive sentiment, and a 0 is considered a neutral sentiment. The subjectivity score is a float between 0 and 1, and represents the degree of personal opinion or judgement rather than factual information. A subjectivity score closer to 0 implies a more objective view, whereas a score closer to 1 represents a more subjective view. The data is generally supplied as a bag-of-words, and after assigning individual scores to each word, the final sentiment is represented by some pooling of all the sentiments. TextBlob has semantic labels that facilitate fine-grained sentiment analysis by recognizing emojis, emoticons, and punctuations such as exclamation marks.

VADER Sentiment Analysis

The VADER library [47] is a lexicon- and rule–based sentiment analyzer of the English language constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon and is available with Python’s Natural Language Toolkit library [77]. VADER does not require any training data and is specifically attuned to semantic constructions in social media texts, which are generally informal, yet can be conveniently applied to multiple domains [77,78]. A Python algorithm using the VADER library was developed to unveil and categorize the sentiment in our data set. The output of the analysis is a percentage score of positivity, negativity, and neutrality, as well as a compound score. The compound score is a normalized weighted composite score (ie, normalized to be

between –1, most extreme negative, and +1, most extreme positive) that is useful for interpreting the sentiment in a text as a single unidimensional measure of sentiment. On the other hand, the positive, negative, and neutral scores are useful to understand the sentiment in a text in multidimensional measures.

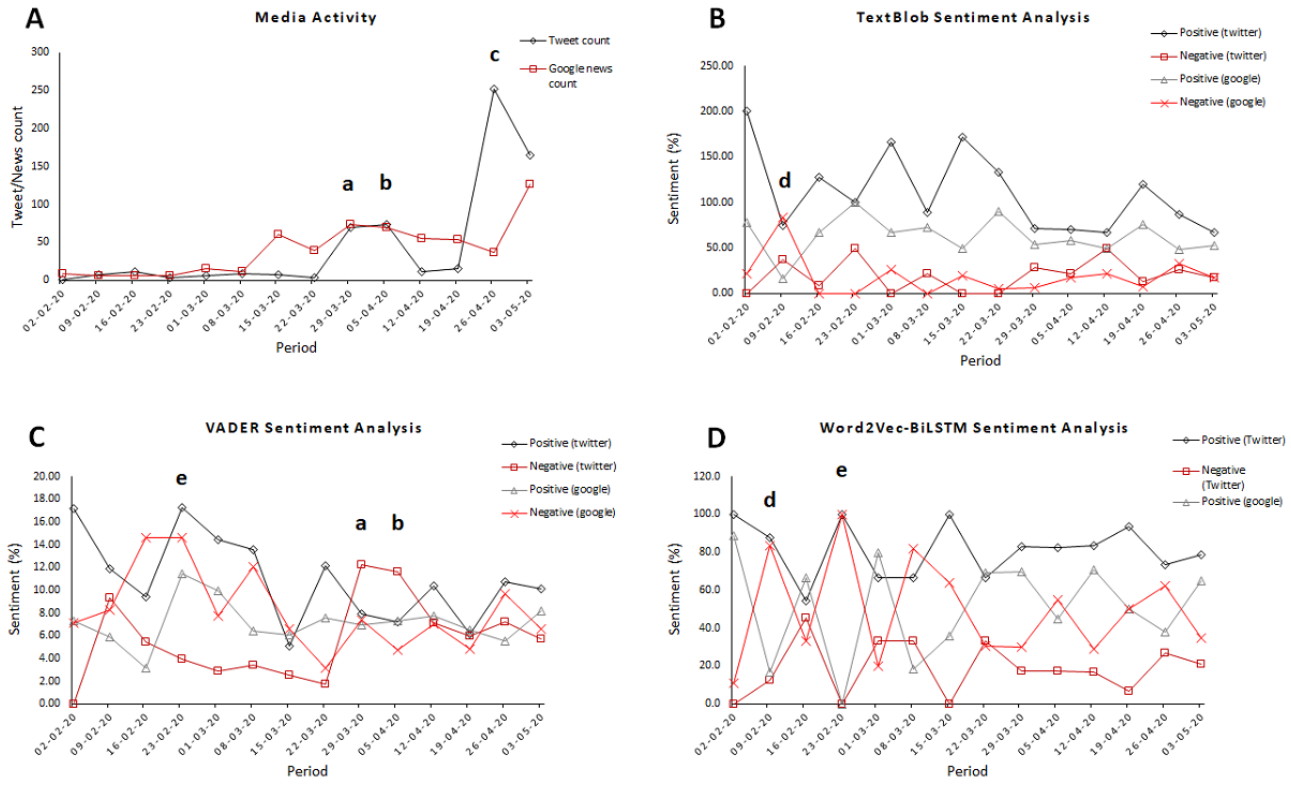
Word2Vec-BiLSTM Sentiment Analysis

Word2Vec-BiLSTM is an advanced model, which combines ML and AI for NLP. Word2Vec is a ML model that projects natural words, groups of words, or phrases into a multidimensional space, creating vector representations of the word or group of words, which are mapped to vectors of real numbers, otherwise called word embeddings. The resultant two-layer NNs are subsequently trained to reconstruct linguistic contexts of the word or group of words that preserve information about the target word or group of words, the overall context of the word, and the crams variable length sequences to fixed-length vectors. Mikolov et al [79-81] describe in detail the underlying principles and algorithms of Word2Vec. The BiLSTM model [82,83] learns two-way dependencies between time steps of time series data (ie, the output layer gets information from forward and backward states simultaneously), which has presented good results in NLP [84,85]. When given adequate training, Word2Vec-BiLSTM models are highly accurate at capturing the ephemeral qualities of human language and guessing the meanings of words within the context of a sentence. The BiLSTM model and related hyperparameters adopted herein with minor modifications is described by Cherniuk [86]. The training data consisted of 75,000 review entries, with a test size of 0.05% of the training data accordingly labeled as a positive or negative review. The model inputs were feature vectors composed of bigrams and trigrams of the cleaned data set. The model architecture consisted of six layers with a sigmoid activation function, Adam optimizer [87], and binary_crossentropy as the loss function. The model batch size was 100 with an epoch of five. After training, the accuracy of the model was 92%.

Results

The results of media activity within the study period is summarized in Figure 1 [88-95]. There was relatively low activity on Twitter and Google News about the query terms within the first 3 weeks of the study period (ie, between February 2-23, 2020). Google News features on the search topics soared from 39 news items in the eighth week (ie, March 22, 2020) to 73 news features by the ninth week (ie, March 29, 2020). Similarly to the trend on Google News, Twitter posts increased from 3 to 70 between the eighth and ninth weeks of the study period.

Figure 1. Media activity and sentiment analysis of tweets and Google News headlines and snippets between February 2 and May 5, 2020. (A) Media activity within the study period, (B) TextBlob sentiment analysis, (C) VADER sentiment analysis, and (D) Word2Vec-BiLSTM sentiment analysis. (a) Two French medical doctors made controversial remarks on COVID-19 vaccine testing in Africa on April 1, (b) the World Health Organization strongly remarked about the comments by the French doctors on April 6, (c) Madagascar announced herbal cure to COVID-19 at a virtual meeting of African presidents on April 20, (d) first reported case of COVID-19 in Africa on February 14, and (e) first reported case of COVID-19 in sub-Saharan Africa on February 28. BiLSTM: bidirectional long short-term memory; VADER: Valence Aware Dictionary and Sentiment Reasoner.



Tables 1 and 2 provide a summary of the sentiment polarities in media communications using computational linguistic models. Zooming in on the results for Google News headlines or snippets, it can be observed that average polarities from the VADER analysis revealed that the news was 7% positive, 8% negative, and 85% neutral, with a compound score of –2, which

indicates overall negativity in the news tonality. The TextBlob analysis showed that the news was 63% positive, 19% negative, and 22% neutral, with an average subjectivity score of 0.5, indicating a general positivity in the news tonality.

Table 1. Sentiment polarities in media communications (Google News headlines or descriptions) analyzed using the TextBlob, VADER, and Word2Vec-BiLSTM models.

Period in 2020	TextBlob				VADER[a]				Word2Vec-BiLSTM[b]	
	Positive	Negative	Neutral	Subjectivity score	Positive	Negative	Neutral	Compound score	Positive	Negative
February 2	77.78	22.22	0.00	0.48	0.07	0.07	0.86	0.02	88.89	11.11
February 9	16.67	83.33	0.00	0.58	0.06	0.08	0.86	–0.01	16.67	83.33
February 16	66.67	0.00	33.33	0.73	0.03	0.15	0.82	–0.35	66.67	33.33
February 23	100.00	0.00	0.00	0.58	0.11	0.15	0.74	–0.11	0.00	100.00
March 1	66.67	26.67	6.67	0.43	0.10	0.08	0.82	0.09	80.00	20.00
March 8	72.73	0.00	27.27	0.64	0.06	0.12	0.82	–0.21	18.18	81.82
March 15	49.18	19.67	31.15	0.40	0.06	0.07	0.87	–0.04	36.07	63.93
March 22	89.74	5.13	5.13	0.37	0.08	0.03	0.89	0.17	69.23	30.77
March 29	53.42	6.85	39.73	0.48	0.07	0.07	0.86	0.02	69.86	30.14
April 5	57.97	17.39	24.64	0.43	0.07	0.05	0.88	0.07	44.93	55.07
April 12	49.09	21.82	29.09	0.34	0.08	0.07	0.85	0.03	70.91	29.09
April 19	75.93	7.41	16.67	0.42	0.06	0.05	0.89	0.10	50.00	50.00
April 26	48.65	32.43	18.92	0.46	0.06	0.10	0.85	–0.14	37.84	62.16
May 3	53.17	17.46	29.37	0.33	0.08	0.07	0.85	0.06	65.08	34.92
Average polarity	62.69	18.60	18.71	0.48	0.07	0.08	0.85	–0.02	51.02	48.98

[a]VADER: Valence Aware Dictionary and Sentiment Reasoner.

[b]BiLSTM: bidirectional long short-term memory.

Table 2. Sentiment polarities in media communications (Twitter posts) analyzed using the TextBlob, VADER, and Word2Vec-BiLSTM models.

Period in 2020	TextBlob				VADER[a]				Word2Vec-BiLSTM[b]	
	Positive	Negative	Neutral	Subjectivity score	Positive	Negative	Neutral	Compound score	Positive	Negative
February 2	100.00	0.00	0.00	0.92	0.17	0.00	0.83	0.70	100.00	0.00
February 9	37.50	37.50	25.00	0.34	0.12	0.09	0.79	0.15	87.50	12.50
February 16	63.64	9.09	27.27	0.27	0.09	0.05	0.85	0.12	54.55	45.45
February 23	50.00	50.00	0.00	0.37	0.17	0.04	0.79	0.48	100.00	0.00
March 1	83.33	0.00	16.67	0.59	0.14	0.03	0.83	0.36	66.67	33.33
March 8	44.44	22.22	33.33	0.32	0.14	0.03	0.83	0.38	66.67	33.33
March 15	85.71	0.00	14.29	0.52	0.05	0.02	0.92	0.16	100.00	0.00
March 22	66.67	0.00	33.33	0.40	0.12	0.02	0.86	0.49	66.67	33.33
March 29	38.57	27.14	34.29	0.37	0.08	0.12	0.80	–0.12	82.86	17.14
April 5	36.49	20.27	43.24	0.35	0.07	0.12	0.81	–0.11	82.43	17.57
April 12	33.33	41.67	25.00	0.49	0.10	0.07	0.82	0.10	83.33	16.67
April 19	66.67	6.67	26.67	0.47	0.06	0.06	0.88	–0.02	93.33	6.67
April 26	48.02	23.41	28.57	0.37	0.11	0.07	0.82	0.11	73.41	26.59
May 3	38.18	15.15	46.67	0.32	0.10	0.06	0.84	0.09	78.79	21.21
Average polarity	56.61	18.08	25.31	0.43	0.11	0.06	0.83	0.21	81.16	18.84

[a]VADER: Valence Aware Dictionary and Sentiment Reasoner.

[b]BiLSTM: bidirectional long short-term memory.

Somewhat similar to the TextBlob results, the Word2Vec-BiLSTM analysis unveiled a slightly more positive sentiment in the news communications, as 51% of the news was positive and 49% was negative. On the other hand, taking a look at the results from Twitter, results showed that, for the VADER analysis, the posts were 11% positive, 6% negative, and 83% neutral. The TextBlob analysis showed that the tweets were 57% positive, 18% negative, and 25% neutral, with a subjectivity score of 0.5, indicating an overall positive sentiment in the tweets. Whereas, the Word2Vec-BiLSTM analysis revealed that 81% of the tweets were positive, while 19% were negative, indicating a strong positivity in Twitter communications. Generally, it can be observed that sentiment results from TextBlob were more positive as compared to results from VADER. Elsewhere [96,97], researchers also made similar observations that VADER is more likely to pick up negative tones as compared to TextBlob.

Discussion

Principal Findings

Media plays a key role in information dissemination and consumption, which in turn influences people's opinions, attitudes, and decisions. Herein, we leveraged on the capability of computers to comprehend human language and describe emotional polarities in large amounts of data to delineate the sentiments in Twitter and Google News communications within the period of February 2 to May 5, 2020, regarding COVID-19 vaccines in the continent of Africa. In examining the media activity within the stated period, as expected, it was observed that media activity on the subject matter was relatively low in the first 3 weeks of the study period (ie, between February 2-23, 2020; Figure 1). This is because COVID-19 was just beginning to gain media attention in Africa during that time. In fact, the first case of COVID-19 in Africa was reported on February 14, 2020, in Egypt [93], while the first confirmed case in sub-Saharan Africa, Nigeria to be specific, was reported on February 28, 2020 [94].

It can be observed that there was a spike in media activity on both Twitter and Google News between March 29 and April 5, 2020, with corresponding increases in negative sentiments based on results from TextBlob and VADER (Figure 1 and Tables 1 and 2). This period coincides with negative public reception of the remarks made by two French medical doctors on COVID-19 testing in Africa [89]. The director-general of the WHO, Dr Tedros Adhanom Ghebreyesus, in a press briefing on April 6, 2020, strongly condemned the remarks by the doctors, comments he termed "racist remarks" and ascribed to a "hangover from colonial mentality" [90]. Interestingly coincidental was a public dissent, which culminated in the incineration of a COVID-19 testing facility in Abidjan, capital city of Côte d'Ivoire the same day, on fears that patients with COVID-19 would be treated at the center, because the facility was too close to residential apartments [98,99]. This hostile public response is reminiscent of psychosocial effects during the Ebola outbreaks in Central and West Africa wherein some health workers were attacked on suspicion that they were spreading the disease in their communities, rather than providing medical care [98]. Although the event in Abidjan is seemingly an isolated situation, such experiences reiterate the importance of knowledge and

information consumption during the time of a medical emergency or crisis.

