Contents

Reviews

Using Machine Learning Technologies in Pressure Injury Management: Systematic Review (e25704)
Mengyao Jiang, Yuxia Ma, Siyi Guo, Liuqi Jin, Lin Lv, Lin Han, Ning An. 4

Physicians’ Use of the Computerized Physician Order Entry System for Medication Prescribing: Systematic Review (e22923)
Asra Mogharbel, Dawn Dowding, John Ainsworth. 150

Original Papers

Electronic Health Record Use in Swiss Nursing Homes and Its Association With Implicit Rationing of Nursing Care Documentation: Multicenter Cross-sectional Survey Study (e22974)
Dietmar Ausserhofer, Lauriane Favez, Michael Simon, Franziska Zühlga. 14

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 (e23951)
Shuma Kanfe, Berhanu Endehabtu, Mohammedjud Ahmed, Nebyu Mengestie, Binyam Tilahun. 28

The Effect of Innovation Capabilities of Health Care Organizations on the Quality of Health Information Technology: Model Development With Cross-sectional Data (e23306)
Moritz Esdar, Ursula Hübner, Johannes Thye, Birgit Babitsch, Jan-David Liebe. 38

A Chatbot for Perinatal Women’s and Partners’ Obstetric and Mental Health Care: Development and Usability Evaluation Study (e18607)
Kyungmi Chung, Hee Cho, Jin Park. 55

Natural Language Processing of Clinical Notes to Identify Mental Illness and Substance Use Among People Living with HIV: Retrospective Cohort Study (e23456)
Jessica Ridgway, Arno Uvin, Jessica Schmitt, Tomasz Oliwa, Ellen Almirol, Samantha Devlin, John Schneider. 72

Hybrid Deep Learning for Medication-Related Information Extraction From Clinical Texts in French: MedExt Algorithm Development Study (e17934)
Jordan Jouffroy, Sarah Feldman, Ivan Lerner, Bastien Rance, Anita Burgun, Antoine Neuraz. 82
Human–Computer Agreement of Electrocardiogram Interpretation for Patients Referred to and Declined for Primary Percutaneous Coronary Intervention: Retrospective Data Analysis Study (e24188)  
Aleeha Iftikhar, Raymond Bond, Victoria McGillian, Stephen Leslie, Charles Knor, James Shand, Adesh Ramsewak, Divyesh Sharma, Anne McShane, Khaled Rjoob, Aaron Peace ................................................................. 93

A Personal Health System for Self-Management of Congestive Heart Failure (HeartMan): Development, Technical Evaluation, and Proof-of-Concept Randomized Controlled Trial (e24501)  
Mitja Luštrek, Marko Bohanec, Carlos Caverio Barca, María Ciancarelli, Els Clays, Amos Dawodu, Jan Derboven, Delphine De Smedt, Erik Dowgan, Jure Lampe, Flavia Marino, Miha Miškar, Giovanni Pioggia, Paolo Puddu, Juan Rodriguez, Michele Schiari, Gašper Slapničar, Karin Slegers, Gennaro Tartarisco, Jakob Vali, Aljoša Vodopija ...................................................... 104

A Clinical Decision Support System (KNOWBED) to Integrate Scientific Knowledge at the Bedside: Development and Evaluation Study (e13182)  
Alicia Martinez-Garcia, Ana Naranjo-Saucedo, Jose Rivas, Antonio Romero Tabares, Ana Marián Cassinello, Anselmo Andrés-Martín, Francisco Sánchez Laguna, Roman Villegas, Francisco Pérez León, Jesús Moreno Conde, Carlos Parra Calderón ............................................................... 123

Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study (e21331)  
Alberto Tena, Francisc Claria, Francesc Solsona, Einar Meister, Monica Povedano ................................................................. 132

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 (e18269)  
Jinfei Ma, Zihao Zou, Emmanuel Pazo, Salissou Moutari, Ye Liu, Feng Jin .................................................................................. 162

Antibiotic Prescription Rates After eVisits Versus Office Visits in Primary Care: Observational Study (e25473)  
Artin Entezarjou, Susanna Calling, Tapomita Bhattacharyya, Veronica Milos Nymberg, Lina Vigren, Ashkan Labaf, Ulf Jakobsson, Patrik Midlöv. 1 7 5

Realistic High-Resolution Body Computed Tomography Image Synthesis by Using Progressive Growing Generative Adversarial Network: Visual Turing Test (e23328)  
Ho Park, Hyun-Jin Bae, Gil-Sun Hong, Minjee Kim, JiHye Yun, Sungwon Park, Won Chung, NamKug Kim .............................................. 188

Machine Learning Approach to Predict the Probability of Recurrence of Renal Cell Carcinoma After Surgery: Prediction Model Development Study (e25635)  
HyungMin Kim, Sun Lee, So Park, In Choi, Sung-Hoo Hong ........................................................................................................... 201

Predictive Modeling of 30-Day Emergency Hospital Transport of German Patients Using a Personal Emergency Response: Retrospective Study and Comparison with the United States (e25121)  
Jorn op den Buijs, Marten Pijl, Andreas Landgraf ................................................................. 213

A Novel Convolutional Neural Network for the Diagnosis and Classification of Rosacea: Usability Study (e23415)  
Zhixiang Zhao, Che-Ming Wu, Shuping Zhang, Fanping He, Fangfen Liu, Ben Wang, Yingxue Huang, Wei Shi, Dan Jian, Hongfu Xie, Chao-Yuan Yeh, Ji Li .................................................................................. 225

Regional Resource Assessment During the COVID-19 Pandemic in Italy: Modeling Study (e18933)  
Pietro Guzzi, Giuseppe Tradigo, Pierangelo Velti ................................................................. 235

Medical Morphology Training Using the Xuexi Tong Platform During the COVID-19 Pandemic: Development and Validation of a Web-Based Teaching Approach (e24497)  
Qinlai Liu, Wenyu Sun, Changqing Du, Leiyin Yang, Na Yuan, Haqing Cui, Wengang Song, Li Ge ................................................................. 244
Systematic Delineation of Media Polarity on COVID-19 Vaccines in Africa: Computational Linguistic Modeling Study (e22916)
Sefater Gbashi, Oluwafemi Adebo, Wesley Doersamy, Patrick Njobeh

Emotional Attitudes of Chinese Citizens on Social Distancing During the COVID-19 Outbreak: Analysis of Social Media Data (e27079)
Lining Shen, Rui Yao, Wenli Zhang, Richard Evans, Guang Cao, Zhiguo Zhang
Using Machine Learning Technologies in Pressure Injury Management: Systematic Review

Mengyao Jiang1*, MSN; Yuxia Ma1*, MSN; Siyi Guo2, BSc; Liuqi Jin2, BSc; Lin Lv3, MSN; Lin Han1,4, PhD; Ning An2, PhD

1Evidence-based Nursing Center, School of Nursing, Lanzhou University, Lanzhou, China
2Key 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
3Wound and Ostomy Center, Outpatient Department, Gansu Provincial Hospital, Lanzhou, China
4Department 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, 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 A5essment 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 PIIs are avoidable [9]. Nurses are primarily responsible for preventing PIIs [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].

**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].

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.
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 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 (195YPYA-4), the Fundamental Research Funds for the Central Universities (Izujbky-2018-c05 and Izujbky-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).
References


**Abbreviations**

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

https://medinform.jmir.org/2021/3/e25704
Electronic Health Record Use in Swiss Nursing Homes and Its Association With Implicit Rationing of Nursing Care Documentation: Multicenter Cross-sectional Survey Study

Dietmar Ausserhofer1,2, PhD; Lauriane Favez2, MA; Michael Simon2,3, PhD; Franziska Zúñiga2, PhD

1College of Health Care-Professions Claudiana, Bolzano-Bozen, Italy
2Nursing Science, Department of Public Health, University of Basel, Basel, Switzerland
3Nursing 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
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.
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.

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 rpiR 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.

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

*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).

<table>
<thead>
<tr>
<th></th>
<th>Total (N=1975), n (%a)</th>
<th>Registered nurses (n=966)</th>
<th>Licensed practical nurses (n=1058)</th>
<th>P valueb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. The electronic health record system allows timely communication between the nursing and therapy teams</td>
<td>1367 (69.96)</td>
<td>667 (71.11)</td>
<td>700 (68.90)</td>
<td>.29</td>
</tr>
<tr>
<td>2. The electronic health record system provides a good overview on the main focus of care and treatment for the individual residents</td>
<td>1507 (76.89)</td>
<td>710 (75.53)</td>
<td>797 (78.14)</td>
<td>.17</td>
</tr>
<tr>
<td>3. The electronic health record system gives a good daily overview on all residents on the care unit</td>
<td>1429 (72.98)</td>
<td>664 (70.78)</td>
<td>765 (75.00)</td>
<td>.04</td>
</tr>
<tr>
<td>4. The electronic health record system guarantees safe care and treatment</td>
<td>1362 (69.42)</td>
<td>645 (68.54)</td>
<td>717 (70.23)</td>
<td>.42</td>
</tr>
<tr>
<td>5. The electronic health record system allows quick access to relevant information on the residents</td>
<td>1535 (78.32)</td>
<td>720 (76.51)</td>
<td>815 (79.98)</td>
<td>.06</td>
</tr>
<tr>
<td>6. On our unit there are sufficient computers to allow timely documentation</td>
<td>914 (46.61)</td>
<td>451 (47.98)</td>
<td>463 (45.35)</td>
<td>.24</td>
</tr>
</tbody>
</table>

*aPercentage agreement (agree and strongly agree).

*bChi-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).

<table>
<thead>
<tr>
<th>Activity not necessary, n (%)</th>
<th>Ne, n (%)</th>
<th>S, n (%)</th>
<th>Som, n (%)</th>
<th>O, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studying care plans at the beginning of the shift</td>
<td>13 (0.67)</td>
<td>478 (24.72)</td>
<td>553 (28.59)</td>
<td>480 (24.82)</td>
</tr>
<tr>
<td>Set up or update residents’ care plans</td>
<td>110 (6.83)</td>
<td>270 (16.77)</td>
<td>424 (26.34)</td>
<td>481 (29.88)</td>
</tr>
<tr>
<td>Documentation of care</td>
<td>4 (0.21)</td>
<td>429 (22.26)</td>
<td>654 (33.94)</td>
<td>561 (29.11)</td>
</tr>
<tr>
<td>Variables</td>
<td>Unit level, ICC1 (95% CI)</td>
<td>Facility level, ICC1 (95% CI)</td>
<td>Mean (SD)</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>---------------------------</td>
<td>-------------------------------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>Rationing of nursing care documentation</td>
<td>0.155 (0.111-0.202)</td>
<td>0.118 (0.076-0.165)</td>
<td>2.38 (0.9)</td>
<td></td>
</tr>
<tr>
<td>Care workers’ perception of the electronic health record system’s usefulness</td>
<td>0.196 (0.152-0.244)</td>
<td>0.111 (0.068-0.156)</td>
<td>3.86 (0.77)</td>
<td></td>
</tr>
<tr>
<td>Work environment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Care workers’ perception of sufficient number of computers</td>
<td>0.221 (0.176-0.269)</td>
<td>0.116 (0.072-0.161)</td>
<td>3.13 (1.33)</td>
<td></td>
</tr>
<tr>
<td>Leadership</td>
<td>0.278 (0.228-0.326)</td>
<td>0.156 (0.104-0.205)</td>
<td>3.18 (0.62)</td>
<td></td>
</tr>
<tr>
<td>Staffing and resources adequacy</td>
<td>0.254 (0.207-0.302)</td>
<td>0.214 (0.151-0.271)</td>
<td>2.67 (0.67)</td>
<td></td>
</tr>
<tr>
<td>Teamwork and safety climate</td>
<td>0.196 (0.152-0.244)</td>
<td>0.111 (0.068-0.156)</td>
<td>3.89 (0.81)</td>
<td></td>
</tr>
</tbody>
</table>

aICC1: 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.

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

<sup>a</sup>Random effect: Facility-level variance (SD)=0.07 (0.27), Unit-level variance (SD)=0.06 (0.25).

<sup>b</sup>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 −.58 to −.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 −.15 to −.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.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Multiple model 1(^a) (without staffing and resources adequacy)</th>
<th>Multiple model 2(^b) (with staffing and resources adequacy)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β (95% CI)</td>
<td>SE</td>
</tr>
<tr>
<td><strong>Explanatory variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Care workers’ perception of the EHR(^b) system’s usefulness</td>
<td>(-.14^{c} (-0.20 to -0.09))</td>
<td>0.03</td>
</tr>
<tr>
<td>Care workers’ perception of whether sufficient numbers of computers were available on their units</td>
<td>(-.12^{c} (-0.15 to -0.09))</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Control variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Facility characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language region</td>
<td>(\ldots)</td>
<td></td>
</tr>
<tr>
<td>Nursing home size</td>
<td>(\ldots)</td>
<td></td>
</tr>
<tr>
<td>Profit status</td>
<td>(\ldots)</td>
<td></td>
</tr>
<tr>
<td>EHR system</td>
<td>(\ldots)</td>
<td></td>
</tr>
<tr>
<td><strong>Unit characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staffing levels</td>
<td>(\ldots)</td>
<td></td>
</tr>
<tr>
<td>Skill mix levels</td>
<td>(\ldots)</td>
<td></td>
</tr>
<tr>
<td><strong>Work environment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership</td>
<td>(-.12^{c} (-0.21 to -0.04))</td>
<td>0.04</td>
</tr>
<tr>
<td>Staffing and resources adequacy</td>
<td>(\ldots)</td>
<td></td>
</tr>
<tr>
<td>Safety and teamwork climate</td>
<td>(-.20^{c} (-0.27 to -0.12))</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>Care workers’ characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>(\ldots)</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>(\ldots)</td>
<td></td>
</tr>
<tr>
<td>Educational background</td>
<td>(-.08^{c} (-0.16 to -0.02))</td>
<td>0.04</td>
</tr>
<tr>
<td>Professional experience</td>
<td>(\ldots)</td>
<td></td>
</tr>
<tr>
<td>Employment level</td>
<td>(\ldots)</td>
<td></td>
</tr>
<tr>
<td>Fixed effects (intercept)</td>
<td>(4.67^{c} (4.39 to 4.94))</td>
<td>0.14</td>
</tr>
</tbody>
</table>

\(^a\)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


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

©Dietmar Ausserhofer, Lauriane Favez, Michael Simon, Franziska Zúñiga. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 02.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction
in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.
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 Kanfe¹, MPH; Berhanu F Endehabtu², MPH; Mohammedjud H Ahmed¹, MPH; Nebyu D Mengestie², MPH; Binyam Tilahun², MSc, MPH, PhD

¹Health Informatics, Mettu University, Metu Zuria, Ethiopia
²Department 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).

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

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.

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

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 2020a.

<table>
<thead>
<tr>
<th>Variable, category</th>
<th>Commitment level</th>
<th>Crude odds ratio</th>
<th>Adjusted odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High commitment, n (%)</td>
<td>Low commitment, n (%)</td>
<td></td>
</tr>
<tr>
<td>Culture of information use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good (n=164)</td>
<td>83 (50.6)</td>
<td>81 (49.4)</td>
<td>1.67 (1.00-2.77)*</td>
</tr>
<tr>
<td>Poor (n=100)</td>
<td>38 (38.0)</td>
<td>62 (62.7)</td>
<td>1^b</td>
</tr>
<tr>
<td>Health needs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes (n=131)</td>
<td>76 (58.0)</td>
<td>55 (42.0)</td>
<td>2.70 (1.64-4.45)</td>
</tr>
<tr>
<td>No (n=133)</td>
<td>45 (33.8)</td>
<td>88 (66.2)</td>
<td>1</td>
</tr>
<tr>
<td>Motivation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High motivation (n=136)</td>
<td>71 (52.2)</td>
<td>65 (47.8)</td>
<td>1.70 (1.04-2.77) *</td>
</tr>
<tr>
<td>Poor motivation (n=128)</td>
<td>50 (39.1)</td>
<td>78 (60.9)</td>
<td>1</td>
</tr>
<tr>
<td>Feedback</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes (n=140)</td>
<td>71 (50.7)</td>
<td>69 (49.3)</td>
<td>1.52 (0.93-2.48)</td>
</tr>
<tr>
<td>No (n=124)</td>
<td>50 (40.3)</td>
<td>74 (59.7)</td>
<td>1</td>
</tr>
<tr>
<td>Supervision</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes (n=141)</td>
<td>84 (59.6)</td>
<td>57 (40.4)</td>
<td>3.42 (2.05-5.71)</td>
</tr>
<tr>
<td>No (n=123)</td>
<td>37 (30.1)</td>
<td>86 (69.9)</td>
<td>1</td>
</tr>
<tr>
<td>Competency</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High (n=133)</td>
<td>76 (57.1)</td>
<td>57 (42.9)</td>
<td>2.54 (1.54-4.19)</td>
</tr>
<tr>
<td>Low (n=131)</td>
<td>45 (34.4)</td>
<td>86 (65.6)</td>
<td>1</td>
</tr>
</tbody>
</table>

^aAll the percentages were calculated for each sociodemographic category.
^bReference.
*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.

