Contents

Reviews

Using Machine Learning Technologies in Pressure Injury Management: Systematic Review (e25704)
Mengyao Jiang, Yuxia Ma, Siyi Guo, Liuji Jin, Lin Lv, Lin Han, Ning An. 3

Physicians’ Use of the Computerized Physician Order Entry System for Medication Prescribing: Systematic Review (e22923)
Asra Mogharbel, Dawn Dowding, John Ainsworth. 121

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 Zühlig. 13

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. 27

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. 37

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. 54

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 Mogililgan, Stephen Leslie, Charles Knoery, James Shand, Adesh Ramsewak, Divyesh Sharma, Anne McShane, Khaled Rjoob, Aaron Peace. 64

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 Bohnec, Carlos Cavero Barca, Maria Giancarenrri, Els Clays, Amos Dawodc, Jan Derboven, Delphine De Smedt, Erik Dovgan, Jure Lampc, Flavia Marinoc, Miha Mlakarc, Giovanni Poggia, Paolo Puaddc, Juan Rodríguezc, Michele Schiartitc, Gašper Slapnicarc, Karin Slegerc, Gennaro Tartrasco, Jakob Vali, Aljoša Vodopija. 75
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 María Cassinello, Anselmo Andrés-Martín, Francisco Sánchez Laguna, Roman Villegas, Francisco Pérez León, Jesús Moreno Conde, Carlos Parra Calderón. .......................................................... 94

Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study (e21331)
Alberto Tena, Francec Claria, Francesc Solsona, Einar Meister, Monica Povedano. .......................................................... 103

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. .......................................................... 133

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. .......................................................... 146

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. .......................................................... 158

Regional Resource Assessment During the COVID-19 Pandemic in Italy: Modeling Study (e18933)
Pietro Guzzi, Giuseppe Tradigo, Pierangelo Veltri. .......................................................... 170
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 ASsessment Tool (PROBAST).

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

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

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

KEYWORDS
pressure injuries; pressure ulcer; pressure sore; pressure damage; decubitus ulcer; decubitus sore; bedsore; artificial intelligence; machine learning; neural network; support vector machine; natural language processing; Naive Bayes; bayesian learning; support
vector; random forest; boosting; deep learning; machine intelligence; computational intelligence; computer reasoning; management; systematic review

**Introduction**

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

Artificial intelligence (AI) has been exerting a positive impact on daily living [17]. Moreover, machine learning (ML) is a way to achieve AI. Over the past two decades, ML has progressed from a laboratory curiosity to practical tools commonly applied in the medical field [18,19]. ML will continue to contribute to improving prognosis and diagnostic accuracies, even potentially taking on some of the work of medical practitioners’ [20,21].

While researchers have developed various novel methods for PI management [22], there is no systematic review to our knowledge that evaluates current ML technologies used in PI management.

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

**Methods**

**Protocol**

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

**Search Strategy**

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

**Textbox 1. Search strategy and search terms used.**

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

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

**Inclusion and Exclusion Criteria**

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

**Study Selection Methods**

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

**Data Extraction**

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

**Quality Appraisal**

The methodological quality of the included studies was assessed independently by two investigators using the Prediction model Risk Of Bias ASeessment Tool (PROBAST) [24].
Disagreements were resolved by discussion. The PROBAST was designed to assess the risk of bias and applicability of diagnostic and prognostic prediction model studies, and it includes 20 signaling questions to judge the risk of bias from four domains (participants, predictors, outcome, and analysis). The risk of bias is judged as low, high, or unclear. If one domain is found to have a high risk of bias, the overall risk of bias is judged as high. Similarly, if one domain is assessed as unclear, the overall risk of bias is judged as unclear even if all other domains are assessed to have a low risk of bias.

Results

Study Process

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

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.

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 bedbound 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.](https://medinform.jmir.org/2021/3/e25704)

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 PI s. 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 PI s in their early stage via microclimate (eg, moisture, temperature, etc), not just via images. The current research is focused on classifying the wound tissue, and it is necessary to combine the percentage of the different tissue with the grading of PI to define the severity of PI. It is better to rely on objective indicators than to rely on human experience.

Future Research

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

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

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

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

Conclusions

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

Acknowledgments

This work was partially supported by the National Nature Science Foundation of China (grants 71363004, 71663002, and 71704071), National Research Training Program of Gansu provincial hospital (19SYPYA-4), the Fundamental Research Funds for the Central Universities (Izuibky-2018-ct05 and Izuibky-2019-58), the National Key R&D Program of China (2018YFB1003204), the Anhui Provincial Key Technologies R&D Program (1804b06020378), and the Program of Introducing Talents of Discipline to Universities (111 program) (B14025).
Conflicts of Interest
None declared.

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

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

References


Abbreviations

AI: artificial intelligence
CDBM: China Biomedical Literature Database
CINAHL: Cumulative Index to Nursing and Allied Health Literature
CNKI: China National Knowledge Infrastructure
DT: decision tree
Using Machine Learning Technologies in Pressure Injury Management: Systematic Review

JMIR Med Inform 2021;9(3):e25704
URL: https://medinform.jmir.org/2021/3/e25704
doi:10.2196/25704
PMID:

©Mengyao Jiang, Yuxia Ma, Siyi Guo, Liuqi Jin, Lin Lv, Lin Han, Ning An. 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.
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.

Data Collection

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

Ethical Aspects

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

Variables and Measures

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

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

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

Data Analyses

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

We report unadjusted (crude) associations and 2 adjusted models: (1) not including staffing and resources adequacy and (2) including staffing and resources adequacy. To compare the models’ relative fits, we used Akaike information criterion; a lower value indicates a better fit. Data analyses were performed with R (version 3.4.2; R Foundation for Statistical Computing, 2017) using the 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(^a), 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></td>
<td></td>
<td></td>
</tr>
<tr>
<td></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>

\(^a\)NH: nursing home.
Variable Result Description

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

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

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

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

<table>
<thead>
<tr>
<th>6 items on care workers’ perceptions of the electronic health record system’s usefulness and sufficiency of the number of computers on the units</th>
<th>Total (N=1975), n (%)</th>
<th>Registered nurses (n=966)</th>
<th>Licensed practical nurses (n=1058)</th>
<th>( P ) value(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>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>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>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>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>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>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>

\( ^a \)Percentage agreement (agree and strongly agree).
\( ^b \)Chi-square test, \( P<.05 \) highlighted in italic.

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

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

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

<table>
<thead>
<tr>
<th>Care activities rationed by care workers in the last 7 days</th>
<th>Activity not necessary, n (%)</th>
<th>Never, n (%)</th>
<th>Seldom, n (%)</th>
<th>Sometimes, n (%)</th>
<th>Often, 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>
<td>410 (21.2)</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>
<td>325 (20.19)</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>
<td>279 (14.48)</td>
</tr>
</tbody>
</table>
Table 4. Characteristics of variables under study (N=1975).

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

<sup>a</sup> ICC1: intraclass correlation coefficient 1.

**Factors Associated With Implicit Rationing of Nursing Care Documentation**

In the crude models (Table 5), as well as models 1 and 2 (Table 6), care workers’ perceptions of both the EHR system’s usefulness and whether a sufficient number of computers were available were significantly associated with implicit rationing of nursing care documentation. More positive care workers’ perceptions of the EHR system’s usefulness ($\beta=-.12; 95\% \text{ CI} -0.17 \text{ to } -0.06$) and of the sufficiency of the number of computers ($\beta=-.09; 95\% \text{ CI} -0.12 \text{ to } -0.06$) 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(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Explanatory variables</strong></td>
<td></td>
</tr>
<tr>
<td>Care workers’ perception of the electronic health record system’s</td>
<td>(\beta) (95% CI)</td>
</tr>
<tr>
<td>usefulness</td>
<td>(-.31) ((-.36) to (-.26))</td>
</tr>
<tr>
<td>Care workers’ perception of whether sufficient numbers of computers were</td>
<td>(-.19) ((-.21) to (-.16))</td>
</tr>
<tr>
<td>available on their units</td>
<td></td>
</tr>
<tr>
<td><strong>Control variables</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Facility characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>Language region</td>
<td>0.18 ((-0.03) to 0.40)</td>
</tr>
<tr>
<td>Nursing home size</td>
<td>(-.03) ((-0.14) to 0.08)</td>
</tr>
<tr>
<td>Profit status</td>
<td>(-.04) ((-0.19) to 0.12)</td>
</tr>
<tr>
<td>Electronic health record system</td>
<td>0.01 ((-0.01)-0.04)</td>
</tr>
<tr>
<td><strong>Unit characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>Staffing levels</td>
<td>0 ((-0.01) to 0.00)</td>
</tr>
<tr>
<td>Skill mix levels</td>
<td>0 ((-0.01) to 0.00)</td>
</tr>
<tr>
<td><strong>Work environment</strong></td>
<td></td>
</tr>
<tr>
<td>Leadership</td>
<td>(-.37) ((-0.44) to (-0.31))</td>
</tr>
<tr>
<td>Staffing and resources adequacy</td>
<td>(-.63) ((-0.69) to (-0.58))</td>
</tr>
<tr>
<td>Safety and teamwork climate</td>
<td>(-.39) ((-0.46) to (-0.34))</td>
</tr>
<tr>
<td><strong>Care workers’ characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>(-.07) ((-0.21) to 0.06)</td>
</tr>
<tr>
<td>Age</td>
<td>0.01 ((-0.02) to 0.03)</td>
</tr>
<tr>
<td>Educational background</td>
<td>(-.08) ((-0.16) to (-0.01))</td>
</tr>
<tr>
<td>Professional experience</td>
<td>0.04 ((-0.01) to 0.08)</td>
</tr>
<tr>
<td>Employment level</td>
<td>(-.04) ((-0.09) to 0.03)</td>
</tr>
<tr>
<td>Fixed effects (intercept)</td>
<td>2.39 (2.32 to 2.47)</td>
</tr>
</tbody>
</table>

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

\(^b\)P value less than .05.

Higher ratings of leadership and safety teamwork climate were significantly associated with lower levels of implicit rationing of nursing care documentation only in model 1 (not accounting for staffing and resource adequacy). In model 2, care worker–perceived staffing and resources adequacy was the strongest explanatory factor, that is, higher ratings for staffing and resources adequacy were associated with lower levels of implicit rationing of nursing care documentation (\(\beta=\)-.52; 95% CI \(-0.58\) to \(-0.45\)). Moreover, care workers’ educational backgrounds were significantly associated with implicit rationing of nursing care documentation in both models (Table 6), with licensed practical nurses in both cases reporting lower levels of rationing of nursing care documentation than RNs (\(\beta=\)-.09; 95% CI \(-0.15\) to \(-0.02\)).
Table 6. Implicit rationing of nursing care documentation regressed on care workers’ perceptions of their electronic health record systems and the sufficiency of the number of computers, along with facility, unit and care worker characteristics and staffing variables, work environment, and teamwork and safety climate.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Multiple model 1&lt;sup&gt;a&lt;/sup&gt; (without staffing and resources adequacy)</th>
<th>Multiple model 2&lt;sup&gt;a&lt;/sup&gt; (with staffing and resources adequacy)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Explanatory variables</strong></td>
<td>β (95% CI)</td>
<td>β (95% CI)</td>
</tr>
<tr>
<td>Care workers’ perception of the EHR&lt;sup&gt;b&lt;/sup&gt; system’s usefulness</td>
<td>−.14&lt;sup&gt;c&lt;/sup&gt; (−0.20 to −0.09)</td>
<td>−.12&lt;sup&gt;c&lt;/sup&gt; (−0.17 to −0.06)</td>
</tr>
<tr>
<td>Care workers’ perception of whether sufficient numbers of computers were available on their units</td>
<td>−.12&lt;sup&gt;c&lt;/sup&gt; (−0.15 to −0.09)</td>
<td>−.09&lt;sup&gt;c&lt;/sup&gt; (−0.12 to −0.06)</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>_&lt;sup&gt;d&lt;/sup&gt;</td>
<td>_</td>
</tr>
<tr>
<td>Nursing home size</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Profit status</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>EHR system</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Unit characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staffing levels</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Skill mix levels</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Work environment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership</td>
<td>−.12&lt;sup&gt;c&lt;/sup&gt; (−0.21 to −0.04)</td>
<td>0.08 (−0.04 to 0.12)</td>
</tr>
<tr>
<td>Staffing and resources adequacy</td>
<td>—</td>
<td>−.52&lt;sup&gt;c&lt;/sup&gt; (−0.58 to −0.45)</td>
</tr>
<tr>
<td>Safety and teamwork climate</td>
<td>−.20&lt;sup&gt;c&lt;/sup&gt; (−0.27 to −0.12)</td>
<td>−.08&lt;sup&gt;c&lt;/sup&gt; (−0.15 to −0.01)</td>
</tr>
<tr>
<td><strong>Care workers’ characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Age</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Educational background</td>
<td>−.08&lt;sup&gt;c&lt;/sup&gt; (−0.16 to −0.02)</td>
<td>−.09&lt;sup&gt;c&lt;/sup&gt; (−0.15 to −0.02)</td>
</tr>
<tr>
<td>Professional experience</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Employment level</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Fixed effects (intercept)</td>
<td>4.67&lt;sup&gt;c&lt;/sup&gt; (4.39 to 4.94)</td>
<td>4.80&lt;sup&gt;c&lt;/sup&gt; (4.53 to 5.05)</td>
</tr>
</tbody>
</table>

<sup>a</sup>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.
<sup>b</sup>EHR: electronic health record.
<sup>c</sup>P value <.05.
<sup>d</sup>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.