Although there was a generally weak decline in positivity from both Twitter and Google News data across the study period, prevailing sentiments were neutral to positive. This gives an indication that generally truer and more evidence-based information regarding COVID-19 vaccines in Africa are circulating on Twitter and Google News as compared to falsehoods. This is because it is expected that, if factual and more science-based information is circulated on the media platforms about the prospects of a viable vaccine to tackle the COVID-19 pandemic in Africa, the communications and corresponding reactions should possess more positivity than negativity. It was observed that there was generally more positivity in Twitter data as compared to Google News data, which might indicate that there was less falsehoods communicated on Twitter as compared to Google News. This trend seems to be in contrast to observations made by other researchers that fake news spreads faster on social media (eg, Twitter, Facebook, Instagram, Snapchat, and WhatsApp) than evidence-based information [100,101]. Notwithstanding, our findings align considerably to those obtained by Pulido et al [67] on public information about COVID-19, wherein the authors noted that for tweets regarding COVID-19, though false information is tweeted more, it had overall less retweets as compared to fact-checking tweets and that science-based evidence tweets captured more engagement than mere facts. Fung et al [24] also made a similar observation, noting that after the declaration of the Ebola outbreak as an emergency in 2014, more accurate information circulated as compared to false information.

These seemingly counterintuitive findings cannot be logically described by chance or normal human behavior alone. Indeed, our observed sentiment patterns seem to be consistent with deliberate and sustained efforts made on various media platforms like Twitter, Facebook, Google, Microsoft, Reddit, and others to limit the spread of misleading and potentially harmful content regarding the COVID-19 pandemic [67,102,103]. For example, in March 2020, Twitter updated its policy guidance to address content sharing on its platform from authoritative sources of global and local health information that goes directly against guidance on COVID-19 [104]. Google, on the other hand, put in place mechanisms to ensure that searches related to COVID-19 on the company's search engine triggered an "SOS Alert," with news from mainstream publications including the WHO, National Public Radio, and the US Centers for Disease Control and Prevention displayed prominently [105].

In consonance with efforts made by the public and social media platforms, other national and international entities and health actors have also intensified efforts to ensure accurate and reliable information to the public regarding COVID-19. For example, the WHO's risk communication team launched a new information platform called WHO Information Network for Epidemics, immediately after it declared COVID-19 a Public Health Emergency of International Concern, devoted to myth-busting, debunking of false information, and sharing of tailored information with specific target groups [13,67]. In response to skepticism regarding COVID-19 vaccines, Media Monitoring Africa, a media organization that aims to promote the development of critical, ethical, and a free and fair media culture in South Africa and the rest of the continent, triggered its Real411 platform, wherein the public can report disinformation regarding COVID-19 to a digital complaints committee [106]. The platform was originally set up during the country's elections to enable reporting of objectionable speech by members of the public. This strategic media response was deemed necessary following a national survey conducted between April 15-23, 2020, among a representative sample of 600 people regarding COVID-19 vaccines in South Africa, which revealed that 21% of South Africans were strongly unwilling to be vaccinated [106].

Conclusion

Information constitutes a critical resource for mitigation and curative measures in the fight against the COVID-19 pandemic. This study assessed prevailing affective inclinations in media submissions regarding COVID-19 vaccines in Africa. Contrary to general perception, the results revealed a more passive to positive sentiment, which we could understand within the context of active media policing in ensuring safe and objective information regarding the COVID-19 pandemic. These findings are consistent with previous studies regarding public information about COVID-19. Our analytical approach attempted to provide a more universal and balanced perspective of the emotive patterns in the data sets by adopting three different NLP models (TextBlob, VADER, and Word2Vec-BiLSTM models) with fundamentally different underlying principles and mechanisms. Of course, sentiment analysis is a complex undertaking because speech communications can be highly subjective. For example, certain words or phrases can have different sentiments (ie, positive or negative) depending on the context of the text. Further to that, some narratives such as ironies and sarcasms are extremely difficult for machines and computers to understand because they cannot be interpreted literally. In this regard, we acknowledge the limitations of our approach (including the scope of the search terms and phrases used to query for the data on the internet); nonetheless, we believe that our study provides a reasonable approximation of the media polarity regarding the study topic. In future undertakings, the Word2Vec-BiLSTM and VADER libraries can be trained using more topic-specific vocabulary to improve their overall accuracies and predictability. Ultimately, building on insights from this study, public health and media actors can be stimulated to develop or re-evaluate relevant policies that promote responsible media use and public consumption to maximize the benefits of health interventions amid the COVID-19 crisis. This does not undermine the efficacy of efforts already made to curb misleading information regarding COVID-19 in the public space.

Acknowledgments

This paper was financially supported via the 2020 University Research Committee funding from the University of Johannesburg, granted to the main author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Google News data.
[XLSX File (Microsoft Excel File), 52 KB - medinform_v9i3e22916_app1.xlsx]

Multimedia Appendix 2
Twitter data.
[XLSX File (Microsoft Excel File), 99 KB - medinform_v9i3e22916_app2.xlsx]

References

1. Driggin E, Madhavan MV, Bikdeli B, Chuich T, Laracy J, Biondi-Zoccai G, et al. Cardiovascular considerations for patients, health care workers, and health systems during the COVID-19 pandemic. J Am Coll Cardiol 2020 May 12;75(18):2352-2371 [FREE Full text] [doi: 10.1016/j.jacc.2020.03.031] [Medline: 32201335]
2. Toure A. COVID-19 (coronavirus) drives sub-Saharan Africa toward first recession in 25 Years. The World Bank. 2020 Apr 09. URL: https://www.worldbank.org/en/news/press-release/2020/04/09/covid-19-coronavirus-drives-sub-saharan-africa-toward-first-recession-in-25-years [accessed 2020-08-08]
3. Nicola M, Alsafi Z, Sohrabi C, Kerwan A, Al-Jabir A, Iosifidis C, et al. The socio-economic implications of the coronavirus pandemic (COVID-19): a review. Int J Surg 2020 Jun;78:185-193 [FREE Full text] [doi: 10.1016/j.ijsu.2020.04.018] [Medline: 32305533]
4. Martin A, Markhvida M, Hallegatte S, Walsh B. Socio-economic impacts of COVID-19 on household consumption and poverty. Econ Disaster Clim Chang 2020 Jul 23:1-27 [FREE Full text] [doi: 10.1007/s41885-020-00070-3] [Medline: 32838120]
5. Buheji M, da Costa Cunha K, Beka G, Mavrić B, Leandro do Carmo de Souza Y, Souza da Costa Silva S, et al. The extent of COVID-19 pandemic socio-economic impact on global poverty. A global integrative multidisciplinary review. Am J Economics 2020 Aug 1;10(4):213-224. [doi: 10.5923/j.economics.20201004.02]
6. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). World Health Organization. 2020 Jan 30. URL: https://tinyurl.com/463gxq7l [accessed 2020-05-08]
7. Impact of the coronavirus (COVID-19) on the African economy. Tralac. 2020. URL: https://www.tralac.org/news/article/14483-impact-of-the-coronavirus-covid-19-on-the-african-economy.html [accessed 2020-05-08]
8. Sumner A, Hoy C, Ortiz-Juarez E. Estimates of the impact of COVID-19 on global poverty. UNU-WIDER. 2020. URL: https://www.wider.unu.edu/sites/default/files/Publications/Working-paper/PDF/wp2020-43.pdf [accessed 2020-10-12]
9. OECD. OECD Employment Outlook 2020: Worker Security and the COVID-19 Crisis. Paris: OECD Publishing; 2020.
10. Boberg S, Quandt T, Schatto-Eckrodt T, Frischlich L. Pandemic populism: facebook pages of alternative news media and the corona crisis--a computational content analysis. arXiv. Preprint posted online April 6, 2020. [FREE Full text]
11. Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, et al. The COVID-19 social media infodemic. arXiv. Preprint posted online March 10, 2020.
12. Hua J, Shaw R. Corona virus (COVID-19) "Infodemic" and emerging issues through a data lens: the case of China. Int J Environ Res Public Health 2020 Mar 30;17(7) [FREE Full text] [doi: 10.3390/ijerph17072309] [Medline: 32235433]
13. Zarocostas J. How to fight an infodemic. Lancet 2020 Feb 29;395(10225):676 [FREE Full text] [doi: 10.1016/S0140-6736(20)30461-X] [Medline: 32113495]
14. Taylor E. New survey reveals lockdown digital usage trends. Stylus. 2020. URL: https://www.stylus.com/new-survey-reveals-lockdown-digital-usage [accessed 2020-10-12]
15. Kemp S. Report: most important data on digital audiences during coronavirus. TNW. 2020. URL: https://thenextweb.com/growth-quarters/2020/04/24/report-most-important-data-on-digital-audiences-during-coronavirus/ [accessed 2020-10-12]
16. Rozenblum R, Bates DW. Patient-centred healthcare, social media and the internet: the perfect storm? BMJ Qual Saf 2013 Mar;22(3):183-186. [doi: 10.1136/bmjqs-2012-001744] [Medline: 23378660]
17. Singh S, Dixit A, Joshi G. Is compulsive social media use amid COVID-19 pandemic addictive behavior or coping mechanism? Asian J Psychiatr 2020 Dec;54:102290 [FREE Full text] [doi: 10.1016/j.ajp.2020.102290] [Medline: 32659658]
18. Sun C, Yang W, Arino J, Khan K. Effect of media-induced social distancing on disease transmission in a two patch setting. Math Biosci 2011 Apr;230(2):87-95 [FREE Full text] [doi: 10.1016/j.mbs.2011.01.005] [Medline: 21296092]

19. Yu W, Liu D, Zheng J, Liu Y, An Z, Rodewald L, et al. Loss of confidence in vaccines following media reports of infant deaths after hepatitis B vaccination in China. Int J Epidemiol 2016 Apr;45(2):441-449. [doi: 10.1093/ije/dyv349] [Medline: 27174834]
20. Jamison AM, Broniatowski DA, Quinn SC. Malicious actors on Twitter: a guide for public health researchers. Am J Public Health 2019 May;109(5):688-692. [doi: 10.2105/AJPH.2019.304969] [Medline: 30896994]
21. Brainard J, Hunter PR. Misinformation making a disease outbreak worse: outcomes compared for influenza, monkeypox, and norovirus. Simulation 2019 Nov 12;96(4):365-374. [doi: 10.1177/0037549719885021]
22. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019 Nov;240:112552 [FREE Full text] [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]
23. Betsch C. Advocating for vaccination in a climate of science denial. Nat Microbiol 2017 Jun 27;2:17106. [doi: 10.1038/nmicrobiol.2017.106] [Medline: 28653682]
24. Fung IC, Fu K, Chan C, Chan BSB, Cheung C, Abraham T, et al. Social media's initial reaction to information and misinformation on Ebola, August 2014: facts and rumors. Public Health Rep 2016;131(3):461-473 [FREE Full text] [doi: 10.1177/003335491613100312] [Medline: 27252566]
25. Kim L, Fast SM, Markuzon N. Incorporating media data into a model of infectious disease transmission. PLoS One 2019;14(2):e0197646 [FREE Full text] [doi: 10.1371/journal.pone.0197646] [Medline: 30716139]
26. Wei R, Lo V, Lu H. Third-person effects of health news: exploring the relationships among media exposure, presumed media influence, and behavioral intentions. Am Behav Scientist 2008 Jul 29;52(2):261-277. [doi: 10.1177/0002764208321355]
27. d'Onofrio A, Manfredi P, Manfredi P. Bifurcation thresholds in an SIR model with information-dependent vaccination. Math Modelling Nat Phenomena 2008 Jun 15;2(1):26-43. [doi: 10.1051/mmnp:2008009]
28. Wright J, Polack C. Understanding variation in measles-mumps-rubella immunization coverage--a population-based study. Eur J Public Health 2006 Apr;16(2):137-142. [doi: 10.1093/eurpub/cki194] [Medline: 16207728]
29. Luman ET, Fiore AE, Strine TW, Barker LE. Impact of thimerosal-related changes in hepatitis B vaccine birth-dose recommendations on childhood vaccination coverage. JAMA 2004 May 19;291(19):2351-2358. [doi: 10.1001/jama.291.19.2351] [Medline: 15150207]
30. Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N, Cook J. Misinformation and its correction: continued influence and successful debiasing. Psychol Sci Public Interest 2012 Dec;13(3):106-131. [doi: 10.1177/1529100612451018] [Medline: 26173286]
31. Danovaro-Holliday MC, Wood AL, LeBaron CW. Rotavirus vaccine and the news media, 1987-2001. JAMA 2002 Mar 20;287(11):1455-1462. [doi: 10.1001/jama.287.11.1455] [Medline: 11903035]
32. Allcott H, Gentzkow M, Yu C. Trends in the diffusion of misinformation on social media. Res Polit 2019 May 09;6(2):1-8. [doi: 10.1177/2053168019848554]
33. Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, et al. The science of fake news. Science 2018 Mar 09;359(6380):1094-1096. [doi: 10.1126/science.aao2998] [Medline: 29590025]
34. Merino JG. Response to Ebola in the US: misinformation, fear, and new opportunities. BMJ 2014 Nov 07;349:g6712. [doi: 10.1136/bmj.g6712] [Medline: 25380659]
35. Tchuenche JM, Bauch CT. Dynamics of an infectious disease where media coverage influences transmission. Int Scholarly Res Notices 2012 Mar 08;2012:1-10. [doi: 10.5402/2012/581274]
36. Liu R, Wu J, Zhu H. Media/psychological impact on multiple outbreaks of emerging infectious diseases. Computational Math Methods Med 2007;8(3):153-164. [doi: 10.1080/17486700701425870]
37. Yadav S, Ekbal A, Saha S, Bhattacharyya P. Medical sentiment analysis using social media: towards building a patient assisted system. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. 2018 Presented at: LREC 2018; May 2018; Miyazaki, Japan.
38. Liu B. Sentiment analysis and opinion mining. In: Hirst G, editor. Synthesis Lectures on Human Language Technologies. San Rafael, CA: Morgan & Claypool Publishers; 2012.
39. Cambria E, Das D, Bandyopadhyay S, Feraco A. Affective computing and sentiment analysis. In: Cambria E, Das D, Bandyopadhyay S, Feraco A, editors. A Practical Guide to Sentiment Analysis. Cham: Springer; 2017:1-10.
40. Natural language processing for sentiment analysis. Expert System Team. 2016. URL: https://expertsystem.com/natural-language-processing-sentiment-analysis/ [accessed 2020-05-08]
41. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. J Computational Sci 2011 Mar;2(1):1-8. [doi: 10.1016/j.jocs.2010.12.007]
42. Ozturk SS, Ciftci K. A sentiment analysis of twitter content as a predictor of exchange rate movements. Rev Econ Analysis 2014;6(2):140.
43. El Ansari O, Zahir J, Mousannif H. Context-based sentiment analysis: a survey. In: Adbelwahed EH, Bellatreche L, Benslimane D, Golfarelli M, Jean S, Mery D, et al, editors. New Trends in Model and Data Engineering: MEDI 2018 International Workshops, DETECT, MEDI4SG, IWCFS, REMEDY, Marrakesh, Morocco, October 24–26, 2018, Proceedings. Cham: Springer; 2018:91-97.