...There is 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...
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=0.03) (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=0.03) was in line with the findings of studies conducted in Ethiopia [25] and Ghana (P=0.01) [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 ($P=0.04$) [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


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]


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

Moritz Esdar¹, MA; Ursula Hübner¹, PhD; Johannes Thye¹, MA; Birgit Babitsch², MPH, PhD; Jan-David Liebe¹,³, PhD

¹Health Informatics Research Group, Faculty of Business Management and Social Sciences, University of Applied Sciences Osnabrueck, Osnabrueck, Germany
²Institute of Health and Education, New Public Health, Osnabrück University, Osnabrueck, Germany
³Institute 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
Capriviistr 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 IQHIT 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 IQHIT 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 IQHIT 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

https://medinform.jmir.org/2021/3/e23306
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.
HIT\textsuperscript{a} quality

<table>
<thead>
<tr>
<th>Domains</th>
<th>Constructs</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIT information management</td>
<td>Professionalism of information management</td>
</tr>
<tr>
<td>HIT workflow support</td>
<td>Workflow composite score including technical descriptors and care processes</td>
</tr>
<tr>
<td>Perceived HIT quality</td>
<td>Perceived HIT workflow support</td>
</tr>
<tr>
<td>Innovation capabilities</td>
<td>Innovation capability: top management team support</td>
</tr>
</tbody>
</table>

\textsuperscript{a}HIT: health information technology.
\textsuperscript{b}IT: information technology.
\textsuperscript{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 the components of the conceptual model. Each step was thereby rooted in findings from studies that suggest individual 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 (ICT TMT), ICs at the IT department level (ICT ITD), and ICs at the organization-wide level (ICT 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 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].
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.

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Exemplary study</th>
</tr>
</thead>
<tbody>
<tr>
<td>The PIM(^{a}) might be linked to HIT(^{b}) workflow support</td>
<td>Ammenwerth et al (2006) [69], Avgar et al (2012) [70], Bradley et al (2012) [71], Winter et al (2011) [72]</td>
</tr>
<tr>
<td>• H1: PIM has a positive effect on the WCS</td>
<td>Cresswell and Sheikh (2013) [12], Liebe et al (2018) [42], Potts et al (2011) [73], Sligo et al (2017) [26]</td>
</tr>
<tr>
<td>Formal participation in terms of the appointment of clinical IT(^{d}) agents might result from PIM practices and might lead to better HIT workflow support</td>
<td>Hadji and Degoulet (2016) [74], Hübner (2015) [75], Yusof et al (2008) [23]</td>
</tr>
<tr>
<td>• H2: The effect of PIM on the WCS is partly mediated by clinical IT agents</td>
<td>Gorla et al (2010) [76], Suki (2012) [77]</td>
</tr>
<tr>
<td>There likely is a direct link between the technical and the perceived quality of HIT workflow support</td>
<td>Abdekhoda et al (2015) [78], Carpenter et al (2004) [79], Laukka et al [80]</td>
</tr>
<tr>
<td>• H3: The WCS has a positive effect on the perceived HIT workflow support</td>
<td>Esdar et al (2018) [38], Paré et al (2020) [56], Leidner et al (2010) [54]</td>
</tr>
<tr>
<td>• H4: The WCS has a positive effect on the overall goodness of information provision</td>
<td>Liebe et al (2017) [83], Watts and Henderson [84]</td>
</tr>
<tr>
<td>A top management team that is capable and willing to innovate might facilitate an innovation-friendly culture throughout the organization, including the IT department</td>
<td>Bradley et al (2012) [71], Liebe et al (2018) [81], Weintraub and McKee (2019) [82]</td>
</tr>
<tr>
<td>• H6: Innovation capability: top management team support has a positive effect on organization-wide innovation capability</td>
<td>Esdar et al (2018) [38], Paré et al (2020) [56], Leidner et al (2010) [54]</td>
</tr>
<tr>
<td>• H7: Innovation capability: top management team support has a positive effect on the innovation capability of the IT department</td>
<td>Liebe et al (2017) [83], Watts and Henderson [84]</td>
</tr>
<tr>
<td>Innovation capabilities of the top management team and the IT department might determine the degree of HIT workflow support</td>
<td>Esdar et al (2018) [38], Paré et al (2020) [56], Leidner et al (2010) [54]</td>
</tr>
<tr>
<td>• H9: Innovation capability: top management team support has a positive effect on the WCS</td>
<td>Liebe et al (2017) [83], Watts and Henderson [84]</td>
</tr>
<tr>
<td>The ability of the IT department to innovate might be linked to information management practices</td>
<td>DesRoches et al (2012) [58], Fadol et al (2015) [87], Kruse et al (2014) [88], Troilo et al (2014) [89]</td>
</tr>
</tbody>
</table>

\(^{a}\)PIM: professionalism of information management.  
\(^{b}\)HIT: health 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 $R^2$ 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>
<thead>
<tr>
<th>Characteristics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Country, n (response rate in %)</strong></td>
<td></td>
</tr>
<tr>
<td>Austria</td>
<td>14 (8.8)</td>
</tr>
<tr>
<td>Germany</td>
<td>205 (18.3)</td>
</tr>
<tr>
<td>Switzerland</td>
<td>13 (11.3)</td>
</tr>
<tr>
<td><strong>Ownership, n (% in sample)</strong></td>
<td></td>
</tr>
<tr>
<td>For-profit</td>
<td>42 (18.1)</td>
</tr>
<tr>
<td>Nonprofit</td>
<td>78 (33.6)</td>
</tr>
<tr>
<td>Public</td>
<td>112 (48.3)</td>
</tr>
<tr>
<td><strong>Teaching status, n (% in sample)</strong></td>
<td></td>
</tr>
<tr>
<td>Major teaching hospital</td>
<td>22 (9.5)</td>
</tr>
<tr>
<td>Minor teaching hospital</td>
<td>101 (43.5)</td>
</tr>
<tr>
<td>Nonteaching hospital</td>
<td>109 (47.0)</td>
</tr>
<tr>
<td><strong>Member of a hospital group, n (% in sample)</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>140 (60.3)</td>
</tr>
<tr>
<td>No</td>
<td>92 (39.7)</td>
</tr>
<tr>
<td><strong>Number of beds, mean (SD)</strong></td>
<td>491.9 (238.5)</td>
</tr>
</tbody>
</table>

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 ($P<0.001$). 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.

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

\(^a\)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 \( \eta^2 \) (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 IQHIT 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.
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 IQHIT 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 IQHIT 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


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]


https://medinform.jmir.org/2021/3/e23306

JMIR Med Inform 2021 | vol. 9 | iss. 3 | e23306 | p.52

(page number not for citation purposes)


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

Kyungmi Chung1,2,3, PhD; Hee Young Cho4, MD, PhD; Jin Young Park1,2,3, MD, PhD

1Department of Psychiatry, Yonsei University College of Medicine, Yongin Severance Hospital, Yonsei University Health System, Yongin-si, Republic of Korea
2Center for Digital Health, Yongin Severance Hospital, Yonsei University Health System, Yongin-si, Republic of Korea
3Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Yonsei University Health System, Seoul, Republic of Korea
4Department 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 (\(\rho=0.26; P=0.36\)), ease of use (\(\rho=0.19; P=0.51\)), satisfaction (\(\rho=0.21; P=0.46\)), or total usability scores (\(\rho=0.32; P=0.24\)). Unlike EOL, all 3 subfactors and the total usability had significant positive associations with each other (all \(\rho>0.80; P<0.001\)). Furthermore, perceived risks exhibited no significant negative associations with perceived benefits (\(\rho=0.29; P=0.30\)) or intention to seek (SEE; \(\rho=-0.28; P=0.32\)) or share (SHA; \(\rho=-0.24; P=0.40\)) 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 (\(\rho=0.94; P<0.001\)) than with SHA (\(\rho=0.70; P=0.004\)).

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.

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 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 sympon (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 ($W_{15}=0.84; P=0.01$) and perceived risks ($W_{15}=0.88; P=0.04$) 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).

<table>
<thead>
<tr>
<th>ID</th>
<th>Age (years)</th>
<th>Sex</th>
<th>Pregnancy stage</th>
<th>Pregnancy/delivery information source</th>
<th>Web-based information search strategy via computer or mobile phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTI-01</td>
<td>34</td>
<td>F</td>
<td>PP</td>
<td>Internet</td>
<td>Keyword search</td>
</tr>
<tr>
<td>UTI-02</td>
<td>35</td>
<td>F</td>
<td>PP</td>
<td>Acquaintances</td>
<td>Keyword search</td>
</tr>
<tr>
<td>UTI-03</td>
<td>35</td>
<td>F</td>
<td>PP</td>
<td>Books</td>
<td>Sentence search</td>
</tr>
<tr>
<td>UTI-04</td>
<td>34</td>
<td>F</td>
<td>PP</td>
<td>Acquaintances</td>
<td>Real-time search</td>
</tr>
<tr>
<td>UTI-05</td>
<td>31</td>
<td>F</td>
<td>FT (8 weeks)</td>
<td>Internet</td>
<td>Keyword search</td>
</tr>
<tr>
<td>UTC-06A</td>
<td>36</td>
<td>F</td>
<td>FT (8 weeks)</td>
<td>Acquaintances</td>
<td>Keyword search</td>
</tr>
<tr>
<td>UTC-07A</td>
<td>38</td>
<td>M</td>
<td>FT (8 weeks)</td>
<td>Books</td>
<td>Keyword search</td>
</tr>
<tr>
<td>UTC-08B</td>
<td>36</td>
<td>F</td>
<td>ST (15 weeks)</td>
<td>Acquaintances</td>
<td>Keyword search</td>
</tr>
<tr>
<td>UTC-09B</td>
<td>40</td>
<td>M</td>
<td>ST (15 weeks)</td>
<td>Books</td>
<td>Keyword search</td>
</tr>
<tr>
<td>UTI-10</td>
<td>43</td>
<td>F</td>
<td>ST (17 weeks)</td>
<td>Internet</td>
<td>Keyword search</td>
</tr>
<tr>
<td>UTI-11</td>
<td>33</td>
<td>F</td>
<td>ST (23 weeks)</td>
<td>Internet</td>
<td>Keyword search</td>
</tr>
<tr>
<td>UTI-12</td>
<td>39</td>
<td>F</td>
<td>ST (24 weeks)</td>
<td>Health professionals</td>
<td>Keyword search</td>
</tr>
<tr>
<td>UTI-13</td>
<td>27</td>
<td>F</td>
<td>TT (32 weeks)</td>
<td>Acquaintances</td>
<td>Keyword search</td>
</tr>
<tr>
<td>UTI-14</td>
<td>31</td>
<td>F</td>
<td>P (3 weeks after birth)</td>
<td>Acquaintances</td>
<td>Keyword search</td>
</tr>
<tr>
<td>UTI-15</td>
<td>32</td>
<td>F</td>
<td>P (3 weeks after birth)</td>
<td>Acquaintances</td>
<td>Keyword search</td>
</tr>
</tbody>
</table>

*aTwo 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.

bF: female; M: male.

cPP: pregnancy preparation (planned natural pregnancy); FT: first trimester; ST: second trimester; TT: third trimester; P: puerperium.

dInternet includes portal/search engines, online communities, blogs, vlogs, etc.

eKeyword search with simple words, search operators, hashtags, etc.

fAcquaintances include friends, colleagues, online community members, experienced mothers in the same postnatal care center, etc.

gBooks include encyclopedias of pregnancy and birth, essays and articles written by medical doctors, magazines, etc.

hSentence search with a single statement/question or multiple statements/questions.

iReal-time search means choosing and looking for attention-capturing content published in real time on the internet.

jHealth 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 $p<0.001$).
Table 2. Descriptive statistics for sex difference in responses to USE questionnaire on the medical chatbot via KakaoTalk (N=15).

<table>
<thead>
<tr>
<th>Sex</th>
<th>Usability subfactors, mean (SD)</th>
<th>Usefulness (USE)</th>
<th>Ease of Use (EOU)</th>
<th>Ease of Learning (EOL)</th>
<th>Satisfaction (SAT)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5.43 (1.21)</td>
<td>6.05 (0.96)</td>
<td>7.00 (0.00)</td>
<td>5.57 (2.02)</td>
<td>6.01 (2.02)</td>
</tr>
<tr>
<td>Men (n=2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women (n=13)</td>
<td></td>
<td>4.78 (1.12)</td>
<td>5.23 (0.67)</td>
<td>6.25 (0.71)</td>
<td>4.80 (1.20)</td>
<td>5.27 (0.82)</td>
</tr>
<tr>
<td>Total (N=15)</td>
<td></td>
<td>4.87 (1.11)</td>
<td>5.34 (0.73)</td>
<td>6.35 (0.71)</td>
<td>4.90 (1.26)</td>
<td>5.37 (0.85)</td>
</tr>
</tbody>
</table>

aUsability 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).

bUSE: usefulness.
cEOU: ease of use.
dEOL: ease of learning.
eSAT: 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

<table>
<thead>
<tr>
<th>Subfactors</th>
<th>USEb</th>
<th>EOUc</th>
<th>EOLd</th>
<th>SATe</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>USE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient (p)</td>
<td>1.00</td>
<td>0.82</td>
<td>0.26</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>P value (2-tailed)</td>
<td></td>
<td>&lt;.001</td>
<td></td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>EOU</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient (p)</td>
<td>0.82</td>
<td>1.00</td>
<td>0.19</td>
<td>0.81</td>
<td>0.89</td>
</tr>
<tr>
<td>P value (2-tailed)</td>
<td>&lt;.001</td>
<td></td>
<td>.51</td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>EOL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient (p)</td>
<td>0.26</td>
<td>0.19</td>
<td>1.00</td>
<td>0.21</td>
<td>0.32</td>
</tr>
<tr>
<td>P value (2-tailed)</td>
<td>.36</td>
<td>.51</td>
<td></td>
<td>.46</td>
<td>.24</td>
</tr>
<tr>
<td><strong>SAT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient (p)</td>
<td>0.98</td>
<td>0.81</td>
<td>0.21</td>
<td>1.00</td>
<td>0.95</td>
</tr>
<tr>
<td>P value (2-tailed)</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.46</td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient (p)</td>
<td>0.97</td>
<td>0.89</td>
<td>0.32</td>
<td>0.95</td>
<td>1.00</td>
</tr>
<tr>
<td>P value (2-tailed)</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.24</td>
<td>&lt;.001</td>
<td></td>
</tr>
</tbody>
</table>

aUsability was measured by the average score of 4 subfactors, which is presented as the “Total” score in this table.
bUSE: usefulness.
cEOU: ease of use.
dEOL: ease of learning.
eSAT: satisfaction.
fNot 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).\(^a\)

<table>
<thead>
<tr>
<th>Sex</th>
<th>Factors, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PB(^b)</td>
</tr>
<tr>
<td>Men (n=2)</td>
<td>6.25 (1.06)</td>
</tr>
<tr>
<td>Women (n=13)</td>
<td>5.19 (1.03)</td>
</tr>
<tr>
<td>Total (N=15)</td>
<td>5.33 (1.06)</td>
</tr>
</tbody>
</table>

\(^a\)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).

<table>
<thead>
<tr>
<th>Factor</th>
<th>PB(^b)</th>
<th>PR(^b)</th>
<th>SEE(^c)</th>
<th>SHA(^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PB</td>
<td>1.00</td>
<td>-0.29</td>
<td>0.94</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>_(^e)</td>
<td>.30</td>
<td>&lt;.001</td>
<td>.004</td>
</tr>
<tr>
<td>PR</td>
<td>-0.29</td>
<td>1.00</td>
<td>-0.28</td>
<td>-0.24</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>_</td>
<td>.32</td>
<td>.40</td>
</tr>
<tr>
<td>SEE</td>
<td>0.94</td>
<td>-0.28</td>
<td>1.00</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>&lt;.001</td>
<td>.32</td>
<td>_</td>
<td>.002</td>
</tr>
<tr>
<td>SHA</td>
<td>0.70</td>
<td>-0.24</td>
<td>0.73</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>.004</td>
<td>.40</td>
<td>.002</td>
<td>_</td>
</tr>
</tbody>
</table>

\(^a\)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

https://medinform.jmir.org/2021/3/e18607
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 usability 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: HI18C09111).

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


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

©Kyungmi Chung, Hee Young Cho, Jin Young Park. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 03.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.
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, MD, MSc; Arno Uvin, BSc; Jessica Schmitt, LCSW; Tomasz Oliwa, PhD; Ellen Almirol, MSc; Samantha Devlin, MSc; John Schneider, MD, MPH

Department of Medicine, University of Chicago, Chicago, IL, United States

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 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 poly-substance 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, 296.36, 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, E950.6, E953.0, V11.0, V40.0, V40.9

- ICD-10 codes

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, 533.3, 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
Table 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; Agoraphoria 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; Depression disorder; Depression 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).

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

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.
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


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]


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

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.

**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 prenotations 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.
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.

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

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.

<table>
<thead>
<tr>
<th>Label</th>
<th>Train</th>
<th>Development</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tokens</td>
<td>Slots</td>
<td>Tokens</td>
</tr>
<tr>
<td>Medication name</td>
<td>1385</td>
<td>1227</td>
<td>146</td>
</tr>
<tr>
<td>Medication class</td>
<td>309</td>
<td>228</td>
<td>38</td>
</tr>
<tr>
<td>Dosage</td>
<td>1366</td>
<td>761</td>
<td>115</td>
</tr>
<tr>
<td>Frequency</td>
<td>1604</td>
<td>600</td>
<td>142</td>
</tr>
<tr>
<td>Duration</td>
<td>161</td>
<td>70</td>
<td>26</td>
</tr>
<tr>
<td>Route</td>
<td>95</td>
<td>85</td>
<td>8</td>
</tr>
<tr>
<td>Condition</td>
<td>192</td>
<td>61</td>
<td>9</td>
</tr>
</tbody>
</table>

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.