References


Abbreviations

EHR: electronic health record
ICC1: intraclass correlation coefficient 1
IT: information technology
NI: 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 without further permission.
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 | vol. 9 | iss. 3 | e23951 | p.27

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><strong>Age</strong></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><strong>Sex</strong></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><strong>Type of facility</strong></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><strong>Educational level</strong></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><strong>Work experience</strong></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><strong>Position at facility</strong></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>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low commitment, n (%)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Female (n=117)</td>
<td>67 (57.3)</td>
</tr>
<tr>
<td>Male (n=147)</td>
<td>76 (51.7)</td>
</tr>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>≤30 years (n=186)</td>
<td>96 (51.6)</td>
</tr>
<tr>
<td>&gt;30 years (n=78)</td>
<td>47 (60.3)</td>
</tr>
<tr>
<td>Type of facilities</td>
<td></td>
</tr>
<tr>
<td>Referral hospital (n=16)</td>
<td>10 (62.5)</td>
</tr>
<tr>
<td>Primary hospitals (n=14)</td>
<td>6 (42.9)</td>
</tr>
<tr>
<td>Health center (n=234)</td>
<td>127 (54.3)</td>
</tr>
<tr>
<td>Educational level</td>
<td></td>
</tr>
<tr>
<td>Master’s degree (n=23)</td>
<td>11 (47.8)</td>
</tr>
<tr>
<td>BSc degree (n=156)</td>
<td>87 (55.8)</td>
</tr>
<tr>
<td>Diploma (n=85)</td>
<td>45 (52.9)</td>
</tr>
<tr>
<td>Position at facility</td>
<td></td>
</tr>
<tr>
<td>Expert position (n=163)</td>
<td>83 (50.9)</td>
</tr>
<tr>
<td>Head position (n=101)</td>
<td>60 (59.4)</td>
</tr>
<tr>
<td>Experience</td>
<td></td>
</tr>
<tr>
<td>≤3 years (n=61)</td>
<td>27 (18.9)</td>
</tr>
<tr>
<td>&gt;4 years (n=203)</td>
<td>116 (81.1)</td>
</tr>
</tbody>
</table>

*All the percentages were calculated for each sociodemographic category.

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

PMT members who received feedback for their DHIS2 data use were 1.85 times (adjusted odds ratio [AOR] 1.85, 95% CI 1.02-3.33) more likely to have a higher commitment level to use DHIS2 data than those who did not receive feedback. PMT members who had regular supervision and managerial support on their daily use of DHIS2 data for decision making were 2.84 times (AOR 2.84, 95% CI 1.50-5.37) more likely to have higher levels of commitment to use DHIS2 data than those who had no supportive supervision. Respondents who were competent to use DHIS2 data for their decision making were 2.41 times (AOR 2.41, 95% CI 1.27-4.55) more likely to have higher levels of commitment to use DHIS2 data than those who were not competent in DHIS2 tasks. PMT members with good culture of information use were 1.92 times (AOR 1.92, 95% CI 1.03-3.59) more likely to have higher levels of commitment to use DHIS2 data for decision making than those who did not have good culture of information use. Similarly, PMT members who inquired for DHIS2 data for health management were 3.96 times (AOR 3.96, 95% CI 2.11-7.41) more likely committed to use DHIS2 data than those who did not need DHIS2 data for health management. PMT members having motivation to use DHIS2 data were 1.80 times (AOR 1.80, 95% CI 1.00-3.25) more likely committed to using DHIS2 data when compared to those who had low motivation to use DHIS2 data for their decision making. These data are presented in Table 3.
Table 3. Factors associated with the level of commitment to use district health information system 2 data among performance monitoring team members at health facilities in the Ilu Aba Bora zone, Oromia region in 2020.

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

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

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

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

**Discussion**

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

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

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

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

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

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

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

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

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

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

Regarding study participants’ motivation to use DHIS2 data, respondents with higher motivation were 1.80 times more likely to have higher levels of commitment when compared to those with lower motivation to use DHIS2 data for their decision making (AOR 1.80, 95% CI 1.00-3.25). This finding (P=.03) was in line with the findings of studies conducted in Ethiopia [25] and Ghana (P=.01) [2].
PMT members with regular supportive supervision visits were 2.84 times more likely to have a higher level of commitment than those who did not have regular supportive supervision (AOR 2.84, 95% CI 1.50-5.37). This result was similar to those reported in studies conducted in Ethiopia [12,25] and Ghana, which showed that the level of commitment to use DHIS2 data was directly associated with the daily managerial supervision ($P=0.04$) [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/5c7003a6e5274a0ce6ed95c5/f176_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
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 $P<0.001$). 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.

Figure 1. Screenshots of (A) Dr. Joy’s persona introduced on the KakaoTalk Channel and examples of (B) 1 Q&A pair and (C) 3 Q&A pairs that can be triggered when user intent matches most closely with them in Dr. Joy’s obstetric and mental health–related Q&A knowledge database.
“pregnancy preparation” was redefined to address all posts excluding the posts on “infertility” and “pregnancy diagnosis,” and any posts irrelevant to the redefined topic were moved to the topic of either “infertility” or “pregnancy diagnosis” to effectively search questions or statements to be updated. Second, the topic of “pregnancy” was redefined to address the posts from the first to the ninth month of pregnancy because the message board about “pregnancy” covered all posts from the first to the last month of pregnancy and that about “labor and delivery” partially included posts at the tenth month of pregnancy.

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

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

**Main Features and User Interface**

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

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

These needs for preventative mobile health care and social supports from the perinatal women’s partners in everyday life were identified through in-depth interviews with 11 patients, 10 women and 1 man in the perinatal period, and a focus group interview with two obstetrician-gynecologist (ob/gyn) groups: (1) 3 ob/gyns at local hospitals and (2) 3 ob/gyns at university hospitals. According to the reports of the interviews, both patients and medical doctors highlighted the importance of the relationship between perinatal women and their partners on the women’s mental health during the prenatal, pregnancy, and postnatal periods. Particularly, the female interviewees and the doctors’ female patients who had experienced depressed symptoms expressed that they had a lack of opportunity to spend time with their partners in common; otherwise, a few women’s partners had cheated on them during pregnancy. By contrast, it was reported that the male interviewee, whose wife had no specific mental problems throughout pregnancy and after birth but who experienced depressed symptoms instead of her, tried to help his wife to overcome postpartum blues by sharing house.
chores, having a talk with her as much as possible, and ventilating her feelings of physical and emotional distress related to the double burden of childcare and housework. However, without their partners’ support, most pregnant women and mothers had difficulty in going out to refresh themselves or to attend a variety of mental health care programs held in local community health centers, local or university hospitals, and postpartum care centers. Although the male partners were also susceptible to the women’s mood fluctuations in the long-term period, both found it difficult to consult with health professionals and others (eg, family members and friends) about emotional or psychiatric problems and to consider using appropriate psychotropic medication about which a concern that it might negatively affect their fetuses might be raised. Furthermore, there has been a limitation in that the accessibility of useful information for effectively treating the women and even their partners was not significantly improved, particularly from the men’s point of view.

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

Figure 2. Screenshots of user interface workflow for a depression screening test using the 10-item Edinburgh Postnatal Depression Scale that can be administered in the prenatal period, followed by the screening test result and therapy suggestions.
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 (W₁₅=0.84; P=.01) and perceived risks (W₁₅=0.88; P=.04) was violated, the Spearman correlation was chosen for the final analysis.

Results

Participant Characteristics

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

When seeking health information on pregnancy or delivery to solve medical problems, all men referred to information sourced from books (2/2, 100%). However, women reported that they referred to multiple information sources, and the main source was acquaintances (7/13, 54%), followed by the internet (4/13, 31%), books (1/13, 8%), and health professionals (1/13, 8%). Particularly when using their personal computers or mobile phones to obtain online information on pregnancy or delivery, the 2 men employed different information search strategies: keyword search (1/2, 50%) and sentence search (1/2, 50%). A majority of women employed keyword search (11/13, 85%), and the others employed sentence search (1/13, 8%) and real-time search (1/13, 8%).
Table 1. Demographic information on the contextual UT participants (N=15).

<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&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Keyword search&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>UTI-02</td>
<td>35</td>
<td>F</td>
<td>PP</td>
<td>Acquaintances&lt;sup&gt;f&lt;/sup&gt;</td>
<td>Keyword search</td>
</tr>
<tr>
<td>UTI-03</td>
<td>35</td>
<td>F</td>
<td>PP</td>
<td>Books&lt;sup&gt;g&lt;/sup&gt;</td>
<td>Sentence search&lt;sup&gt;h&lt;/sup&gt;</td>
</tr>
<tr>
<td>UTI-04</td>
<td>34</td>
<td>F</td>
<td>PP</td>
<td>Acquaintances</td>
<td>Real-time search&lt;sup&gt;i&lt;/sup&gt;</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>Sentence 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&lt;sup&gt;j&lt;/sup&gt;</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>

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

<sup>b</sup>F: female; M: male.

<sup>c</sup>PP: pregnancy preparation (planned natural pregnancy); FT: first trimester; ST: second trimester; TT: third trimester; P: puerperium.

<sup>d</sup>Internet includes portal/search engines, online communities, blogs, vlogs, etc.

<sup>e</sup>Keyword search with simple words, search operators, hashtags, etc.

<sup>f</sup>Acquaintances include friends, colleagues, online community members, experienced mothers in the same postnatal care center, etc.

<sup>g</sup>Books include encyclopedias of pregnancy and birth, essays and articles written by medical doctors, magazines, etc.

<sup>h</sup>Sentence search with a single statement/question or multiple statements/questions.

<sup>i</sup>Real-time search means choosing and looking for attention-capturing content published in real time on the internet.

<sup>j</sup>Health professionals include medical doctors, nurses, etc. An acquaintance who was a medical doctor was included in health professionals.

Quantitative Data Analysis

The results from the USE questionnaire are shown in Table 2. Among the psychometric aspects of usability, the mean score of EOL was the highest, followed by the EOU, satisfaction, and usefulness scores in this sample. Even though the number of participants was insufficient to determine statistical significance of the difference in all 4 subfactors and total usability scores across sex, male participants showed higher mean scores than female ones. Both men and women had a tendency to rate the scores of usefulness and satisfaction lower than those of EOU and EOL; these trends were also identified within the total scores of usability and its subfactors. Despite the higher mean score of EOL in the entire participant group, the results of the Spearman correlation indicated that there were no significant associations with usefulness, EOU, satisfaction, or total usability scores (Table 3). Unlike EOL, the total usability and other 3 subfactors had significant positive associations with each other (all p>0.80; P<.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>USE b</th>
<th>EOU c</th>
<th>EOL d</th>
<th>SAT e</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men (n=2)</td>
<td>Total</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>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>USE b</th>
<th>EOU c</th>
<th>EOL d</th>
<th>SAT e</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>USE</td>
<td>Correlation coefficient (p)</td>
<td>1.00</td>
<td>0.82</td>
<td>0.26</td>
<td>0.98</td>
</tr>
<tr>
<td>P value (2-tailed)</td>
<td>_f</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.36</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>EOU</td>
<td>Correlation coefficient (p)</td>
<td>0.82</td>
<td>1.00</td>
<td>0.19</td>
<td>0.81</td>
</tr>
<tr>
<td>P value (2-tailed)</td>
<td>&lt;.001</td>
<td>—</td>
<td>.51</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>EOL</td>
<td>Correlation coefficient (p)</td>
<td>0.26</td>
<td>0.19</td>
<td>1.00</td>
<td>0.21</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>SAT</td>
<td>Correlation coefficient (p)</td>
<td>0.98</td>
<td>0.81</td>
<td>0.21</td>
<td>1.00</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>Total</td>
<td>Correlation coefficient (p)</td>
<td>0.97</td>
<td>0.89</td>
<td>0.32</td>
<td>0.95</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>PBb</th>
<th>PRc</th>
<th>SEEd</th>
<th>SHAe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men (n=2)</td>
<td>6.25 (1.06)</td>
<td>3.00 (0.71)</td>
<td>6.17 (0.47)</td>
<td>5.00 (1.41)</td>
</tr>
<tr>
<td>Women (n=13)</td>
<td>5.19 (1.03)</td>
<td>2.54 (1.64)</td>
<td>5.01 (1.21)</td>
<td>4.98 (1.30)</td>
</tr>
<tr>
<td>Total (N=15)</td>
<td>5.33 (1.06)</td>
<td>2.60 (1.54)</td>
<td>5.17 (1.20)</td>
<td>4.98 (1.26)</td>
</tr>
</tbody>
</table>

aAll scales were rated from 1 (strongly disagree) to 7 (strongly agree).
bPB: perceived benefits.
cPR: perceived risks.
dSEE: intention to seek health information.
eSHA: 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>PBb</th>
<th>PRc</th>
<th>SEEd</th>
<th>SHAe</th>
</tr>
</thead>
<tbody>
<tr>
<td>PB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient (ρ)</td>
<td>1.00</td>
<td>-0.29</td>
<td>0.94</td>
<td>0.70</td>
</tr>
<tr>
<td>P value (2-tailed)</td>
<td>_e</td>
<td>.30</td>
<td>&lt;.001</td>
<td>.004</td>
</tr>
<tr>
<td>PR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient (ρ)</td>
<td>-0.29</td>
<td>1.00</td>
<td>-0.28</td>
<td>-0.24</td>
</tr>
<tr>
<td>P value (2-tailed)</td>
<td>.30</td>
<td>—</td>
<td>.32</td>
<td>.40</td>
</tr>
<tr>
<td>SEE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient (ρ)</td>
<td>0.94</td>
<td>-0.28</td>
<td>1.00</td>
<td>0.73</td>
</tr>
<tr>
<td>P value (2-tailed)</td>
<td>&lt;.001</td>
<td>.32</td>
<td>—</td>
<td>.002</td>
</tr>
<tr>
<td>SHA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient (ρ)</td>
<td>0.70</td>
<td>-0.24</td>
<td>0.73</td>
<td>1.00</td>
</tr>
<tr>
<td>P value (2-tailed)</td>
<td>.004</td>
<td>.40</td>
<td>.002</td>
<td>—</td>
</tr>
</tbody>
</table>

aPB: perceived benefits.
bPR: perceived risks.
cSEE: Intention to seek health information
dSHA: Intention to seek health information
eNot 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.
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?

(18) What is the chance of having a girl after a boy?

(19) Can I tell the sex of my baby by my belly shape?

(20) What is the possibility that the baby’s sex will change after the ultrasound scan?

(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.

Theme 1-1: Chatbot Identity as a Social Agent
Although Dr. Joy was a text-based Q&A chatbot whose weakness was the lack of ability to understand what users were saying and to interact with them in a natural manner, it was found that our participants tended to consider Dr. Joy as a social actor as follows:

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

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

Theme 1-2: Chatbot Identity as a Male-Friendly Agent
Even though the facilitator gave them no specific instruction on what to ask, male participants raised questions about themselves as well as their wives, and female participants also did so about their husbands as well as themselves. They asked the following: “Can men have morning sickness?”; “Should men take folic acid?”; “Is there postpartum depression for fathers?”; “Does fathers’ medication affect pregnancy?”; “Husband is really having a hard time”; “What age is considered advanced paternal age?”

Theme 2-1: Strengths in Chatbot’s Utilitarian and Hedonic Values
According to the reports of all user utterance data, participants tried to view other given Q&A sets rather than press the negative feedback button or produce negative utterances on all Q&A sets. Regarding the strengths of this newly developed chatbot, a response that participants had in common was that Dr. Joy had both utilitarian and hedonic values (Textbox 1, quotes 4-9).