44. Kumar A, Garg G. Systematic literature review on context-based sentiment analysis in social multimedia. Multimedia Tools Applications 2019 Feb 23;79(21-22):15349-15380. [doi: 10.1007/s11042-019-7346-5]
45. Nankani H, Dutta H, Shrivastava H, Krishna PVNSR, Mahata D, Shah RR. Multilingual sentiment analysis. In: Agarwal B, Nayak R, Mittal N, Patnaik S, editors. Deep Learning-Based Approaches for Sentiment Analysis. Singapore: Springer; 2020:193-236.
46. Katz G, Ofek N, Shapira B. ConSent: context-based sentiment analysis. Knowledge-Based Syst 2015 Aug;84:162-178. [doi: 10.1016/j.knosys.2015.04.009]
47. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. 2014 Presented at: Eighth International AAAI Conference on Weblogs and Social Media; June 1-4, 2014; Ann Arbor, MI URL: http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf
48. Loria S. TextBlob: Simplified Text Processing. 2020. URL: https://textblob.readthedocs.io/en/dev/ [accessed 2020-05-09]
49. Zhang D, Xu H, Su Z, Xu Y. Chinese comments sentiment classification based on word2vec and SVMperf. Expert Syst Applications 2015 Mar;42(4):1857-1863. [doi: 10.1016/j.eswa.2014.09.011]
50. Bilyk V. Why business applies sentiment anaylsis? 5 successful examples. The App Solutions. 2019. URL: https://theappsolutions.com/blog/development/sentiment-analysis-for-business/ [accessed 2020-10-12]
51. Roccetti M, Marfia G, Salomoni P, Prandi C, Zagari RM, Gningaye Kengni FL, et al. Attitudes of Crohn's disease patients: infodemiology case study and sentiment analysis of Facebook and Twitter posts. JMIR Public Health Surveill 2017 Aug 09;3(3):e51 [FREE Full text] [doi: 10.2196/publichealth.7004] [Medline: 28793981]
52. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res 2013 Nov 01;15(11):e239 [FREE Full text] [doi: 10.2196/jmir.2721] [Medline: 24184993]
53. Abualigah L, Alfar HE, Shehab M, Hussein AMA. Sentiment analysis in healthcare: a brief review. In: Elaziz MA, Al-qaness MAA, Ewees AA, Dahou A, editors. Recent Advances in NLP: The Case of Arabic Language. Cham: Springer; 2020:129-141.
54. García-Díaz JA, Cánovas-García M, Valencia-García R. Ontology-driven aspect-based sentiment analysis classification: an infodemiological case study regarding infectious diseases in Latin America. Future Gener Comput Syst 2020 Nov;112:641-657 [FREE Full text] [doi: 10.1016/j.future.2020.06.019] [Medline: 32572291]
55. Afyouni S, Fetit A, Arvanitis T. #DigitalHealth: exploring users' perspectives through social media analysis. Stud Health Technol Inform 2015;213:243-246. [Medline: 26153005]
56. Monnier J, Laken M, Carter CL. Patient and caregiver interest in internet-based cancer services. Cancer Pract 2002;10(6):305-310. [doi: 10.1046/j.1523-5394.2002.106005.x] [Medline: 12406053]
57. Melzi S, Abdaoui A, Azé J, Bringay S, Poncelet P, Galtier F. Patient's rationale: patient Knowledge retrieval from health forums. 2014 Presented at: eTELEMED: eHealth, Telemedicine, and Social Medicine; 2014; Barcelone, Spain URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-01130720/document
58. Yang F, Lee AJ, Kuo S. Mining health social media with sentiment analysis. J Med Syst 2016 Nov;40(11):236. [doi: 10.1007/s10916-016-0604-4] [Medline: 27663246]
59. Barkur G, Vibha, Kamath GB. Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: evidence from India. Asian J Psychiatr 2020 Jun;51:102089 [FREE Full text] [doi: 10.1016/j.ajp.2020.102089] [Medline: 32305035]
60. Xue J, Chen J, Chen C, Zheng C, Li S, Zhu T. Public discourse and sentiment during the COVID 19 pandemic: using latent Dirichlet allocation for topic modeling on Twitter. PLoS One 2020;15(9):e0239441 [FREE Full text] [doi: 10.1371/journal.pone.0239441] [Medline: 32976519]
61. Imran AS, Daudpota SM, Kastrati Z, Batra R. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. IEEE Access 2020;8:181074-181090. [doi: 10.1109/access.2020.3027350]
62. Sayce D. The number of tweets per day in 2019. David Sayce. URL: https://www.dsayce.com/social-media/tweets-day/ [accessed 2020-05-08]
63. Genç Ö. The basics of NLP and real time sentiment analysis with open source tools. Towards Data Science. 2019. URL: https://towardsdatascience.com/real-time-sentiment-analysis-on-social-media-with-open-source-tools-f864ca239afe [accessed 2020-05-08]
64. Understanding the infodemic and minsinformation in the fight against COVID-19. IRIS PAHO. 2020. URL: https://iris.paho.org/bitstream/handle/10665.2/52052/Factsheet-infodemic_eng.pdf?sequence=14 [accessed 2020-07-09]
65. Haselmayer M, Jenny M. Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. Qual Quant 2017;51(6):2623-2646 [FREE Full text] [doi: 10.1007/s11135-016-0412-4] [Medline: 29070915]
66. Garvey C, Maskal C. Sentiment analysis of the news media on artificial intelligence does not support claims of negative bias against artificial intelligence. OMICS 2020 May;24(5):286-299. [doi: 10.1089/omi.2019.0078] [Medline: 31313979]
67. Pulido CM, Villarejo-Carballido B, Redondo-Sama G, Gómez A. COVID-19 infodemic: more retweets for science-based information on coronavirus than for false information. Int Sociol 2020 Apr 15;35(4):377-392. [doi: 10.1177/0268580920914755]
68. von Nordheim G, Boczek K, Koppers L. Sourcing the sources: an analysis of the use of Twitter and Facebook as a journalistic source over 10 years in The New York Times, The Guardian, and Süddeutsche Zeitung. Digital Journalism 2018 Sep 20;6(7):807-828. [doi: 10.1080/21670811.2018.1490658]

69. Castillo C, Mendoza M, Poblete B. Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web. 2011 Presented at: WWW '11; March 2011; Hyderabad, India p. 675-684. [doi: 10.1145/1963405.1963500]
70. Mendoza M, Poblete B, Castillo C. Twitter under crisis: can we trust what we RT? In: Proceedings of the First Workshop on Social Media Analytics. 2010 Presented at: SOMA '10; July 2010; Washington DC p. 71-79. [doi: 10.1145/1964858.1964869]
71. Veenstra AS, Iyer N, Hossain MD, Park J. Time, place, technology: Twitter as an information source in the Wisconsin labor protests. Comput Hum Behav 2014 Feb;31:65-72. [doi: 10.1016/j.chb.2013.10.011]
72. Hu H. GoogleNews 1.5.5. The Python Package Index. 2020. URL: https://pypi.org/project/GoogleNews/ [accessed 2020-05-06]
73. Sahni T, Chandak C, Chedeti NR, Singh M. Efficient Twitter sentiment classification using subjective distant supervision. 2017 Presented at: 9th International Conference on Communication Systems and Networks (COMSNETS); January 4-8, 2017; Bangalore URL: https://arxiv.org/pdf/1701.03051.pdf [doi: 10.1109/comsnets.2017.7945451]
74. Salas-Zárate MDP, Medina-Moreira J, Lagos-Ortiz K, Luna-Aveiga H, Rodríguez-García M, Valencia-García R. Sentiment analysis on tweets about diabetes: an aspect-level approach. Comput Math Methods Med 2017;2017:5140631. [doi: 10.1155/2017/5140631] [Medline: 28316638]
75. Hasan A, Moin S, Karim A, Shamshirband S. Machine learning-based sentiment analysis for Twitter accounts. Math Computational Applications 2018 Feb 27;23(1):11. [doi: 10.3390/mca23010011]
76. Kumar Singh A, Kumar Gupta D, Mohan Singh R. Sentiment analysis of Twitter user data on Punjab Legislative Assembly Election, 2017. Int J Modern Education Computer Sci 2017 Sep 08;9(9):60-68. [doi: 10.5815/ijmecs.2017.09.07]
77. Pandey P. Simplifying sentiment analysis using VADER in Python (on social media text). Medium. 2018. URL: https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f [accessed 2020-05-09]
78. Elbagir S, Yang J. Twitter sentiment analysis using natural language toolkit and VADER sentiment. In: Proceedings of the International MultiConference of Engineers and Computer Scientists. 2019 Presented at: IMECS; March 13-15, 2019; Hong Kong URL: https://pdfs.semanticscholar.org/74a2/7879b6c245d9ff7d9c4b41175ffd84b79d73.pdf
79. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv. Preprint posted online January 16, 2013. [FREE Full text]
80. Mikolov T, Sutskever I, Chen K, Corrado G, Dean T. Distributed representations of words and phrases and their compositionality. 2013 Presented at: Twenty-seventh Conference on Neural Information Processing Systems; December 5-10, 2013; Lake Tahoe, NV URL: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
81. Mikolov T, Yih WT, Zweig G. Linguistic regularities in continuous space word representations. 2013 Presented at: 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2013; Atlanta, GA URL: https://www.aclweb.org/anthology/N13-1090.pdf
82. Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Trans Signal Processing 1997;45(11):2673-2681. [doi: 10.1109/78.650093]
83. Sharfuddin AA, Tihami MN, Islam MS. A deep recurrent neural network with bilstm model for sentiment classification. 2018 Presented at: 2018 International Conference on Bangla Speech and Language Processing; September 21-22, 2018; Sylhet, Bangladesh. [doi: 10.1109/icbslp.2018.8554396]
84. Rhanoui M, Mikram M, Yousfi S, Barzali S. A CNN-BiLSTM model for document-level sentiment analysis. Machine Learning Knowledge Extraction 2019 Jul 25;1(3):832-847. [doi: 10.3390/make1030048]
85. Ceraj T, Kliman I, Kutnjak M. Redefining cancer treatment: comparison of Word2vec embeddings using deep BiLSTM classification model. University of Zagreb. 2019. URL: https://www.fer.unizg.hr/_download/repository/TAR-2019-ProjectReports.pdf#page=16 [accessed 2020-08-09]
86. Cherniuk A. Kaggle. 2019. URL: https://www.kaggle.com/alexcherniuk/imdb-review-word2vec-bilstm-99-acc [accessed 2020-05-09]
87. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint posted online December 22, 2014.
88. Coronavirus: France racism row over doctors' Africa testing comments. BBC News. 2020. URL: https://www.bbc.com/news/world-europe-52151722 [accessed 2020-07-13]
89. Bekker L, Mizrahi V. COVID-19 research in Africa. Science 2020 May 29;368(6494):919. [doi: 10.1126/science.abc9528] [Medline: 32467365]
90. COVID-19 virtual press conference - 6 April, 2020. World Health Organization. 2020. URL: https://www.who.int/docs/default-source/coronaviruse/transcripts/who-audio-emergencies-coronavirus-press-conference-full-06apr2020-final.pdf?sfvrsn=7753b813_2 [accessed 2020-06-16]
91. Fisayo-Bambi J. Madagascar president with herbal remedy for COVID-19. Africanews. 2020. URL: https://www.africanews.com/2020/04/21/madagascar-president-with-herbal-remedy-for-covid-19-morning-call/ [accessed 2020-06-15]

92. Reihani H, Ghassemi M, Mazer-Amirshahi M, Aljohani B, Pourmand A. Non-evidenced based treatment: an unintended cause of morbidity and mortality related to COVID-19. Am J Emerg Med 2021 Jan;39:221-222 [FREE Full text] [doi: 10.1016/j.ajem.2020.05.001] [Medline: 32402498]
93. Egypt announces first Coronavirus infection. Egypt Today. 2020. URL: https://www.egypttoday.com/Article/1/81641/Egypt-announces-first-Coronavirus-infection [accessed 2020-06-15]
94. Coronavirus: Nigeria confirms first case in sub-Saharan Africa. BBC News. 2020. URL: https://www.bbc.com/news/world-africa-51671834 [accessed 2020-06-15]
95. Paintsil E. COVID-19 threatens health systems in sub-Saharan Africa: the eye of the crocodile. J Clin Invest 2020 Jun 01;130(6):2741-2744. [doi: 10.1172/JCI138493] [Medline: 32224550]
96. White B. Sentiment analysis: VADER or TextBlob? Toward Data Science. 2020. URL: https://towardsdatascience.com/sentiment-analysis-vader-or-textblob-ff25514ac540 [accessed 2020-06-16]
97. VADER, IBM Watson or TextBlob: which is better for unsupervised sentiment analysis? Intellica.AI. 2020. URL: https://medium.com/@Intellica.AI/vader-ibm-watson-or-textblob-which-is-better-for-unsupervised-sentiment-analysis-db4143a39445 [accessed 2020-06-16]
98. Coronavirus: Ivory Coast protesters target testing centre. BBC News. 2020. URL: https://www.bbc.com/news/world-africa-52189144 [accessed 2020-06-17]
99. Crowd in Ivory Coast destroys coronavirus testing centre in residential area. France 24. 2020. URL: https://www.france24.com/en/20200406-crowd-in-ivory-coast-destroys-coronavirus-testing-centre [accessed 2020-06-17]
100. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018 Mar 09;359(6380):1146-1151. [doi: 10.1126/science.aap9559] [Medline: 29590045]
101. Fox M. Fake News: Lies spread faster on social media than truth does. NBC News. 2018. URL: https://www.nbcnews.com/health/health-news/fake-news-lies-spread-faster-social-media-truth-does-n854896 [accessed 2020-07-14]
102. Hancock JR. Las redes sociales y Google intentan contener la desinformación y el pánico sobre el coronavirus. Verne en El País. 2020. URL: https://verne.elpais.com/verne/2020/02/26/articulo/1582728106_118621.html [accessed 2020-06-16]
103. Statt N. Major tech platforms say they're 'jointly combating fraud and misinformation' about COVID-19. The Verge. 2020. URL: https://www.theverge.com/2020/3/16/21182726/coronavirus-covid-19-facebook-google-twitter-youtube-joint-effort-misinformation-fraud [accessed 2020-06-16]
104. Roth Y, Pickles N. Updating our approach to misleading information. Twitter Blog. 2020. URL: https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html [accessed 2020-06-15]
105. Newton C. Google has been unusually proactive in fighting COVID-19 misinformation. The Verge. 2020. URL: https://www.theverge.com/interface/2020/3/11/21173135/google-coronavirus-misinformation-youtube-covid-19-twitter-manipulated-media-biden [accessed 2020-06-16]
106. Khumalo J. Poll reveals significant skepticism over possible Covid-19 vaccine. News24. 2020. URL: https://www.news24.com/citypress/Special-Report/Covid-19_Survey/poll-reveals-significant-skepticism-over-possible-covid-19-vaccine-20200427 [accessed 2020-06-15]

Abbreviations

AI: artificial intelligence
API: application processing interface
BiLSTM: bidirectional long short-term memory
CSV: comma-separated values
HepB: hepatitis B
ML: machine learning
NLP: natural language processing
NN: neural network
VADER: Valence Aware Dictionary and Sentiment Reasoner
WHO: World Health Organization

Edited by G Eysenbach; submitted 27.07.20; peer-reviewed by A Chang, M Antoniou; comments to author 28.08.20; revised version received 20.10.20; accepted 08.12.20; published 16.03.21.