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

\(^a\)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).

**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 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].

When using rule-based systems, we showed that the combination of expert knowledge and a deep learning system increased the F-measure, 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].
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 redefine 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


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.


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]


43. Mikolov T, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017 Dec;5:135-146. [doi: 10.1162/tacl_a_00051]


Abbreviations

AP-HP: Assistance Publique–Hôpitaux de Paris

IOB: inside, outside, beginning

©Jordan Jouffroy, Sarah F Feldman, Ivan Lerner, Bastien Rance, Anita Burgun, Antoine Neuraz. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 16.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.
Human–Computer Agreement of Electrocardiogram Interpretation for Patients Referred to and Declined for Primary Percutaneous Coronary Intervention: Retrospective Data Analysis Study

Aleeha Iftikhar¹, MS; Raymond Bond¹, BSc, PhD; Victoria Mcgilligan², BSc, PhD; Stephen J Leslie³, BSc, MBChB, PhD, FRCP; Charles Knoery³, MBChB; James Shand³, MBBSMD, FRCP; Adesh Ramsewak³, MBBS, FRCP; Divyesh Sharma⁴, MBBS, MSc, FRCP; Anne McShane⁵, MSc; Khaled Rjoob¹, MSc; Aaron Peace⁵, MBChB, MRCPE, PhD

¹Computing Engineering and Build Environment, Ulster University, Belfast, United Kingdom
²Centre for Personalised Medicine, Ulster University, Londonderry, United Kingdom
³Cardiac Unit, Raigmore Hospital, Inverness, United Kingdom
⁴Department of Cardiology, Altnagelvin Hospital, Western Health and Social Care Trust, Londonderry, United Kingdom
⁵Letterkenny 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 $\kappa$ showed poor agreement ($\kappa=0.08$, $P=.001$). 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.
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] 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

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

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.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Odds ratio (95% CI)</th>
<th>SE</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out of hours (true/false)</td>
<td>1.41 (0.95-2.11)</td>
<td>0.012</td>
<td>.09</td>
</tr>
<tr>
<td>Age</td>
<td>1.07 (1.05-1.09)</td>
<td>0.434</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Chest pain (true)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.26 (0.73-2.16)</td>
<td>0.012</td>
<td>.39</td>
</tr>
<tr>
<td>Activator nurse diagnosis suggesting acute myocardial infarction (true)</td>
<td>1.30 (0.78-2.17)</td>
<td>0.013</td>
<td>.31</td>
</tr>
<tr>
<td>Computer diagnosis suggesting acute myocardial infarction (true)</td>
<td>0.97 (0.47-2.03)</td>
<td>0.011</td>
<td>.95</td>
</tr>
</tbody>
</table>

<sup>a</sup>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.

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

<sup>a</sup>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.

<table>
<thead>
<tr>
<th>Classification and rank&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Computer</th>
<th>Activator nurse</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Interpretation term</td>
<td>Patients, n (%)</td>
</tr>
<tr>
<td>Suggests acute myocardial infarction&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>“acute myocardial infarction”</td>
<td>337 (47)</td>
</tr>
<tr>
<td>2</td>
<td>“inferior infarct”</td>
<td>23 (3)</td>
</tr>
<tr>
<td>3</td>
<td>“anterior injury”</td>
<td>34 (5)</td>
</tr>
<tr>
<td>Does not suggest acute-myocardial infarction&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>“abnormal ECG”</td>
<td>382 (51)</td>
</tr>
<tr>
<td>2</td>
<td>“LBBB”</td>
<td>108 (15)</td>
</tr>
<tr>
<td>3</td>
<td>“borderline ECG”</td>
<td>41 (5.5)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Terms with low frequencies (1 or 2) are not included.

<sup>b</sup>n=722 for Computer; n=634 for Activator nurse.

<sup>c</sup>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=0.001). However, the activator nurse interpretation 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=0.001). 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<0.001). 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=0.001). 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=0.001). 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<0.001 and P=0.001, respectively) and 1 year (P<0.001 and P=0.008, 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 infarction. 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

4. Diagnosis-heart attack. UK National Health Service. URL: https://www.nhs.uk/conditions/heart-attack/diagnosis/ [accessed 2020-04-02]


**Abbreviations**

- CathLab: catheterization laboratory
- ECG: electrocardiogram
- OR: odds ratio
A Personal Health System for Self-Management of Congestive Heart Failure (HeartMan): Development, Technical Evaluation, and Proof-of-Concept Randomized Controlled Trial

Mitja Luštrek¹, PhD; Marko Bohanec², PhD; Carlos Cavero Barca³, BSc; Maria Costanza Ciancarelli⁴, MSc; Els Clays⁵, PhD; Amos Adeyemo Dawodu⁴, MD; Jan Derbove⁶, PhD; Delphine De Smedt⁵, PhD; Erik Dovgan¹, PhD; Jure Lampe⁷, BSc; Flavia Marino⁷, PhD; Miha Mlakar¹, PhD; Giovanni Pioggia⁸, PhD; Paolo Emilio Puddu⁴, PhD, MD; Juan Mario Rodríguez³, BSc; Michele Schiariti⁴, PhD, MD; Gašper Slapničar¹, MSc; Karin Slegers⁹, PhD; Gennaro Tartarisco⁸, PhD; Jakob Valič¹, MTh; Aljoša Vodopija¹, MSc

¹Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia
²Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
³Health Unit, Atos Research and Innovation (ARI), Atos Spain S.A., Madrid, Spain
⁴Department of Cardiovascular, Respiratory, Nephrological, Anesthesiological and Geriatric Sciences, Sapienza University of Rome, Rome, Italy
⁵Department of Public Health and Primary Care, Ghent University, Ghent, Belgium
⁶Meaningful Interactions Lab, KU Leuven, Leuven, Belgium
⁷SenLab d.o.o., Ljubljana, Slovenia
⁸Institute for Biomedical Research and Innovation, National Research Council of Italy, Messina, Italy
⁹Department 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 ($P=0.02$), whereas depression and anxiety rates were significantly reduced ($P<0.001$), as were perceived sexual problems ($P=0.01$). 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.

https://medinform.jmir.org/2021/3/e24501
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

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—for 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.
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].

**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.

**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.

**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>
<thead>
<tr>
<th>Psychological profile</th>
<th>Festinger principle</th>
<th>Example message</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motivated profile</td>
<td>Cognitive consequences of forced compliance</td>
<td>I should perform physical exercise to obtain benefits similar to those from medications</td>
</tr>
<tr>
<td>Anxious profile</td>
<td>Free choice</td>
<td>Walking for 10 min and watching TVa are two ways to relax. Walking improves your heart health, whereas TV does not</td>
</tr>
<tr>
<td>Depressed profile</td>
<td>Effort justification</td>
<td>Walking for 10 min will bring benefits similar to those obtained from medication</td>
</tr>
</tbody>
</table>

aTV: television.

### Mindfulness

Mindfulness exercises enhance patients’ awareness of their present condition and help them dissociate (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.

### 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.

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.
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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>MAE of systolic blood pressure (mm Hg)</th>
<th>MAE of diastolic blood pressure (mm Hg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline dummy (mean)</td>
<td>11.4</td>
<td>8.9</td>
</tr>
<tr>
<td>Decision tree</td>
<td>13.1</td>
<td>10.1</td>
</tr>
<tr>
<td>k-nearest neighbors</td>
<td>10.6</td>
<td>7.5</td>
</tr>
<tr>
<td>Support vector regression</td>
<td>11.3</td>
<td>8.5</td>
</tr>
<tr>
<td>Random forest</td>
<td>9.0</td>
<td>7.0</td>
</tr>
</tbody>
</table>

*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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Classification accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline dummy (majority)</td>
<td>37.9</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>79.7</td>
</tr>
<tr>
<td>Multilayer perceptron</td>
<td>75.1</td>
</tr>
<tr>
<td>Random forest</td>
<td>62.6</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>88.6</td>
</tr>
</tbody>
</table>
Table 4. Precision, recall, and F-measures of the psychological profile detection.

<table>
<thead>
<tr>
<th>Psychological profile</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-measure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motivated profile</td>
<td>93</td>
<td>94</td>
<td>94</td>
</tr>
<tr>
<td>Anxious profile</td>
<td>86</td>
<td>83</td>
<td>85</td>
</tr>
<tr>
<td>Depressed profile</td>
<td>84</td>
<td>87</td>
<td>86</td>
</tr>
</tbody>
</table>

Table 5. Precision, recall, and F-measure of the physical activity recognition.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-measure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rest</td>
<td>84</td>
<td>89</td>
<td>87</td>
</tr>
<tr>
<td>Standing</td>
<td>48</td>
<td>32</td>
<td>38</td>
</tr>
<tr>
<td>Walking</td>
<td>75</td>
<td>86</td>
<td>80</td>
</tr>
<tr>
<td>Nordic walking</td>
<td>67</td>
<td>78</td>
<td>72</td>
</tr>
<tr>
<td>Running</td>
<td>74</td>
<td>62</td>
<td>67</td>
</tr>
<tr>
<td>Exercise</td>
<td>72</td>
<td>77</td>
<td>74</td>
</tr>
<tr>
<td>Eating</td>
<td>62</td>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>Washing</td>
<td>73</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td>Chores</td>
<td>84</td>
<td>81</td>
<td>82</td>
</tr>
<tr>
<td>Hand activities</td>
<td>67</td>
<td>65</td>
<td>66</td>
</tr>
<tr>
<td>Macro average</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
</tbody>
</table>

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.

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=0.02) 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<0.001) 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


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]


21. Heart failure info. URL: https://m.apkpure.com/heart-failure-info/com.programmingisfun.heartfailure [accessed 2021-02-04]


23. HF Path. URL: https://m.apkpure.com/hf-path/com.wellness.aha [accessed 2021-02-04]


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]


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

©Mitja Luštrek, Marko Bohanec, Carlos Cavero Barca, Maria Costanza Ciancarelli, Els Clays, Amos Adeyemo Dawodu, Jan Derboven, Delphine De Smedt, Erik Dovygan, Jure Lampe, Flavia Marino, Miha Mlakar, Giovanni Pioggia, Paolo Emilio Puddu, Juan Mario Rodríguez, Michele Schiariti, Gašper Slapničar, Karin Slegers, Gennaro Tartarisco, Jakob Valič, Aljoša Vodopija. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 05.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR.
Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.
A Clinical Decision Support System (KNOWBED) to Integrate Scientific Knowledge at the Bedside: Development and Evaluation Study

Alicia Martinez-Garcia¹, PhD; Ana Belén Naranjo-Saucedo¹, MSc; Jose Antonio Rivas¹, BSc; Antonio Romero Tabares², PhD; Ana María Cassinello³, BSc; Anselmo Andrés-Martín³, PhD; Francisco José Sánchez Laguná⁴, MSc; Roman Villegas¹, MSc; Francisco De Paula Pérez León¹, BSc; Jesús Moreno Conde¹, MSc; Carlos Luis Parra Calderón¹,⁵, MSc

¹Group of Research and Innovation in Biomedical Informatics, Biomedical Engineering and Health Economy, Institute of Biomedicine of Seville, IBiS / Virgen del Rocio University Hospital / CSIC / University of Seville, Seville, Spain
²Publications Department, Andalusian Institute of Emergencies and Public Safety, Seville, Spain
³Paediatrics Unit, Virgen Macarena University Hospital, Seville, Spain
⁴Department of Information Systems Coordination, Andalusian Health Service, Seville, Spain
⁵Department of Innovation Technology, Virgen del Rocio 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 Rocio 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

https://medinform.jmir.org/2021/3/e13182
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).
- Synoposes: 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 mobile 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.
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 Pediatrics 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 Pediatrics 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.
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.
**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.
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 less 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


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

©Alicia Martinez-Garcia, Ana Belén Naranjo-Saucedo, Jose Antonio Rivas, Antonio Romero Tabares, Ana Marín Cassinello, Anselmo Andrés-Martín, Francisco José Sánchez Laguna, Roman Villegas, Francisco De Paula Pérez León, Jesús Moreno Conde, Carlos Luis Parra Calderón. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 10.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.
Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study

Alberto Tena\textsuperscript{1}, MSc; Francec Claria\textsuperscript{2}, PhD; Francesc Solsona\textsuperscript{2}, PhD; Einar Meister\textsuperscript{3}, PhD; Monica Povedano\textsuperscript{4}, PhD

\textsuperscript{1}Information and Communication Technologies Group, International Centre for Numerical Methods in Engineering, Barcelona, Spain
\textsuperscript{2}Department of Computer Science, Universitat de Lleida, Lleida, Spain
\textsuperscript{3}Institute of Cybernetics, Tallinn University of Technology, Tallinn, Estonia
\textsuperscript{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.

\textit{(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.

García-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), Naive Bayes, Random Forest (RF), and Stacked Generalization (Stacking) which is one of the most frequently used models for classification purposes. The scores of the participants with ALS who did not present this involvement ranged from 6 to 46 (mean 23.1, SD 9.8), and the scores of participants with ALS who did not present this involvement ranged from 6 to 46 (mean 31.3, SD 8.6; 3 patients’ scores were reported as not available). Within the 45 participants in this study ranged from 6 to 46 (mean 31.3, 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 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 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 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 Audacity, an open-source application.

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 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; 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.

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):

\[
T_i \text{ is the duration of the } i\text{th cycle and } N \text{ is the total number of cycles. If } T_i \text{ or } T_{i+1} \text{ is outside the floor and ceiling periods, or if } T_i \text{ or } T_{i+1} \text{ is greater than the maximum period factor, the term is not counted in the sum, and } N \text{ is lowered by 1 (if } N \text{ ends up being less than 2, the result of the computation becomes “undefined”). }
\]

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:

\[
\text{Where } A_i \text{ is the extracted peak-to-peak amplitude data and } N \text{ is the number of extracted fundamental periods.}
\]

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:

where \( r(t) \) is the normalized autocorrelation function, \( r(t = \tau) \) is the second local maximum of the normalized autocorrelation and \( \tau \) 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].
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 parameters 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(beat), 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: $X = USV^T$, 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: \( (x_i - \bar{x}) / \sigma \), where \( x_i \) is the feature vector, \( \bar{x} \) is the mean, and \( \sigma \) 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 package [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 $\alpha=.05$ was established for all tests. The $P$ values obtained by performing the tests with values below $\alpha=.05$ 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.
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%.

<table>
<thead>
<tr>
<th>Model and metrics</th>
<th>Classification performance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C\textsuperscript{a} vs B\textsuperscript{b}</td>
</tr>
<tr>
<td><strong>Random forest</strong></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>93.6</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>91.1</td>
</tr>
<tr>
<td>Specificity</td>
<td>95.5</td>
</tr>
<tr>
<td><strong>Naive Bayes</strong></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>91.0</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>89.2</td>
</tr>
<tr>
<td>Specificity</td>
<td>93.2</td>
</tr>
<tr>
<td><strong>Logistic regression</strong></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>93.8</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>92.5</td>
</tr>
<tr>
<td>Specificity</td>
<td>94.8</td>
</tr>
<tr>
<td><strong>Linear discriminant analysis</strong></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>94.3</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>95.6</td>
</tr>
<tr>
<td>Specificity</td>
<td>90.0</td>
</tr>
<tr>
<td><strong>Neuronal network</strong></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>94.8</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>91.7</td>
</tr>
<tr>
<td>Specificity</td>
<td>97.2</td>
</tr>
<tr>
<td><strong>Support vector machine</strong></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>95.8</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>91.4</td>
</tr>
<tr>
<td>Specificity</td>
<td>99.3</td>
</tr>
</tbody>
</table>

\textsuperscript{a}C: control group.
\textsuperscript{b}B: patients with amyotrophic lateral sclerosis (ALS) with bulbar involvement.
\textsuperscript{c}NB: patients with ALS without bulbar involvement.
\textsuperscript{d}ALS: all patients with ALS.
Table 2. Classification performance of the supervised models with the classification threshold set at 95%.