Theme 2-2: Strengths in Chatbot’s Informational Support
In addition to these strong points, some participants mentioned the benefits from health-related information sourced from Dr. Joy (Textbox 1, quotes 10-12).

Theme 3: Weaknesses in Chatbot’s Content Coverage
The most frequently reported weak point was that Dr. Joy failed to meet all user intents and to cover a much broader range of content domains because we focused more on helping perinatal women to prevent and solve their own mental and physical problems than on offering them answers to baby-oriented questions (Textbox 1, quotes 13-16).

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

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

As productivity has been considered as the main motivation for chatbot use [1], this “always-on” Q&A chatbot for offering 24/7 digital support to perinatal women and their partners can be an easier and more efficient way to obtain credible information and be more intuitive to the target users than conventional means (eg, books, internet search, acquaintances, and health professionals). In line with this study, previous studies tried to expand the Q&A chatbots’ own knowledge databases to ensure content quality and improve response capability. Chung and his colleagues [19,34] applied (1) the expert-based approach to create rules for the provision of the medical information and (2) the data-based approach to provide customized information based on the already established medical knowledge database for the chatbot-based health care service, thus increasing reliability. In order to create a domain-specific or context-based chatbot to provide optimal, up-to-date answers immediately,
the high-quality chatbot knowledge was extracted from social networking services such as Twitter [20], online discussion forums as web communities [35], and messengers [19]. Similar to our approach, Jeong and Seo [20] proposed a keyword matching–based answer retrieval technique based on the collection of Q&A sets from Twitter by utilizing the tweet-and-reply and the tweet-and-mention pairs and the refinement of the newly collected pairs by adding them to the existing Q&A knowledge database. As these related works focused on developing the Q&A chatbots’ answer retrieval technique to provide more accurate and flexible answers to their users, the response appropriateness of each chatbot based on quantitative data such as self-report questionnaire [20] or recall and precision measurement [18,35] was evaluated. While these proposed knowledge databases and answer retrieval techniques for Q&A chatbots were appropriate to be applied to general health care or lifecare services whose target users and content coverage were not specified, there have also been a variety of informative chatbots designed for the specific purposes of supporting pregnant women and mothers or families with young children in emergency situations [6] and providing low-cost accessible fertility and preconception health education for perinatal women [36] or breastfeeding education for community health workers and mothers in under-developed areas [37].

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

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

Another finding was that the negative association between the perceived benefits and risks of using Dr. Joy was not significantly strong enough to influence behavioral intention in a negative direction. Furthermore, Dr. Joy led its users to perceive a low level of risks that discussing health-related information on this medical chatbot via KakaoTalk would confront them with unwanted problems or that the expected benefits of doing so would not materialize. With a low possibility of trade-off between benefit and risk, the different intentions to seek and share health information on Dr. Joy were significantly associated only with the perceived benefits, not with the perceived risks. The more its users think Dr. Joy can benefit them, the more likely they are to seek and share information from it. Compared to women, who scored SEE and SHA at similar level, the men had more intention to seek health information on medical chatbot via KakaoTalk than the women. This might be because the male partners have comparatively less opportunity to access information sources than perinatal women, who have tended to seek medical help from multiple informal and formal sources [25]. As pregnant women’s partners, our male participants, whose main source of pregnancy or delivery information was books such as encyclopedias of
pregnancy and birth or essays written by medical doctors, were less likely to show the tendency to double-check information from other sources by sharing Dr. Joy’s relatively more credible information verified by health professionals. In line with the findings of our previous study [23], it can be explained that female participants, who reported relying more on multiple word-of-mouth sources of information and less on health professionals, were highly likely to share many concerns that they were reluctant to discuss with their doctors in the outpatient clinic, particularly with this KakaoTalk chatbot with a humanlike medical doctor persona.

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

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

**Limitations and Future Direction**

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

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

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

Finally, considering that a full-term pregnancy lasts 38 weeks or longer, a 7-day study period is insufficient to assess whether Dr. Joy can improve the participants’ knowledge, answer their questions effectively, or be useful for certain tasks, even if the participants provided positive usability and UX ratings in this study. To answer these research questions, which remain open for future studies, a perinatal and mental health–related variable should be directly adopted in the short-term study period, or a more longitudinal evaluation should be performed. Taken together, future studies will benefit from addressing these limitations.
Conclusions
In sum, this study provides the potential for the uptake of this newly developed Q&A knowledge database–based KakaoTalk chatbot for perinatal women’s and their partners’ obstetric and mental health care. As Dr. Joy has quality contents, which are positively linked with both utilitarian and hedonic value, its male and female users can be encouraged to adopt and use medical chatbots in a convenient, easy-to-use, and pleasant manner. To boost their intention to continue use of Dr. Joy, its Q&A sets should be periodically updated to satisfy more user intent by monitoring both male and female user utterances.

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

Conflicts of Interest
None declared.

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

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

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

 References


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

Jessica P Ridgway1, MD, MSc; Arno Uvin1, BSc; Jessica Schmitt1, LCSW; Tomasz Oliwa2, PhD; Ellen Almirol1, MSc; Samantha Devlin1, MSc; John Schneider1, MD, MPH

1Department of Medicine, University of Chicago, Chicago, IL, United States
2Center for Research Informatics, University of Chicago, Chicago, IL, United States

Abstract

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

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

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

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

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

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

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

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

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

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

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

Methods

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

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

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

https://medinform.jmir.org/2021/3/e23456
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, E950.3, E950.4, E950.9, 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, 305.0, 305.01, 305.02, 305.03, 305.2, 305.21, 305.22, 305.23, 305.4, 305.5, 305.51, 305.52, 305.53, 305.6, 305.61, 305.62, 305.63, 305.7, 305.91, 305.93, 425.5, 535.1, 571.1, 648.33, 965.01, 970.81, E850.0, E850.1, E850.2, E854.8, E860.0, E935.0

- ICD-10 codes
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 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
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 Shand4, MBBSMD, FRCP; Adesh Ramsewak5, MBBS, FRCP; Divyesh Sharma4, MBBS, MSc, FRCP; Anne McShane5, MSc; Khaled Rjoob1, MSc; Aaron Peace4, 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 κ showed poor agreement (κ=0.08, P=.001). Only the age (odds ratio [OR] 1.07, 95% CI 1.05-1.09) and chest pain (OR 0.59, 95% CI 0.39-0.89) independent variables were statistically significant (P=.008) in predicting mortality after 30 days and 1 year. The odds for mortality within 1 year of referral were lower in patients with chest pain compared to those patients without chest pain. A referral being out of hours was a trending variable (OR 1.41, 95% CI 0.95-2.11, P=.09) for predicting the odds of 1-year mortality.

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

Introduction

Background

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

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

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

Prior Work

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

Study Goals

Having summarized the research to date, we have identified that ECG interpretation is challenging for both humans and computers, and there is a need to better understand the characteristics of the patients who are declined for primary percutaneous coronary intervention, especially given that there are a number of likely false negatives (patients who are declined but needed an emergency intervention).
We aimed to analyze agreement between computer and human (activator nurses) ECG interpretations for patients who were referred 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%; \( P=0.001 \)). 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%; \( P=0.001 \)). 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%; \( P=0.001 \)).
Figure 1. (a) Activator nurse and (b) computer interpretations of acute myocardial infarction rate by the hour. AMI: acute myocardial infarction; MI: myocardial infarction.

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

Figure 2. Activator nurse and computer agreement of (a) acute myocardial infarction and (b) not acute myocardial infarction. AMI: acute myocardial infarction; MI: myocardial infarction.
Figure 3. Activator nurse and computer agreement by the hour.

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

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

Figure 4. (a) Activator nurse interpretation suggesting acute myocardial infarction and computer disagreed; (b) computer interpretation suggesting acute myocardial infarction and activator nurse disagreed.
Computer Suggested Acute Myocardial Infarction

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

Analysis of Other Variables

Patients With Chest Pain

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

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

Figure 5. Proportion of patients with chest pain by the hour.
Table 1. Odds ratios of variables derived from multiple logistic regression where the response variable was mortality after 1 year.

<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)†</td>
<td>0.59 (0.39-0.89)</td>
<td>0.012</td>
<td>.008</td>
</tr>
<tr>
<td>Activator nurse diagnosis suggesting acute myocardial infarction (true)</td>
<td>1.26 (0.73-2.16)</td>
<td>0.012</td>
<td>.39</td>
</tr>
<tr>
<td>Computer 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>Activator nurse–computer acute myocardial infarction agreement (true)</td>
<td>0.97 (0.47-2.03)</td>
<td>0.011</td>
<td>.95</td>
</tr>
</tbody>
</table>

†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)†</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>

†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†</th>
<th>Computer Interpretation term</th>
<th>Patients, n (%)</th>
<th>Activator nurse Interpretation term</th>
<th>Patients, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suggests acute myocardial infarction‡</td>
<td>“acute myocardial infarction”</td>
<td>337 (47)</td>
<td>“Ste”</td>
<td>159 (25)</td>
</tr>
<tr>
<td>1</td>
<td>“inferior infarct”</td>
<td>23 (3)</td>
<td>“St depression”</td>
<td>125 (20)</td>
</tr>
<tr>
<td>2</td>
<td>“anterior injury”</td>
<td>34 (5)</td>
<td>“twi”</td>
<td>129 (20)</td>
</tr>
<tr>
<td>Does not suggest acute-myocardial infarction§</td>
<td>“abnormal ECG”</td>
<td>382 (51)</td>
<td>“nil acute”</td>
<td>377 (45)</td>
</tr>
<tr>
<td>1</td>
<td>“LBBB”</td>
<td>108 (15)</td>
<td>“high take-off”</td>
<td>187 (23)</td>
</tr>
<tr>
<td>2</td>
<td>“borderline ECG”</td>
<td>41 (5.5)</td>
<td>“RBBB”</td>
<td>59 (7)</td>
</tr>
</tbody>
</table>

†Terms with low frequencies (1 or 2) are not included.
‡n=722 for Computer; n=634 for Activator nurse.
§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=0.01). However, the activator nurse interpreted that ECGs suggested acute myocardial infarction more often during the middle of the night (12 AM to 6 AM; mean 53%, SD 5.3%) than in daytime hours (mean 41%, SD 6.6%; P=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 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=0.01). 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 Derboven⁶, PhD; Delphine De Smedt⁷, PhD; Erik Dovgan¹, PhD; Jure Lampe⁸, BSc; Flavia Marino⁹, 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=.02), whereas depression and anxiety rates were significantly reduced (P<.001), as were perceived sexual problems (P=.01). According to the Unified Theory of Acceptance and Use of Technology questionnaire, a positive attitude toward HeartMan was seen among end users, resulting in increased awareness, self-monitoring, and empowerment.
Conclusions: The HeartMan project combined a range of advanced technologies with human-centered design to develop a complex system that was shown to help patients with CHF. More psychological than physical benefits were observed.

Trial Registration: ClinicalTrials.gov NCT03497871; https://clinicaltrials.gov/ct2/history/NCT03497871.

International Registered Report Identifier (IRRID): RR2-10.1186/s12872-018-0921-2

(JMIR Med Inform 2021;9(3):e24501) doi:10.2196/24501

KEYWORDS
congestive heart failure; personal health system; mobile application; mobile phone; wearable electronic devices; decision support techniques; psychological support; human centered design

Introduction

Background and Motivation

Congestive heart failure (CHF) is a disease in which the heart cannot pump enough blood to supply oxygen and nutrients to the body. The main symptoms are shortness of breath (dyspnea), diminished ability to exercise, fatigue, and swelling in the feet and legs (edema). The lifetime risk of developing CHF ranges from 20% to 33%, and only approximately half of patients survive for more than 5 years after diagnosis [1]. As CHF is frequently the end stage of various conditions that affect left ventricular function and cannot be cured, the focus of the treatment is to prevent deterioration, manage symptoms, and maintain a good quality of life [2].

The management of CHF includes multiple medications, appropriate exercise, diet (paying particular attention to fluids and salt), management of body weight, and abstaining from alcohol and smoking. As the average age at CHF diagnosis is 74 (SD 14) years [3], 25% to 80% of the patients are affected by cognitive impairment [4], a third of them have depression or anxiety [5], and other comorbidities are also common, they often find it difficult to manage the disease on their own [6]. Cardiac rehabilitation programs are either not available or poorly attended—participation in Europe is approximately 20% [7]. Therefore, the relevant alternatives are technological solutions to support the management of CHF.

Approximately 64 million people live with CHF globally [1], and the economic burden of their disease amounts to more than 100 billion US $ annually [8]. This is a strong incentive to improve CHF management. In addition to medications, implantable devices (mainly pacemakers and defibrillators) are already established treatment options [9]. Another option is telemonitoring, but its benefits in CHF are uncertain [9]. Another option is mobile health (mHealth) solutions, whose benefits in CHF are poorly explored (see the Related Work section) but have a strong backing of the market: the mHealth market in 2019 was US $46 billion and grew by 22% annually [10] (compared with the telemonitoring market of US $2 billion with 13% growth [11] and the more mature implantable devices market of US $23 billion and 8% growth [12]).

In the HeartMan project, we developed a comprehensive personal health system for the self-management of physical and psychological aspects of CHF. The first step was to analyze evidence-based medical requirements and—following the human-centered design process—to elicit requirements related to everyday management of CHF from the patients themselves. We then developed a mobile app comprising a decision support system (DSS) and several intelligent data analysis modules. A web application for medical professionals has also been developed. Finally, the system was evaluated in a proof-of-concept trial that assessed both its effectiveness and patient perception.

Related Work

In 2018, a systematic review was devoted to mobile apps supporting the self-management of CHF [13]. The authors surveyed 10 leading paper repositories for papers on interventions that used a mobile platform, evaluated them with a randomized controlled trial or a similar design, and provided usability or efficacy results. Papers on telecare and structured telephone support were excluded. In total, 18 papers meeting the inclusion and exclusion criteria were included in the review. The authors also searched Google Play and Apple App Store for health care apps by including “heart failure” as a keyword. After excluding apps that track only blood pressure (BP) and/or heart rate, a total of 26 apps were downloaded and evaluated with respect to the quality of self-management components included in the apps and quality of the user experience provided by the apps.