Please cite as:
Gbashi S, Adebo OA, Doorsamy W, Njobeh PB
Systematic Delineation of Media Polarity on COVID-19 Vaccines in Africa: Computational Linguistic Modeling Study
JMIR Med Inform 2021;9(3):e22916
URL: https://medinform.jmir.org/2021/3/e22916
doi:10.2196/22916
PMID:33667172

Original Paper

Emotional Attitudes of Chinese Citizens on Social Distancing During the COVID-19 Outbreak: Analysis of Social Media Data

Lining Shen[1,2,3*], PhD; Rui Yao[1*], MS; Wenli Zhang[1], MS; Richard Evans[4], PhD; Guang Cao[1], MS; Zhiguo Zhang[1,2], PhD

[1]School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science & Technology, Wuhan, China

[2]Hubei Provincial Research Center for Health Technology Assessment, Wuhan, China

[3]Institute of Smart Health, Huazhong University of Science & Technology, Wuhan, China

[4]College of Engineering, Design and Physical Sciences, Brunel University London, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Lining Shen, PhD
School of Medicine and Health Management
Tongji Medical College
Huazhong University of Science & Technology
No 13 Hangkong Road
Wuhan, 430030
China
Phone: 86 02783692730
Email: sln2008@hust.edu.cn

Abstract

Background: Wuhan, China, the epicenter of the COVID-19 pandemic, imposed citywide lockdown measures on January 23, 2020. Neighboring cities in Hubei Province followed suit with the government enforcing social distancing measures to restrict the spread of the disease throughout the province. Few studies have examined the emotional attitudes of citizens as expressed on social media toward the imposed social distancing measures and the factors that affected their emotions.

Objective: The aim of this study was twofold. First, we aimed to detect the emotional attitudes of different groups of users on Sina Weibo toward the social distancing measures imposed by the People's Government of Hubei Province. Second, the influencing factors of their emotions, as well as the impact of the imposed measures on users' emotions, was studied.

Methods: Sina Weibo, one of China's largest social media platforms, was chosen as the primary data source. The time span of selected data was from January 21, 2020, to March 24, 2020, while analysis was completed in late June 2020. Bi-directional long short-term memory (Bi-LSTM) was used to analyze users' emotions, while logistic regression analysis was employed to explore the influence of explanatory variables on users' emotions, such as age and spatial location. Further, the moderating effects of social distancing measures on the relationship between user characteristics and users' emotions were assessed by observing the interaction effects between the measures and explanatory variables.

Results: Based on the 63,169 comments obtained, we identified six topics of discussion—(1) delaying the resumption of work and school, (2) travel restrictions, (3) traffic restrictions, (4) extending the Lunar New Year holiday, (5) closing public spaces, and (6) community containment. There was no multicollinearity in the data during statistical analysis; the Hosmer-Lemeshow goodness-of-fit was 0.24 (χ^2_8=10.34, P>.24). The main emotions shown by citizens were negative, including anger and fear. Users located in Hubei Province showed the highest amount of negative emotions in Mainland China. There are statistically significant differences in the distribution of emotional polarity between social distancing measures (χ^2_{20}=19,084.73, P<.001), as well as emotional polarity between genders (χ^2_4=1784.59, P<.001) and emotional polarity between spatial locations (χ^2_4=1659.67, P<.001). Compared with other types of social distancing measures, the measures of *delaying the resumption of work and school* or *travel restrictions* mainly had a positive moderating effect on public emotion, while *traffic restrictions* or *community containment* had a negative moderating effect on public emotion.

Conclusions: Findings provide a reference point for the adoption of epidemic prevention and control measures, and are considered helpful for government agencies to take timely actions to alleviate negative emotions during public health emergencies.

(JMIR Med Inform 2021;9(3):e27079) doi:10.2196/27079

KEYWORDS

COVID-19; Sina Weibo; social distancing measures; emotional analysis; machine learning; moderating effects; deep learning; social media; emotion; attitude; infodemiology; infoveillance

Introduction

Background

In late 2019, COVID-19 began to spread rapidly throughout Hubei Province, China, creating devastating consequences for citizens and organizations and heavily burdening the provision of public health care in the province. The unknown pneumonia strain led to a major public health emergency; it has been deemed to be the disease with the fastest transmission rate, the greatest infection rate, and the most difficult to prevent since the establishment of New China [1]. The World Health Organization declared COVID-19 a public health emergency of international concern on January 31, 2020, and a pandemic on March 11, 2020 [2]. To prevent further outbreaks and disease transmission, the Chinese government adopted a series of measures to restrict the transmission and infection of the disease during the Lunar New Year holiday [3]. Although the 2020 Chinese Lunar New Year holiday is from January 24–31, 2020, to prevent and control the epidemic, the General Office of the State Council extended the holiday to February 2, 2020. At the epicenter of the epidemic, the People's Government of Hubei Province extended the holiday to February 13, 2020. The most widely imposed measure was to increase physical distancing between people [4], including traffic restrictions [5], delaying the resumption of work and school [6], extending the Lunar New Year holiday [7], travel restrictions [8], and community containment and closing of public spaces [9]. Such measures that aim to reduce exposure to the disease, by reducing contact between people, is known as "social distancing" [10]. The United States Centers for Disease Control and Prevention defined social distancing as the restriction of close face-to-face contact with others and considers it the best method for reducing the spread of COVID-19 [11].

Since the advent of the internet, social media has become an indispensable part of citizens' lives, greatly enriching the way people share feelings and exchange opinions. As of October 2020, there were approximately 4.66 billion active internet users worldwide, including 4.14 billion active social media users, accounting for 59% and 51% of the global population, respectively [12]. According to the 46th China Statistical Report on Internet Development, revised in June 2020, the number of Chinese internet users has reached 940 million [13]. International social media users tend to express their opinions on Twitter, due to its quick release and acceptance of information [14]. As an alternative to Twitter in Mainland China, Sina Weibo [15] has played an important role during many public health emergencies in recent years [16]. Studies have shown that through the analysis of content published on social media platforms, citizens' views and attitudes toward an event can be tracked and discovered [17,18]. During the COVID-19 pandemic, netizens expressed significant views on the imposed social distancing measures for disease prevention and control in Hubei Province. Through the collection of text related to social distancing on the internet, we can gain a better understanding of Sina Weibo users' emotional attitudes toward the imposed measures, so as to provide data for government agencies to implement measures in a timely fashion.

Related Work

Social Distancing During Public Health Emergencies

In public health emergencies, social distancing has become an important course of action to prevent the spread of diseases. For example, social distancing measures imposed during the influenza pandemic drastically reduced the infection rate [19,20]. One study into the severe acute respiratory syndrome (SARS) epidemic found that the prevention and control of the epidemic, through the implementation of social distancing measures, had a certain effect in Canada [21]. During the COVID-19 pandemic, government agencies have also encouraged the adoption of social distancing measures. Zhang et al [22] studied the impact of social distancing measures on the spread of COVID-19 by establishing a disease transmission model. Other studies have confirmed that by closing schools and universities, there has been a significant reduction in the spread of the disease [23-25]. Self-quarantine measures also helped reduce the transmission rates of influenza in Texas in 2009, as well as for SARS in Singapore in 2003 [26,27]. The sanitary cordon helped set back an epidemic of Ebola in 1995 in the city of Kikwit, Zaire [28]. Workplace social distancing guidelines delayed and reduced the peak influenza attack rate [29]. In addition, travel restrictions and the canceling of mass gatherings have also been effective strategies for reducing the burden of COVID-19 on health care providers [30,31].

However, the use of social distancing measures by local governments during the COVID-19 pandemic may have a negative impact on citizens' mental health and well-being [32,33]. Therefore, it is important to study the psychology of citizens and their behaviors in the context of social distancing measures.

Emotional Expression During Public Health Emergencies

Emotional polarity analysis is often used to examine the emotional tendencies expressed in text and to discover the emotional attitudes of users. Some studies have found that the pandemic has led to negative emotions being expressed by internet users [34,35]. For example, Ogoina et al [36] found that health care workers demonstrated varying degrees of fear related to the Ebola epidemic in Nigeria in 2014, through self-administered questionnaires and documented observations. Other scholars have explored the emotional attitudes of citizens toward epidemic diseases through questionnaires [37,38]. However, all studies have only explored the change in citizens'

emotions through cross-sectional data, and the samples examined were not large.

As an important medium for public communication, social media provides a large-scale corpus for emotional analysis. Emotion lexicons and machine learning are two methods commonly used for analytical analysis. In terms of emotion lexicons, there are relatively mature lexical resources available, such as SentiWordNet and the National Research Council of Canada (NRC) word-emotion lexicon [39,40]. Das et al [41] used the NRC word-emotion lexicon to analyze the emotions of Twitter users about COVID-19 in India. In the field of machine learning, Du et al [42] used a convolutional neural network (CNN) classifier to conduct an emotional analysis on Twitter data during the measles outbreak. Ji et al [43] also found that multinomial naive Bayes (NB) was effective in analyzing the emotion of Twitter users facing epidemics. Although a classification method based on emotion lexicons is effective, it relies heavily on the scale and frequency of updating the lexicon, while a classification method based on machine learning avoids this limitation and is widely used today. For example, Behera et al [44] studied the emotional classification of tweets related to three diseases—malaria, swine flu, and cancer—through emotion lexicon and NB methods. Their results showed that the performance of the NB algorithm was better than that of the emotion lexicon.

Objectives

Social distancing measures are essential for controlling the spread of infectious diseases during pandemics. However, based on existing research, it is evident that current studies into social distancing have mainly focused on the impact of social distancing measures on the spread of infection rates and mortality. To date, there are few studies that have explored citizens' emotions related to specific social distancing measures. By detecting the emotional attitudes of Sina Weibo users toward the imposed social distancing measures adopted by Hubei Province, the epicenter of the COVID-19 outbreak in Mainland China, government agencies can take timely action to calm citizens' emotions. In this study, the following four research questions (RQs) are identified:

- RQ1: what were the emotional attitudes of Sina Weibo users toward the various social distancing measures imposed by Hubei Province?
- RQ2: what were the changing trends in Sina Weibo users' emotions over time?
- RQ3: what was the impact of user characteristics of social media on their emotions?
- RQ4: what was the moderating effect of social distancing measures on the relationship between the explanatory variables and the explained variable?

Methods

Data Collection and Preprocessing

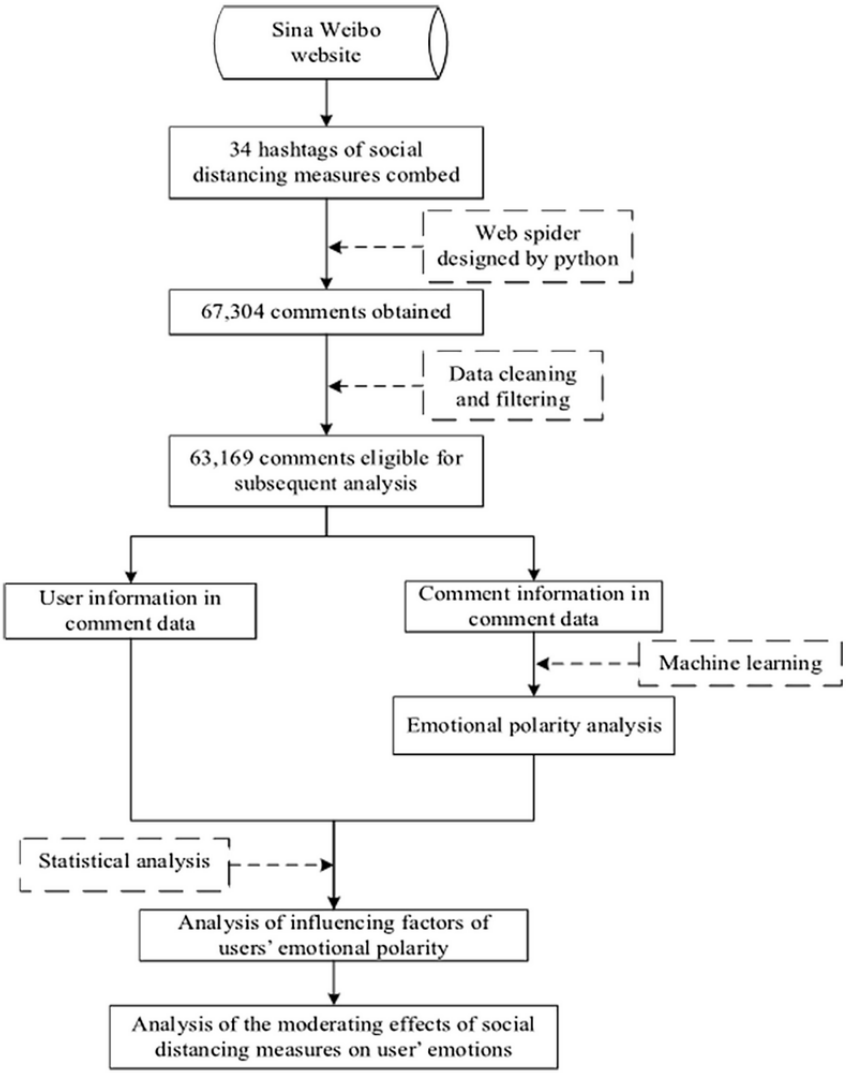
Due to the severity of the COVID-19 outbreak in Hubei Province, the government of Hubei Province issued the "Notice of the People's Government of Hubei Province on strengthening the prevention and control of the new coronavirus pneumonia" on January 21, 2020 [45], which stipulated strict restrictions against large-scale activities and personal movement; this led to the successful employment of social distancing measures to reduce the risk of infection among citizens. Subsequently, due to improvements in the epidemic situation, the government of Hubei Province announced on March 24, 2020, that it would gradually lift restrictions in the province. Therefore, the time span of the selected data in this study range from January 21 to March 24, 2020.

In addition, since Sina Weibo has gradually become the main social media platform for Chinese citizens, this study selected this platform as the data source. After combing hashtags on Sina Weibo, which is an efficient method for identifying trending topics [46], a total of 34 hashtags related to social distancing measures adopted by Hubei Province were identified. These hashtags were then categorized under the six social distancing measures, as shown in Table 1. To ensure data quality, comments about posts published by the official news channels of government agencies, certified by the Sina Weibo platform, were crawled using Python 3.7, retrieving a total of 67,304 comments published by 58,996 netizens. After removing duplicate data and comments without textual expression, 63,169 comments were identified. The main data collected included username, gender, spatial location, account registration year, and comment content. The process of subsequent analysis is shown in Figure 1.

Table 1. Hashtags on social distancing measures related to COVID-19 on Sina Weibo.

Measures categories	Hashtags
Delaying the resumption of work and school	• #The start of the school year for Hubei elementary and middle schools has been postponed# • #Colleges and universities in Hubei have postponed the start of classes# • #Schools in Hubei have postponed the start of the school year# • #Hubei has postponed the start of school# • #Hubei continues to delay the resumption of work and school# • #The opening of schools in Hubei Province has been postponed# • #Hubei continues to delay the start of school# • #Hubei issued a notice to continue to delay the resumption of work and schools# • #Various enterprises in Hubei will resume work no earlier than 24:00 on February 20# • #Enterprises in Hubei Province will resume work no earlier than 24:00 on March 10#
Travel restrictions	• #Wuhan canceled all tour groups# • #All tour groups in Wuhan will be cancelled# • #Hubei travel agencies have suspended business activities# • #Hubei province has suspended the operations of travel agencies across the province#
Traffic restrictions	• #Wuhan's public transportation and subway operations have been suspended# • #Buses and subways in Wuhan were suspended# • #Traffic in Wuhan was suspended# • #Long-distance passenger transportation of Wuhan bus, subway and ferry will be suspended from the 23rd# • #The Wuhan exit route was temporarily closed# • #Huanggang railway station was closed# • #Wuhan airport, railway station, and other exit routes were temporarily closed# • #Wuhan closed the river-crossing tunnel#
Extending the Lunar New Year holiday	• #The Spring Festival holiday was extended to February 2# • #Hubei Province will appropriately extend the Spring Festival holiday# • #Hubei Province extended the Spring Festival holiday until February 13#
Closing public spaces	• #Cinemas throughout Wuhan were temporarily closed# • #Wuhan cultural and entertainment venues were temporarily closed# • #All star hotel activities in A-level scenic spots in Hubei Province will be cancelled# • #All non-essential public places in Hubei were closed#
Community containment	• #Communities in Hubei Province are under closed management# • #All residential communities in Wuhan are under closed management# • #Wuhan community adopted closed management# • #The communities in Hubei Province are most strictly closed 24 hours a day# • #The closed management of villages and community in the Wuhan will continue 24 hours a day#

Figure 1. Flowchart for obtaining data from Sina Weibo for subsequent analysis.