<table>
<thead>
<tr>
<th>Model and metrics</th>
<th>Classification performance (%)</th>
<th>C(^a) vs B(^b)</th>
<th>C vs NB(^c)</th>
<th>B vs NB</th>
<th>C vs ALS(^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random forest</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td>58.3</td>
<td>56.1</td>
<td>68.8</td>
<td>75.1</td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td>4.8</td>
<td>30.4</td>
<td>0.0</td>
<td>65.6</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>98.8</td>
</tr>
<tr>
<td><strong>Naïve Bayes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td>82.3</td>
<td>68.8</td>
<td>72.8</td>
<td>75.1</td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td>64.7</td>
<td>54.6</td>
<td>15.8</td>
<td>65.6</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>96.1</td>
<td>93.3</td>
<td>98.6</td>
<td>98.8</td>
</tr>
<tr>
<td><strong>Logistic regression</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td>92.8</td>
<td>77.7</td>
<td>74.1</td>
<td>76.0</td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td>84.8</td>
<td>65.1</td>
<td>16.7</td>
<td>66.4</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>99.0</td>
<td>99.6</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>Linear discriminant analysis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td>88.1</td>
<td>70.6</td>
<td>71.7</td>
<td>71.1</td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td>72.7</td>
<td>53.5</td>
<td>0.9</td>
<td>59.5</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>Neuronal network</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td>92.6</td>
<td>84.8</td>
<td>73.1</td>
<td>86.8</td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td>83.2</td>
<td>76.1</td>
<td>20.5</td>
<td>81.6</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>100.0</td>
<td>100.0</td>
<td>96.8</td>
<td>99.8</td>
</tr>
<tr>
<td><strong>Support vector machine</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td>86.3</td>
<td>71.1</td>
<td>70.7</td>
<td>71.1</td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td>68.8</td>
<td>54.3</td>
<td>6.1</td>
<td>59.4</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

\(\text{aC: control group.}\)
\(\text{bB: patients with amyotrophic lateral sclerosis (ALS) with bulbar involvement.}\)
\(\text{cNB: patients with ALS without bulbar involvement.}\)
\(\text{dALS: all patients with ALS.}\)

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 up-sampling 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 measurements 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


35. Tena et al. JMIR Medical Informatics 2021 | vol. 9 | iss. 3 | e21331 | p.147 https://medinform.jmir.org/2021/3/e21331 (page number not for citation purposes)
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]


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: naive Bayes
NN: neuronal network
PCA: principal component
RF: random forest
SVD: singular value decomposition
SVM: support vector machine
TN: true negative
TP: true positive
Tena et al

Detecting 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

©Alberto Tena, Francec Claria, Francesc Solsona, Einar Meister, Monica Povedano. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 10.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.
Review

Physicians’ Use of the Computerized Physician Order Entry System for Medication Prescribing: Systematic Review

Asra Mogharbel1, BA, MSc; Dawn Dowding2, PhD; John Ainsworth1, MSc, PhD

1Division 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
2Division 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
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

https://medinform.jmir.org/2021/3/e22923
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 DeIone 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.

<table>
<thead>
<tr>
<th>Group A: type of system</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Medication alert systems</td>
</tr>
<tr>
<td>2. Computerized provider order entry</td>
</tr>
<tr>
<td>3. Computerized physician order entry</td>
</tr>
<tr>
<td>4. CPOE</td>
</tr>
<tr>
<td>5. Electronic prescription</td>
</tr>
<tr>
<td>6. Prescription decision support system</td>
</tr>
<tr>
<td>7. Computerized prescriber order entry</td>
</tr>
<tr>
<td>8. Pharmaceutical decision-support systems</td>
</tr>
<tr>
<td>9. Pharmacy information system</td>
</tr>
<tr>
<td>10. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group B: usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Use</td>
</tr>
<tr>
<td>2. Actual usage</td>
</tr>
<tr>
<td>3. System use</td>
</tr>
<tr>
<td>4. Utilization</td>
</tr>
<tr>
<td>5. Acceptance</td>
</tr>
<tr>
<td>6. Adoption</td>
</tr>
<tr>
<td>7. Usage</td>
</tr>
<tr>
<td>8. 1 or 2 or 3 or 4 or 5 or 6 or 7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group C: factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Factors</td>
</tr>
<tr>
<td>2. Determinants</td>
</tr>
<tr>
<td>3. 1 or 2</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Theme, construct, and factor</th>
<th>Studies, n</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Individual factors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance expectancy: perception that using CPOE will improve the physician’s job performance [18]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived usefulness</td>
<td>1</td>
<td>[29]</td>
</tr>
<tr>
<td>Relative advantage</td>
<td>1</td>
<td>[30]</td>
</tr>
<tr>
<td>Effect on quality of care and/or patient outcomes</td>
<td>3</td>
<td>[25,26,32]</td>
</tr>
<tr>
<td>Effects on productivity</td>
<td>2</td>
<td>[25,34]</td>
</tr>
<tr>
<td>Effects on safety</td>
<td>1</td>
<td>[24]</td>
</tr>
<tr>
<td>Performance outcomes</td>
<td>1</td>
<td>[25]</td>
</tr>
<tr>
<td><strong>Effort expectancy: belief that the CPOE is easy to use [18]</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ease of use</td>
<td>3</td>
<td>[28,29,32]</td>
</tr>
<tr>
<td>User-friendliness</td>
<td>1</td>
<td>[31]</td>
</tr>
<tr>
<td>Difficult to use</td>
<td>2</td>
<td>[24,25]</td>
</tr>
<tr>
<td>Complexity</td>
<td>1</td>
<td>[30]</td>
</tr>
<tr>
<td>Social influence: perceived importance of others’ (e.g., leaders, colleagues) opinions that the physician should or should not use the system [18]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>External normative beliefs</td>
<td>1</td>
<td>[25]</td>
</tr>
<tr>
<td><strong>Organizational factors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facilitating conditions: available resources, facilities, and infrastructure that facilitate using CPOE [18]</td>
<td>4</td>
<td>[24,25,33,34]</td>
</tr>
<tr>
<td>Training</td>
<td>4</td>
<td>[24,25,33,34]</td>
</tr>
<tr>
<td>Availability of technical support</td>
<td>4</td>
<td>[25,27,31,32]</td>
</tr>
<tr>
<td>Compatibility</td>
<td>1</td>
<td>[30]</td>
</tr>
<tr>
<td>Computer skills</td>
<td>1</td>
<td>[34]</td>
</tr>
<tr>
<td>Time constraints</td>
<td>3</td>
<td>[24,25,27]</td>
</tr>
<tr>
<td>Availability of hardware</td>
<td>2</td>
<td>[25,27]</td>
</tr>
<tr>
<td>Lack of awareness of the availability of certain features</td>
<td>1</td>
<td>[33]</td>
</tr>
<tr>
<td>Management support</td>
<td>1</td>
<td>[25]</td>
</tr>
<tr>
<td>User involvement</td>
<td>1</td>
<td>[25]</td>
</tr>
<tr>
<td><strong>Technological factors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information quality: relevance, accuracy, comprehensiveness, understandability, prevalence, timeliness, and usability of the outputs or content [19]</td>
<td>1</td>
<td>[32]</td>
</tr>
<tr>
<td>Usefulness of error messages</td>
<td>1</td>
<td>[31]</td>
</tr>
<tr>
<td>Clarity and brevity of the reminders</td>
<td>1</td>
<td>[31]</td>
</tr>
<tr>
<td>Confidentiality, privacy, and security of patients’ records</td>
<td>1</td>
<td>[25]</td>
</tr>
<tr>
<td>System quality: reliability, functionality, flexibility, ease of use, integration, and response time of the system [19]</td>
<td>2</td>
<td>[28,32]</td>
</tr>
<tr>
<td>Clarity</td>
<td>2</td>
<td>[28,32]</td>
</tr>
<tr>
<td>Layout</td>
<td>1</td>
<td>[31]</td>
</tr>
<tr>
<td>Technical problems causing delays during prescribing</td>
<td>1</td>
<td>[31]</td>
</tr>
<tr>
<td>System’s speed</td>
<td>3</td>
<td>[31,32,34]</td>
</tr>
<tr>
<td>Software barriers</td>
<td>1</td>
<td>[25]</td>
</tr>
<tr>
<td>Reliability</td>
<td>1</td>
<td>[32]</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Theme, construct, and factor</th>
<th>Studies, n</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customization to individual departments</td>
<td>2</td>
<td>[25,33]</td>
</tr>
<tr>
<td>Functionality of the tools in the system</td>
<td>1</td>
<td>[34]</td>
</tr>
<tr>
<td>Locating items on the system</td>
<td>1</td>
<td>[32]</td>
</tr>
<tr>
<td>Retrieval of radiology data</td>
<td>1</td>
<td>[32]</td>
</tr>
<tr>
<td>Usability</td>
<td>1</td>
<td>[24]</td>
</tr>
<tr>
<td>System’s efficiency</td>
<td>2</td>
<td>[24,26]</td>
</tr>
<tr>
<td>Availability of reference materials</td>
<td>1</td>
<td>[32]</td>
</tr>
<tr>
<td>Alert fatigue</td>
<td>2</td>
<td>[24,33]</td>
</tr>
</tbody>
</table>

CPOE: computerized physician order entry.
Resistance to use was reported in both reviews [12,14], as identified other factors that were not found in this study. Nevertheless, these reviews 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. Customization 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.
Multimedia Appendix 3

References


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
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 Ma1*, MD, PhD; Zihao Zou1*, MD, MMed; Emmanuel Eric Pazo2, MD, MSc, PhD; Salissou Moutari3, PhD; Ye Liu1, BSc; Feng Jin1, MD, PhD

1Department of Breast Surgery, The First Affiliated Hospital of China Medical University, Shenyang, China
2Department of Ophthalmology, He Eye Hospital, Shenyang, China
3Mathematical 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 administering 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.
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 (FACT) 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
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 χ² tests. In all analyses, \( P < 0.05 \) (2-tailed) was considered indicative of statistically significant differences (α = 0.05). As such an analysis is considered an exploratory 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.

<table>
<thead>
<tr>
<th>Patient characteristics</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Menstrual status</strong></td>
<td></td>
</tr>
<tr>
<td>Premenopause</td>
<td>133 (37.6)</td>
</tr>
<tr>
<td>Perimenopause</td>
<td>107 (30.2)</td>
</tr>
<tr>
<td>Postmenopause</td>
<td>114 (32.2)</td>
</tr>
<tr>
<td><strong>Level of education completed</strong></td>
<td></td>
</tr>
<tr>
<td>Primary</td>
<td>59 (16.7)</td>
</tr>
<tr>
<td>Secondary</td>
<td>161 (45.5)</td>
</tr>
<tr>
<td>Tertiary</td>
<td>134 (37.9)</td>
</tr>
<tr>
<td><strong>Marital status</strong></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>16 (4.5)</td>
</tr>
<tr>
<td>Married</td>
<td>338 (95.5)</td>
</tr>
<tr>
<td><strong>Region</strong></td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>165 (46.6)</td>
</tr>
<tr>
<td>Urban</td>
<td>189 (53.4)</td>
</tr>
<tr>
<td><strong>Treatment</strong></td>
<td></td>
</tr>
<tr>
<td>Neoadjuvant therapy</td>
<td>244 (68.9)</td>
</tr>
<tr>
<td>Adjuvant therapy</td>
<td>110 (31.1)</td>
</tr>
</tbody>
</table>
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).

<table>
<thead>
<tr>
<th>NFBSI-16a items</th>
<th>Paper-based patient score</th>
<th>Web-based patient score</th>
<th>P value</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>Median (IQR)</td>
<td>Mean’ (SD)</td>
<td>Median (IQR)</td>
</tr>
<tr>
<td>Disease-Related Symptoms Subscale – Physical (DRS-P)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GP1 (item 1)</td>
<td>2.31 (0.92)</td>
<td>2 (2-3)</td>
<td>2.32 (0.90)</td>
<td>2 (2-3)</td>
</tr>
<tr>
<td>GP4 (item 2)</td>
<td>2.19 (0.90)</td>
<td>2 (2-3)</td>
<td>2.17 (0.88)</td>
<td>2 (2-3)</td>
</tr>
<tr>
<td>GP6 (item 3)</td>
<td>2.29 (1.13)</td>
<td>2 (1-3)</td>
<td>2.30 (1.15)</td>
<td>2 (1-3)</td>
</tr>
<tr>
<td>B1 (item 4)</td>
<td>2.01 (0.89)</td>
<td>2 (1-3)</td>
<td>2.00 (0.88)</td>
<td>2 (1-3)</td>
</tr>
<tr>
<td>GP3 (item 5)</td>
<td>2.07 (0.98)</td>
<td>2 (1-3)</td>
<td>2.00 (0.91)</td>
<td>2 (1-3)</td>
</tr>
<tr>
<td>HI7 (item 6)</td>
<td>2.59 (1.02)</td>
<td>2 (2-3)</td>
<td>2.59 (1.06)</td>
<td>2 (2-3)</td>
</tr>
<tr>
<td>BP1 (item 7)</td>
<td>1.88 (0.93)</td>
<td>2 (1-2)</td>
<td>1.90 (0.93)</td>
<td>2 (1-2)</td>
</tr>
<tr>
<td>GF5 (item 8)</td>
<td>2.59 (1.18)</td>
<td>2 (2-3)</td>
<td>2.55 (1.17)</td>
<td>2 (2-3)</td>
</tr>
<tr>
<td>Disease-Related Symptoms Subscale – Emotional (DRS-E)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE6 (item 9)</td>
<td>2.00 (1.04)</td>
<td>2 (1-2)</td>
<td>2.01 (1.05)</td>
<td>2 (1-2)</td>
</tr>
<tr>
<td>Treatment Side Effects Subscale (TSE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GP2 (item 10)</td>
<td>2.20 (1.15)</td>
<td>2 (1-3)</td>
<td>2.25 (1.10)</td>
<td>2 (1-3)</td>
</tr>
<tr>
<td>N6 (item 11)</td>
<td>1.87 (0.98)</td>
<td>2 (1-2)</td>
<td>1.85 (0.93)</td>
<td>2 (1-2)</td>
</tr>
<tr>
<td>GP5 (item 12)</td>
<td>2.77 (1.01)</td>
<td>3 (2-3)</td>
<td>2.75 (1.00)</td>
<td>3 (2-3)</td>
</tr>
<tr>
<td>B5 (item 13)</td>
<td>2.98 (1.35)</td>
<td>3 (2-4)</td>
<td>2.98 (1.33)</td>
<td>3 (2-4)</td>
</tr>
<tr>
<td>Function and Well-Being Subscale (FWB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GF1 (item 14)</td>
<td>2.52 (1.04)</td>
<td>2 (2-3)</td>
<td>2.55 (1.01)</td>
<td>2.5 (2-3)</td>
</tr>
<tr>
<td>GF3 (item 15)</td>
<td>2.82 (1.12)</td>
<td>3 (2-4)</td>
<td>2.81 (1.08)</td>
<td>3 (2-4)</td>
</tr>
<tr>
<td>GF7 (item 16)</td>
<td>2.82 (1.19)</td>
<td>3 (2-4)</td>
<td>2.85 (1.21)</td>
<td>3 (2-4)</td>
</tr>
<tr>
<td>Total score</td>
<td>NFBSI-16 score 37.92 (7.79)</td>
<td>38 (32-42.5)</td>
<td>37.88 (7.71)</td>
<td>38 (32.75-42)</td>
</tr>
</tbody>
</table>


bStatistically significant difference.
Table 3. Parallel test-retest reliability of subscales (Wilcoxon test).

<table>
<thead>
<tr>
<th>NFBSI-16(^a) subscale</th>
<th>Paper-based patient outcome</th>
<th>Web-based patient outcome</th>
<th>(P) value</th>
<th>(\Delta) [Mean–Mean]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>Median (IQR)</td>
<td>Mean (SD)</td>
<td>Median (IQR)</td>
</tr>
<tr>
<td>Disease-Related Symptoms Subscale–Physical</td>
<td>35.90 (9.59)</td>
<td>36 (30-42)</td>
<td>35.64 (9.58)</td>
<td>34 (10-42)</td>
</tr>
<tr>
<td></td>
<td>32.00 (16.54)</td>
<td>32 (16-32)</td>
<td>32.05 (16.84)</td>
<td>32 (16-32)</td>
</tr>
<tr>
<td>Function and Well-Being Subscale</td>
<td>43.37 (13.37)</td>
<td>42.67 (32-53.33)</td>
<td>43.62 (13.72)</td>
<td>42.67 (32-53.33)</td>
</tr>
</tbody>
</table>

\(^a\)NFBSI-16: National Comprehensive Cancer Network–Functional Assessment of Cancer Therapy–Breast Cancer Symptom Index.

Test of Internal Consistency

Table 4 shows the Spearman \(\rho\) 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 \(\tau\) analysis between the two versions. In all items, the rank correlation was high as the Kendall \(\tau\) 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).

<table>
<thead>
<tr>
<th>Items</th>
<th>Spearman ρ</th>
<th>P value</th>
<th>Kendall τ</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disease-Related Symptoms Subscale – Physical (DRS-P)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GP1 (item 1)</td>
<td>0.89</td>
<td>&lt;.001</td>
<td>0.877</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>GP4 (item 2)</td>
<td>0.84</td>
<td>&lt;.001</td>
<td>0.810</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>GP6 (item 3)</td>
<td>0.86</td>
<td>&lt;.001</td>
<td>0.804</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>B1 (item 4)</td>
<td>0.90</td>
<td>&lt;.001</td>
<td>0.87</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>GP3 (item 5)</td>
<td>0.85</td>
<td>&lt;.001</td>
<td>0.825</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>H17 (item 6)</td>
<td>0.85</td>
<td>&lt;.001</td>
<td>0.813</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>BP1 (item 7)</td>
<td>0.89</td>
<td>&lt;.001</td>
<td>0.856</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>GF5 (item 8)</td>
<td>0.84</td>
<td>&lt;.001</td>
<td>0.796</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Subscale total</td>
<td>0.93</td>
<td>&lt;.001</td>
<td>0.827</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Disease-Related Symptoms Subscale – Emotional (DRS-E)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE6 (item 9)</td>
<td>0.85</td>
<td>&lt;.001</td>
<td>0.826</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Subscale total</td>
<td>0.85</td>
<td>&lt;.001</td>
<td>0.882</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Treatment Side Effects Subscale (TSE)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GP2 (item 10)</td>
<td>0.88</td>
<td>&lt;.001</td>
<td>0.830</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>N6 (item 11)</td>
<td>0.89</td>
<td>&lt;.001</td>
<td>0.857</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>GP5 (item 12)</td>
<td>0.83</td>
<td>&lt;.001</td>
<td>0.795</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>B5 (item 13)</td>
<td>0.84</td>
<td>&lt;.001</td>
<td>0.788</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Subscale total</td>
<td>0.95</td>
<td>&lt;.001</td>
<td>0.882</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Function and Well-Being Subscale (FWB)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GF1 (item 14)</td>
<td>0.82</td>
<td>&lt;.001</td>
<td>0.787</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>GF3 (item 15)</td>
<td>0.86</td>
<td>&lt;.001</td>
<td>0.821</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>GF7 (item 16)</td>
<td>0.83</td>
<td>&lt;.001</td>
<td>0.79</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Subscale total</td>
<td>0.91</td>
<td>&lt;.001</td>
<td>0.825</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Total score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score</td>
<td>0.94</td>
<td>&lt;.001</td>
<td>0.823</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
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).