According to the authors of the review [13], most apps are poorly designed and do not include all the necessary components for the self-management of CHF. Indeed, only 2 apps—Heart Failure Storylines [14] and HeartMapp [15]—include exercise interventions, which is one of the most important aspects of CHF management. The Heart Failure Storylines app is perhaps the most complete one that can be currently found in the market. It provides medication reminders, a symptom tracker, keeps a record of vital signs, and tracks physical activity and daily moods. Nevertheless, the interventions provided by the app are poorly personalized (except for medication reminders) because the app does not consider the patients’ psychophysical state, making the usefulness of such interventions questionable [9,16]. The HeartMapp app provides personalized interventions, but it is quite basic and is not adapted to the patients’ psychophysical state. The app was tested in a randomized controlled trial with only 18 participants (intervention group, n=9) [17].

We searched the Google Play and Apple App Store for apps that were not included in the review. We found 6 apps that were published after the review. Five of these apps include only educational materials [18-23], whereas 1 app provides only guidance on medication therapy [24]. In short, no new apps provide a comprehensive solution for CHF management.
Methods

Collection of Requirements and Human-Centered Design

Medical Requirements
The first step in designing the HeartMan system was to study the state-of-the-art medical knowledge on CHF self-management. A systematic review of the available literature was performed to identify parameters that predict the hard outcomes of mortality and hospitalization in patients with CHF as well as variables that affect the patient-reported outcome of quality of life in this patient group [25]. We further selected those parameters that are modifiable by self-care behaviors that the HeartMan system can recommend. These modifiable parameters are primarily clinical parameters (eg, body mass index, BP, heart rate), physical capacity, medication use, characteristics of CHF (eg, fluid retention), and mental health (eg, depression, anxiety). We then screened relevant medical guidelines for CHF, focusing on nonpharmacological recommendations and lifestyle advice, to identify the best approaches for modifying these parameters [26] and incorporated these into the HeartMan DSS. We designed an exercise training and nutrition program (including diet and fluid intake restrictions) to influence physical capacity, clinical parameters, and fluid retention. Medication adherence is expected to be enhanced through DSS, providing reminders, disease education, and self-monitoring. Finally, cognitive behavioral therapy and mindfulness exercises were included to improve mental health and self-management. Management guidelines for comorbidities were also taken into account, as many patients with CHF have conditions such as diabetes, atrial fibrillation, and chronic obstructive pulmonary disease.

An additional source for developing the medical requirements was data from the Chiron project [27], a previous telemonitoring study in patients with CHF focusing on short-term outcomes of subjective well-being on a daily basis. Data mining analysis suggested environmental parameters, that is, ambient conditions such as temperature and humidity, to play a role in predicting day-to-day changes in perceived health. This was incorporated into an additional module of DSS.

User Requirements
As our goal was not only to provide medically relevant advice but also to design the HeartMan system to be useful and well accepted by the patients, we adopted a human-centered design [28]. This approach involves users throughout the design process, focusing on their perspective and needs. In our case, it consisted of a thorough analysis of patients’ context of use, which took place in three stages in Belgium and Italy. The first stage was a diary study, in which patients kept a diary for a period of 10-14 days (n=19 in Belgium; n=18 in Italy). The diary contained questions and assignments related to everyday activities and habits, such as patients’ experience, disease management, and their social network. The second stage was a follow-up interview study conducted with most patients who participated in the diary study (n=14 in Belgium; n=15 in Italy). In this interview study, patients participated in semistructured interviews in which the output of the diary study was discussed in detail. This analysis resulted in a rich, qualitative description of patient characteristics as well as the patient experience regarding disease management, the challenges related to therapy adherence, lifestyle changes as a result of being a CHF patient, and relationships with caregivers. These insights were translated into concrete user requirements for the HeartMan system, which served, together with the medical requirements, as the starting point for the third stage: the design and evaluation of a series of prototypes with both patients and caregivers. In this process, several design trade-offs were made regarding patient autonomy, technology appropriation, and patient well-being [29]. The main patient characteristics that were found to impact the design of the HeartMan system were the patient’s digital literacy, perception of empowerment, and existing therapy adherence habits.

For medical professionals, a web portal was developed, allowing them to follow up on the patients’ data gathered by the HeartMan system. This prototype was developed and evaluated using a separate human-centered design process. In this process, various stakeholders (including cardiologists, nurses, dieticians, psychologists, and physiotherapists) offered insights into the needs and requirements related to the follow-up of patients with CHF based on the HeartMan monitoring data.

System Overview
In the HeartMan system designed as described in the previous section, sensing devices collect information about the patient, patient monitoring methods further interpret some of this information, and a DSS recommends actions based on the (interpreted) information. The recommendations are presented to the patient via a mobile app, and medical professionals have access to the system via a web application. A diagram presenting an overview of the system is presented in Figure 1.
The sensing devices (yellow in Figure 1) are custom sensing wristbands, off-the-shelf BP monitors, weight scales, and environmental sensors that measure temperature and humidity. According to the medical requirements, heart rate (obtained from the photoplethysmogram [PPG] signal), BP, weight, and ambient temperature and humidity are important determinants of the health and well-being of patients with CHF. As it would be relevant to monitor BP more frequently than once per day (which can be expected with a regular BP monitor), we developed a method to estimate BP continuously from the PPG signal (green). Owing to the importance of psychological support for patients with CHF, we also developed a method to recognize their psychological profile from the heart rate, heart-rate variability, and voice recorded with the smartphone. Finally, the accelerometer in the wristband is used to recognize the patient’s physical activities, which allows the initiation of psychological interventions at the appropriate moment. As the accelerometer provides the greatest volume of data of all the sensors, this last method is implemented on the smartphone, whereas the previous 2 reside in the cloud.

All patient information is fed into the DSS and stored in the cloud (blue in Figure 1). The DSS has three components, the first of which is an expert system that helps patients manage their physical health (exercise, nutrition, medications, and self-monitoring). The second is another expert system that provides psychological support (elements of cognitive behavioral therapy and mindfulness). The third uses predictive models (based on the previously mentioned Chiron data) to recommend actions related to temperature and humidity that are expected to improve patients' well-being. The first 2 components rely on expert knowledge because it is well established how the aspects of the CHF management they address should be tackled. The last one relies on data and predictive modeling because we had relevant data available, but there is little expert knowledge on the effect of the environment on the well-being of patients with CHF.

The recommendations provided by the DSS are shown in the mobile app (purple in Figure 1), which also collects inputs from the patients. Medical professionals can use a web application to view information collected from sensing devices as well as the patients’ adherence to recommendations. Although the content of recommendations is mostly based on the medical requirements, the way information is presented via the 2 applications was heavily influenced by the users’ inputs obtained during the human-centered design process.

Patient-Monitoring Methods

The HeartMan Wristband

The wristband used by the system includes a PPG sensor, which provides information on the heart rate and beat-to-beat intervals in addition to the raw data, tri-axial accelerometer, and temperature sensor. It communicates with the HeartMan app via Bluetooth Low Energy 4.1. Its battery life is sufficient for a full day of operation, while continuously streaming sensor data to the phone. It features a liquid crystal display and vibration motor, which can be used to deliver urgent notifications to the user, such as about too high or low heart rate during exercise.

BP Estimation

Continuous BP estimation is well researched when 2 signals, typically ECG and PPG, are available, as the pulse transit time between 2 points on the body is highly correlated with the BP [30,31]. In HeartMan, we aimed to use a single wristband PPG sensor [32,33], as this is the most convenient for the patients. However, such a sensor typically has a modest sampling frequency, the sensor-to-skin contact is often compromised due to movement, and the wrist area exhibits less pulsatility compared with a fingertip, making this approach challenging.

To obtain high-quality parts of the PPG waveform, the signal was preprocessed. The first step was zero-mean unit-variance normalization. Outlier samples above 3 SDs from the local
median (10-sample window) were removed using a Hampel filter. Afterward, the signal was filtered using a fourth-order Butterworth band-pass (0.5-4.0 Hz) filter. Then, a transformation based on the first-order derivative was used to detect systolic peaks and diastolic valleys in between. Once the valleys were detected, the signal was traversed with a sliding window, and a template was created as the average of all cycles in a window. Following this, each individual cycle was compared with the template using several metrics. This allowed for the detection of segments where the signal was stable with only a few artifacts, while also allowing for individual bad cycles in an otherwise good segment to be discarded [34].

After preprocessing, per-cycle temporal features describing the cycle shape were computed based on related work [35] and further expanded with some features from the frequency domain. The latter were computed from a window centered on a cycle and extending 5 seconds before the cycle start point and 5 seconds after the end point. Most of the temporal morphologic features rely on high-quality waveform, exhibiting a clear systolic and diastolic peak, as they were designed for fingertip PPG devices in a controlled setting. The HeartMan wristband signal is generally of lower quality, so we focused on frequency domain features, which are more robust, as they are computed from longer windows and not on a per-cycle basis. In addition, as some morphological features were infeasible to compute from the HeartMan wristband data, we additionally leveraged information from the accelerometer, which tells us about the person’s physical activity. We considered some commonly used features computed from the three-axis accelerometer, which are known to work well in separating a person’s activities [36]. We decided on this because having information about a person’s activity might prove useful for BP estimation, as the cardiovascular response of the body changes during intense physical activity compared with the state. This fact differentiates this work from previous work dealing with similar problems, as related work often focuses on PPG signals without considering the person’s activity, which can be reflected in the accelerometer signal [37]. Finally, heart rate was also used as a feature to inform us about a person’s cardiac activity. All these features were fed into regression models that estimated systolic BP (SBP) and diastolic BP (DBP). Several algorithms implemented in the Scikit-learn toolbox [38] were used to train the models, some of which are compared in the Results section.

**Psychological Profile Recognition**

The development of technological interventions for behavior changes as well as growing interest in affective computing have resulted in various attempts to recognize psychological states from sensor data. Some authors [39] used mobile phones to analyze user voices and classify their emotions (happy, sad, fear, anger, and neutral). Others have focused on stress, dementia, and cognitive dysfunctions, relying more on wearable devices that sense the heart rate, electrodermal activity, skin temperature, and acceleration [40,41].

The HeartMan system combines the patient’s voice obtained during a structured weekly phone interview with an informal caregiver with heart rate features, which can be obtained from the HeartMan wristband. The speech data were preprocessed to normalize the different acoustic properties, such as higher volume and background noise, using standard techniques [42]. The features extracted from the speech are the fundamental frequency (pitch), mel-frequency cepstral coefficients, and the smoothed energy. The mean, SD, range, maximum, and minimum were computed for each base speech feature. In addition, the heart rate and heart rate variability represented by the root mean square of successive differences between heartbeats were extracted. The features are then fed into a machine learning model that recognizes motivated, anxious, and depressed psychological profiles. All the data were preprocessed and analyzed using MATLAB and R software.

**Physical Activity Recognition**

Physical activity recognition is a relatively mature field, although the requirements of HeartMan present some challenges. As the purpose was to initiate psychological interventions, it was most relevant to recognize eating and to distinguish resting from walking and more intense activities. Eating recognition is quite difficult and rarely addressed in the literature, whereas wrist—being able to move independently from the body—is not the best location for recognizing the intensity of activity.

Similar to the previous 2 patient-monitoring methods, this method also uses machine learning. The stream of acceleration data is first low-pass filtered to remove noise and then band-pass filtered to remove the gravitational component, retaining the component due to dynamic human motion. The stream was then segmented into 2-second windows. In each window, the low-pass filtered data are used to compute features related to the orientation of the sensor, whereas the band-pass filtered data are used to compute the features related to the motion of the sensor. A total of 90 features were extracted [37]. Some describe the intensity and shape of the acceleration signal, such as the mean, variance, skewness, and kurtosis. Others have a physics-based interpretation, such as changes in velocity and kinetic energy. The rest are based on expert knowledge, such as the number of peaks in the signal and the number of times the signal crosses its mean value. The features are fed into a machine learning model that returns one of the following activities: rest, standing, walking, Nordic walking, running, other exercise, eating, washing hands or face, household chores (whole-body movement), and light hand activities (hand movement). The model was built using the random forest algorithm implemented in the Weka toolkit [43].

**DSS**

**Expert System for Physical Health Management**

**Exercise**

The HeartMan DSS administers a comprehensive exercise program [44] according to the established medical guidelines [16]. Before starting the exercise program, the patients were expected to perform a cardiopulmonary exercise (cycloergometry) or a 6-min walking test to assess their physical capacity. On this basis, the physical capacity of each patient is assessed as *low* or *normal*, which affects the exercise planning.
Weekly Exercise Planning

The DSS proposes a weekly exercise plan for each patient, consisting of endurance and resistance exercises. The DSS suggests the frequency (times per week), intensity, and duration of each exercise type. The suggestions are based on the patient’s physical capacity, the number of active weeks in the program, and the current frequency and intensity. They are also based on the patient’s psychological profile: the difficulty increases more gradually for depressed patients, which is in line with the shaping technique suitable for this profile. For instance, low-capacity patients start with very light 10- to 15-min endurance exercises twice per week. According to the patient’s progress, these parameters may change with time, typically by increasing the frequency and intensity of exercises, if the patient agrees. The planning process is governed by an expert system that consists of 2 rule-based models, developed using a qualitative multicriteria method decision expert [45] and described in more detail in our earlier work [44].

Exercise Sessions

Before the start of each exercise session, the HeartMan DSS checks whether the patient’s BP and heart rate are in a safe range and whether the patient feels well enough to exercise. If the exercise is allowed, a list of exercises is shown to the patient, who can then select the preferred exercise. This is illustrated in Figure 2. Typical endurance exercises involve walking and cycling, whereas resistance exercises aim to strengthen the patient’s arms, legs, and body. After selecting the exercise, a detailed description (text or graphical) was provided. During the exercise, the heart rate and SBP were continuously measured using the wristband. Patients are advised to stop the exercise in cases of symptoms or measurements outside a safe range. During endurance exercises, the DSS uses the wristband display to suggest an increase or decrease in pace based on the heart rate. After completing the exercise, the patients can rate their feeling of intensity, which is used in the weekly planning to decide whether to increase the intensity.

Figure 2. Exercise-related screens of the HeartMan app: the main screen, blood pressure input before the exercise, health check before the exercise, and exercise list.

Nutrition

To provide appropriate nutrition advice, the DSS requires the following medical information: the patient’s BMI, whether the patient has diabetes, and the prescribed amount of liquid intake. Next, the DSS creates a personalized questionnaire to be answered by the patient; it includes general questions about healthy nutrition and specific questions about the patient’s eating and drinking behavior. On this basis, the DSS assesses the level to which topics (about breakfast, lunch, dinner, fat and cholesterol, fluid intake, salt, diabetes, and medication) are understood by the patient. Finally, the patient received feedback in terms of positive reinforcement messages (for well-understood topics), educational statements (for misunderstood general topics), and advice on how to modify the diet to make it healthier (for misunderstood eating behavior topics).