Emotional Polarity Analysis

Since the introduction of text sentiment analysis in 2008, machine learning methods have achieved consistently good results [47,48]. Therefore, we chose machine learning algorithms to perform emotional analysis on the corpus.

First, corpora marking was completed. The current research on emotional analysis mainly divides emotional types into three or five categories [49,50]. In addition, the Ortony-Clore-Collins (OCC) model, proposed by Ortony et al [51] in 1988, is a refined emotion classification model that offers a rule-based emotion export mechanism, which has been widely applied to studies that examine emotion classification of social media users [52]. Based on the three evaluation criteria of the OCC model, consequences of events, action of agents, and aspects of objects, as well as different intensities or inducing causes, this study constructed "anger," "fear," "neutral," "encouragement," and "hope" emotions to analyze Sina Weibo users' emotions. The emotions of "anger" and "encouragement" are related to the action of agents, emotions of "fear" and "hope" are related to the consequences of events, and neutral emotions indicate objective facts. Then, according to the above rules of emotion classification, we programed in Python; the first and second authors used traditional labeling methods to label the corpus [53]. We randomly chose more than 5000 corpora and marked them with one of the five emotional polarity. Then, the Kappa coefficient was used to evaluate the consistency of the corpus marking [54]. The Kappa value was 0.95 (P<.01), indicating that the marked corpora had strong consistency [55].

We then predicted the emotional polarity of the comments collected. First, we selected four representative training classifiers of machine learning algorithms: support vector machine (SVM), convolutional neural networks (CNN), long short-term memory (LSTM), and bi-directional long short-term memory (Bi-LSTM). Based on the marked corpora, we divided them into training sets, validation sets, and test sets, according to a 6:2:2 ratio, and then input them into different classifier models for testing. These tests were carried out in September 2020. Three indicators—precision, recall, and F1-score—were used for the model evaluation [56]. The precision rate reflects the ability of the classifier to determine the whole sample; the recall rate intuitively reflects the proportion of positive samples that are correctly identified; and the F1-score can be interpreted as a weighted average of precision and recall. As shown in Table 2, the Bi-LSTM classifier exhibited the best performance for testing emotional polarity for the remaining corpora [57].

Subsequently, the Bi-LSTM classifier was used to predict the emotional polarity of all corpora, and this process was performed using Python.

Table 2. The performance of the emotional polarity classification model.

Model	Precision	Recall	F1-score
SVM[a]	0.512	0.506	0.509
CNN[b]	0.619	0.602	0.607
LSTM[c]	0.664	0.664	0.658
Bi-LSTM[d]	0.701	0.699	0.701

[a]SVM: support vector machine.

[b]CNN: convolutional neural network.

[c]LSTM: long short-term memory.

[d]Bi-LSTM: bidirectional long short-term memory.

Research Hypotheses and Statistical Analysis

Theoretical Background and Hypotheses Design

To study the emotional attitudes of Sina Weibo users, we examined the impact of user characteristics and social distancing measures on the emotional tendency of users. We then proposed hypotheses.

First, according to the Media Dependency Theory [58], the effects of the media are due to the media meeting the needs of specific audiences in a specific way in a specific society. Obviously, audiences' use of media platforms determines the media's influence, namely, personal media dependence depends on personal factors [59]. Therefore, based on the Media Dependence Theory and the data collected on Sina Weibo users, this study hypothesizes that users' age, spatial location, and social media registration year affects their emotional expression [60,61]. Accordingly, we proposed the following hypotheses:

- H1a: the age of Sina Weibo users has a significant impact on their emotional expression;
- H1b: the spatial location of Sina Weibo users has a significant impact on their emotional expression;
- H1c: the registration year of Sina Weibo users has a significant impact on their emotional expression.

Second, according to the Risk Information Seeking and Processing Model, proposed by Griffin [62], relevant channel beliefs affect individuals' information processing methods differently [63]. In addition, the 5W Theory (Who, When, Where, What, and How), proposed by Lasswell [64], reveals the communication elements in the communication process. Among them, the first element "who" is played by different types of users on the Sina Weibo platform [65-67]. Therefore, this study measured the emotional response of different user characteristics from the perspective of social media, and proposed the following hypotheses:

- H2a: the number of fans of Sina Weibo users has a significant impact on their emotional expression;
- H2b: the number of follows of Sina Weibo users has a significant impact on their emotional expression;
- H2c: the number of posts shared by Sina Weibo users has a significant impact on their emotional expression.

Third, the Agenda-Setting Theory, formally proposed by McCombs and Shaw [68], posits that mass media can effectively influence users' attention to certain facts by providing information and arranging related issues. The theory further discusses that records of public discussion on public affairs can be obtained from social media for observation and analysis [69,70]. Based on this theory, we discussed the impact of different social distancing measures on the emotional expression of different types of Sina Weibo users, and proposed the following hypothesis:

- H3: The categories of social distancing measures have a moderating effect on the relationship between the user characteristics of different types of Sina Weibo users and their emotions.

Statistical Analysis of Influencing Factors of Emotional Polarity

Based on the above hypotheses, we selected the factors that affect users' emotions for statistical analysis. Given that the P value of the test of parallel lines is less than .001, we analyzed the factors that influence Sina Weibo users' emotions using multinomial logistic regression, performed by SPSS 21.0 (IBM Corp); our variables are shown in Table 3. Among the variables, based on the existing research results [71], gender was set as the control variable. Based on related data [72,73], the spatial locations of Sina Weibo users were divided into Hubei Province, emigrant provinces with the largest population influx from Hubei Province before the Wuhan lockdown (including Guangdong, Beijing, Shanghai, Henan, Anhui, Jiangxi, Sichuan, Hunan, and Chongqing), and other provinces. In addition, the effects of various explanatory variables on anger and fear emotions, as well as hope and encouragement emotions, are consistent. Therefore, we merged the five categories of emotions into three categories (positive, neutral, and negative) for statistical analysis. Further, the moderator variables included the six categories of social distancing measures adopted by Hubei Province. In the subsequent analysis, each category was set as the reference group to observe the interaction effect between it and the explanatory variables, which can judge the moderating effect of social distancing measures [74]. From the above, we established the following model:

$$Logit(P\_Emo) = \beta_0 + \beta_1 Sex + \beta_2 Age + \beta_3 Space1 + \beta_4 Space2 + \beta_5 Ln(N\_Fan) + \beta_6 Ln(N\_Follow) + \beta_7 Ln(N\_Post) + \beta_8 Time\_Reg + \beta_9 T\_SDM + \beta_{10} Age \times T\_SDM + \beta_{11} Space1 \times T\_SDM + \beta_{12} Space2 \times T\_SDM + \beta_{13} Ln(N\_Fan) \times T\_SDM + \beta_{14} Ln(N\_Follow) \times T\_SDM + \beta_{15} Ln(N\_Post) \times T\_SDM + \beta_{16} Time\_Reg \times T\_SDM + \varepsilon$$

Table 3. Description of variables that influence Sina Weibo users' emotions.

Variable	Variable symbol	Description and coding
Explained variable		
Emotional polarity	P_Emo	Sina Weibo users' emotional polarity
Control variable		
Sex	Sex	The sex of Sina Weibo users. The coding is as follows: 0=female and 1=male (reference group)
Explanatory variables		
Age	Age	The age of Sina Weibo users (range 16-65 years)
Spatial location	Space1, Space2	The spatial locations of Sina Weibo users. We used "other provinces" as the reference group. The dummy variables are as follows: • Space0= {other provinces, when Space1=0 and Space2=0} (reference group) • Space1= {1=Hubei Province; 0=others} • Space2={1=emigrant provinces; 0=others}
Number of fans	Ln(N_Fan)	The number of fans of Sina Weibo users; smoothed logarithmically
Number of follows	Ln(N_Follow)	The number of other users that Sina Weibo users follow; smoothed logarithmically
Number of post	Ln(N_Post)	The number of posts shared by Sina Weibo users; smoothed logarithmically
Registration year	Time_Reg	The registration year of Sina Weibo users' accounts
Moderator variable		
Social distancing measures	T_SDM	The types of social distancing measures. The coding is listed as follows: • SDM1=delaying the resumption of work and school • SDM2=travel restrictions • SDM3= traffic restrictions • SDM4=closing public spaces • SDM5=community containment • SDM6=extending the Lunar New Year holiday (reference group)

In total, 21,395 comments were identified for statistical analysis after removing some records with missing age, since disclosing age is not required for user registration.

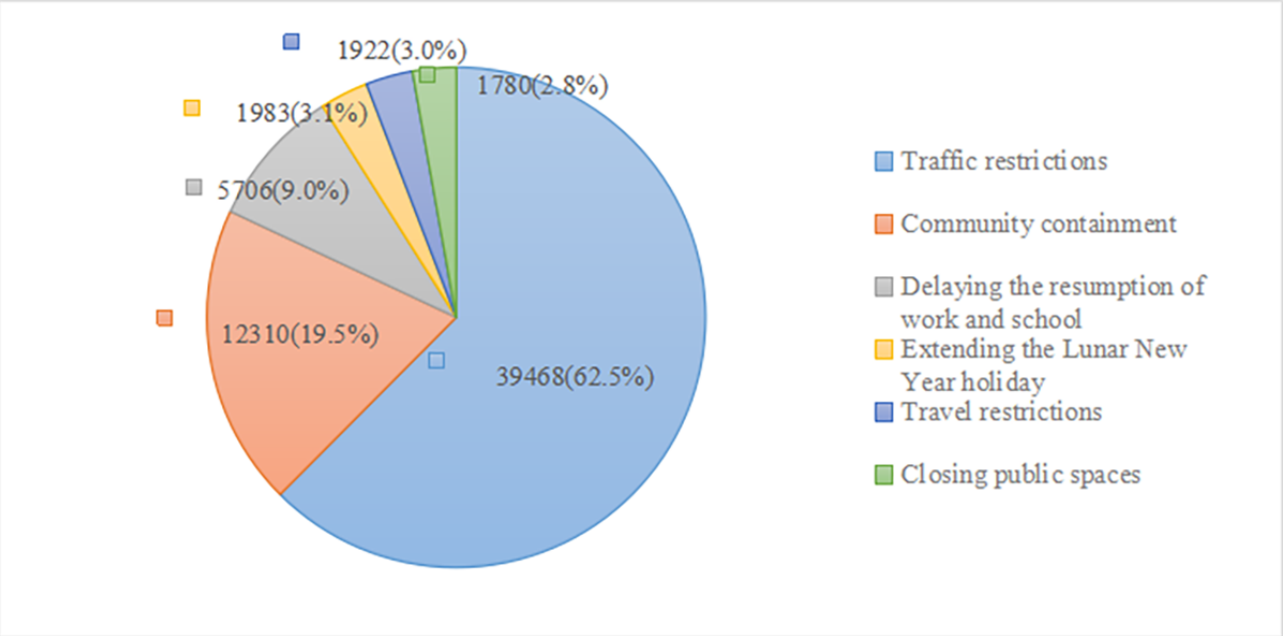
Results

Basic Description of Comments and Users' Emotional Polarities

Distribution of Comments

From January 21 to March 24, 2020, Sina Weibo users discussed the social distancing measures imposed by the People's Government of Hubei Province. The three measures of *traffic restrictions*, *community containment*, and *delaying the resumption of work and school* attracted high attention from users, while *travel restrictions*, *extending the Lunar New Year holiday*, and *closing public places* attracted less attention; further details are shown in Figure 2.

Figure 2. Proportion of comments related to each social distancing measure (N=63,169).



To analyze the data of various provinces in China, the data of users registered as “other” and “overseas” were removed, yielding 46,431 comments. Local netizens in Hubei Province paid the greatest attention to social distancing measures, followed by those residing in the Guangdong, Beijing, and Shanghai. These are areas where citizens from Hubei Province relocated to before the lockdown was imposed in Wuhan City. The number of comments from users in other provinces was relatively small, as shown in Figure 3.

Figure 3. Spatial distribution of comments from Sina Weibo users (N=46,431).

Distribution of Users' Emotional Polarity Across Gender

The findings show that women held a higher proportion of positive emotions than men (Table 4). Further, the chi-square test result (χ^2_4=1784.59, *P*<.001) shows that the emotional expression of different genders varied.

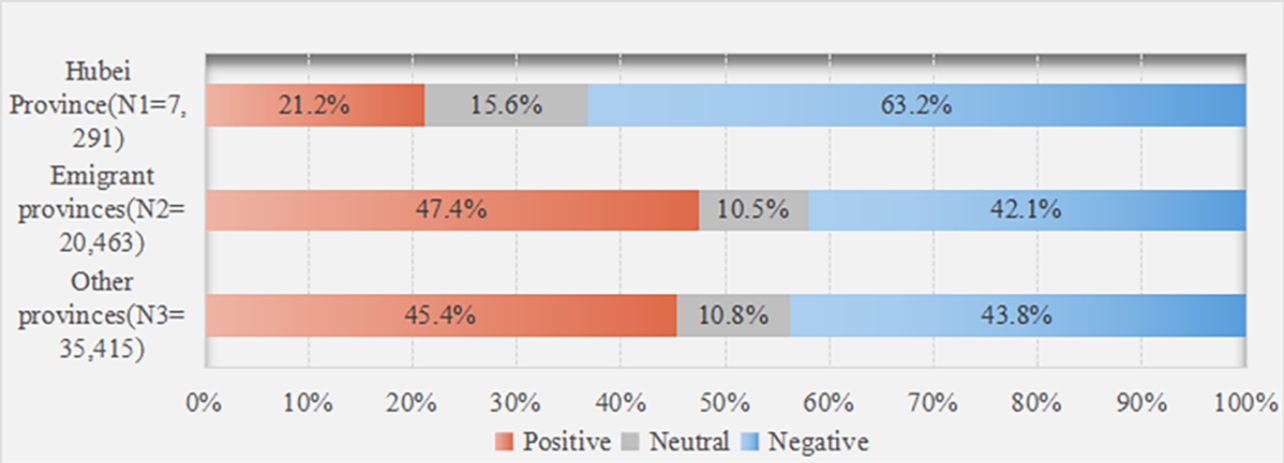
Table 4. Proportion of users' emotional polarities across different genders (N=63,169).

Sex	Positive		Neutral	Negative	
	Hope	Encouragement		Fear	Anger
Male (n=19,016), n (%)	1512 (8.0)	4505 (23.7)	2896 (15.2)	4467 (23.5)	5636 (29.6)
Female (n=44,153), n (%)	4383 (9.9)	16,936 (38.4)	4207 (9.5)	9601 (21.7)	9026 (20.4)

Distribution of Emotional Polarity Among Users From Different Spatial Locations

As shown in Figure 4, for users in Hubei Province, the proportion of negative emotions (4605/7291, 63.2%) is significantly different from that of positive emotions (1545/7291, 21.2%). Sina Weibo users in Hubei Province showed the least amount of positive emotions while residents of emigrant provinces expressed the most positive emotions. There was little difference between the emigrant provinces and other provinces, but the proportion of negative emotions was slightly lower than that of other provinces. In addition, the chi-square test result (χ^2_4=1659.67, *P*<.001) shows that the emotional expression of users in different spatial locations is indeed different.