<table>
<thead>
<tr>
<th>Items</th>
<th>$R_c^a$</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease-Related Symptoms Subscale – Physical (DRS-P)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GP1 (item 1)</td>
<td>0.92</td>
<td>0.90-0.94</td>
</tr>
<tr>
<td>GP4 (item 2)</td>
<td>0.85</td>
<td>0.82-0.88</td>
</tr>
<tr>
<td>GP6 (item 3)</td>
<td>0.86</td>
<td>0.83-0.88</td>
</tr>
<tr>
<td>B1 (item 4)</td>
<td>0.9</td>
<td>0.88-0.71</td>
</tr>
<tr>
<td>GP3 (item 5)</td>
<td>0.86</td>
<td>0.83-0.89</td>
</tr>
<tr>
<td>H17 (item 6)</td>
<td>0.86</td>
<td>0.83-0.89</td>
</tr>
<tr>
<td>BP1 (item 7)</td>
<td>0.88</td>
<td>0.87-0.91</td>
</tr>
<tr>
<td>GF5 (item 8)</td>
<td>0.85</td>
<td>0.82-0.88</td>
</tr>
<tr>
<td>Subscale total</td>
<td>0.94</td>
<td>0.93-0.95</td>
</tr>
<tr>
<td>Disease-Related Symptoms Subscale – Emotional (DRS-E)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE6 (item 9)</td>
<td>0.84</td>
<td>0.81-0.87</td>
</tr>
<tr>
<td>Subscale total</td>
<td>0.84</td>
<td>0.81-0.87</td>
</tr>
<tr>
<td>Treatment Side Effects Subscale (TSE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GP2 (item 10)</td>
<td>0.87</td>
<td>0.85-0.90</td>
</tr>
<tr>
<td>N6 (item 11)</td>
<td>0.88</td>
<td>0.86-0.91</td>
</tr>
<tr>
<td>GP5 (item 12)</td>
<td>0.86</td>
<td>0.83-0.89</td>
</tr>
<tr>
<td>B5 (item 13)</td>
<td>0.83</td>
<td>0.80-0.86</td>
</tr>
<tr>
<td>Subscale total</td>
<td>0.96</td>
<td>0.95-0.97</td>
</tr>
<tr>
<td>Function and Well-Being Subscale (FWB)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GF1 (item 14)</td>
<td>0.85</td>
<td>0.82-0.88</td>
</tr>
<tr>
<td>GF3 (item 15)</td>
<td>0.86</td>
<td>0.83-0.89</td>
</tr>
<tr>
<td>GF7 (item 16)</td>
<td>0.84</td>
<td>0.81-0.87</td>
</tr>
<tr>
<td>Subscale total</td>
<td>0.91</td>
<td>0.89-0.93</td>
</tr>
<tr>
<td>Total score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score</td>
<td>0.94</td>
<td>0.93-0.95</td>
</tr>
</tbody>
</table>

$R_c$: 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.

<table>
<thead>
<tr>
<th>Patient preference</th>
<th>Observed, n</th>
<th>Expected, n</th>
<th>Residual</th>
<th>Chi-square (df)</th>
<th>Asymptotic significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preferred paper-based questionaire</td>
<td>98</td>
<td>177</td>
<td>−79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preferred web-based questionaire</td>
<td>256</td>
<td>177</td>
<td>79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>354</td>
<td></td>
<td></td>
<td>70.5³ (1)</td>
<td>.001¹</td>
</tr>
</tbody>
</table>

°0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 177.0.

¹Statistically 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
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 FACT 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. 2013225385).

Conflicts of Interest

None declared.
Multimedia Appendix 1

References


Abbreviations

ANOVA: analysis of variance
CCC: concordance correlation coefficient
DRS-E: Disease-Related Symptoms Subscale–Emotional
DRS-P: Disease-Related Symptoms Subscale–Physical
FACT: 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

©Jinfei Ma, Zihao Zou, Emmanuel Eric Pazo, Salissou Moutari, Ye Liu, Feng Jin. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 02.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.
Antibiotic Prescription Rates After eVisits Versus Office Visits in Primary Care: Observational Study

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).

https://medinform.jmir.org/2021/3/e25473

JMIR Med Inform 2021 | vol. 9 | iss. 3 | e25473 | p.177
(page number not for citation purposes)
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.

<table>
<thead>
<tr>
<th>Chief complaint</th>
<th>Age in years, mean (SD)</th>
<th>P value for difference</th>
<th>Sex, male, n (%)</th>
<th>P value for difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sore throat (n=1110)</td>
<td>—&lt;sup&gt;a&lt;/sup&gt;</td>
<td>&lt;.001</td>
<td>—</td>
<td>.03</td>
</tr>
<tr>
<td>DIGIT&lt;sup&gt;b&lt;/sup&gt; (n=798)</td>
<td>35.1 (11.5)</td>
<td>—</td>
<td>262 (32.8)</td>
<td>—</td>
</tr>
<tr>
<td>PHYSIT&lt;sup&gt;c&lt;/sup&gt; (n=312)</td>
<td>44.5 (17.5)</td>
<td>—</td>
<td>81 (26.0)</td>
<td>—</td>
</tr>
<tr>
<td>Respiratory (n=1975)</td>
<td>—&lt;sup&gt;a&lt;/sup&gt;</td>
<td>&lt;.001</td>
<td>—</td>
<td>.28</td>
</tr>
<tr>
<td>DIGIR&lt;sup&gt;d&lt;/sup&gt; (n=1724)</td>
<td>42.8 (14.5)</td>
<td>—</td>
<td>637 (36.9)</td>
<td>—</td>
</tr>
<tr>
<td>PHYSIR&lt;sup&gt;e&lt;/sup&gt; (n=251)</td>
<td>60.0 (16.2)</td>
<td>—</td>
<td>84 (33.5)</td>
<td>—</td>
</tr>
<tr>
<td>Dysuria (n=1521)</td>
<td>—&lt;sup&gt;a&lt;/sup&gt;</td>
<td>&lt;.001</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>DIGIU&lt;sup&gt;f&lt;/sup&gt; (n=1325)</td>
<td>42.1 (15.4)</td>
<td>—</td>
<td>0 (0.0)</td>
<td>—</td>
</tr>
<tr>
<td>PHYSIU&lt;sup&gt;g&lt;/sup&gt; (n=196)</td>
<td>60.0 (18.9)</td>
<td>—</td>
<td>0 (0.0)</td>
<td>—</td>
</tr>
</tbody>
</table>

<sup>a</sup>Not applicable.<br />
<sup>b</sup>DIGIT: eVisits with a chief complaint of sore throat.<br />
<sup>c</sup>PHYSIT: Office visits with a chief complaint of sore throat.<br />
<sup>d</sup>DIGIR: eVisits with a chief complaint of common cold/influenza or cough.<br />
<sup>e</sup>PHYSIR: Office visits with a chief complaint of common cold/influenza or cough.<br />
<sup>f</sup>DIGIU: eVisits with a chief complaint of dysuria.<br />
<sup>g</sup>PHYSIU: 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 DIGIT compared with 33.3% (104/312) of PHYSIT. Viral upper and lower respiratory diagnoses were recorded among 61.3% (1057/1724) of DIGIR compared with 48.6% (122/251) of PHYSIR.

A total of 0.7% (19/2522) recorded diagnoses were for pneumonia across DIGIT and DIGIR compared with 2.3% (13/563) across PHYSIT and PHYSIR. Peritonsillar abscess was recorded in 0.8% (6/798) of DIGIT compared with 0.6% (2/312) of PHYSIT. There was one recorded diagnosis of pyelonephritis among PHYSIU and none among DIGIU.

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 PHYSIT visit compared with DIGIT-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.

<table>
<thead>
<tr>
<th>Chief complaint</th>
<th>Antibiotic prescription within 3 days of visit, n (%)</th>
<th>P value for difference</th>
<th>Guideline-recommended antibiotics, n (%)</th>
<th>P value for difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sore throat (n=1110)</td>
<td>____a.</td>
<td>&lt;.001</td>
<td>—</td>
<td>.39</td>
</tr>
<tr>
<td>DIGI-Tb (n=798)</td>
<td>169 (21.2)</td>
<td>—</td>
<td>163 (96.4)</td>
<td>—</td>
</tr>
<tr>
<td>PHYSI-Tc (n=312)</td>
<td>124 (39.7)</td>
<td>—</td>
<td>117 (94.4)</td>
<td>—</td>
</tr>
<tr>
<td>Respiratory (n=1975)</td>
<td>—</td>
<td>&lt;.001</td>
<td>—</td>
<td>.02</td>
</tr>
<tr>
<td>DIGI-Rd (n=1724)</td>
<td>27 (1.6)</td>
<td>—</td>
<td>26 (96.3)</td>
<td>—</td>
</tr>
<tr>
<td>PHYSI-Re (n=251)</td>
<td>50 (19.9)</td>
<td>—</td>
<td>37 (74.0)</td>
<td>—</td>
</tr>
<tr>
<td>Dysuria (n=1521)</td>
<td>—</td>
<td>.25</td>
<td>—</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>DIGI-Uf (n=1325)</td>
<td>1016 (76.7)</td>
<td>—</td>
<td>1009 (99.3)</td>
<td>—</td>
</tr>
<tr>
<td>PHYSI-Ug (n=196)</td>
<td>143 (73.0)</td>
<td>—</td>
<td>135 (94.4)</td>
<td>—</td>
</tr>
</tbody>
</table>

aNot applicable.
bDIGI-T: eVisits with a chief complaint of sore throat.
cPHYSI-T: Office visits with a chief complaint of sore throat.
dDIGI-R: eVisits with a chief complaint of common cold/influenza or cough.
ePHYSI-R: Office visits with a chief complaint of common cold/influenza or cough.
fDIGI-U: eVisits with a chief complaint of dysuria.
gPHYSI-U: Office visits with a chief complaint of dysuria.

Table 3. Regression models for antibiotic prescription for office visits compared with eVisits.

<table>
<thead>
<tr>
<th>Chief complaint</th>
<th>Antibiotic prescription within 3 days of office visits vs eVisits, UORa (95% CI)</th>
<th>P value</th>
<th>Antibiotic prescription within 3 days of office visits vs eVisits, AORbc (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sore throat (n=1110)</td>
<td>2.46 (1.85-3.26)</td>
<td>&lt;.001</td>
<td>2.94 (1.99-4.33)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Respiratory (n=1975)</td>
<td>15.63 (9.58-25.53)</td>
<td>&lt;.001</td>
<td>11.57 (5.50-24.32)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Dysuria (n=1521)</td>
<td>0.82 (0.58-1.15)</td>
<td>.25</td>
<td>1.01 (0.66-1.53)</td>
<td>.98</td>
</tr>
</tbody>
</table>

aUOR: unadjusted odds ratio.
bAOR: adjusted odds ratio.
cEach 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.
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’ reluctance 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 for

https://medinform.jmir.org/2021/3/e25473

JMIR Med Inform 2021 | vol. 9 | iss. 3 | e25473 | p.183

(page number not for citation purposes)
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


38. Rekommendationer-antibiotika-vid-faryngotonsillit-i-oppenvard.pdf [accessed 2021-03-01]


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

©Artin Entezarjou, Susanna Callings, Tapomita Bhattacharyya, Veronica Milos Nymberg, Lina Vigren, Ashkan Labaf, Ulf Jakobsson, Patrik Midløv. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 15.03.2021. This is

Ho Young Park¹, MD; Hyun-Jin Bae², PhD; Gil-Sun Hong¹, MD; Minjee Kim³, BSc; JiHye Yun¹, PhD; Sungwon Park⁴, MD; Won Jung Chung⁴, MD; NamKug Kim¹,⁵, PhD

¹Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine & Asan Medical Center, Seoul, Republic of Korea
²Department of Medicine, University of Ulsan College of Medicine & Asan Medical Center, Seoul, Republic of Korea
³Department 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
⁴Department of Health Screening and Promotion Center, University of Ulsan College of Medicine & Asan Medical Center, Seoul, Republic of Korea
⁵Department 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 (κ=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 (64x64 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 (512x512 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 (512x512 pixels) synthetic body CT images. Unlike PGGAN, previous GAN models such as deep convolutional GANs were able to generate relatively low-resolution (256x256 pixels) synthetic images [27]. However, PGGANs have demonstrated that high-resolution images (1024x1024 pixels) can be generated by applying progressive growing techniques [28]. Because CT images are acquired in high resolutions (512x512 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 portable 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 $4 \times 4$ to $512 \times 512$ (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 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.

<table>
<thead>
<tr>
<th>Groups, readers (R)</th>
<th>Accuracy (%)\textsuperscript{a}</th>
<th>Sensitivity (%)\textsuperscript{b}</th>
<th>Specificity (%)\textsuperscript{c}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group I\textsuperscript{d}</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R01</td>
<td>56.7</td>
<td>67.3</td>
<td>46.0</td>
</tr>
<tr>
<td>R05</td>
<td>48.3</td>
<td>53.3</td>
<td>43.3</td>
</tr>
<tr>
<td>R09</td>
<td>61.0</td>
<td>70.7</td>
<td>51.3</td>
</tr>
<tr>
<td>R10</td>
<td>66.0</td>
<td>70.7</td>
<td>61.3</td>
</tr>
<tr>
<td><strong>Group II\textsuperscript{e}</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R02</td>
<td>43.7</td>
<td>50.0</td>
<td>37.3</td>
</tr>
<tr>
<td>R06</td>
<td>73.0</td>
<td>68.7</td>
<td>77.3</td>
</tr>
<tr>
<td>R07</td>
<td>61.3</td>
<td>65.3</td>
<td>57.3</td>
</tr>
<tr>
<td>R08</td>
<td>64.0</td>
<td>77.3</td>
<td>50.7</td>
</tr>
<tr>
<td><strong>Group III\textsuperscript{f}</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R03</td>
<td>65.3</td>
<td>86.0</td>
<td>44.7</td>
</tr>
<tr>
<td>R04</td>
<td>54.3</td>
<td>58.7</td>
<td>50.0</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Mean (95% CI) accuracy: 59.4 (56.9-61.8), P=.36. \(P\) value was determined by generalized estimating equations.

\textsuperscript{b}Mean (95% CI) sensitivity: 66.8 (63.9-69.5), \(P=.04\).

\textsuperscript{c}Mean (95% CI) specificity: 51.9 (48.4-55.5), \(P=.02\).

\textsuperscript{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).

\textsuperscript{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).

\textsuperscript{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).
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 ($\kappa=0.11$) and for the 3 CT subsets (chest-L, chest-M, and AP-CT; $\kappa=0.04-0.13$). Interreader agreement was higher for the thoracoabdominal junction subset than for the other anatomical regions ($\kappa=0.31$ vs 0.03-0.14) (Table 2).
Table 2. Interreader agreement of the 10 readers with respect to the imaging subgroups.

<table>
<thead>
<tr>
<th>Image type, subsets</th>
<th>Kappa values</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire image set</td>
<td>0.11</td>
<td>0.09 to 0.13</td>
</tr>
<tr>
<td><strong>Image subsets</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chest-L(^a)</td>
<td>0.04</td>
<td>0.01 to 0.07</td>
</tr>
<tr>
<td>Chest-M(^b)</td>
<td>0.13</td>
<td>0.10 to 0.15</td>
</tr>
<tr>
<td>AP-CT(^c)</td>
<td>0.11</td>
<td>0.08 to 0.14</td>
</tr>
<tr>
<td><strong>Chest-L</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper lung</td>
<td>0.04</td>
<td>−0.01 to 0.09</td>
</tr>
<tr>
<td>Middle lung</td>
<td>0.01</td>
<td>−0.04 to 0.07</td>
</tr>
<tr>
<td>Lower lung</td>
<td>0.06</td>
<td>0.00 to 0.12</td>
</tr>
<tr>
<td><strong>Chest-M and AP-CT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thorax</td>
<td>0.03</td>
<td>−0.01 to 0.06</td>
</tr>
<tr>
<td>Thoracoabdominal junction</td>
<td>0.31</td>
<td>0.25 to 0.36</td>
</tr>
<tr>
<td>Abdomen</td>
<td>0.14</td>
<td>0.10 to 0.18</td>
</tr>
<tr>
<td>Pelvis</td>
<td>0.03</td>
<td>−0.02 to 0.08</td>
</tr>
</tbody>
</table>

\(^a\)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.