Self-Monitoring and Medication

Patients with CHF are required to measure their BP, heart rate, and daily weight. The HeartMan system reminds them of this and warns if the measurements are outside the safe ranges. It also reminds the patients to take their medications and helps them fill the weekly pillbox (if they use one). It periodically asks the patient about the number of pills remaining in the pillbox and assesses medication adherence based on the deviation from the expected number.

Expert System for Psychological Support

In most cases, CHF diagnosis requires substantial changes in daily life and habits, such as dietary modifications and increased physical activity. Combined with psychological distress, which also often follows the diagnosis, patients can face an intrusion of distorted beliefs and negative automated thoughts that cause them to feel unable to pursue a goal [46]. Sometimes a vicious circle called cognitive dissonance is triggered: a conflict between their desire to be healthy on one hand and practicing unhealthy behaviors for short-term comfort on the other hand. In the long run, this results in poor adherence to self-management guidelines as well as psychological discomfort [47].
The psychological DSS is designed to select the appropriate strategy to improve patients’ psychological well-being and adherence to physical exercise and dietary guidelines. The strategy is adapted to the user’s psychological profile, as discussed in the section on psychological profile recognition. The DSS provides cognitive behavioral interventions and mindfulness exercises that are modified according to a weekly plan. These exercises are suggested daily, at a time when the user engaged in a physical activity expected to make them receptive to the suggestion. The relevant activities are eating, walking, and sitting, as discussed in the Physical Activity Recognition section.

**Cognitive Behavioral Therapy**

This is a combination of behavioral and cognitive techniques developed to reduce anxiety and depressive symptoms, which tend to make patients less motivated, tired, and less energetic. The DSS provides specially designed messages intended to align the patients’ actions with their desires, as shown in the examples in Table 1. These messages are formulated according to the principles by Festinger [48] of *cognitive consequences of forced compliance* for the motivated profile, *free choice* for the anxious profile, and *effort justification* for the depressed profile.

<table>
<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 TV 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>

*TV: television.

**Mindfulness**

Mindfulness exercises enhance patients’ awareness of their present condition and help them disassociate (unhealthy) emotional and behavioral responses from physical sensations and thoughts. Mindfulness exercises consisted of the following:

- Games to deal with intrusive thoughts (eg, loss of independence, feeling restricted in daily activities), as shown in Figure 3.

**Figure 3.** Mindful game "World Sense".

**Predictive Models for Environment Management**

Unlike the DSS approaches used for physical health management and mental support, which mainly rely on expert knowledge, a data-based approach was developed for environment management. We used data from the Chiron project [27], which consists of features describing the patient’s situation and their self-reported feeling of health. The features are physiological (eg, heart rate, BP) and environmental (eg, temperature, humidity) and very similar to those available to the HeartMan system.

In the first step, we built a machine learning model that could predict the feeling of health from the features. We used the random forest algorithm implemented in the Weka toolkit [43].

The accuracy of distinguishing between good and bad feelings of health was 83.2%. We also divided the features into modifiable, correlated (with modifiable), and uncorrelated. We build linear regression models that can predict each correlated feature from the modifiable ones.

In the second step, we set up a multi-objective optimization problem, where we searched for minimal modifications of modifiable features that change the feeling of health from bad to good. For each solution, the correlated features were predicted using linear models, and the admissibility of the solution was checked using the feeling-of-health model. The objectives were the sum of the volumes of modifications needed and the number of...
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.
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(^a) 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>

\(^a\)MAE: mean absolute error.

An example segment of the DBP estimates is shown in Figure 8. The results show that BP estimation is feasible; however, most state-of-the-art methods are highly dependent on high signal quality to obtain precise morphological features on a per-cycle basis, which is difficult to achieve with an affordable wristband.

Figure 8. Segment of example estimates and ground truth diastolic blood pressure from the HeartMan trial. DBP: diastolic blood pressure.

Psychological Profile Recognition

To test the psychological profile recognition, we collected a data set from 30 healthy subjects (mean age 68, SD 2 years, 6 women and 23 men). The subjects used the HeartMan mobile app for psychophysiological data collection. Leave-one-subject-out evaluation was conducted, and classification accuracy into depressed, anxious, and motivated profiles was used as the evaluation metric. Classification models trained with four machine learning algorithms were compared against a baseline dummy model, which always returned the majority class.

As shown in Table 3, the support vector machine (SVM) model performed best, achieving a fairly high accuracy, especially considering that this is a subject-independent result. In Table 4, we can see the results in terms of precision, recall, and F1-score for the SVM model. The percentages of the confusion matrix as a result of the cross-validation procedure showed that SVM can classify all 3 classes with precisions of 93\%, 86\%, and 84\%, respectively. From the results, it can be observed that the motivated profile was recognized most accurately, whereas most of the misclassifications came from the anxious and depressed profiles, which are sometimes very similar.

Table 3. Classification accuracies of the psychological profile detection.

<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>Naïve 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>

Physical Activity Recognition

The model for physical activity recognition was built and evaluated on recordings of 10 healthy subjects (mean age 59, SD 5 years, 6 women and 4 men). The subjects performed a scenario consisting of all the activities to be recognized with several variations: walking at different speeds, uphill and carrying a burden, eating various foods, and performing a wide range of chores (cooking, sweeping floor, gardening tasks, etc) and hand activities (writing, using a computer, knitting, etc). Similar to the previous cases, a leave-one-subject-out evaluation was conducted. Precision (the fraction of the instances recognized as a certain activity that in fact belong to that activity), recall (the fraction of the instances belonging to a certain activity that are recognized as such), and F-measure (harmonic mean of precision and recall) were used as the evaluation metrics.

Table 5 shows that most of the activities can be recognized reliably. Standing has the smallest F-measure, because it is often misclassified as rest. This is understandable because in both cases, the hand with the wristband does not move much and is not overly problematic because most people rarely stand still for a long time. The second largest problem is confusing eating with hand activities, which is also understandable but makes accurately triggering psychological interventions during eating difficult.

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

<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>

General Effectiveness of the System

A proof-of-concept trial was set up to evaluate the effects of the HeartMan intervention on health-related quality of life and disease management (self-care) as primary endpoints [54]. The secondary endpoints we targeted were clinical parameters, illness perception, and mental and sexual health. The clinical trial was registered on NCT03497871 on 2018-04-13. It was implemented in two countries: three hospitals were involved in Belgium, and one hospital and a local health authority participated in Italy. A randomized controlled design was used with a 1:2 ratio of the control and intervention groups. Eligible patients were recruited by the treating cardiologist or general practitioner at the time of regular consultation. After providing informed consent, participants underwent a baseline data collection, containing medical record data registration, questionnaire assessments, and some clinical assessments, including a 6-min walking test. Patients were then randomly assigned to either the control group receiving the usual care or the intervention condition additionally receiving the HeartMan personal health system that they used in their home setting for a period of 3-6 months. All outcome measurements were repeated in both the intervention and control groups at the end of the trial.

The intervention effects were evaluated in a final sample of 56 patients (ie, 34 in the intervention group and 22 in the control group). Trial results showed that the HeartMan system was successful in improving self-care behavior, resulting in a higher quality of disease management, as indicated by the significant ($P=.02$) improvement of 11% in the Self-Care of Heart Failure Index [55]. No such effect was observed on health-related quality of life, as assessed with the Minnesota Living with Heart Failure Questionnaire [56]. Regarding secondary endpoints, using HeartMan significantly ($P<.001$) improved psychological outcomes, that is, intervention patients decreased their level of depression (Beck Depression Inventory II [57]) and anxiety (State Trait Anxiety Inventory Form Y [58]) by 15%, and these reductions were even higher in the patients who had used the
mental support module in the app more intensely. The HeartMan intervention also significantly ($P=.01$) reduced the experience of sexual problems, that is, by 26% on the Sexual Adjustment Scale [59]. No effects were shown for illness perception or clinical outcome of exercise capacity. However, additional data available in a subgroup of the trial sample showed a significant ($P=.04$) improvement of 11% in the left ventricular ejection fraction. A more extensive publication of trial results is pending.

Patients' Perception of the System

The user experience of HeartMan was investigated both qualitatively and quantitatively in the intervention group. Quantitatively, the Unified Theory of Acceptance and Use of Technology (UTAUT) questionnaire was used [60], adapted to the objectives of the HeartMan system and to the population of older adult users [61]. This questionnaire assesses users’ intentions to use the HeartMan system and their usage behavior. The UTAUT questionnaire pointed out that HeartMan users’ attitude toward the system was generally positive, with low scores on technology anxiety related to this positive attitude and relatively high-performance expectancy (“the degree to which the user expects that using the system will help him or her to attain gains in job performance” [60]).

Qualitatively, semistructured interviews were performed with 10 patients (7 men and 3 women) and their informal caregivers after having participated in the trial for 3–4 months [62]. The results of an in-depth analysis of sociotechnical complexities in home-based health monitoring systems [63] showed some potential for the HeartMan system as a tool for self-management. Although stressful for some participants, collecting health data such as weight and BP in the HeartMan trial generally raised awareness among the patients of their lifestyle and health. Monitoring their health parameters enabled them to be more aware of their bodies, intervene, and ask for help in a timely manner. The evaluations also showed that the HeartMan system positively affected patients’ dietary knowledge and that they felt stimulated to engage in physical activities. This suggests that self-monitoring and empowerment goals are generally achieved. Some weaknesses were also found, such as the need for increased flexibility regarding the interface and interactions with the system.

Discussion

Technology

The HeartMan system is complex, spanning sensing devices, a mobile app, and the cloud; combining diverse technologies; and featuring extensive content to comprehensively address CHF management. The challenge of integrating all this was tackled by an architecture with independent components connected through the IoTool middleware as well as the interface and interoperability layer. A lesson learned was that there is a tradeoff between too tight integration, which makes changes difficult, and too many layers between components, which makes integration testing difficult.

Individual components largely performed as expected. BP estimation from PPG proved the most difficult, as this is a difficult research problem even in ideal conditions, when high-quality PPG signals from a clinical or research device are available. Thus, this technology is not yet sufficiently mature for everyday use by patients. In the DSS, we mainly relied on expert knowledge, and only recommendations regarding temperature and humidity were provided by data-based methods. Although we believe data-based decisions will play a greater role in health management in the future, the amount of raw data currently available to support the range of decisions needed to manage a disease such as CHF cannot yet rival the expert knowledge available in the literature and medical practice. Although that knowledge is ultimately based on data, these data are simply not available in one place (and possibly not at all in some cases).

Medical Perspective

Although the use of telemonitoring systems in cardiac patients has increased tremendously, evidence regarding their effectiveness in managing patients with CHF remains to be mixed [64]. HeartMan, however, is different from most telemonitoring systems: it focuses on empowering patients to properly manage their disease, rather than remote monitoring by health care professionals. It mainly aims to improve the quality of life and self-management in patients by integrating several intervention modalities in the domains of physical health management and psychological support. The trial results showed that the obtained beneficial effects were mostly psychological, more than physical, which is in line with the predefined primary outcomes. A possible explanation is that the system did not achieve sufficient adherence to the advanced and gradually progressive exercise program, which would probably be the most effective way to improve physical health. Nonetheless, before drawing definite conclusions, we need to investigate the effectiveness of the HeartMan system in a wider context, that is, in a larger sample of patients with CHF over a longer intervention period.

User Perspective

As early as during the analysis of the patients’ context of use, the HeartMan concept was presented to patients and their initial reactions were captured. Several insights gathered in this phase remained relevant during later evaluation phases and applied to patient-monitoring systems in general. One of the most important such insights was the fact that patients tend to have high and not necessarily correct expectations of automatic patient-monitoring systems such as HeartMan. Patients tend to expect their caregivers to be continuously aware of what the system detects. Although this can lead to a positive motivation to monitor health parameters, it can also lead to a false sense of safety. In addition, while many patients were motivated to monitor these health parameters, they were closely related to lifestyle choices, such as nutrition and physical exercise. We learned that several patients disliked the fact that HeartMan monitors these lifestyle choices and are concerned about a possible loss of control and autonomy in this respect.

These observations lead to a nuanced view of the patients’ perspective on self-monitoring technology, with both perceived benefits (feeling of reassurance, increased awareness) and drawbacks (false perception of safety and loss of autonomy). This view suggests that patient empowerment truly is the correct
goal, in the sense that patients should not rely on the supervision of caregivers (as it may not be available) and should also not feel judged and controlled by the system (but should be making healthy lifestyle choices for themselves). We also observe that although HeartMan started on the way to this goal, further improvements can still be made.

On a more practical level, we learned that a distinction between patients regarding digital literacy can be useful [29]. The patients with high literacy received a full explanation of HeartMan functionality at the beginning of the trial. They were encouraged to be proactive and to navigate through the various functions of the application, which was empowering. Such use was feasible because the interface, particularly the information hierarchy of the application, was designed, tested, and refined in collaboration with the patients. Patients with lower digital literacy were asked to react primarily to notifications in the app. In this way, they were able to cope with the app that, even though it was designed to be simple, it was still relatively complex for some users.

Conclusions

We developed HeartMan, a personal health system for the comprehensive self-management of CHF. It uses a wristband and other sensing devices to obtain information on the patient’s BP, physical activity, and psychological profile by means of machine learning as well as some other parameters by more mundane means. All this information is fed into a DSS, which provides recommendations on physical health and psychological support. These translate into a detailed physical exercise program, mindfulness exercises, games, and other forms of support for the patient. This is adapted to the patient’s physical capacity, current activity, and psychological profile. A web application for medical professionals is also a part of the system. Patients with CHF were involved throughout the development of the system to ensure the system meets their needs. The final prototype was evaluated in a proof-of-concept trial in 56 patients, showing significantly improved disease management while reducing depression, anxiety, and sexual problems. Although illness perception and exercise capacity did not improve, a significant improvement in left ventricular ejection fraction was observed in a subgroup. Overall, the patients’ perception of the system was positive.

The HeartMan system was designed with both patients and medical professionals. It works best when integrated with a hospital information system to have access to the users’ up-to-date health records and to provide information on the users to their treating clinicians. As such, it bridges the gap between user-friendly mHealth solutions and medical devices, but it can only be offered to patients through a health provider. Therefore, we are also working on a simplified version of the system that will not be a medical device from a regulatory perspective and will not require connection to a hospital or any kind of backend. This will make it easily deployable via mobile app stores and widely accessible to patients with CHF.