Figure 4. Proportion of users' emotional polarity from different spatial locations (N=63,169).

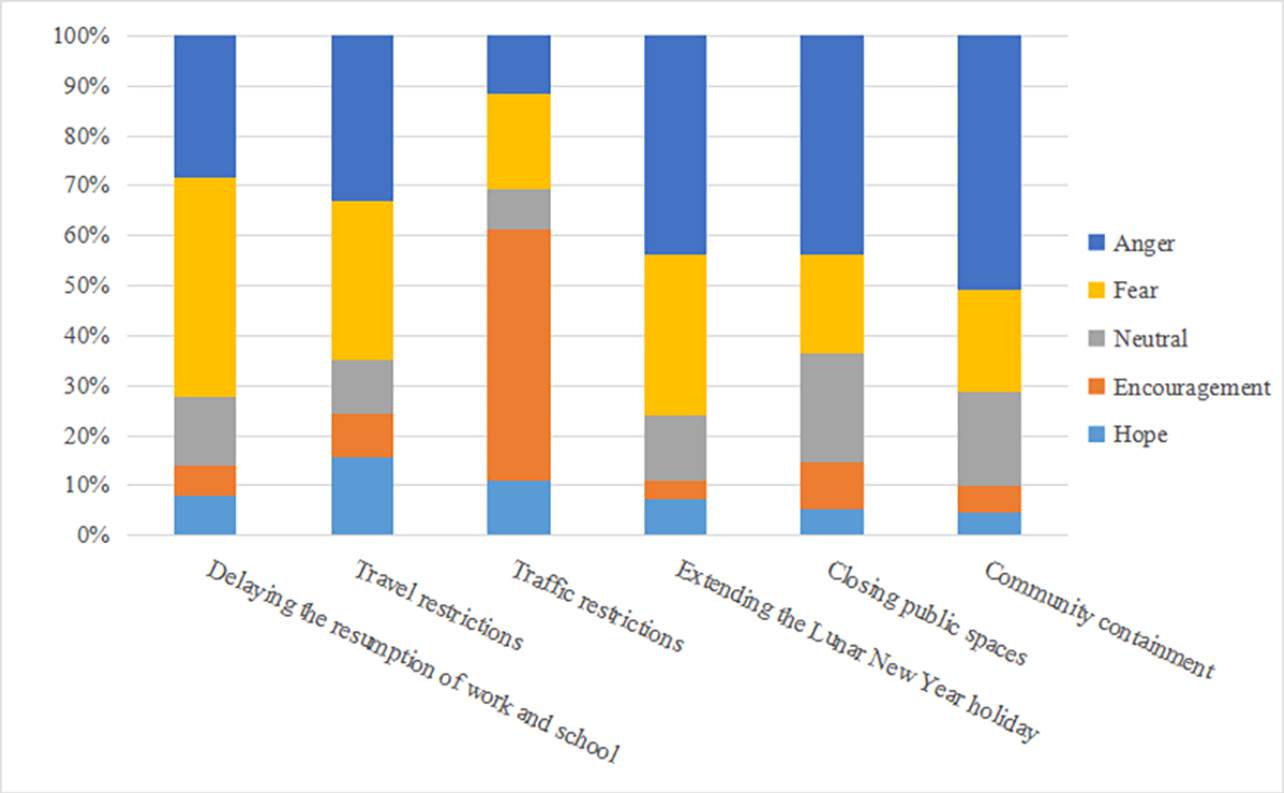


Evolution of Emotional Polarity

Distribution of Users' Emotional Polarity Based on Different Social Distancing Measures

Except for the measure of *traffic restrictions*, netizens' emotions regarding the other five measures were all negatively inclined (Figure 5). Specifically, for the measure of *traffic restrictions*, netizens' positive emotions make up the largest proportion (24,143/39,468, 61.2%) among all the measures. For *delaying the resumption of work and schoo*l, netizens mostly expressed negative emotions (4129/5706, 72.4%), such as fear and anger. For *travel restrictions*, emotions surrounding hope accounted for the highest proportion (472/1922, 15.6%). In terms of *extending the Lunar New Year holiday*, *closing public spaces*, and *community containment*, negative emotions accounted for a high proportion of total comments. For *community containment*, anger was expressed more frequently compared to all other measures. The chi-square test result (χ^2_{20}=19,084.73, *P*<.001) showed that Sina Weibo users have varying emotional expressions for different topics.

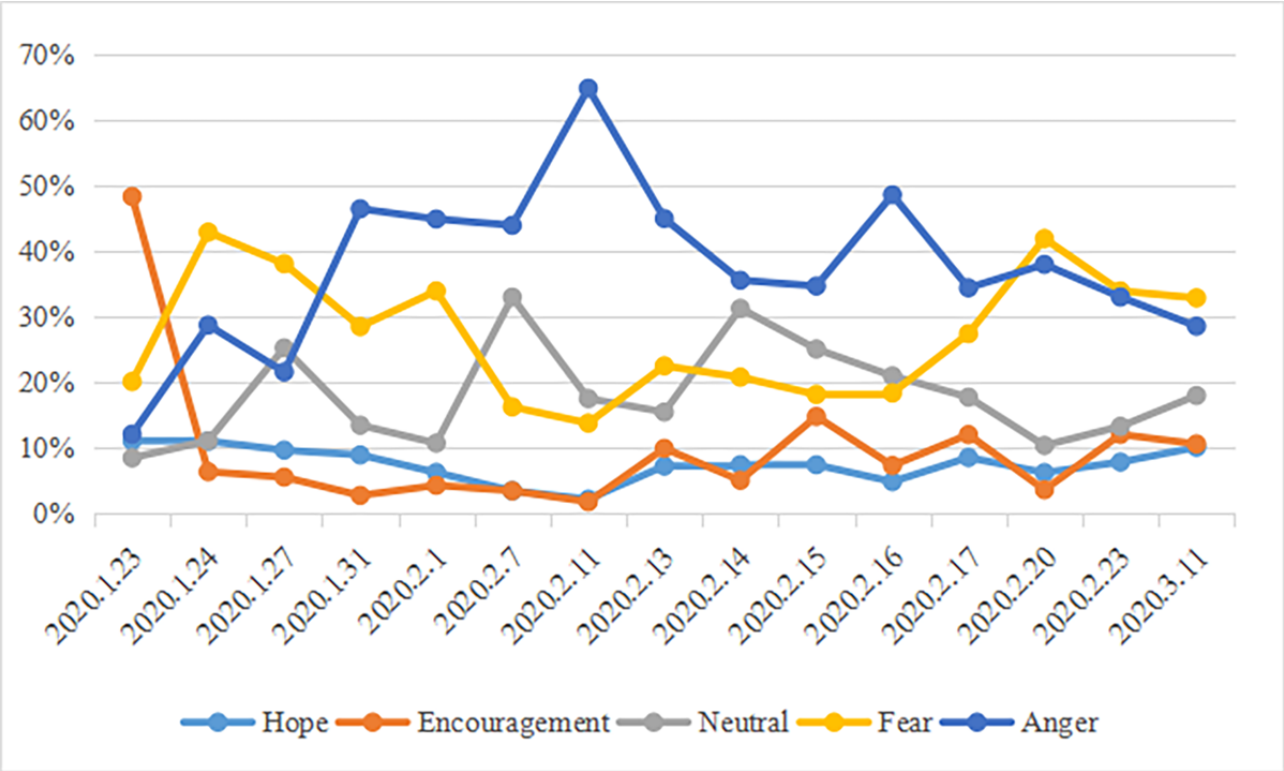
Figure 5. Distribution of users' emotional polarity under various social distancing measures (N=63,169).



Evolution Process of Users' Emotional Polarity Over Time

As shown in Figure 6, on January 23, 2020, when the city of Wuhan was locked down, users showed a high degree of encouragement. However, since January 24, 2020, negative emotions gradually increased with users expressing fear and anger toward the imposed social distancing measures employed by Hubei Province. Then, after January 27, 2020, users' negative emotions centered on anger toward the social distancing measures imposed in Hubei Province.

Figure 6. Evolution process of emotional polarity over time (N=63,169).



Moderating Effects of Social Distancing Measures on the Influencing Factors of Emotional Polarity

The influence of explanatory variables on the explained variable is shown in Table 5. The reference group of emotional polarity is neutral emotions, and there was no multicollinearity in the data. Older citizens were more inclined to express neutral attitudes. Users in Hubei Province were more likely to express negative emotions, although there was no statistically significant difference in emotional tendency between emigrant provinces and other provinces. Further, users with accounts registered for longer periods of time were inclined to express negative emotions. Similarly, users with more follows and posts were inclined to express positive attitudes, while users with more fans were inclined to express neutral attitudes. Therefore, the explanatory variables do have an impact on the explained variable. Accordingly, hypotheses H1a-c and H2a-c were supported.

Table 5. Analysis of the influence of explanatory variables on explained variables (N=21,395).

Variables	Negative, OR[a] (95% CI)	Positive, OR (95% CI)
Sex (reference group: male)	1.18 (1.05-1.32[b])	2.21 (1.96-2.49[b])
Age	0.99 (0.98-0.99[b])	0.97 (0.98-0.98[b])
Space0 (reference group)	—[c]	—
Space1	1.18 (1.02-1.38[b])	0.52 (0.44-0.62[b])
Space2	1.09 (0.97-1.22)	1.10 (0.98-1.24)
Time_Reg[d]	1.02 (0.99-1.04)	0.90 (0.88-0.93[b])
Ln(N_Fan)[e]	0.97 (0.94-1.01)	0.98 (0.94-0.99[b])
Ln(N_Follow)[f]	1.01 (1.02-1.08[b])	1.05 (0.99-1.12)
Ln(N_Post)[g]	0.96 (0.93-0.99[b])	1.06 (1.03-1.10[b])
SDM1[h]	0.49 (0.31-0.78[b])	7.66 (4.99-11.77[b])
SDM2[i]	0.46 (0.28-0.75[b])	6.75 (4.14-11.00[b])
SDM3[j]	0.59 (0.38-0.92[b])	1.08 (0.73-1.59)
SDM4[k]	0.92 (0.46-1.81)	0.70 (0.38-1.28)
SDM5[l]	0.56 (0.36-0.88[b])	8.63 (5.68-13.11[b])
SDM6[m] (reference group)	—	—

[a]OR: odds ratio.
[b]*P*<.05.
[c]Not applicable.
[d]Time_Reg: registration year.
[e]Ln(N_Fan): logarithm of fan numbers.
[f]Ln(N_Follow): logarithm of follow numbers.
[g]Ln(N_Post): logarithm of post numbers.
[h]SDM1: delaying the resumption of work and school.
[i]SDM2: travel restrictions.
[j]SDM3: traffic restrictions.
[k]SDM4: closing public spaces.
[l]SDM5: community containment.
[m]SDM6: extending the Lunar New Year holiday.

To examine the effects of the moderating variable on the relationship between users' characteristics and emotions (positive or negative), interactions between social distancing measures and explanatory variables were observed through binary logistic regression analysis. During this process, 64 outliers were removed, based on the residuals analysis, making the results more meaningful. The omnibus tests of model coefficients were statistically significant (χ^2_{48}=4994.56, *P*<.001), and the Hosmer-Lemeshow goodness-of-fit was 0.24 (χ^2_8=10.34, *P*>.24), indicating that our model has a good goodness of fit. In review of the interaction term results for each social distancing measure of the moderating variable and the explanatory variable, including the odds ratio and 95% CI, it was found that compared with other social distancing measures, some measures had a positive or negative emotion regulation

effect on the relationship between the explanatory variable and emotional attitude.

Table 6 summarizes the statistically significant emotional tendency of each explanatory variable under each social distancing measure (for the details of interaction effect, see Multimedia Appendix 1). In general, for these explanatory variables, compared with other types of social distancing measures, the measures of delaying the resumption of work and school (SDM1) or travel restrictions (SDM2) mainly had a positive moderating effect on public emotion, while the measures of traffic restrictions (SDM3) or community containment (SDM5) mainly had a negative moderating effect on public emotion. For the explanatory variable Ln(N_Post), the emotional regulation effects of SDM3 and SDM5 were positive. Different social distancing measures had differing moderating effects on the relationship between user characteristics and user emotions. Therefore, hypothesis H3 is valid.

Table 6. Statistics on the emotional tendency of explanatory variables under the moderating effect of various social measures.

Variable	The measures make the explanatory variables have a positive tendency toward emotions	The measures make the explanatory variables have a negative tendency toward emotions
Age	SDM1[a], SDM2[b], SDM3[c]	SDM4[d], SDM5[e], SDM6[f]
Space1	SDM1, SDM2	SDM3, SDM4, SDM5
Time_Reg[g]	SDM1, SDM2, SDM3, SDM6	SDM5
Ln(N_Fan)[h]	SDM2	SDM1, SDM3
Ln(N_Follow)[i]	SDM1	SDM2, SDM3
Ln(N_Post)[j]	SDM1, SDM3, SDM5	SDM2, SDM6

[a]SDM1: delaying the resumption of work and school.

[b]SDM2: travel restrictions.

[c]SDM3: traffic restrictions.

[d]SDM4: closing public spaces.

[e]SDM5: community containment.

[f]SDM6: extending the Lunar New Year holiday.

[g]Time_Reg: registration year.

[h]Ln(N_Fan): logarithm of fan numbers.

[i]Ln(N_Follow): logarithm of follow numbers.

[j]Ln(N_Post): logarithm of post numbers.

Discussion

In this study, the comments shared by Sina Weibo users related to the social distancing measures imposed by the People's Government of Hubei Province during the COVID-19 pandemic were examined. Machine learning and statistical analysis were used to reveal the emotional attitudes of users toward social distancing measures, as well as the effect of different user characteristics and social distancing measures on the users' emotions, especially the moderating effect of social distancing measures on the relationship between user characteristics and users' emotions.

Sina Weibo Users' Attention to Social Distancing Measures Varied

Sina Weibo users paid varying attention to the social distancing measures imposed by the People's Government of Hubei Province. The three measures of *traffic restrictions*, *community containment*, and *delaying resumption of work and school* attracted public attention. The main reason for this is that these measures are closely related to the transportation and daily lives of citizens in Hubei Province. When the epidemic situation became more serious, citizens paid greater attention to the daily travel of medical staff and the guaranteeing of living materials for citizens in quarantine. Although the measures of *extending the Lunar New Year holiday*, *travel restrictions*, and *closing public places* are necessary for epidemic prevention and control, they are less relevant to the daily lives of citizens living in Hubei Province. Therefore, these measures were less concerning for users.