<table>
<thead>
<tr>
<th>Configuration, artificial features</th>
<th>Images (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abnormal vascular configuration⁶</td>
<td></td>
</tr>
<tr>
<td>Hepatic vessel (portal vein and hepatic vein)</td>
<td>13</td>
</tr>
<tr>
<td>Gastric vessel</td>
<td>3</td>
</tr>
<tr>
<td>Mesenteric vessel</td>
<td>2</td>
</tr>
<tr>
<td>Pulmonary vessel</td>
<td>2</td>
</tr>
<tr>
<td>Others (peripancreatic, coronary, rectal, axillary vessel)</td>
<td>4</td>
</tr>
<tr>
<td>Abnormal contour or structure⁶</td>
<td></td>
</tr>
<tr>
<td>Stomach</td>
<td>3</td>
</tr>
<tr>
<td>Pancreas</td>
<td>2</td>
</tr>
<tr>
<td>Heart</td>
<td>2</td>
</tr>
<tr>
<td>Mediastinal fat around the heart</td>
<td>2</td>
</tr>
<tr>
<td>Small bowel</td>
<td>1</td>
</tr>
</tbody>
</table>

⁶Ill-defined vascular margin, bizarre vascular course, or abnormal vascular diameter.

⁶Blurred 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 (512x512 pixels) whole-body CT images can be synthesized by GAN. GAN was trained with whole-body CT images (512x512 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


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

©Ho Young Park, Hyun-Jin Bae, Gil-Sun Hong, Minjee Kim, JiHye Yun, Sungwon Park, Won Jung Chung, NamKug Kim. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 17.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.
Machine Learning Approach to Predict the Probability of Recurrence of Renal Cell Carcinoma After Surgery: Prediction Model Development Study

HyungMin Kim¹,², MSc; Sun Jung Lee¹,², BSc; So Jin Park¹,², MSc; In Young Choi¹,²*, PhD; Sung-Hoo Hong³*, MD, PhD

¹Department of Medical Informatics, College of Medicine, The Catholic University, Seoul, Republic of Korea
²Department of Biomedicine & Health Sciences, College of Medicine, The Catholic University, Seoul, Republic of Korea
³Department 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 $P=0.05$, 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 $\geq$60 years, Fuhrman nuclear grade $\geq$3, and pathological stage $\geq$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).
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.

<table>
<thead>
<tr>
<th></th>
<th>Training set (n=2251)</th>
<th>Test set (n=563)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recurrence group, n (%)</td>
<td>Nonrecurrence group, n (%)</td>
</tr>
<tr>
<td>Before</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recurrence</td>
<td>226 (10.04)</td>
<td>511 (9.076)</td>
</tr>
<tr>
<td>Nonrecurrence</td>
<td>2025 (89.96)</td>
<td>2025 (50.00)</td>
</tr>
<tr>
<td>After</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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/m² and 24.7 kg/m², 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 \( \geq pT2 \) 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 \( \geq 3 \); 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).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Recurrence group (n=278)</th>
<th>Nonrecurrence group (n=2536)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>58.4 (11.9)</td>
<td>55.4 (12.7)</td>
</tr>
<tr>
<td>BMI (kg/m^2), mean (SD)</td>
<td>23.6 (3.2)</td>
<td>24.7 (3.3)</td>
</tr>
<tr>
<td>Gender, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>212 (76.3)</td>
<td>1811 (71.4)</td>
</tr>
<tr>
<td>Female</td>
<td>66 (23.7)</td>
<td>725 (28.6)</td>
</tr>
<tr>
<td>Smoking, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsmoker</td>
<td>207 (74.5)</td>
<td>2026 (79.9)</td>
</tr>
<tr>
<td>Current smoker</td>
<td>71 (25.5)</td>
<td>510 (20.1)</td>
</tr>
<tr>
<td>Pathological tumor stage, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a</td>
<td>50 (18.0)</td>
<td>1663 (65.6)</td>
</tr>
<tr>
<td>1b</td>
<td>60 (21.6)</td>
<td>487 (19.2)</td>
</tr>
<tr>
<td>2a</td>
<td>30 (10.8)</td>
<td>106 (4.2)</td>
</tr>
<tr>
<td>2b</td>
<td>12 (4.3)</td>
<td>29 (1.1)</td>
</tr>
<tr>
<td>3a</td>
<td>82 (29.5)</td>
<td>201 (7.9)</td>
</tr>
<tr>
<td>3b</td>
<td>34 (12.2)</td>
<td>36 (1.4)</td>
</tr>
<tr>
<td>3c</td>
<td>1 (0.4)</td>
<td>3 (0.1)</td>
</tr>
<tr>
<td>4</td>
<td>9 (3.2)</td>
<td>11 (0.4)</td>
</tr>
<tr>
<td>Histologic type, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear cell</td>
<td>242 (87.1)</td>
<td>2243 (88.4)</td>
</tr>
<tr>
<td>Papillary</td>
<td>14 (5.0)</td>
<td>44 (1.7)</td>
</tr>
<tr>
<td>Chromophobe</td>
<td>4 (1.4)</td>
<td>180 (7.1)</td>
</tr>
<tr>
<td>Collecting duct</td>
<td>5 (1.8)</td>
<td>4 (0.2)</td>
</tr>
<tr>
<td>Unclassified</td>
<td>5 (1.8)</td>
<td>15 (0.6)</td>
</tr>
<tr>
<td>Multilocular cystic</td>
<td>0 (0.0)</td>
<td>19 (0.7)</td>
</tr>
<tr>
<td>Mixed</td>
<td>6 (2.2)</td>
<td>24 (0.9)</td>
</tr>
<tr>
<td>Xp11.2 translocation</td>
<td>1 (0.4)</td>
<td>3 (0.1)</td>
</tr>
<tr>
<td>Clear cell papillary</td>
<td>1 (0.4)</td>
<td>4 (0.2)</td>
</tr>
<tr>
<td>Necrosis, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>143 (51.4)</td>
<td>2272 (89.6)</td>
</tr>
<tr>
<td>Microscopic</td>
<td>30 (10.8)</td>
<td>126 (5.0)</td>
</tr>
<tr>
<td>Macroscopic</td>
<td>105 (37.8)</td>
<td>138 (5.4)</td>
</tr>
<tr>
<td>Lymphovascular invasion, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>200 (71.9)</td>
<td>2436 (96.1)</td>
</tr>
<tr>
<td>Yes</td>
<td>78 (28.1)</td>
<td>100 (3.9)</td>
</tr>
<tr>
<td>Capsular invasion, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>148 (53.2)</td>
<td>2114 (83.4)</td>
</tr>
<tr>
<td>Yes</td>
<td>130 (46.8)</td>
<td>422 (16.6)</td>
</tr>
<tr>
<td>Fuhrman nuclear grade, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5 (1.8)</td>
<td>108 (4.3)</td>
</tr>
<tr>
<td>2</td>
<td>57 (20.5)</td>
<td>1293 (51.0)</td>
</tr>
<tr>
<td>3</td>
<td>141 (50.7)</td>
<td>1008 (39.7)</td>
</tr>
<tr>
<td>4</td>
<td>75 (27.0)</td>
<td>127 (5.0)</td>
</tr>
</tbody>
</table>
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 $\alpha$ value—a parameter in Laplace smoothing (Table 4).
Table 3. Diagnostic performance of machine learning algorithms for the prediction of renal cell carcinoma recurrence.

<table>
<thead>
<tr>
<th>Algorithm (parameter name) and parameter value (in 5 years, in 10 years)</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>AUROC&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel SVM&lt;sup&gt;b,c&lt;/sup&gt;</td>
<td>0.733</td>
<td>0.673</td>
<td>0.805</td>
<td>0.853</td>
</tr>
<tr>
<td>Logistic regression&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.644</td>
<td>0.692</td>
<td>0.839</td>
<td>0.816</td>
</tr>
<tr>
<td>Decision tree&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.533</td>
<td>0.442</td>
<td>0.866</td>
<td>0.869</td>
</tr>
<tr>
<td>KNN&lt;sup&gt;d&lt;/sup&gt; (n-neighbors)</td>
<td>0.556</td>
<td>0.519</td>
<td>0.905</td>
<td>0.898</td>
</tr>
<tr>
<td>(100, 100)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.467</td>
<td>0.426</td>
<td>0.947</td>
<td>0.928</td>
</tr>
<tr>
<td>(10, 10)</td>
<td>0.511</td>
<td>0.461</td>
<td>0.931</td>
<td>0.922</td>
</tr>
<tr>
<td>(200, 200)</td>
<td>0.556</td>
<td>0.481</td>
<td>0.899</td>
<td>0.902</td>
</tr>
<tr>
<td>NB&lt;sup&gt;e&lt;/sup&gt; (alpha)</td>
<td>0.822</td>
<td>0.731</td>
<td>0.850</td>
<td>0.828</td>
</tr>
<tr>
<td>(10, 100)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.578</td>
<td>0.500</td>
<td>0.858</td>
<td>0.853</td>
</tr>
<tr>
<td>Random forest (number of trees)</td>
<td>0.511</td>
<td>0.423</td>
<td>0.866</td>
<td>0.861</td>
</tr>
<tr>
<td>(5, 5)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.511</td>
<td>0.442</td>
<td>0.875</td>
<td>0.861</td>
</tr>
<tr>
<td>(10, 10)</td>
<td>0.511</td>
<td>0.462</td>
<td>0.864</td>
<td>0.861</td>
</tr>
<tr>
<td>(50, 50)</td>
<td>0.511</td>
<td>0.462</td>
<td>0.864</td>
<td>0.861</td>
</tr>
<tr>
<td>(100, 100)</td>
<td>0.511</td>
<td>0.462</td>
<td>0.864</td>
<td>0.861</td>
</tr>
<tr>
<td>AdaBoost (number of trees)</td>
<td>0.733</td>
<td>0.692</td>
<td>0.815</td>
<td>0.810</td>
</tr>
<tr>
<td>(50, 200)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.600</td>
<td>0.577</td>
<td>0.895</td>
<td>0.845</td>
</tr>
<tr>
<td>(10, 10)</td>
<td>0.733</td>
<td>0.673</td>
<td>0.815</td>
<td>0.824</td>
</tr>
<tr>
<td>(50, 50)</td>
<td>0.711</td>
<td>0.692</td>
<td>0.835</td>
<td>0.802</td>
</tr>
<tr>
<td>(100, 100)</td>
<td>0.711</td>
<td>0.692</td>
<td>0.837</td>
<td>0.810</td>
</tr>
<tr>
<td>(200, 200)</td>
<td>0.711</td>
<td>0.692</td>
<td>0.837</td>
<td>0.810</td>
</tr>
<tr>
<td>Gradient boost (number of trees)</td>
<td>0.688</td>
<td>0.635</td>
<td>0.819</td>
<td>0.826</td>
</tr>
<tr>
<td>(50, 100)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.756</td>
<td>0.596</td>
<td>0.667</td>
<td>0.849</td>
</tr>
<tr>
<td>(10, 10)</td>
<td>0.688</td>
<td>0.615</td>
<td>0.819</td>
<td>0.826</td>
</tr>
<tr>
<td>(50, 100)</td>
<td>0.555</td>
<td>0.635</td>
<td>0.823</td>
<td>0.826</td>
</tr>
<tr>
<td>(200, 200)</td>
<td>0.533</td>
<td>0.558</td>
<td>0.848</td>
<td>0.832</td>
</tr>
</tbody>
</table>

<sup>a</sup>AUROC: area under the receiver operating characteristic curve.

<sup>b</sup>SVM: support vector machine.

<sup>c</sup>Final algorithms selected by adjusting parameters.

<sup>d</sup>KNN: k-nearest neighbor.

<sup>e</sup>NB: naive Bayes.
Table 4. Performance according to the $\alpha$ value in the naïve Bayes model.

<table>
<thead>
<tr>
<th>$\alpha$ value</th>
<th>Sensitivity 5-year</th>
<th>Sensitivity 10-year</th>
<th>Specificity 5-year</th>
<th>Specificity 10-year</th>
<th>Accuracy 5-year</th>
<th>Accuracy 10-year</th>
<th>AUROC$^a$ 5-year</th>
<th>AUROC$^a$ 10-year</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (no smoothing)</td>
<td>0.800</td>
<td>0.731</td>
<td>0.848</td>
<td>0.828</td>
<td>0.844</td>
<td>0.819</td>
<td>0.824</td>
<td>0.779</td>
</tr>
<tr>
<td>1</td>
<td>0.822</td>
<td>0.731</td>
<td>0.848</td>
<td>0.828</td>
<td>0.846</td>
<td>0.819</td>
<td>0.835</td>
<td>0.779</td>
</tr>
<tr>
<td>10</td>
<td>0.822</td>
<td>0.731</td>
<td>0.850</td>
<td>0.834</td>
<td>0.848</td>
<td>0.824</td>
<td>0.836</td>
<td>0.782</td>
</tr>
<tr>
<td>20</td>
<td>0.800</td>
<td>0.731</td>
<td>0.850</td>
<td>0.834</td>
<td>0.846</td>
<td>0.824</td>
<td>0.825</td>
<td>0.782</td>
</tr>
<tr>
<td>30</td>
<td>0.800</td>
<td>0.731</td>
<td>0.852</td>
<td>0.834</td>
<td>0.848</td>
<td>0.824</td>
<td>0.826</td>
<td>0.782</td>
</tr>
<tr>
<td>100</td>
<td>0.800</td>
<td>0.731</td>
<td>0.854</td>
<td>0.840</td>
<td>0.850</td>
<td>0.828</td>
<td>0.827</td>
<td>0.784</td>
</tr>
<tr>
<td>200</td>
<td>0.756</td>
<td>0.692</td>
<td>0.860</td>
<td>0.845</td>
<td>0.852</td>
<td>0.831</td>
<td>0.807</td>
<td>0.769</td>
</tr>
</tbody>
</table>

$^a$AUROC: area under the receiver operating characteristic curve.

For predictions within 5 years, the AUROC was found to be 0.836 when $\alpha=10$, which was the highest performance compared with that before smoothing was applied ($\alpha=0$, AUROC 0.824). For predictions within 10 years, the AUROC was 0.784 when $\alpha=100$, which was the highest performance compared with that before smoothing was applied ($\alpha=0$, 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.
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**


Abbreviations

**AUROC:** area under the receiver operating characteristic curve  
**KNN:** k-nearest neighbor  
**KORCC:** KOran 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

©HyungMin Kim, Sun Jung Lee, So Jin Park, In Young Choi, Sung-Hoo Hong. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 01.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.
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 Buijs¹, MSc, PhD; Marten Pijl¹, PhD; Andreas Landgraf², PhD

¹Philips Research, Eindhoven, Netherlands
²Philips 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].

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>
<thead>
<tr>
<th>Classification example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case type</strong></td>
<td></td>
</tr>
<tr>
<td>Incident</td>
<td>Case where the subscriber is in need of help</td>
</tr>
<tr>
<td>Accidental</td>
<td>Subscriber accidentally pushed the help button</td>
</tr>
<tr>
<td>Test</td>
<td>Test call by subscriber</td>
</tr>
<tr>
<td><strong>Case reason</strong></td>
<td></td>
</tr>
<tr>
<td>Fall</td>
<td>Subscriber fell</td>
</tr>
<tr>
<td>Breathing problems</td>
<td>Subscriber has breathing problems</td>
</tr>
<tr>
<td>Heart problems</td>
<td>Subscriber has heart problems</td>
</tr>
<tr>
<td><strong>Case outcome</strong></td>
<td></td>
</tr>
<tr>
<td>Nurse</td>
<td>Nurse visit</td>
</tr>
<tr>
<td>Ambulance transport</td>
<td>Emergency medical services dispatched to bring subscriber to hospital</td>
</tr>
<tr>
<td>No assistance required</td>
<td>Subscriber did not require further assistance</td>
</tr>
</tbody>
</table>

**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.

**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 (i.e., 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.

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 (i.e., 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 the German 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.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Germany predictive model (this study)</th>
<th>US predictive model (from [9])</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training set</td>
<td>Test set</td>
</tr>
<tr>
<td># of unique PERS* users, n (%)</td>
<td>4644 (80)</td>
<td>116 (20)</td>
</tr>
<tr>
<td># of prediction windows, n (%)</td>
<td>96,273 (79.5)</td>
<td>24,847 (20.5)</td>
</tr>
<tr>
<td>Age in years, mean (SD)</td>
<td>84.0 (8.2)</td>
<td>83.6 (8.3)</td>
</tr>
<tr>
<td>Female gender, n (%)</td>
<td>2997 (64.5)</td>
<td>743 (64.0)</td>
</tr>
<tr>
<td>Years on PERS service, n (%)</td>
<td>2598 (55.9)</td>
<td>630 (54.3)</td>
</tr>
<tr>
<td>2-4</td>
<td>1079 (23.2)</td>
<td>264 (22.7)</td>
</tr>
<tr>
<td>4 or more</td>
<td>967 (20.8)</td>
<td>267 (23.0)</td>
</tr>
<tr>
<td>Number of PERS self-reported medical conditions, n (%)</td>
<td>265 (5.7)</td>
<td>73 (6.3)</td>
</tr>
<tr>
<td>1-2</td>
<td>1325 (28.5)</td>
<td>329 (28.3)</td>
</tr>
<tr>
<td>3-4</td>
<td>1397 (30.1)</td>
<td>370 (31.9)</td>
</tr>
<tr>
<td>5 or more</td>
<td>1657 (35.7)</td>
<td>389 (33.5)</td>
</tr>
<tr>
<td>30-day emergency hospital transport (% of prediction windows)</td>
<td>1506 (1.6)</td>
<td>350 (1.4)</td>
</tr>
</tbody>
</table>

*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 in 30 days. 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.