In summary, the HeartMan project combined a range of advanced technologies with human-centered design to develop a complex system that was shown to help patients with CHF. Its benefits were psychological more than physical, which may be because the system did not manage to cause difficult behavioral changes such as increased exercise. The reason for this may be that the system was designed to be more supportive than persuasive. Thus, a key area for future development should be behavior change techniques. Nevertheless, the system is ready to be used, and we are pursuing multiple paths to the market.

Acknowledgments

The HeartMan project received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement no. 689660. The project partners are Jožef Stefan Institute, Sapienza University, Ghent University, National Research Council, Atos Spain SA, SenLab, KU Leuven, Bittium Biosignals Ltd, and European Heart Network. The authors acknowledge financial support from the Slovenian Research Agency (research core funding no. P2-0209).

Conflicts of Interest

None declared.

References


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
FHII: fast health care interoperability resource
MAE: mean absolute error
mHealth: mobile health
PPG: photoplethysmogram
SBP: systolic blood pressure
SVM: support vector machine
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by C Lovis; submitted 02.10.20; peer-reviewed by H Papadopoulos, P Athilingam; comments to author 15.11.20; revised version received 30.12.20; accepted 11.01.21; published 05.03.21.

Please cite as:
A Personal Health System for Self-Management of Congestive Heart Failure (HeartMan): Development, Technical Evaluation, and Proof-of-Concept Randomized Controlled Trial
JMIR Med Inform 2021;9(3):e24501
URL: https://medinform.jmir.org/2021/3/e24501
doi:10.2196/24501
PMID:33666562

©Mitja Luštrek, Marko Bohanec, Carlos Cavero Barca, Maria Costanza Ciancarelli, Els Clays, Amos Adeyemo Dawodu, Jan Derboven, Delphine De Smedt, Erik Dovgan, Jure Lampe, Flavia Marino, Miha Mikar, Giovanni Pioggia, Paolo Emilio Puddu, Juan Mario Rodríguez, Michele Schiariti, Gašper Šlapničar, 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 Laguna⁴, 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 Rocío 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 Rocío University Hospital, Seville, Spain

Corresponding Author:
Alicia Martinez-Garcia, PhD
Group of Research and Innovation in Biomedical Informatics, Biomedical Engineering and Health Economy
Institute of Biomedicine of Seville
IBiS / Virgen del Rocío University Hospital / CSIC / University of Seville
Av Manuel Siurot
Seville, 41013
Spain
Phone: 34 955 01 36 16
Email: alicia.martinez.garcia@juntadeandalucia.es

Abstract

Background: The evidence-based medicine (EBM) paradigm requires the development of health care professionals’ skills in the efficient search of evidence in the literature, and in the application of formal rules to evaluate this evidence. Incorporating this methodology into the decision-making routine of clinical practice will improve the patients’ health care, increase patient safety, and optimize resources use.

Objective: The aim of this study is to develop and evaluate a new tool (KNOWBED system) as a clinical decision support system to support scientific knowledge, enabling health care professionals to quickly carry out decision-making processes based on EBM during their routine clinical practice.

Methods: Two components integrate the KNOWBED system: a web-based knowledge station and a mobile app. A use case (bronchiolitis pathology) was selected to validate the KNOWBED system in the context of the Paediatrics Unit of the Virgen Macarena University Hospital (Seville, Spain). The validation was covered in a 3-month pilot using 2 indicators: usability and efficacy.

Results: The KNOWBED system has been designed, developed, and validated to support clinical decision making in mobility based on standards that have been incorporated into the routine clinical practice of health care professionals. Using this tool, health care professionals can consult existing scientific knowledge at the bedside, and access recommendations of clinical protocols established based on EBM. During the pilot project, 15 health care professionals participated and accessed the system for a total of 59 times.

Conclusions: The KNOWBED system is a useful and innovative tool for health care professionals. The usability surveys filled in by the system users highlight that it is easy to access the knowledge base. This paper also sets out some improvements to be made in the future.

(JMIR Med Inform 2021;9(3):e13182) doi:10.2196/13182
KEYWORDS
evidence-based medicine; clinical decision support system; scientific knowledge integration

Introduction

Currently, in developed countries, the concept of evidence-based medicine (EBM) is part of medicine itself. In the beginning, the EBM meant a paradigm change in the way that clinical practice was accomplished, leaving a process regarding learning and practice based on static knowledge and authority. However, the EBM concept assumes that the scientific-medical knowledge must emerge from clinical experimentation, and must be used, criticized, and qualitatively interpreted with the best available methodology. Consequently, this knowledge must be essentially dynamic. In the EBM general approach, this knowledge, in conjunction with the clinical experience and the patient’s preferences and data, should directly influence the clinical decision-making process at all the levels of care, considering that the goal of EBM is to improve the patient’s health care quality through enhanced clinical practice [1,2].

Clinical practice is carried out at many complexity levels, so the necessary knowledge to perform it according to the EBM concept must adapt to the real conditions to use the highest quality information possible. The EBM knowledge sources are categorized according to the usability that allows them to be incorporated into the clinical decision-making process at any level in which this process takes place. The usability of the knowledge source shows a direct relationship with the complexity of its methodology, and is therefore better assimilated by the decision process. As a result, the products with detailed information are also more difficult to be incorporated in the health system environment.

Although the EBM supposed a change of attitude in clinical systems, ensuring efficient support to the clinical decisions that must be taken in the patient–doctor relationship context (where it is not easy to consult nor perform an in-depth reading of the original research) was always difficult. In parallel with the EBM conceptual consolidation, some clinical researchers proposed systematic methodologies to achieve products based on the knowledge, to reduce the distance between the research and the practice, thereby saving time for health care professionals in the critical interpretation of the evidence during decision making.

These products were hierarchized in models in 3 proposals. The first one, in 2001, developed a 4-level classification [3]. The second, in 2007, proposed a classification of 5 levels [4]. And the more recent one, in 2009, developed a 6-level classification [5], and has been recently used in relevant research [6,7].

More concretely, the first proposal [3] defined a classification of the following 4 levels:

- Studies: original papers published in journals.
- Syntheses: recompilation of the existing evidence about a specific issue (eg, systematic revisions).
- Synopses: the most relevant elements of a set of evaluated primary studies, including evaluating the methodological quality (eg, the ACP Journal Club).
- Systems: integrate information about the rest of the levels with electronic health records (EHRs).

Comparing this model with the more recent proposal, the 6-level model [5], the main differences between these are (1) the synopses repositories on systematic studies published in scientific reviews that some institutions maintain, and (2) the editorial products that integrate the best practice, in terms of efficacy, according to the explicit and rigorous methods, such as clinical practice guidelines (CPGs) or evidence-based manuals. Probably, CPGs have been the most relevant attempt to inform about the quotidian clinical decisions, and their institutional adoption and individual use are currently accepted criteria for good practice. However, CPGs are complex so their adoption and use are difficult, even in the best of circumstances.

At the top of the pyramid in the 3 proposals, the clinical decision support systems (CDSSs) appear. The CDSSs are the clinical information systems in charge of integrating and summarizing the relevant information about the clinical problems, actualizing, and connecting this information with the patient’s situation. The generalization of the EHR makes possible knowledge integration and records management, allowing the habitual use of the evidence at the patient’s bedside. Incorporating the CDSS in the EHR is a tough challenge that is not solved yet [8,9].

Adopting an EHR by a health care organization involves making organizational decisions to register and maintain patients’ health data, including changes. However, this adoption also makes possible the approach of other types of choices, such as integrating the evidence-based decision support [10].

Consequently, EBM-based interventions improve patient safety. Any clinical intervention must comply with the beneficence principle to the patient, and it is an obligation not to add damage that exceeds the initial clinical condition. Effectiveness and safety are the 2 dimensions that determine the degree of quality of the interventions because no intervention should be assumed to be ineffective even if its cost is zero. The context in which patient care is practiced—the health system—requires improving the effectiveness of interventions and optimizing the efficiency of resources, because health care, whatever its nature, offers a balance between benefits, risks, inconveniences, and costs. Areas such as public health, nursing, and even health policies (called evidence-based health care) have been incorporated into the EBM to ensure the optimal functioning of health systems. It is necessary to extend the dissemination of systematic reviews and clinical guidelines to include electronic access to EHR for all devices, including smartphones [11,12].

EBM contributes to the knowledge in all these dimensions to increase the quality of the intervention. This knowledge is dispersed in many CPGs applied to generalize those actions of proven effectiveness within a specialty or a clinical condition. Despite this, a substantial variation in the provision of services and patient management is documented, between institutions and between professionals of the same institution. The result is known as variability in clinical practice, which can compromise...
the quality of the services themselves beyond the health care professionals’ actions and the resource allocation equity [13-15]. EBM tends to reduce this variability, promoting the adoption of the most effective, safe, and efficient practices. CDSSs to support translational medicine have been proposed by some researchers [16,17].

The scientific knowledge integration at the bedside with a mobile platform enables health care professionals to make faster and more effective decisions based on validated clinical practice experience. In this sense, a CDSS called the KNOWBED system [18] has been designed, which provides to the health care professional clinically relevant questions concerning the pathology of interest. These questions are associated with recommendations at the bedside, based on the scientific evidence in different existing knowledge bases (eg, massive reference bases, CPGs, systematic reviews). The global architecture of the KNOWBED system is designed as a secure, scalable, standards-based, and EBM service–oriented architecture. Regarding scalability, the infrastructure in which the KNOWBED system has been developed supports more than a dozen similar projects, so it is prepared to receive an even more significant number of users, providing service to all physicians who want to use it within a health system. The fact that the KNOWBED system generates and indexes a set of recommendations from existing scientific evidence, offering intelligent assistance for health professionals, makes it a system based on EBM.

This paper aims to disseminate the KNOWBED project results, highlighting the benefits identified using a CDSS to integrate scientific knowledge at the bedside, encouraging the scientific community to use this kind of system.

The paper is structured as follows: after this introduction, where a brief review of relevant EBM work is presented, we expose the methods carried out. Then, the study results obtained are discussed. Finally, the discussion and conclusions are presented.

**Methods**

**Overview of System Components**

Functionally, 2 components integrate the KNOWBED system: a web-based knowledge station and a mobile app.

The knowledge station’s actors are the knowledge managers who use the system to manage all the information shown in the mobile app. In other words, the knowledge managers collect the information coming from the existing scientific knowledge in the bibliography and, at the same time, index the clinical recommendations and questions that usually arise throughout the clinical practice, which will be accessible by context based on the HL7 Infobutton standard [19,20]. OpenInfobutton service, from the University of Utah, was used for this task. This service uses contextual information (based on the HL7 Infobutton standard) about the patient, user, clinical setting, and EHR task to anticipate clinicians’ and patients’ information needs. Furthermore, this service retrieves information from online provider reference and patient education resources that may help meet their information needs. This web service exposes an endpoint that receives all the previously detailed information, and returns a JSON format response. This response is processed to offer access through links to the different sources of information provided by it. The effort to deal with the conflict between recommendations from different sources is made by the knowledge manager technologically supported by the knowledge station.

The mobile app allows health care professionals to have access to scientific knowledge from their mobiles device—both smartphones and tablets—as it provides access to questions and clinical recommendations to follow regarding patients’ diagnosis, admission, treatment, etc., indexed by knowledge managers.

**System Technological Architecture**

From a technological point of view, the KNOWBED system is based on the development and deployment of 2 different modules (Figure 1): the knowledge station and the question manager.

The knowledge station is a web application that allows access to health professionals from the health care centers through their workstations. This web application will be responsible for visualizing, managing, and maintaining the information associated with the knowledge bases defined in the KNOWBED system. For the development of this web application, the Angular 2 framework has been used which, through HTML5, Sass, and TypeScript, allows the development of web applications based on the SPA paradigm (Single-Page-Application). The PostgreSQL relational database engine supported storage and knowledge management, which stores the information associated with questions, recommendations, and suggestions defined for each of the knowledge areas established in the KNOWBED system. The communication between the application and the server uses the HTTP protocol utilizing the Angular 2 built-in HTTP library. The system has communications security based on tokens generated on the server, and managed in the application through the JWT library; these tokens are renewed in each new connection or after a while.

The question manager offers a multipurpose 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 Paediatrics Unit of the Virgen Macarena University Hospital (Seville).

Bronchiolitis is a common viral infection of the lower respiratory tract that affects children under 2 years of age. This pathology is characterized by acute infection and inflammation of the small airways in the lungs [21,22]. It is the most frequent cause of non-elective admission to the intensive care unit [23,24]. Other researchers have performed studies to improve bronchiolitis management using the technology [25,26].

Based on these considerations, and considering this pathology has a greater incidence during the winter months [27], the pilot was carried out between December and February. In this way, the system was more frequently used and more useful for health care professionals’ clinical decision making.

**System Evaluation**

The system was evaluated using 2 indicators: usability and efficacy.

The KNOWBED system usability was assessed to evaluate the health care professionals’ acceptance, using an ad hoc survey asking users regarding the functionalities (Multimedia Appendix 1). The survey recorded sociodemographic information (sex, date of birth, and job title) as well as 13 items that were answered with a 10-point Likert scale (1=strongly disagree; 10=strongly agree) [28]. The survey was administered in 2 phases: phase 1, before using the technological system to know their expectations of the system before interacting with it; and phase 2, to learn about their experience after using the mobile app. Likewise, when new health care professionals joined the Paediatrics Unit, they were informed about the mobile app and were invited to use it.

By contrast, the system’s efficacy was studied by analyzing the percentage of acceptance of the recommendations generated by the system. This acceptance was studied by launching the following question when leaving the system: “You are going to leave the App, was this App useful to you?”

**Results**

**Development and Evaluation of the KNOWBED System**

The KNOWBED system has been developed to incorporate scientific evidence into daily clinical practice, improving patient care and providing health care professionals with recommendations based on up-to-date and relevant scientific knowledge.

A screenshot of the KNOWBED knowledge station is shown in Figure 2, which presents the section to add new recommendations, specifying the type, the source, the date, and possible observations.
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.