In addition, Sina Weibo users in different spatial locations expressed differing levels of attention to social distancing measures imposed in Hubei Province. Users residing in Hubei Province are at the center of the epidemic and felt more deeply about the relevant measures. They are also the group most concerned about social distancing measures. Secondly, for Sina Weibo users in emigrant provinces, such as Guangdong and Beijing, where people in Hubei Province went after the lockdown of Wuhan, these people mainly included students and workers who went to school and work in Hubei Province and, therefore, became worried about the safety of their long-term residence. In particular, Guangdong confirmed the first case of the coronavirus on January 19, 2020, becoming the first province in Mainland China to have a confirmed case of COVID-19 outside of Hubei Province [75]. Sina Weibo users in other provinces paid less attention to Hubei Province since they reside far away from the region. Further, results of our emotional polarity analyses show that users from Hubei Province expressed

strong negative emotions, while users in emigrant and other provinces expressed stronger positive emotions, indicating that people at the center of the outbreak have strong negative emotions [61]. As for gender, women shared stronger and more positive emotions than men [71].

Sina Weibo Users' Attitudes Toward Social Distancing Measures Were Mainly Negative

The emotional polarity analysis revealed the attitudes of Sina Weibo users toward the social distancing measures imposed in Hubei Province. The emotional analysis results of the whole corpus showed that the emotions of users were mainly negative, which indicated that although social distancing measures had brought great benefits to the prevention and control of the epidemic in Hubei Province, they still inevitably affected citizens' lives in a negative way [76]. As for the social distancing measures imposed, the measure of *traffic restrictions* was promulgated at the initial stage of the epidemic, together with the notice of the Wuhan lockdown, with users residing outside of Hubei Province supporting and encouraging this measure. However, after the measure of *extending the Lunar New Year holiday* was imposed on January 27, 2020, citizens gradually realized the threat to their lives brought on by COVID-19. At the same time, a series of measures had been imposed to disturb citizens daily lives and, therefore, the negative emotions of fear and anger remained high. For example, under the measures of *delaying the resumption of work and school* and *extending the Lunar New Year holiday*, citizens began to worry about their studies, job security, and future family income [77]. After the traffic ban was imposed, people worried about travel outside of their homes, especially medical staff. For *community containment*, people were forced to stay at home, resulting in a reduction in life satisfaction, which led to an increased feeling of loneliness and psychological distress. Further, loneliness, which is likely to be exacerbated through greater fear and confusion, was seen to increase the risk of mental and physical diseases [78]. For this, more remote assistance, such as psychological counseling for families, was needed.

Sina Weibo Users With Different Social Media Characteristics Have Varying Emotional Tendencies

The results of the logistic regression analysis supported the results of our emotional polarities analysis. Women were indeed more likely to express stronger and more positive emotions. Users in Hubei Province, who were at the center of the epidemic, were more inclined to express negative emotions. For users of different ages, older citizens were inclined to send neutral comments, possibly because younger users do not have enough life experiences to distinguish right from wrong and tend to accept all information without question. This is consistent with the research of Holmes [79], who found that the COVID-19 pandemic had a significant negative impact on social groups, especially young people. In addition, users who had been using Sina Weibo for a longer period of time were more likely to share negative comments, possibly because they have a higher risk perception than those who seldom use social media [80]. Further, users with more fans and follows and those who shared a greater number of posts were inclined to express neutral or positive attitudes. For this result, we understand that users with ample follows can obtain information from multiple sources to identify relatively positive information and disseminate it. Similarly, users with a large number of fans are considered to be somewhat influential in their community [81] and hold an attitude of being responsible to their fans and, therefore, tend to post objective and positive comments to prevent fans from panicking due to excessive processing of information. Therefore, users with many fans can play a guiding role by posting objective facts or words of encouragement, so as to reduce the negative emotions experienced by fans. Social media providers should also give full attention to the role that their platform plays in the lives of citizens, providing comprehensive and accurate information to citizens [82]. They should also send appropriate notifications to prevent negative mental health.

Attention Should Be Paid to the Moderating Effect of Social Distancing Measures

The moderating effect of social distancing measures on the relationship between different user characteristics and emotions of users varied. In public health emergencies, for all six social distancing measures, the measures of *delaying the resumption of work and school* or *travel restrictions* may be met with more acceptance by citizens. *Travel restrictions* during the epidemic are an inevitable measure, especially since the epidemic occurred during the Lunar New Year holiday. At the time, many residents left Wuhan to travel to see relatives, making the spread of the disease a more serious issue [83]. Tourism is not very important in the face of personal safety, so people are more accepting of travel restrictions. Further, the measures of *traffic restrictions* or *community containment* are easily accepted by citizens. The emotion results show that these measures were associated with the most positive emotions; this is because when these measures were first promulgated, people believed their implementation could cut off routes of viral transmission and be effective for epidemic prevention and control. However, continuous traffic restrictions have brought significant inconvenience to citizens' travel plans, especially those related to employment. In response to this situation, government departments should take timely measures to resolve this problem, such as arranging specialized personnel or calling on more volunteers to provide convenience to those who need to travel, so as to alleviate the negative emotions of people. In particular, the measure of *community containment* is shown to increase health anxiety, financial worry, and loneliness [84,85], and may lead to depression and anxiety among older citizens. For this, governments must take action to strengthen interactions within local communities [86] and provide remote assistance, such as counseling for families, to reduce the psychological burden of citizens.

Implications and Limitations

This study has a good degree of theoretical value. We explored the relevant characteristics that affect users' emotions from the perspective of the Media Dependence Theory, 5W Theory, and Agenda-Setting Theory, and then analyzed the influence relationship between user characteristics and users' emotions. This study has enriched the research directions of these three theories from a new perspective and has created a certain reference value for future studies.

In addition, our results have a degree of practical significance. We found that Sina Weibo users' views and attitudes toward social distancing measures imposed by the People's Government of Hubei Province varied. Users with different characteristics also had different emotional tendencies. In particular, social distancing measures had a moderating effect on the relationship between user characteristics and users' emotions. These results are helpful for government agencies to uncover, in a timely manner, citizens' emotions pertaining to measure implementations. Our findings also provide guidelines for social media platforms to push targeted content to users.

This study has several limitations. First, due to restrictions by Sina Weibo, the secondary comment data below the posts was not obtained, which may affect the results and discussion presented. Second, the training data labels of emotion classifiers were mainly marked manually through the establishment of labeling guidelines. Further, the expression of emotion is highly subjective, and manual labeling may not reflect the real emotions of users adequately. These factors may affect the classification effect of emotional polarity classifiers to a certain extent. Finally, this study only analyzed the emotions in the text without considering emoticons, which may have a certain influence on the results of emotion classification. Future research should consider the effect of emoticons and punctuation on emotional intensity.

Conclusions

This study combined machine learning and statistical analysis to explore the emotional attitudes of citizens toward social distancing measures imposed by the People's Government of Hubei Province, as well as these emotions' influencing factors. The results of our emotional analysis show that Sina Weibo users have different attitudes toward the six types of social distancing measures implemented, but they are mainly inclined to express negative emotions. In addition, users' emotional attitudes vary across gender and spatial locations. The logistic regression analysis show that users of different ages, spatial locations, account registration year, number of fans, number of follows, and number of posts have different attitudes toward the imposed social distancing measures. Most importantly, this study found that social distancing measures have a moderating effect on the relationship between different user characteristics and users' emotions. The results obtained allow government agencies to better understand the views of citizens toward related events and can help government agencies take timely actions to alleviate negative emotions during public health emergencies.

Acknowledgments

This study was supported by the Fundamental Research Funds for the Central Universities, HUST (#2019WKYXZX011). The authors would like to thank all anonymous reviewers for their valuable comments and input on this research.

Authors' Contributions

RY, the co–first author, designed the study and contributed to data collection and the writing of the manuscript. LS, the co–first author and corresponding author, designed and conducted the study and finalized the manuscript draft. WZ contributed to the discussion and writing of the manuscript draft. RE contributed to the writing of the manuscript and final proofreading. GC contributed to the writing of the manuscript draft. ZZ contributed to the statistical analysis. All authors contributed to the preparation and approval of the final accepted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The interaction effect of the social distancing measures and explanatory variables.
[DOC File , 209 KB - medinform_v9i3e27079_app1.doc]

References

1. The voice of the war epidemic – National Health Commission talks about the new crown pneumonia epidemic-major public health emergency since the founding of New China. China Central Television. 2020 Feb 28. URL: http://tv.cctv.com/2020/02/28/VIDE0JG9u43VKoQ4dVUAQWQh200228.shtml [accessed 2021-01-02]
2. Pan SL, Cui M, Qian J. Information resource orchestration during the COVID-19 pandemic: A study of community lockdowns in China. Int J Inf Manage 2020 Oct;54:102143 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2020.102143] [Medline: 32394997]
3. Chen S, Yang J, Yang W, Wang C, Bärnighausen T. COVID-19 control in China during mass population movements at New Year. The Lancet 2020 Mar 07;395(10226):764-766 [FREE Full text] [doi: 10.1016/S0140-6736(20)30421-9] [Medline: 32105609]
4. Fong MW, Gao H, Wong JY, Xiao J, Shiu EYC, Ryu S, et al. Nonpharmaceutical Measures for Pandemic Influenza in Nonhealthcare Settings-Social Distancing Measures. Emerg Infect Dis 2020 May;26(5):976-984 [FREE Full text] [doi: 10.3201/eid2605.190995] [Medline: 32027585]

5. Wuhan novel coronavirus infection prevention and control command announcement (No. 1). China Government Network. 2020 Jan 23. URL: http://www.gov.cn/xinwen/2020-01/23/content_5471751.htm [accessed 2021-01-02]
6. Notice of novel coronavirus pneumonia epidemic prevention and control headquarters in Hubei Province. People's Government of Hubei Province. 2020. URL: https://baijiahao.baidu.com/s?id=1658419579329417941&wfr=spider&for=pc [accessed 2021-01-02]
7. Notice of the General Office of the Hubei Provincial People's Government on extending the Spring Festival Holiday in 2020. People's Government of Hubei Province. 2020. URL: https://baijiahao.baidu.com/s?id=1657340867892589871&wfr=spider&for=pc [accessed 2021-01-02]
8. Notice on doing a good job in responding to pneumonia caused by new coronavirus infection. Hubei Department of Culture and Tourism. 2020 Jan 30. URL: http://wlt.hubei.gov.cn/bmdt/ztzl/xxgzbd/202001/t20200130_2016384.shtml [accessed 2021-01-02]
9. Notice by the People's Government of Hubei Province on further strengthening COVID-19 prevention and control. People's Government of Hubei Province. 2020. URL: https://baijiahao.baidu.com/s?id=1658701503675498305&wfr=spider&for=pc [accessed 2021-01-02]
10. Caley P, Philp DJ, McCracken K. Quantifying social distancing arising from pandemic influenza. J R Soc Interface 2008 Jun 6;5(23):631-639 [FREE Full text] [doi: 10.1098/rsif.2007.1197] [Medline: 17916550]
11. Social distancing: What is social distancing? Centers for Disease Control and Prevention. 2020 Nov 17. URL: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html [accessed 2021-01-02]
12. Global digital population as of October 2020. Statista. 2020 Oct. URL: https://www.statista.com/statistics/617136/digital-population-worldwide/ [accessed 2021-01-02]
13. The 46th China Statistical Report on Internet Development. China Internet Network Information Center. 2020 Sep. URL: http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/202009/P020200929546215182514.pdf [accessed 2021-01-02]
14. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 2010 Nov 29;5(11):e14118 [FREE Full text] [doi: 10.1371/journal.pone.0014118] [Medline: 21124761]
15. Sina Weibo. URL: http://weibo.com/ [accessed 2021-01-02]
16. Chen Z, Su K. Research on the media role of Weibo in public health emergencies—take the 'novel coronavirus pneumonia' as an example. Today's Mass Media 2020 May 5;28(5):12-15 [FREE Full text]
17. Giunti G, Claes M, Dorronzoro Zubiete E, Rivera-Romero O, Gabarron E. Analysing Sentiment and Topics Related to Multiple Sclerosis on Twitter. Stud Health Technol Inform 2020 Jun 16;270:911-915. [doi: 10.3233/SHTI200294] [Medline: 32570514]
18. Bai H, Yu G. A Weibo-based approach to disaster informatics: incidents monitor in post-disaster situation via Weibo text negative sentiment analysis. Nat Hazards 2016 May 30;83(2):1177-1196. [doi: 10.1007/s11069-016-2370-5]
19. Hatchett RJ, Mecher CE, Lipsitch M. Public health interventions and epidemic intensity during the 1918 influenza pandemic. Proc Natl Acad Sci U S A 2007 May 01;104(18):7582-7587 [FREE Full text] [doi: 10.1073/pnas.0610941104] [Medline: 17416679]
20. Flu pandemic study supports social distancing. National Institutes of Health. 2011 Jun 6. URL: https://www.nih.gov/news-events/nih-research-matters/flu-pandemic-study-supports-social-distancing [accessed 2021-01-02]
21. Bondy SJ, Russell ML, Laflèche JM, Rea E. Quantifying the impact of community quarantine on SARS transmission in Ontario: estimation of secondary case count difference and number needed to quarantine. BMC Public Health 2009 Dec 24;9:488 [FREE Full text] [doi: 10.1186/1471-2458-9-488] [Medline: 20034405]
22. Zhang J, Litvinova M, Liang Y, Wang Y, Wang W, Zhao S, et al. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. Science 2020 Jun 26;368(6498):1481-1486 [FREE Full text] [doi: 10.1126/science.abb8001] [Medline: 32350060]
23. Brown ST, Tai JHY, Bailey RR, Cooley PC, Wheaton WD, Potter MA, et al. Would school closure for the 2009 H1N1 influenza epidemic have been worth the cost?: a computational simulation of Pennsylvania. BMC Public Health 2011 May 20;11:353 [FREE Full text] [doi: 10.1186/1471-2458-11-353] [Medline: 21599920]
24. Milne GJ, Kelso JK, Kelly HA, Huband ST, McVernon J. A small community model for the transmission of infectious diseases: comparison of school closure as an intervention in individual-based models of an influenza pandemic. PLoS One 2008;3(12):e4005 [FREE Full text] [doi: 10.1371/journal.pone.0004005] [Medline: 19104659]
25. Chin TDY, Foley JF, Doto IL, Gravelle CR, Weston J. Morbidity and mortality characteristics of Asian strain influenza. Public Health Rep 1960 Feb;75(2):149-158 [FREE Full text] [Medline: 19316351]
26. Teh B, Olsen K, Black J, Cheng AC, Aboltins C, Bull K, et al. Impact of swine influenza and quarantine measures on patients and households during the H1N1/09 pandemic. Scand J Infect Dis 2012 Apr;44(4):289-296. [doi: 10.3109/00365548.2011.631572] [Medline: 22106922]
27. Tan C. SARS in Singapore--key lessons from an epidemic. Ann Acad Med Singap 2006 May;35(5):345-349 [FREE Full text] [Medline: 16830002]
28. Hoffmann RK, Hoffmann K. Ethical considerations in the use of cordons sanitaires. Clinical Correlations. 2015 Feb 19. URL: https://www.clinicalcorrelations.org/2015/02/19/ethical-considerations-in-the-use-of-cordons-sanitaires/ [accessed 2021-01-02]