<table>
<thead>
<tr>
<th>Performance metric and threshold (percentile)</th>
<th>Germany predictive model (this study), % (95% CI)</th>
<th>US predictive model (adapted from [9]), % (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PPV</strong>&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>5.7 (4.9-6.4)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>9.4 (8.9-9.8)</td>
</tr>
<tr>
<td>95%</td>
<td>7.5 (6.3-8.8)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>13.6 (12.9-14.3)</td>
</tr>
<tr>
<td>99%</td>
<td>15.0 (10.7-19.3)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>26.2 (23.7-28.5)</td>
</tr>
<tr>
<td><strong>Sensitivity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>40.3 (35.1-45.4)</td>
<td>41.9 (40.0-43.9)</td>
</tr>
<tr>
<td>95%</td>
<td>26.9 (22.3-31.1)</td>
<td>30.3 (28.6-32.0)</td>
</tr>
<tr>
<td>99%</td>
<td>10.6 (7.4-14.0)</td>
<td>11.7 (10.5-13.0)</td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>90.4 (90.1-90.8)</td>
<td>90.8 (90.6-91.0)</td>
</tr>
<tr>
<td>95%</td>
<td>95.3 (95.1-95.6)</td>
<td>95.6 (95.6-95.7)</td>
</tr>
<tr>
<td>99%</td>
<td>99.1 (99.0-99.2)</td>
<td>99.3 (99.2-99.3)</td>
</tr>
<tr>
<td><strong>AUC</strong>&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.749 (0.721-0.777)</td>
<td>0.778 (0.769-0.788)</td>
</tr>
</tbody>
</table>

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

<sup>b</sup>Nonoverlapping 95% CI between Germany and US predictive models.

<sup>c</sup>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 $R^2=0.9935$ and $R^2=0.9992$ 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.

**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>
<thead>
<tr>
<th>Predictor category</th>
<th>Number of predictors</th>
<th>Total gain, %</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Germany</td>
<td>US</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case data-based predictors</td>
<td>48</td>
<td>62</td>
<td>72.9</td>
<td>87.7</td>
</tr>
<tr>
<td>Self-reported medical condi-</td>
<td>43</td>
<td>44</td>
<td>9.6</td>
<td>3.7</td>
</tr>
<tr>
<td>tions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other predictors</td>
<td>7</td>
<td>15</td>
<td>17.5</td>
<td>8.7</td>
</tr>
<tr>
<td>Total</td>
<td>98</td>
<td>121</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

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.
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


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

©Jorn op den Buijs, Marten Pijl, Andreas Landgraf. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 08.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.
Original Paper

A Novel Convolutional Neural Network for the Diagnosis and Classification of Rosacea: Usability Study

Zhixiang Zhao\textsuperscript{1,2*}, MD, PhD; Che-Ming Wu\textsuperscript{3*}, MEng; Shuping Zhang\textsuperscript{1,2}, MD, PhD; Fanping He\textsuperscript{1,2}, MD, PhD; Fangfen Liu\textsuperscript{1,2}, MD, PhD; Ben Wang\textsuperscript{1,2}, MD, PhD; Yingxue Hu\textsuperscript{1,2}, MD, PhD; Wei Shi\textsuperscript{1,2}, MD, PhD; Dan Jian\textsuperscript{1,2}, MD, PhD; Hongfu Xie\textsuperscript{1,2}, MD, PhD; Chao-Yuan Yeh\textsuperscript{1,2}, MD; Ji Li\textsuperscript{1,2,4,5*}, MD, PhD

\textsuperscript{1}Department of Dermatology, Xiangya Hospital of Central South University, Changsha, China
\textsuperscript{2}Hunan Key Laboratory of Aging Biology, Xiangya Hospital of Central South University, Changsha, China
\textsuperscript{3}aetherAI, Co Ltd, Taipei, Taiwan, China
\textsuperscript{4}National Clinical Research Center for Geriatric Disorders, Xiangya Hospital of Central South University, Changsha, China
\textsuperscript{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

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.

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, 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, angioiectasis, 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].

\[ y = \begin{cases} 
  0, & \text{if } x < 0 \\
  x, & \text{if } x \geq 0 
\end{cases} \]

y=x, if x≥0
y=0, 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.

**Figure 1.** Examples of 3 photos taken from 3 different angles for each patient.

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).

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 of 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.](https://medinform.jmir.org/2021/3/e23415)

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.
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).

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 grassroots 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.

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


Abbreviations

AUROC: area under the receiver operating characteristic curve
CNN: convolutional neural network
ETR: erythematotelangiectatic rosacea
PhR: phymatous rosacea
PPR: papulopustular rosacea

©Zhixiang Zhao, Che-Ming Wu, Shuping Zhang, Fanping He, Fangfen Liu, Ben Wang, Yingxue Huang, Wei Shi, Dan Jian, Hongfu Xie, Chao-Yuan Yeh, Ji Li. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 15.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.
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 infections.
of patients who were infected. All countries reported difficulties in estimating 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 patients 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 hospitals dedicated to patients with COVID. In Italy, one 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.

<table>
<thead>
<tr>
<th>Region</th>
<th>ICU beds, n</th>
<th>Population, n</th>
<th>ICU beds per citizen (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lombardia</td>
<td>1067</td>
<td>10,060,574</td>
<td>0.0106</td>
</tr>
<tr>
<td>Lazio</td>
<td>590</td>
<td>5,879,082</td>
<td>0.0100</td>
</tr>
<tr>
<td>Campania</td>
<td>350</td>
<td>5,801,692</td>
<td>0.0060</td>
</tr>
<tr>
<td>Sicilia</td>
<td>346</td>
<td>4,999,891</td>
<td>0.0069</td>
</tr>
<tr>
<td>Veneto</td>
<td>498</td>
<td>4,905,854</td>
<td>0.0102</td>
</tr>
<tr>
<td>Emilia Romagna</td>
<td>539</td>
<td>4,459,477</td>
<td>0.0121</td>
</tr>
<tr>
<td>Piemonte</td>
<td>320</td>
<td>4,356,406</td>
<td>0.0073</td>
</tr>
<tr>
<td>Puglia</td>
<td>210</td>
<td>4,029,053</td>
<td>0.0052</td>
</tr>
<tr>
<td>Toscana</td>
<td>450</td>
<td>3,729,641</td>
<td>0.0121</td>
</tr>
<tr>
<td>Calabria</td>
<td>110</td>
<td>1,947,131</td>
<td>0.0056</td>
</tr>
<tr>
<td>Sardegna</td>
<td>150</td>
<td>1,639,591</td>
<td>0.0091</td>
</tr>
<tr>
<td>Liguria</td>
<td>70</td>
<td>1,550,640</td>
<td>0.0045</td>
</tr>
<tr>
<td>Marche</td>
<td>108</td>
<td>1,525,271</td>
<td>0.0071</td>
</tr>
<tr>
<td>Abruzzo</td>
<td>73</td>
<td>1,311,580</td>
<td>0.0056</td>
</tr>
<tr>
<td>Friuli Venezia Giulia</td>
<td>80</td>
<td>1,215,220</td>
<td>0.0066</td>
</tr>
<tr>
<td>Trentino Alto Adige</td>
<td>71</td>
<td>1,072,276</td>
<td>0.0066</td>
</tr>
<tr>
<td>Umbria</td>
<td>30</td>
<td>882,015</td>
<td>0.0034</td>
</tr>
<tr>
<td>Basilicata</td>
<td>49</td>
<td>562,869</td>
<td>0.0087</td>
</tr>
<tr>
<td>Molise</td>
<td>30</td>
<td>305,617</td>
<td>0.0098</td>
</tr>
<tr>
<td>Valle D’Aosta</td>
<td>15</td>
<td>125,666</td>
<td>0.0119</td>
</tr>
<tr>
<td>Italy (total)</td>
<td>5156</td>
<td>60,359,546</td>
<td>0.0085</td>
</tr>
</tbody>
</table>

aICU: 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 $y = a \cdot e^{bx}$, 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.

<table>
<thead>
<tr>
<th>Observed ICU(^b) beds(^a), n</th>
<th>Predicted ICU beds(^c), n</th>
<th>Percentage increase (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>106</td>
<td>172</td>
<td>62</td>
</tr>
<tr>
<td>127</td>
<td>146</td>
<td>15</td>
</tr>
<tr>
<td>167</td>
<td>201</td>
<td>20</td>
</tr>
<tr>
<td>209</td>
<td>262</td>
<td>25</td>
</tr>
<tr>
<td>244</td>
<td>371</td>
<td>52</td>
</tr>
<tr>
<td>309</td>
<td>380</td>
<td>23</td>
</tr>
<tr>
<td>359</td>
<td>552</td>
<td>54</td>
</tr>
<tr>
<td>399</td>
<td>491</td>
<td>23</td>
</tr>
<tr>
<td>440</td>
<td>537</td>
<td>22</td>
</tr>
<tr>
<td>466</td>
<td>521</td>
<td>12</td>
</tr>
<tr>
<td>560</td>
<td>892</td>
<td>59</td>
</tr>
<tr>
<td>605</td>
<td>700</td>
<td>16</td>
</tr>
<tr>
<td>650</td>
<td>841</td>
<td>29</td>
</tr>
<tr>
<td>732</td>
<td>893</td>
<td>22</td>
</tr>
<tr>
<td>767</td>
<td>930</td>
<td>21</td>
</tr>
</tbody>
</table>

\(^a\)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


Abbreviations

ICU: intensive care unit

©Pietro H Guzzi, Giuseppe Tradigo, Pierangelo Velttri. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 09.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.
**Medical Morphology Training Using the Xuexi Tong Platform During the COVID-19 Pandemic: Development and Validation of a Web-Based Teaching Approach**

Qinlai Liu¹, PhD; Wenping Sun¹, MD; Changqing Du², MD; Leiyiing Yang¹, MD; Na Yuan¹, MD; Haiqing Cui², MD; Wengang Song³*, PhD; Li Ge²*, PhD

¹Department of Pathology, Shandong First Medical University & Shandong Academy of Medical Sciences, Tai’an, China
²Department of Histology and Embryology, Shandong First Medical University & Shandong Academy of Medical Sciences, Tai’an, China
³Department 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.

https://medinform.jmir.org/2021/3/e24497
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 delivering 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) pulmonary edema, (4) interstitial pneumonia, (5) chronic bronchitis, (6) chronic obstructive pulmonary emphysema, (7) silicois, (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 for 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 \( P < 0.05 \) 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.](image_url)

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.

高等医学院校首届大学生形态学绘图作品选

学校：山东第一医科大学 姓名：刘熙瑞 班级：2018级临床医学本科4班

Monkey Liver HE stain 400X

5. Interlobular artery 6. Interlobular vein 7. Interlobular bile duct
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 a 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.

<table>
<thead>
<tr>
<th>Survey question and course</th>
<th>Responses, n (%)(^a)</th>
<th>Very satisfied</th>
<th>A greater degree of satisfaction</th>
<th>Generally satisfied</th>
<th>Less satisfied</th>
<th>Dissatisfied</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What do you think of the current teaching mode? Are you satisfied?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology and Embryology</td>
<td>64 (26.2)</td>
<td>110 (45.1)</td>
<td>57 (23.4)</td>
<td>7 (2.9)</td>
<td>6 (2.5)</td>
<td></td>
</tr>
<tr>
<td>Pathology</td>
<td>78 (31.7)</td>
<td>125 (50.8)</td>
<td>36 (14.6)</td>
<td>3 (1.2)</td>
<td>4 (1.6)</td>
<td></td>
</tr>
<tr>
<td>2. Is the web-based platform satisfactory?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology and Embryology</td>
<td>63 (25.8)</td>
<td>110 (45.1)</td>
<td>58 (23.8)</td>
<td>8 (3.3)</td>
<td>5 (2.0)</td>
<td></td>
</tr>
<tr>
<td>Pathology</td>
<td>85 (34.6)</td>
<td>112 (45.5)</td>
<td>36 (14.6)</td>
<td>8 (3.3)</td>
<td>5 (2.0)</td>
<td></td>
</tr>
<tr>
<td>3. Do you think the current teaching goals are clear?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology and Embryology</td>
<td>73 (29.9)</td>
<td>109 (44.7)</td>
<td>51 (20.9)</td>
<td>7 (2.9)</td>
<td>4 (1.6)</td>
<td></td>
</tr>
<tr>
<td>Pathology</td>
<td>92 (37.4)</td>
<td>113 (45.9)</td>
<td>36 (14.6)</td>
<td>3 (1.2)</td>
<td>2 (0.8)</td>
<td></td>
</tr>
<tr>
<td>4. Do you think the current teaching arrangements are reasonable?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology and Embryology</td>
<td>71 (29.1)</td>
<td>114 (46.7)</td>
<td>50 (20.5)</td>
<td>7 (2.9)</td>
<td>2 (0.8)</td>
<td></td>
</tr>
<tr>
<td>Pathology</td>
<td>102 (41.5)</td>
<td>116 (47.2)</td>
<td>24 (9.8)</td>
<td>2 (0.8)</td>
<td>2 (0.8)</td>
<td></td>
</tr>
<tr>
<td>5. What is your attitude toward the quality of the web-based teaching microvideos?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology and Embryology</td>
<td>53 (21.7)</td>
<td>114 (46.7)</td>
<td>67 (27.5)</td>
<td>7 (2.9)</td>
<td>3 (1.2)</td>
<td></td>
</tr>
<tr>
<td>Pathology</td>
<td>88 (35.8)</td>
<td>111 (45.1)</td>
<td>37 (15.0)</td>
<td>7 (2.9)</td>
<td>3 (1.2)</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)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.

<table>
<thead>
<tr>
<th>Survey content and course</th>
<th>Responses, n (%)</th>
<th>Very helpful, better than traditional classroom teaching</th>
<th>Helpful, similar to traditional classroom teaching</th>
<th>Helpful, but not as good as traditional classroom teaching</th>
<th>Not helpful</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. In terms of improving learning efficiency and quality and solving difficult problems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology and Embryology</td>
<td>47 (19.3)</td>
<td>96 (39.3)</td>
<td>95 (38.9)</td>
<td>6 (2.5)</td>
<td></td>
</tr>
<tr>
<td>Pathology</td>
<td>57 (23.2)</td>
<td>97 (39.4)</td>
<td>88 (35.8)</td>
<td>4 (1.6)</td>
<td></td>
</tr>
<tr>
<td>2. In terms of mobilizing students’ learning enthusiasm and initiative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology and Embryology</td>
<td>48 (19.7)</td>
<td>93 (38.1)</td>
<td>94 (38.5)</td>
<td>9 (3.7)</td>
<td></td>
</tr>
<tr>
<td>Pathology</td>
<td>58 (23.6)</td>
<td>90 (36.6)</td>
<td>89 (36.2)</td>
<td>9 (3.7)</td>
<td></td>
</tr>
<tr>
<td>3. In terms of knowledge summary, integration, application, and systematic understanding and memory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology and Embryology</td>
<td>45 (18.4)</td>
<td>97 (39.8)</td>
<td>94 (38.5)</td>
<td>8 (3.3)</td>
<td></td>
</tr>
<tr>
<td>Pathology</td>
<td>57 (23.1)</td>
<td>100 (40.7)</td>
<td>84 (34.1)</td>
<td>5 (2.0)</td>
<td></td>
</tr>
<tr>
<td>4. For unity and cooperation, language and communication skills, and problem-solving</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology and Embryology</td>
<td>46 (18.9)</td>
<td>96 (39.3)</td>
<td>92 (37.7)</td>
<td>10 (4.1)</td>
<td></td>
</tr>
<tr>
<td>Pathology</td>
<td>61 (24.8)</td>
<td>92 (37.4)</td>
<td>84 (34.1)</td>
<td>9 (3.7)</td>
<td></td>
</tr>
</tbody>
</table>

aPercentages 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.