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.
Original Paper

Detection of Bulbar Involvement in Patients With Amyotrophic Lateral Sclerosis by Machine Learning Voice Analysis: Diagnostic Decision Support Development Study

Alberto Tena¹, MSc; Frances Claria², PhD; Francesc Solsona², PhD; Einar Meister³, PhD; Monica Povedano⁴, PhD

¹Information and Communication Technologies Group, International Centre for Numerical Methods in Engineering, Barcelona, Spain
²Department of Computer Science, Universitat de Lleida, Lleida, Spain
³Institute of Cybernetics, Tallinn University of Technology, Tallinn, Estonia
⁴Motoneuron Functional Unit, Hospital Universitari de Bellvitge, Barcelona, Spain

Corresponding Author:
Francesc Solsona, PhD
Department of Computer Science
Universitat de Lleida
Jaume II, 69
Lleida
Spain
Phone: 34 973702735
Email: francesc.solsona@udl.cat

Abstract

Background: Bulbar involvement is a term used in amyotrophic lateral sclerosis (ALS) that refers to motor neuron impairment in the corticobulbar area of the brainstem, which produces a dysfunction of speech and swallowing. One of the earliest symptoms of bulbar involvement is voice deterioration characterized by grossly defective articulation; extremely slow, laborious speech; marked hypernasality; and severe harshness. Bulbar involvement requires well-timed and carefully coordinated interventions. Therefore, early detection is crucial to improving the quality of life and lengthening the life expectancy of patients with ALS who present with this dysfunction. Recent research efforts have focused on voice analysis to capture bulbar involvement.

Objective: The main objective of this paper was (1) to design a methodology for diagnosing bulbar involvement efficiently through the acoustic parameters of uttered vowels in Spanish, and (2) to demonstrate that the performance of the automated diagnosis of bulbar involvement is superior to human diagnosis.

Methods: The study focused on the extraction of features from the phonatory subsystem—jitter, shimmer, harmonics-to-noise ratio, and pitch—from the utterance of the five Spanish vowels. Then, we used various supervised classification algorithms, preceded by principal component analysis of the features obtained.

Results: To date, support vector machines have performed better (accuracy 95.8%) than the models analyzed in the related work. We also show how the model can improve human diagnosis, which can often misdiagnose bulbar involvement.

Conclusions: The results obtained are very encouraging and demonstrate the efficiency and applicability of the automated model presented in this paper. It may be an appropriate tool to help in the diagnosis of ALS by multidisciplinary clinical teams, in particular to improve the diagnosis of bulbar involvement.

(JMIR Med Inform 2021;9(3):e21331) doi:10.2196/21331

KEYWORDS
amyotrophic lateral sclerosis; bulbar involvement; voice; diagnosis; machine learning

Introduction

Background

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease with an irregular and asymmetric progression, characterized by a progressive loss of both upper and lower motor neurons that leads to muscular atrophy, paralysis, and death, mainly from respiratory failure. The life expectancy of patients with ALS is between 3 and 5 years from the onset of symptoms. ALS produces muscular weakness and difficulties...
of mobility, communication, feeding, and breathing, making the patient heavily dependent on caregivers and relatives and generating significant social costs. Currently, there is no cure for ALS, but early detection can slow the disease progression [1].

The disease is referred to as spinal ALS when the first symptoms appear in the arms and legs (limb or spinal onset; 80% of cases) and bulbar ALS when it begins in cranial nerve nuclei (bulbar onset; 20% of cases). Patients with the latter form tend to have a shorter life span because of the critical nature of the bulbar muscle function that is responsible for speech and swallowing. However, 80% of all patients with ALS experience dysarthria, or unclear, difficult articulation of speech [2]. On average, speech remains adequate for approximately 18 months after the first bulbar symptoms appear [3]. These symptoms usually become noticeable at the beginning of the disease in bulbar ALS or in later stages of spinal ALS. Early identification of bulbar involvement in people with ALS is critical for improving diagnosis and prognosis and may be the key to effectively slowing progression of the disease. However, there are no standardized diagnostic procedures for assessing bulbar dysfunction in ALS.

Speech impairment may begin up to 3 years prior to diagnosis of ALS [3], and as ALS progresses over time there is significant deterioration in speech [4]. Individuals with ALS with severe dysarthria present specific speech production characteristics [5-7]. However, it is possible to detect early, often imperceptible, changes in speech and voice through objective measurements, as suggested in previous works [8-11]. The authors concluded that phonatory features may be well suited to early ALS detection.

**Related Work**

Previous speech production studies have revealed significant differences in specific acoustic parameters in patients with ALS. Carpenter et al [7] studied the articulatory subsystem of individuals with ALS and found different involvement of articulators—that is, the tongue function was more involved than the jaw function. In a recent study, Shellkerié et al [5] found that the maximum speed of tongue movements and their duration were only significantly different at an advanced stage of bulbar ALS compared with the healthy control group. Connaghan et al [12] used a smartphone app to identify and track speech decline. Lee et al [6] obtained acoustic patterns for vowels in relation to the severity of the dysarthria in patients with ALS.

Other works have demonstrated the efficiency of features obtained from the phonatory subsystem for detecting early deterioration in ALS [8-11,13-15]. Studies have shown significant differences between jitter, shimmer, and the harmonics-to-noise ratio (HNR) in patients with ALS [8,10,11]. More specifically, Silbergleit et al [8] obtained these features from a steady portion of sustained vowels that provided information regarding changes in the vocal signal that are believed to reflect physiologic changes of the vocal folds. Alternative approaches used formant trajectories to classify the ALS condition [13], correlating formants with articulatory patterns [14], fractal jitter [15], Mel Frequency Cepstral Coefficients (MFCCs) [16], or combined acoustic and motion-related features [9] at the expense of introducing more invasive measurements to obtain data. Besides, the findings revealed significant differences in motion-related features only at an advanced stage of bulbar ALS.

Other related studies, such as one by Frid et al [17], used speech formants and their ratios to diagnose neurological disorders. Teixeira et al [18] and Mekyska et al [19] suggested jitter, shimmer, and HNR as good parameters to be used in intelligent diagnosis systems for dysphonia pathologies. Garcia-Gancedo et al [20] demonstrated the feasibility of a novel digital platform for remote data collection of digital speech characteristics, among other parameters, from patients with ALS.

In the literature, classification models are widely used to test the performance of acoustic parameters in the analysis of pathological voices. Norel et al [21] identified acoustic speech features in naturalistic contexts and machine learning models developed for recognizing the presence and severity of ALS using a variety of frequency, spectral, and voice quality features. Wang et al [9] explored the classification of the ALS condition using the same features with support vector machine (SVM) and neuronal network (NN) classifiers. Rong et al [22] used SVMs with two feature selection techniques (decision tree and gradient boosting) to predict the intelligible speaking rate from speech acoustic and articulatory samples. Suhas et al [16] implemented SVMs and deep neuronal networks (DNNs) for automatic classification by using MFCCs. An et al [23] used convolutional neuronal networks (CNNs) to compare the intelligible speech produced by patients with ALS to that of healthy individuals. Gutz et al [24] merged SVM and feature filtering techniques (SelectKBest). In addition, Vashkevich et al [25] used linear discriminant analysis (LDA) to verify the suitability of the sustain vowel phonation test for automatic detection of patients with ALS.

Among feature extraction techniques, principal component analysis (PCA) [26] shows good performance in a wide range of domains [27,28]. Although PCA is an unsupervised technique, it can efficiently complement a supervised classifier in order to achieve the objective of the system. In fact, any classifier can be used in conjunction with PCA because it does not make any kind of assumption about the subsequent classification model.

**Hypothesis**

Based on previous works, our paper suggests that the acoustic parameters obtained through automated signal analysis from a steady portion of sustained vowels may be used efficiently as predictors for the early detection of bulbar involvement in patients with ALS. For that purpose, the main objectives (and contributions) of this research were (1) to design a methodology for diagnosing bulbar involvement efficiently through the acoustic parameters of uttered vowels in Spanish; and (2) to demonstrate that the performance of the automated diagnosis of bulbar involvement is superior to human diagnosis.

To fulfill these objectives, 45 Spanish patients with ALS and 18 control subjects took part in the study. They were recruited by a neurologist, and the five Spanish vowel segments were
elicited from each participant. The study focused on the extraction of features from the phonatory subsystem—jitter, shimmer, HNR, and pitch—from the utterance of each Spanish vowel.

Once the features were obtained, we used various classification algorithms to perform predictions based on supervised classification. In addition to traditional SVMs [9, 16, 21, 22, 24], NNs [9, 16, 23], and LDA [25], we used logistic regression (LR), which is one of the most frequently used models for classification purposes [29, 30]; random forest (RF) [31], which is an ensemble method in machine learning that involves the construction of multiple tree predictors that are classic predictive analytic algorithms [22]; and naive Bayes (NB), which is still a relevant topic [32] and is based on applying Bayes’ theorem. Prior to feeding the models, PCA was applied to the features obtained due to the good performance observed of this technique in a wide range of domains.

Methods
Participants
The study was approved by the Research Ethics Committee for Biomedical Research Projects (CEIm) at the Bellvitge University Hospital in Barcelona, Spain. A total of 45 participants with ALS (26 males and 19 females) aged from 37 to 84 (mean 57.8, SD 11.8) years and 18 control subjects (9 males and 9 females) aged from 21 to 68 (mean 45.2, SD 12.2) years took part in this transversal study. All participants with ALS were diagnosed by a neurologist.

Bulbar involvement was diagnosed by following subjective clinical approaches [33], and the neurologist made the diagnosis of whether a patient with ALS had bulbar involvement. Of the 45 participants with ALS, 5 reported bulbar onset and 40 reported spinal onset, but at the time of the study 14 of them presented bulbar symptoms.

To summarize, of the 63 participants in the study, 14 were diagnosed with ALS with bulbar involvement (3 males and 11 females; aged from 38 to 84 years, mean 56.8 years, SD 12.3 years); 31 were diagnosed with ALS but did not display this dysfunction (23 males and 8 females, aged from 37 to 81 years, mean 58.3 years, SD 11.7 years); and 18 were control subjects (9 males and 9 females; aged from 21 to 68 years, mean 45.2 years, SD 12.2 years).

The severity of ALS and its bulbar presentation also varied among participants, as assessed by the ALS Functional Rating Scale-Revised (ALSFRS-R). The ALSFRS-R score (0-48) was obtained from 12 survey questions that assess the degree of functional impairment, with the score of each question ranging from 4 (least impaired) to 0 (most impaired). The scores of the 45 participants in this study ranged from 6 to 46 (mean 31.3, SD 8.6; 3 patients’ scores were reported as not available). Within the subgroups, the scores of patients diagnosed with bulbar involvement ranged from 6 to 46 (mean 23.1, SD 9.8), and the scores of participants with ALS who did not present this dysfunction ranged from 17 to 46 (mean 30.2, SD 8.0; 3 patients’ scores reported as not available).

The main clinical records of the participants with ALS are summarized in Multimedia Appendix 1.

Vowel Recording
The Spanish phonological system includes five vowel segments—a, e, i, o, and u. These were obtained and analyzed from each patient with ALS and each control participant, all of whom were Spanish speakers.

Sustained samples of the Spanish vowels a, e, i, o, and u were elicited under medium vocal loudness conditions for 3–4 s. The recordings were made in a regular hospital room using a USB GXT 252 Emita Streaming Microphone (Trust International BV) connected to a laptop. The speech signals were recorded at a sampling rate of 44.100 Hz and 32-bit quantization using Audacity, an open-source application [34].

Feature Extraction
Each individual phonation was cut out and anonymously labeled. The boundaries of the speech segments were determined with an oscillogram and a spectrogram using the Praat manual [35] and were audibly checked. The starting point of the boundaries was established as the onset of the periodic energy in the waveform observed in the oscillogram and checked by the apparition of the formants in the spectrogram. The end point was established as the end of the periodic oscillation when a marked decrease in amplitude in the periodic energy was observed. It was also identified by the disappearance of the waveform in the oscillogram and the formants in the spectrogram.

Acoustic analysis was done by taking into account the following features: jitter, shimmer, HNR, and pitch. Once the phonations of each participant had been segmented, the parameters were obtained from each vowel through the standard methods used in Praat [35]; they are explained in detail in this section and consist of a short-term spectral analysis and an autocorrelation method for periodicity detection.

Jitter and shimmer are acoustic characteristics of voice signals. Jitter is defined as the periodic variation from cycle to cycle of the fundamental period, and shimmer is defined as the fluctuation of the waveform amplitudes of consecutive cycles. Patients with lack of control of the vibration of the vocal folds tend to have higher values of jitter. A reduction of glottal resistance causes a variation in the magnitude of the glottal period correlated with breathiness and noise emission, causing an increase in shimmer [18].

To compute jitter parameters, some optional parameters in Praat were established. Period floor and period ceiling, defined as the minimum and maximum durations of the cycles of the waveform that were considered for the analysis, were set at 0.002 s and 0.025 s, respectively. The maximum period factor—the largest possible difference between two consecutive cycles—was set at 1.3. This means that if the period factor—the ratio of the duration of two consecutive cycles—was greater than 1.3, this pair of cycles was not considered in the computation of jitter.

The methods used to determine shimmer were almost identical to those used to determine jitter, the main difference being that
jitter considers periods and shimmer takes into account the maximum peak amplitude of the signal.

Once the previous parameters had been established, jitter and shimmer were obtained by the formulas shown below [35].

Jitter(absolute) is the cycle-to-cycle variation of the fundamental period (ie, the average absolute difference between consecutive periods):

\[
\text{Jitter(absolute)} = \frac{1}{N} \sum_{i=1}^{N} |T_i - T_{i-1}|
\]

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

Jitter(relative) is the average absolute difference between consecutive periods divided by the average period. It is expressed as a percentage:

\[
\text{Jitter(relative)} = \frac{1}{\bar{T}} \sum_{i=1}^{N} |T_i - \bar{T}|
\]

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:

\[
\text{Jitter(rap)} = \frac{1}{\bar{T}} \sum_{i=1}^{N} |T_i - \frac{1}{2}(T_{i-1} + T_{i+1})|
\]

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:

\[
\text{Shimmer(dB)} = 20 \log_{10} \left( \frac{A_{\text{max}} - A_{\text{min}}}{2} \right)
\]

where \(A_i\) is the extracted peak-to-peak amplitude data and \(N\) is the number of extracted fundamental periods.