29. Ahmed F, Zviedrite N, Uzicanin A. Effectiveness of workplace social distancing measures in reducing influenza transmission: a systematic review. BMC Public Health 2018 Apr 18;18(1):518 [FREE Full text] [doi: 10.1186/s12889-018-5446-1] [Medline: 29669545]
30. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science 2020 Apr 24;368(6489):395-400 [FREE Full text] [doi: 10.1126/science.aba9757] [Medline: 32144116]
31. Islam N, Sharp SJ, Chowell G, Shabnam S, Kawachi I, Lacey B, et al. Physical distancing interventions and incidence of coronavirus disease 2019: natural experiment in 149 countries. BMJ 2020 Jul 15;370:m2743 [FREE Full text] [doi: 10.1136/bmj.m2743] [Medline: 32669358]
32. Ao B. Social distancing can strain mental health. The Philadelphia Inquirer. 2020 Mar 19. URL: https://www.inquirer.com/health/coronavirus/coronavirus-mental-health-social-distancing-20200319.html [accessed 2021-01-02]
33. Cao W, Fang Z, Hou G, Han M, Xu X, Dong J, et al. The psychological impact of the COVID-19 epidemic on college students in China. Psychiatry Res 2020 May;287:112934 [FREE Full text] [doi: 10.1016/j.psychres.2020.112934] [Medline: 32229390]
34. Chen HY, Wang JP, Xie W, Chen W. A study of Internet user's mood and its relevant variables under the crisis of SARS. Chinese Journal of Applied Psychology 2003 Dec 30(4):7-13. [doi: 10.3969/j.issn.1006-6020.2003.04.002]
35. Qian M, Ye D, Dong W, Huang Z, Zhang L, Liu X. Behaviour, cognition and emotion of the public in Beijing towards SARS. Chinese Mental Health Journal 2003 Aug 15(8):515-520. [doi: 10.3321/j.issn:1000-6729.2003.08.001]
36. Ogoina D, Oyeyemi AS, Ayah O, Onabor A, Midia A, Olomo WT, et al. Preparation and Response to the 2014 Ebola Virus Disease Epidemic in Nigeria-The Experience of a Tertiary Hospital in Nigeria. PLoS One 2016 Oct 27;11(10):e0165271 [FREE Full text] [doi: 10.1371/journal.pone.0165271] [Medline: 27788191]
37. Khalid I, Khalid TJ, Qabajah MR, Barnard AG, Qushmaq IA. Healthcare Workers Emotions, Perceived Stressors and Coping Strategies During a MERS-CoV Outbreak. Clin Med Res 2016 Mar;14(1):7-14 [FREE Full text] [doi: 10.3121/cmr.2016.1303] [Medline: 26847480]
38. Jiao WY, Wang LN, Liu J, Fang SF, Jiao FY, Pettoello-Mantovani M, et al. Behavioral and Emotional Disorders in Children during the COVID-19 Epidemic. J Pediatr 2020 Jun;221:264-266.e1 [FREE Full text] [doi: 10.1016/j.jpeds.2020.03.013] [Medline: 32248989]
39. Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. 2010 Presented at: The 2010 International Conference on Language Resources and Evaluation; May 17-23; Valletta, Malta URL: https://www.aclweb.org/anthology/L10-1531/
40. Mohammad S, Turney P. Crowdsourcing a word-emotion association lexicon. Comput Intell 2012;29(3):436-465. [doi: 10.1111/j.1467-8640.2012.00460.x]
41. Das S, Dutta A. Characterizing public emotions and sentiments in COVID-19 environment: A case study of India. J Hum Behav Soc Environ 2020 Jul 14:1-14. [doi: 10.1080/10911359.2020.1781015]
42. Du J, Tang L, Xiang Y, Zhi D, Xu J, Song H, et al. Public Perception Analysis of Tweets During the 2015 Measles Outbreak: Comparative Study Using Convolutional Neural Network Models. J Med Internet Res 2018 Jul 09;20(7):e236 [FREE Full text] [doi: 10.2196/jmir.9413] [Medline: 29986843]
43. Ji X, Chun S, Geller J. Monitoring public health concerns using Twitter sentiment classifications. 2013 Presented at: The 2013 IEEE International Conference on Healthcare Informatics; Sept 9-11; Philadelphia, PA p. 335-344 URL: https://ieeexplore.ieee.org/abstract/document/6680494 [doi: 10.1109/ichi.2013.47]
44. Pondora Naresh Behera, Suneetha Eluri. Analysis of Public Health Concerns using Two-step Sentiment Classification. IJERT 2015 Sep 24;V4(09):606-610. [doi: 10.17577/ijertv4is090641]
45. Notice of the People's Government of Hubei Province on strengthening the prevention and control work of novel Coronavirus infection. People's Government of Hubei Province. 2020 Jan 21. URL: http://www.hubei.gov.cn/zfwj/ezf/202002/t20200220_2142431.shtml [accessed 2021-01-02]
46. Hu CL, Tang JT, Wang T. Topical relevance analysis of hashtags in Chinese microblogging environment. Computer Science 2013 Nov 15;40(S2):235-237. [doi: 10.3969/j.issn.1002-137X.2013.z2.058]
47. Pang B, Lee L. Opinion Mining and Sentiment Analysis. Ithaca, NY: Foundations and Trends in Information Retrieval; 2008.
48. Mukhtar N, Khan MA, Chiragh N. Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains. Telematics and Informatics 2018 Dec;35(8):2173-2183. [doi: 10.1016/j.tele.2018.08.003]
49. Shrestha H, Dhasarathan C, Munisamy S, Jayavel A. Natural Language Processing Based Sentimental Analysis of Hindi (SAH) Script an Optimization Approach. Int J Speech Technol 2020 Jul 09;23(4):757-766. [doi: 10.1007/s10772-020-09730-x]
50. Kumar A, Srinivasan K, Cheng W, Zomaya AY. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. Information Processing & Management 2020 Jan;57(1):102141. [doi: 10.1016/j.ipm.2019.102141]
51. Ortony A, Clore G, Collins A. The Cognitive Structure of Emotions. New York, NY: Cambridge University Press; 1988.

52. Wu P, Li X, Shen S, He D. Social media opinion summarization using emotion cognition and convolutional neural networks. Int J Inf Manage 2020 Apr;51:101978. [doi: 10.1016/j.ijinfomgt.2019.07.004]
53. Xia F, Yetisgen-Yildiz M. Clinical corpus annotation: challenges and strategies. 2012 Presented at: The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining; May 26; Istanbul, Turkey URL: http://www.nactem.ac.uk/biotxtm2012/presentations/Yetisgen-Yildiz-pres.pdf
54. Carletta J. Assessing agreement on classification tasks: the kappa statistic. Comput Linguist 1996 Jun;22:249-254 [FREE Full text]
55. Rigby AS. Statistical methods in epidemiology. v. Towards an understanding of the kappa coefficient. Disabil Rehabil 2000 May 20;22(8):339-344. [doi: 10.1080/096382800296575] [Medline: 10896093]
56. Hu J. Automated Detection of Driver Fatigue Based on AdaBoost Classifier with EEG Signals. Front Comput Neurosci 2017 Aug 3;11:72 [FREE Full text] [doi: 10.3389/fncom.2017.00072] [Medline: 28824409]
57. Bi Q, Shen L, Evans R, Zhang Z, Wang S, Dai W, et al. Determining the Topic Evolution and Sentiment Polarity for Albinism in a Chinese Online Health Community: Machine Learning and Social Network Analysis. JMIR Med Inform 2020 May 29;8(5):e17813 [FREE Full text] [doi: 10.2196/17813] [Medline: 32469320]
58. Ball-Rokeach SJ. The origins of individual media system dependency: a sociological framework. Communication Research 2016 Jun 30;12(4):485-510. [doi: 10.1177/009365085012004003]
59. Wu B. Research on Microblog Emotional Expression Based on Media System Dependence Perspective—Taking “Hangzhou Nanny Arson” as an Example. ASS 2018;07(10):1693-1701. [doi: 10.12677/ass.2018.710253]
60. de Souza LC, Bertoux M, de Faria Â, Corgosinho LTS, Prado ACDA, Barbosa IG, et al. The effects of gender, age, schooling, and cultural background on the identification of facial emotions: a transcultural study. Int Psychogeriatr 2018 Dec;30(12):1861-1870. [doi: 10.1017/S1041610218000443] [Medline: 29798733]
61. van Lent LG, Sungur H, Kunneman FA, van de Velde B, Das E. Too Far to Care? Measuring Public Attention and Fear for Ebola Using Twitter. J Med Internet Res 2017 Dec 13;19(6):e193 [FREE Full text] [doi: 10.2196/jmir.7219] [Medline: 28611015]
62. Griffin RJ, Dunwoody S, Neuwirth K. Proposed model of the relationship of risk information seeking and processing to the development of preventive behaviors. Environ Res 1999 Feb;80(2 Pt 2):S230-S245. [doi: 10.1006/enrs.1998.3940] [Medline: 10092438]
63. Yang ZJ, Aloe AM, Feeley TH. Risk Information Seeking and Processing Model: A Meta-Analysis. J Commun 2014 Jan 07;64(1):20-41. [doi: 10.1111/jcom.12071]
64. Lasswell HD. The structure and function of communication in society. In: Bryson L, editor. The Communication of Ideas. New York, NY: Harper and Row; 1948:37-51.
65. Liao HH, Wang YF. Public opinion dissemination over social media: case study of Sina Weibo and 8/12 Tianjin explosion. Data Analysis and Knowledge Discovery 2016 Dec 25:12-93 [FREE Full text]
66. Chen J, Liu YP, Deng SL. Research on user reviews of government rumor-refuting information and factors influencing their emotional tendencies. Information Science 2017 Dec 5;35(12):61-65. [doi: 10.13833/j.issn.1007-7634.2017.12.011]
67. Lin C. Analysis on evaluation indexes of information dissemination impact of micro-blog Individual. Library and Information Service 2014 Feb 25;58(1):40-43. [doi: 10.13266/j.issn.0252-3116.2014.01.006]
68. McCombs M, Shaw D. The agenda-setting function of mass media. Public Opinion Quarterly 1972 Jan 1;36(2):176-187. [doi: 10.1086/267990]
69. McCombs ME, Shaw DL, Weaver DH. New Directions in Agenda-Setting Theory and Research. Mass Communication and Society 2014 Nov 24;17(6):781-802. [doi: 10.1080/15205436.2014.964871]
70. Vargo CJ, Guo L, McCombs M, Shaw DL. Network Issue Agendas on Twitter During the 2012 U.S. Presidential Election. J Commun 2014 Mar 24;64(2):296-316. [doi: 10.1111/jcom.12089]
71. Kucuktunc O, Cambazoglu B, Weber I, Ferhatosmanoglu H. A large-scale sentiment analysis for Yahoo! Answers. 2012 Feb Presented at: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining; Feb 8-12; Seattle, WA p. 633-642. [doi: 10.1145/2124295.2124371]
72. Popular places to relocate during the Spring Festival travel season (destination). Baidu Migration. URL: http://qianxi.baidu.com/ [accessed 2021-01-02]
73. Chen Z, Zhang Q, Lu Y, Guo Z, Zhang X, Zhang W, et al. Distribution of the COVID-19 epidemic and correlation with population emigration from Wuhan, China. Chin Med J (Engl) 2020 May 05;133(9):1044-1050 [FREE Full text] [doi: 10.1097/CM9.0000000000000782] [Medline: 32118644]
74. Jaccard J. Interaction effects in logistic regression, 1st edition. London, UK: Sage Publications, Inc; 2001.
75. Jiang Y. The National Health Commission confirmed the first confirmed case of novel coronavirus infection in Guangdong province. China News. 2020 Jan 20. URL: http://www.chinanews.com/sh/2020/01-20/9064733.shtml [accessed 2021-01-02]
76. Thu TPB, Ngoc PNH, Hai NM, Tuan LA. Effect of the social distancing measures on the spread of COVID-19 in 10 highly infected countries. Sci Total Environ 2020 Nov 10;742:140430 [FREE Full text] [doi: 10.1016/j.scitotenv.2020.140430] [Medline: 32623158]
77. Baum NM, Jacobson PD, Goold SD. "Listen to the people": public deliberation about social distancing measures in a pandemic. Am J Bioeth 2009 Nov;9(11):4-14. [doi: 10.1080/15265160903197531] [Medline: 19882444]

78. Beutel ME, Klein EM, Brähler E, Reiner I, Jünger C, Michal M, et al. Loneliness in the general population: prevalence, determinants and relations to mental health. BMC Psychiatry 2017 Mar 20;17(1):97 [FREE Full text] [doi: 10.1186/s12888-017-1262-x] [Medline: 28320380]
79. Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, Arseneault L, et al. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. The Lancet Psychiatry 2020 Jun 15;7(6):547-560 [FREE Full text] [doi: 10.1016/S2215-0366(20)30168-1] [Medline: 32304649]
80. Ju Y, You M. The Outrage Effect of Personal Stake, Dread, and Moral Nature on Fine Dust Risk Perception Moderated by Media Use. Health Commun 2020 Feb 05:1-11. [doi: 10.1080/10410236.2020.1723046] [Medline: 32024391]
81. Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media? 2010 Presented at: Proceedings of the 19th International World Wide Web Conference; Apr 26-30; Raleigh, NC p. 591-600. [doi: 10.1145/1772690.1772751]
82. Sadah SA, Shahbazi M, Wiley MT, Hristidis V. Demographic-Based Content Analysis of Web-Based Health-Related Social Media. J Med Internet Res 2016 Jun 13;18(6):e148 [FREE Full text] [doi: 10.2196/jmir.5327] [Medline: 27296242]
83. Kraemer MUG, Yang C, Gutierrez B, Wu C, Klein B, Pigott DM, Open COVID-19 Data Working Group, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science 2020 May 01;368(6490):493-497 [FREE Full text] [doi: 10.1126/science.abb4218] [Medline: 32213647]
84. González-Sanguino C, Ausín B, Castellanos MA, Saiz J, López-Gómez A, Ugidos C, et al. Mental health consequences during the initial stage of the 2020 Coronavirus pandemic (COVID-19) in Spain. Brain Behav Immun 2020 Jul;87:172-176 [FREE Full text] [doi: 10.1016/j.bbi.2020.05.040] [Medline: 32405150]
85. Tull MT, Edmonds KA, Scamaldo KM, Richmond JR, Rose JP, Gratz KL. Psychological Outcomes Associated with Stay-at-Home Orders and the Perceived Impact of COVID-19 on Daily Life. Psychiatry Res 2020 Jul 12;289:113098 [FREE Full text] [doi: 10.1016/j.psychres.2020.113098] [Medline: 32434092]
86. Block P, Hoffman M, Raabe IJ, Dowd JB, Rahal C, Kashyap R, et al. Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world. Nat Hum Behav 2020 Jun;4(6):588-596. [doi: 10.1038/s41562-020-0898-6] [Medline: 32499576]

Abbreviations

Bi-LSTM: bi-directional long short-term
CNN: convolutional neural network
LSTM: long short-term memory
NB: naive Bayes
NRC: National Research Council of Canada
OCC: Ortony-Clore-Collins model
RQ: research question
SARS: severe acute respiratory syndrome
SVM: support vector machine

Edited by C Lovis; submitted 10.01.21; peer-reviewed by C González-Sanguino, D Huang; comments to author 31.01.21; revised version received 19.02.21; accepted 27.02.21; published 16.03.21.

Please cite as:
Shen L, Yao R, Zhang W, Evans R, Cao G, Zhang Z
Emotional Attitudes of Chinese Citizens on Social Distancing During the COVID-19 Outbreak: Analysis of Social Media Data
JMIR Med Inform 2021;9(3):e27079
URL: https://medinform.jmir.org/2021/3/e27079
doi:10.2196/27079
PMID:33724200

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org