<table>
<thead>
<tr>
<th>Survey item and course</th>
<th>Responses, n (%)</th>
<th>Very helpful, better than web-based classroom teaching</th>
<th>Helpful, similar to web-based classroom teaching</th>
<th>Helpful, but inferior to web-based classroom teaching</th>
<th>Not helpful</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. In terms of improving learning efficiency and quality and solving difficult problems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology and Embryology</td>
<td>118 (46.6)</td>
<td>105 (41.5)</td>
<td>25 (9.9)</td>
<td>5 (2.0)</td>
<td></td>
</tr>
<tr>
<td>Pathology</td>
<td>115 (46)</td>
<td>95 (38.0)</td>
<td>39 (15.6)</td>
<td>1 (0.4)</td>
<td></td>
</tr>
<tr>
<td>2. In terms of mobilizing students’ learning enthusiasm and initiative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology and Embryology</td>
<td>129 (51.0)</td>
<td>99 (39.1)</td>
<td>20 (7.9)</td>
<td>5 (2.0)</td>
<td></td>
</tr>
<tr>
<td>Pathology</td>
<td>121 (48.4)</td>
<td>93 (37.2)</td>
<td>35 (14.0)</td>
<td>1 (0.4)</td>
<td></td>
</tr>
<tr>
<td>3. In terms of knowledge summary, integration, application, and systematic understanding and memory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology and Embryology</td>
<td>123 (48.6)</td>
<td>95 (37.5)</td>
<td>27 (10.7)</td>
<td>8 (3.2)</td>
<td></td>
</tr>
<tr>
<td>Pathology</td>
<td>115 (46.0)</td>
<td>92 (36.8)</td>
<td>38 (15.2)</td>
<td>5 (2.0)</td>
<td></td>
</tr>
<tr>
<td>4. For unity and cooperation, language and communication skills, problem-solving</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology and Embryology</td>
<td>138 (54.5)</td>
<td>89 (35.2)</td>
<td>18 (7.1)</td>
<td>8 (3.2)</td>
<td></td>
</tr>
<tr>
<td>Pathology</td>
<td>131 (52.4)</td>
<td>89 (35.6)</td>
<td>29 (11.6)</td>
<td>1 (0.4)</td>
<td></td>
</tr>
</tbody>
</table>

aPercentages 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.

![Pie chart showing proportions of teaching methods]

**Figure 8.** The proportions of students who chose different teaching methods of pathology after the COVID-19 epidemic.

![Pie chart showing proportions of teaching methods]
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.
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.

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.
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.

References


https://medinform.jmir.org/2021/3/e24497

JMIRED Inf Med 2021 | vol. 9 | iss. 3 | e24497 | p.258

(page number not for citation purposes)


Abbreviations

MOOC: massive open online course

©Qinlai Liu, Wenping Sun, Changqing Du, Leiying Yang, Na Yuan, Haiqing Cui, Wengang Song, Li Ge. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 15.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.
Original Paper

Systematic Delineation of Media Polarity on COVID-19 Vaccines in Africa: Computational Linguistic Modeling Study

Sefater Gbashi1*, PhD; Oluwafemi Ayodeji Adebo1*, PhD; Wesley Doorsamy2*, PhD; Patrick Berka Njobeh1*, PhD

1Faculty of Science, University of Johannesburg, Johannesburg, South Africa
2Institute 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, 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

https://medinform.jmir.org/2021/3/e22916 JMIR Med Inform 2021 | vol. 9 | iss. 3 | e22916 | p.263 (page number not for citation purposes)
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.

**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.

<table>
<thead>
<tr>
<th>Period in 2020</th>
<th>TextBlob</th>
<th>VADER&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Word2Vec-BiLSTM&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
<td>Neutral</td>
</tr>
<tr>
<td>February 2</td>
<td>77.78</td>
<td>22.22</td>
<td>0.00</td>
</tr>
<tr>
<td>February 9</td>
<td>16.67</td>
<td>83.33</td>
<td>0.00</td>
</tr>
<tr>
<td>February 16</td>
<td>66.67</td>
<td>0.00</td>
<td>33.33</td>
</tr>
<tr>
<td>February 23</td>
<td>100.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>March 1</td>
<td>66.67</td>
<td>26.67</td>
<td>6.67</td>
</tr>
<tr>
<td>March 8</td>
<td>72.73</td>
<td>0.00</td>
<td>27.27</td>
</tr>
<tr>
<td>March 15</td>
<td>49.18</td>
<td>19.67</td>
<td>31.15</td>
</tr>
<tr>
<td>March 22</td>
<td>89.74</td>
<td>5.13</td>
<td>5.13</td>
</tr>
<tr>
<td>March 29</td>
<td>53.42</td>
<td>6.85</td>
<td>39.73</td>
</tr>
<tr>
<td>April 5</td>
<td>57.97</td>
<td>17.39</td>
<td>24.64</td>
</tr>
<tr>
<td>April 12</td>
<td>49.09</td>
<td>21.82</td>
<td>29.09</td>
</tr>
<tr>
<td>April 19</td>
<td>75.93</td>
<td>7.41</td>
<td>16.67</td>
</tr>
<tr>
<td>April 26</td>
<td>48.65</td>
<td>32.43</td>
<td>18.92</td>
</tr>
<tr>
<td>May 3</td>
<td>53.17</td>
<td>17.46</td>
<td>29.37</td>
</tr>
<tr>
<td>Average polarity</td>
<td>62.69</td>
<td>18.60</td>
<td>18.71</td>
</tr>
</tbody>
</table>

<sup>a</sup>VADER: Valence Aware Dictionary and Sentiment Reasoner.
<sup>b</sup>BiLSTM: bidirectional long short-term memory.
### Table 2. Sentiment polarities in media communications (Twitter posts) analyzed using the TextBlob, VADER, and Word2Vec-BiLSTM models.

<table>
<thead>
<tr>
<th>Period in 2020</th>
<th>TextBlob</th>
<th>VADER(^a)</th>
<th>Word2Vec-BiLSTM(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
<td>Neutral</td>
</tr>
<tr>
<td>February 2</td>
<td>100.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>February 9</td>
<td>37.50</td>
<td>37.50</td>
<td>25.00</td>
</tr>
<tr>
<td>February 16</td>
<td>63.64</td>
<td>9.09</td>
<td>27.27</td>
</tr>
<tr>
<td>February 23</td>
<td>50.00</td>
<td>50.00</td>
<td>0.00</td>
</tr>
<tr>
<td>March 1</td>
<td>83.33</td>
<td>0.00</td>
<td>16.67</td>
</tr>
<tr>
<td>March 8</td>
<td>44.44</td>
<td>22.22</td>
<td>33.33</td>
</tr>
<tr>
<td>March 15</td>
<td>85.71</td>
<td>0.00</td>
<td>14.29</td>
</tr>
<tr>
<td>March 22</td>
<td>66.67</td>
<td>0.00</td>
<td>33.33</td>
</tr>
<tr>
<td>March 29</td>
<td>38.57</td>
<td>27.14</td>
<td>34.29</td>
</tr>
<tr>
<td>April 5</td>
<td>36.49</td>
<td>20.27</td>
<td>43.24</td>
</tr>
<tr>
<td>April 12</td>
<td>33.33</td>
<td>41.67</td>
<td>25.00</td>
</tr>
<tr>
<td>April 19</td>
<td>66.67</td>
<td>6.67</td>
<td>26.67</td>
</tr>
<tr>
<td>April 26</td>
<td>48.02</td>
<td>23.41</td>
<td>28.57</td>
</tr>
<tr>
<td>May 3</td>
<td>38.18</td>
<td>15.15</td>
<td>46.67</td>
</tr>
<tr>
<td>Average polarity</td>
<td>56.61</td>
<td>18.08</td>
<td>25.31</td>
</tr>
</tbody>
</table>

\(^a\)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 14, 2020, in Egypt [93], while the first confirmed case in sub-Saharan Africa, Nigeria to be specific, was reported on February 14, 2020 [93].

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, 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 falsehoods. This is because it is expected that, if factual and more evidence-based 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.

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


62. Sayce D. The number of tweets per day in 2019. David Sayce. URL: https://dsayce.wordpress.com/ [accessed 2020-05-08]


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-db413a39445 [accessed 2020-06-16]


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
Emotional Attitudes of Chinese Citizens on Social Distancing During the COVID-19 Outbreak: Analysis of Social Media Data

Lining Shen1,2,3*, PhD; Rui Yao1*, MS; Wenli Zhang1, MS; Richard Evans4, PhD; Guang Cao1, MS; Zhiguo Zhang1,2, PhD

1School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science & Technology, Wuhan, China
2Hubei Provincial Research Center for Health Technology Assessment, Wuhan, China
3Institute of Smart Health, Huazhong University of Science & Technology, Wuhan, China
4College 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 (χ² = 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 (χ² = 19,084.73, P < .001), as well as emotional polarity between genders (χ² = 1784.59, P < .001) and emotional polarity between spatial locations (χ² = 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.
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>
<thead>
<tr>
<th>Measures categories</th>
<th>Hashtags</th>
</tr>
</thead>
</table>
| 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# |
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.512</td>
<td>0.506</td>
<td>0.509</td>
</tr>
<tr>
<td>CNN&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.619</td>
<td>0.602</td>
<td>0.607</td>
</tr>
<tr>
<td>LSTM&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.664</td>
<td>0.664</td>
<td>0.658</td>
</tr>
<tr>
<td>Bi-LSTM&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.701</td>
<td>0.699</td>
<td>0.701</td>
</tr>
</tbody>
</table>

<sup>a</sup>SVM: support vector machine.
<sup>b</sup>CNN: convolutional neural network.
<sup>c</sup>LSTM: long short-term memory.
<sup>d</sup>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 Dependency 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(\(P_{\text{Emo}}\)) = \(\beta_0 + \beta_1 \text{Sex} + \beta_2 \text{Age} + \beta_3 \text{Space1} + \beta_4 \text{Space2} + \beta_5 \text{Ln}(N_{\text{Fan}}) + \beta_6 \text{Ln}(N_{\text{Follow}}) + \beta_7 \text{Ln}(N_{\text{Post}}) + \beta_8 \text{Time}_\text{Reg} + \beta_9 \text{Age} \times T_{\text{SDM}} + \beta_{10} \text{Space1} \times T_{\text{SDM}} + \beta_{11} \text{Space2} \times T_{\text{SDM}} + \beta_{12} \text{Ln}(N_{\text{Fan}}) \times T_{\text{SDM}} + \beta_{13} \text{Ln}(N_{\text{Follow}}) \times T_{\text{SDM}} + \beta_{14} \text{Ln}(N_{\text{Post}}) \times T_{\text{SDM}} + \beta_{15} \text{Time}_\text{Reg} \times T_{\text{SDM}} + \epsilon

Table 3. Description of variables that influence Sina Weibo users’ emotions.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable symbol</th>
<th>Description and coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explained variable</td>
<td></td>
<td>Sina Weibo users’ emotional polarity</td>
</tr>
<tr>
<td>Emotional polarity</td>
<td>P_\text{Emo}</td>
<td></td>
</tr>
<tr>
<td>Control variable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>\text{Sex}</td>
<td>The sex of Sina Weibo users. The coding is as follows: 0=female and 1=male (reference group)</td>
</tr>
<tr>
<td>Explanatory variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>\text{Age}</td>
<td>The age of Sina Weibo users (range 16-65 years)</td>
</tr>
</tbody>
</table>
| Spatial location  | \text{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    | \text{Ln}(N_{\text{Fan}}) | The number of fans of Sina Weibo users; smoothed logarithmically                        |
| Number of follows | \text{Ln}(N_{\text{Follow}}) | The number of other users that Sina Weibo users follow; smoothed logarithmically        |
| Number of post    | \text{Ln}(N_{\text{Post}}) | The number of posts shared by Sina Weibo users; smoothed logarithmically                |
| Registration year | \text{Time}_\text{Reg} | The registration year of Sina Weibo users’ accounts                                    |
| Moderator variable| \text{T}_{\text{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.
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 ($\chi^2 = 1784.59, P < .001$) shows that the emotional expression of different genders varied.

Table 4. Proportion of users’ emotional polarities across different genders (N=63,169).

<table>
<thead>
<tr>
<th>Sex</th>
<th>Positive</th>
<th>Neutral</th>
<th>Negative</th>
<th>Anger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male (n=19,016), n (%)</td>
<td>1512 (8.0)</td>
<td>4505 (23.7)</td>
<td>2896 (15.2)</td>
<td>4467 (23.5)</td>
</tr>
<tr>
<td>Female (n=44,153), n (%)</td>
<td>4383 (9.9)</td>
<td>16,936 (38.4)</td>
<td>4207 (9.5)</td>
<td>9601 (21.7)</td>
</tr>
</tbody>
</table>

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 ($\chi^2 = 1659.67, P < .001$) 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).](image)

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 school, 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 ($\chi^2 = 19,084.73, P < .001$) showed that Sina Weibo users have varying emotional expressions for different topics.
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).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Negative, OR(^a) (95% CI)</th>
<th>Positive, OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex (reference group: male)</td>
<td>1.18 (1.05-1.32(^b))</td>
<td>2.21 (1.96-2.49(^b))</td>
</tr>
<tr>
<td>Age</td>
<td>0.99 (0.98-0.99(^b))</td>
<td>0.97 (0.98-0.98(^b))</td>
</tr>
<tr>
<td>Space0 (reference group)</td>
<td>_(^c)</td>
<td>_</td>
</tr>
<tr>
<td>Space1</td>
<td>1.18 (1.02-1.38(^b))</td>
<td>0.52 (0.44-0.62(^b))</td>
</tr>
<tr>
<td>Space2</td>
<td>1.09 (0.97-1.22)</td>
<td>1.10 (0.98-1.24)</td>
</tr>
<tr>
<td>Time_Reg(^d)</td>
<td>1.02 (0.99-1.04)</td>
<td>0.90 (0.88-0.93(^b))</td>
</tr>
<tr>
<td>Ln(N_Fan)(^e)</td>
<td>0.97 (0.94-1.01)</td>
<td>0.98 (0.94-0.99(^b))</td>
</tr>
<tr>
<td>Ln(N_Follow)(^f)</td>
<td>1.01 (1.02-1.08(^b))</td>
<td>1.05 (0.99-1.12)</td>
</tr>
<tr>
<td>Ln(N_Post)(^g)</td>
<td>0.96 (0.93-0.99(^b))</td>
<td>1.06 (1.03-1.10(^b))</td>
</tr>
<tr>
<td>SDM1(^h)</td>
<td>0.49 (0.31-0.78(^b))</td>
<td>7.66 (4.99-11.77(^b))</td>
</tr>
<tr>
<td>SDM2(^i)</td>
<td>0.46 (0.28-0.75(^b))</td>
<td>6.75 (4.14-11.00(^b))</td>
</tr>
<tr>
<td>SDM3(^j)</td>
<td>0.59 (0.38-0.92(^b))</td>
<td>1.08 (0.73-1.59)</td>
</tr>
<tr>
<td>SDM4(^k)</td>
<td>0.92 (0.46-1.81)</td>
<td>0.70 (0.38-1.28)</td>
</tr>
<tr>
<td>SDM5(^l)</td>
<td>0.56 (0.36-0.88(^b))</td>
<td>8.63 (5.68-13.11(^b))</td>
</tr>
<tr>
<td>SDM6(^m) (reference group)</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

\(^a\)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 (\(\chi^2\)\(_{48}\)=4994.56, P<.001), and the Hosmer-Lemeshow goodness-of-fit was 0.24 (\(\chi^2\)\(_{8}\)=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. 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.

<table>
<thead>
<tr>
<th>Variable</th>
<th>The measures make the explanatory variables have a positive tendency toward emotions</th>
<th>The measures make the explanatory variables have a negative tendency toward emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>SDM1&lt;sup&gt;a&lt;/sup&gt;, SDM2&lt;sup&gt;b&lt;/sup&gt;, SDM3&lt;sup&gt;c&lt;/sup&gt;</td>
<td>SDM4&lt;sup&gt;d&lt;/sup&gt;, SDM5&lt;sup&gt;e&lt;/sup&gt;, SDM6&lt;sup&gt;f&lt;/sup&gt;</td>
</tr>
<tr>
<td>Space1</td>
<td>SDM1, SDM2</td>
<td>SDM3, SDM4, SDM5</td>
</tr>
<tr>
<td>Time_Reg&lt;sup&gt;g&lt;/sup&gt;</td>
<td>SDM1, SDM2, SDM3, SDM6</td>
<td>SDM5</td>
</tr>
<tr>
<td>Ln(N_Fan)&lt;sup&gt;h&lt;/sup&gt;</td>
<td>SDM2</td>
<td>SDM1, SDM3</td>
</tr>
<tr>
<td>Ln(N_Follow)&lt;sup&gt;i&lt;/sup&gt;</td>
<td>SDM1</td>
<td>SDM2, SDM3</td>
</tr>
<tr>
<td>Ln(N_Post)&lt;sup&gt;j&lt;/sup&gt;</td>
<td>SDM1, SDM3, SDM5</td>
<td>SDM2, SDM6</td>
</tr>
</tbody>
</table>

<sup>a</sup>SDM1: delaying the resumption of work and school.
<sup>b</sup>SDM2: travel restrictions.
<sup>c</sup>SDM3: traffic restrictions.
<sup>d</sup>SDM4: closing public spaces.
<sup>e</sup>SDM5: community containment.
<sup>f</sup>SDM6: extending the Lunar New Year holiday.
<sup>g</sup>Time_Reg: registration year.
<sup>h</sup>Ln(N_Fan): logarithm of fan numbers.
<sup>i</sup>Ln(N_Follow): logarithm of follow numbers.
<sup>j</sup>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.

References


Abbreviations

**Bi-LSTM**: bi-directional long short-term memory

**CNN**: convolutional neural network

**LSTM**: long short-term memory

**NB**: naïve Bayes

**NRC**: National Research Council of Canada

**OCC**: Ortony-Clore-Collins model

**RQ**: research question

**SARS**: severe acute respiratory syndrome

**SVM**: support vector machine

©Lining Shen, Rui Yao, Wenli Zhang, Richard Evans, Guang Cao, Zhiguo Zhang. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 16.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on http://medinform.jmir.org/, as well as this copyright and license information must be included.