Shimmer(relative) is defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude, expressed as a percentage:

\[
\text{Shimmer(relative)} = \frac{1}{\bar{A}} \sum_{i=1}^{N} |A_i - \bar{A}|
\]

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:

\[
\text{Shimmer(apq3)} = \frac{1}{\bar{A}} \sum_{i=1}^{N} |A_i - \frac{1}{3}(A_{i-1} + A_{i} + A_{i+1})|
\]

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:

\[
\text{Shimmer(apq5)} = \frac{1}{\bar{A}} \sum_{i=1}^{N} |A_i - \frac{1}{5}(A_{i-2} + A_{i-1} + A_i + A_{i+1} + A_{i+2})|
\]

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:

\[
\text{Shimmer(apq11)} = \frac{1}{\bar{A}} \sum_{i=1}^{N} |A_i - \frac{1}{11}(A_{i-5} + A_{i-4} + \ldots + A_{i+4} + A_{i+5})|
\]

The HNR provides an indication of the overall periodicity of the voice signal by quantifying the ratio between the periodic (harmonics) and aperiodic (noise) components. The HNR was computed using Praat [35], based on the second maximum of normalized autocorrelation function detection, which is used in the following equation:

\[
\text{HNR} = \frac{r(t = \tau)}{r(t = 0)}
\]

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].
Textbox 1. Algorithm for obtaining the features (jitter, shimmer, harmonics-to-noise ratio [HNR], and pitch) for acoustic analysis.

1. Each individual phonation of each vowel was cut out and anonymously labeled to define the boundaries of the speech segments.
2. The values for the optional 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(rap), jitter(ppq5), shimmer(dB), shimmer(relative), shimmer(apq3), shimmer(apq5), shimmer(apq11)—using the configuration parameters established and then obtain the mean of each of these parameters for each vowel.
4. Compute HNR using the configuration parameters established and then obtain the mean (HNR[mean]) and standard deviation (HNR[SD]) values.
5. Compute pitch using the configuration parameters established and then obtain the mean (pitch[mean]), standard deviation (pitch[SD]), minimum (pitch[min]), and maximum (pitch[max]) values.
6. Obtain a data set with the 15 features computed.

PCA

The PCA technique [37], a ranking feature extraction approach, was implemented in R [38] using the Stats package [38]. PCA was used to decompose the original data set into principal components (PCs) to obtain another data set whose data were linearly independent and therefore uncorrelated. It was performed by means of singular value decomposition (SVD) [39].

Prior to applying PCA, given that the mean age of control subjects was approximately 12 years younger than patients with ALS, we removed the age effects by using the data from the control subjects and applying the correction to all the participants as in the study by Norel et al [21]. We fitted the features extracted for healthy people and their age linearly. Then, the “normal aging” of each single feature of each participant was obtained by multiplying the age of the participants by the slope parameter obtained from the linear fit. Finally, the computed “normal aging” was removed from the features. Afterward, a standardized data set was obtained by subtracting the mean and centering the age-adjusted features at 0.

Then, by applying SVD to the standardized data set, a decomposition was obtained: $X = USV^T$, 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 involvement (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: \( \frac{x_i - \mu}{\sigma} \), where \( x_i \) is the feature vector, \( \mu \) 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 sigmoid 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.
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&lt;sup&gt;a&lt;/sup&gt; vs B&lt;sup&gt;b&lt;/sup&gt;</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>Naïve 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>

<sup>a</sup>C: control group.  
<sup>b</sup>B: patients with amyotrophic lateral sclerosis (ALS) with bulbar involvement.  
<sup>c</sup>NB: patients with ALS without bulbar involvement.  
<sup>d</sup>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\textsuperscript{a} vs B\textsuperscript{b}</th>
<th>C vs NB\textsuperscript{c}</th>
<th>B vs NB</th>
<th>C vs ALS\textsuperscript{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>58.3</td>
<td>56.1</td>
<td>68.8</td>
<td>75.1</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>4.8</td>
<td>30.4</td>
<td>0.0</td>
<td>65.6</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>98.8</td>
<td></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>82.3</td>
<td>68.8</td>
<td>72.8</td>
<td>75.1</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>64.7</td>
<td>54.6</td>
<td>15.8</td>
<td>65.6</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>96.1</td>
<td>93.3</td>
<td>98.6</td>
<td>98.8</td>
<td></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>92.8</td>
<td>77.7</td>
<td>74.1</td>
<td>76.0</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>84.8</td>
<td>65.1</td>
<td>16.7</td>
<td>66.4</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>99.0</td>
<td>99.6</td>
<td>100.0</td>
<td>100.0</td>
<td></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>88.1</td>
<td>70.6</td>
<td>71.7</td>
<td>71.1</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>72.7</td>
<td>53.5</td>
<td>0.9</td>
<td>59.5</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td></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>92.6</td>
<td>84.8</td>
<td>73.1</td>
<td>86.8</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>83.2</td>
<td>76.1</td>
<td>20.5</td>
<td>81.6</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>100.0</td>
<td>100.0</td>
<td>96.8</td>
<td>99.8</td>
<td></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>86.3</td>
<td>71.1</td>
<td>70.7</td>
<td>71.1</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>68.8</td>
<td>54.3</td>
<td>6.1</td>
<td>59.4</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td></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.

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%. 

https://medinform.jmir.org/2021/3/e21331
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 vocal 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 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 patients 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 perturbation in their voices could still not be appreciated by human hearing.

Figure 1 also indicates that the means of the features of the patients in the NB group were between the means of the features of the C and B groups, thus corroborating these assumptions.

Limitations

This study has some limitations. First, using machine learning on small sample sizes makes it difficult to fully evaluate the significance of the findings. The sample size of this study was heavily influenced by the fact that ALS is a rare disease. At the time of the study, 14 of the patients with ALS presented bulbar symptoms. The relatively small size of this group was because ALS is a very heterogeneous disease and not all patients with ALS present the same symptomatology. Additionally, the control subjects were approximately 12 years younger than the patients with ALS. Vocal quality changes with age, and comparing younger control subjects’ vocalic sounds with those of older participants with ALS might introduce additional variations.
Although upsampling techniques were used in this study to correct the bias and age adjustments have been applied to correct the vocal quality changes due to the age difference, it would be necessary in future studies to increase the number of participants, especially of patients with ALS with bulbar involvement and control participants of older ages, to draw definitive conclusions.

Second, the variability inherent in establishing the boundaries of the speech segments on spectrograms manually makes replicability challenging. Speakers will differ in their production, and even the same speaker in the same context will not produce two completely identical utterances. In this study, the recorded speech was processed manually in the uniform approach detailed in the Methods section. Automatic instruments have been developed, but unfortunately these methods are not yet accurate enough and require manual correction.

**Comparison with Prior Work**

The PCA biplot charts indicated that shimmer and jitter were the most important features for group separation in the 2-PC model for ALS classification; however, they also revealed pitch and HNR parameters as good variables for this purpose. These results are consistent with those of Vashkevich et al [25], who demonstrated significant differences in jitter and shimmer in patients with ALS. They are also consistent with Mekyska et al [19] and Teixeira et al [18], who mentioned pitch, jitter, shimmer, and HNR values as the most popular features describing pathological voices. Finally, Silbergleit et al [8] suggested that the shimmer, jitter, and HNR parameters are sensitive indicators of early laryngeal deterioration in ALS.

Concerning the classification models, Norel et al [21] recently implemented SVM classifiers to recognize the presence of speech impairment in patients with ALS. They identified acoustic speech features in naturalistic contexts, achieving 79% accuracy (sensitivity 78%, specificity 76%) for classification of males and 83% accuracy (sensitivity 86%, specificity 78%) for classification of females. The data used did not originate from a clinical trial or contrived study nor was it collected under laboratory conditions. Wang et al [9] implemented SVM and NN using acoustic features and adding articulatory motion information (from tongue and lips). When only acoustic data were used to fit the SVM, the overall accuracy was slightly higher than the level of chance (50%). Adding articulatory motion information further increased the accuracy to 80.9%. The results using NN were more promising, with accuracies of 91.7% being obtained using only acoustic features and increasing to 96.5% with the addition of both lip and tongue data. Adding motion 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.

[16] Tena et al.
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


36. Voice features extraction. Alberto Tena. URL: https://github.com/atnad/greco [accessed 2021-02-01]


47. Supervised classification models for automated detection of bulbar involvement in als patients. Alberto Tena. URL: https://github.com/atenad/greco [accessed 2021-02-01]

48. 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 analysis
PC: principal component
RF: random forest
SVD: singular value decomposition
SVM: support vector machine
TN: true negative
TP: true positive
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
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 DeLong 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.

**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><strong>Performance expectancy: perception that using CPOE will improve the physician’s job performance</strong></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</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><strong>Social influence: perceived importance of others’ (eg, leaders, colleagues) opinions that the physician should or should not use the system</strong></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><strong>Facilitating conditions: available resources, facilities, and infrastructure that facilitate using CPOE</strong></td>
<td></td>
<td></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><strong>Information quality: relevance, accuracy, comprehensiveness, understandability, prevalence, timeliness, and usability of the outputs or content</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usefulness of error messages</td>
<td>1</td>
<td>[32]</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><strong>System quality: reliability, functionality, flexibility, ease of use, integration, and response time of the system</strong></td>
<td></td>
<td></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 usage was associated with physicians’ effort to learn and use the CPOE system [24,25,27].
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 CPOE 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. Customized refers to tailoring

**Technological Factors**

Technological factors included the technical and design aspects of CPOE in terms of the system’s quality; information quality; and its reliability, functionality, flexibility, ease of use, integration, and response time [19]. Evidence from 8 of the included studies [24-26,28,31-34] indicated that the factors related to CPOE were the most relevant for affecting its use by physicians. A total of 17 factors were reported (Table 1). The system’s efficiency was the most cited factor [31,32,34], specifically the quick prescribing process [31], fast data retrieval, response time [32], and the system’s speed, in terms of entering patient data [34]. Furthermore, studies that reported the system’s speed as an influential factor in its use by physicians were conducted shortly after the implementation phase, that is, halfway through the intervention year (about 6 months later), shortly after implementation (not clear), and 3 months after implementation. This finding suggests that because the system was newly implemented, the processing speed was significant for physicians’ performance of tasks.

The findings indicate that ease of use, the effect of using CPOE on quality of care, training, availability of technical support, time, and the system’s speed were the factors with the strongest influence on the use of CPOE for medication prescribing among all the studies.

**Discussion**

**Principal Findings and Comparisons With Other Works**

CPOE for medication prescribing can serve physicians as a tool to enhance patient quality of care. However, this has not led to a rapid uptake of the system by health organizations and clinicians to use it [6,14]. A key factor in the slow adoption of CPOE by health care organizations is attributed to the costs associated with installing the system and the costs of sustaining it [6]. The first CPOE was installed in the United States in 1971 [36]. Although that was long ago, the adoption rate in health organizations is still rare to moderate, with a percentage of 15.7% [13]. This low adoption rate has been reported in other countries [8,9].

Despite many years of implementation of CPOE for medication prescription, development, and research, the issue of low adoption postimplementation remains. This study focuses on the usage of the user—the physician—after the system has been implemented. We identified factors that were related to the users (physicians), organization, and technological aspects of CPOE that influence the actual use of CPOE by physicians for medication prescribing, rather than intention to use a CPOE system.

The findings of this study are consistent with those of Van Dort et al [14] and Gagnon et al [12]. Nevertheless, these reviews identified other factors that were not found in this study. Resistance to use was reported in both reviews [12,14], as a factor that negatively affected the usage of the system by physicians for medication prescribing. CDS systems embedded in the CPOE system for medication prescribing were examined in Van Dort et al [14]. As CDS systems are known to offer suggestions and recommendations, user resistance was present as the physicians reported concerns that the information presented might not be reliable [14].

In addition to resistance to using CPOE, Gagnon et al [12] described how the system could negatively affect the patient-clinician relationship and identified financial issues as another influential factor, neither of which was detected in this study. This inconsistency might be because of the focus of this study on the actual use of CPOE after the system had been installed and used and resistance is no longer an issue.

This study showed that technological factors related to the system were the most frequently reported factors that influenced how a physician used the CPOE system for medication prescribing. This finding is consistent with the results reported by Gagnon et al [12]. As their findings suggest, technical and design concerns were the most frequently identified factors limiting the system’s use [12].

One of the principal findings of this study is that among the 3 main themes, 5 factors were cited most frequently (any factor cited 3 or more times was considered frequently cited), indicating that it was significant in the physicians’ decisions about using the CPOE system. Quality of care, ease of use, training, availability of technical support, time constraints, and system speed were key factors in the use of CPOE by physicians. A similar pattern of results has been reported in an extensive body of literature [12,14,37,38]. One unexpected finding was that the effect of alert fatigue, as a factor in the use of CPOE, was identified in only 2 studies [24,33]. Alert fatigue is the receipt of a massive amount of reminders or warnings that cost time and effort and is eventually ignored [39].

This finding contradicts the observation that alert fatigue has previously been found to be associated with the usage of CPOE for medication prescribing. In their review, Gagnon et al [12] showed that alert fatigue was associated with the use of an electronic prescription system in 5 studies. In addition, Van Dort et al [14] showed that too many irrelevant alerts were related to the uptake of medication-related CPOE 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. Customized 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.
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
Original Paper

Comparative Analysis of Paper-Based and Web-Based Versions of the National Comprehensive Cancer Network-Functional Assessment of Cancer Therapy-Breast Cancer Symptom Index (NFBSI-16) Questionnaire in Breast Cancer Patients: Randomized Crossover Study

Jinfei 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 administrating the questionnaire; however, the reliability of the NFBSI-16 in electronic format has not been assessed.

Objective: This study aimed to assess the reliability of a web-based NFBSI-16 questionnaire in breast cancer patients undergoing systematic treatment with a prospective open-label randomized crossover study design.

Methods: We recruited random patients with breast cancer under systematic treatment from the central hospital registry to complete both paper- and web-based versions of the questionnaires. Both versions of the questionnaires were self-assessed. Patients were randomly assigned to group A (paper-based first and web-based second) or group B (web-based first and paper-based second). A total of 354 patients were included in the analysis (group A: n=177, group B: n=177). Descriptive sociodemographic characteristics, reliability and agreement rates for single items, subscales, and total score were analyzed using the Wilcoxon test. The Lin concordance correlation coefficient (CCC) and Spearman and Kendall \(\tau\) 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 \(\rho\)) were documented for all 4 subscales—Disease-Related Symptoms Subscale–Physical (\(\rho=0.93\)), Disease-Related Symptoms Subscale–Emotional (\(\rho=0.85\)), Treatment Side Effects Subscale (\(\rho=0.95\)), and Function and Well-Being Subscale (\(\rho=0.91\))—and total NFBSI-16 score (\(\rho=0.94\)). 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 version 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]. Gwalney et al’s meta-analysis on assessing the equivalence of computer versus paper versions of PROMs showed “a high overall level of agreement between paper and computerized measures” [28]. The review encompassed the fields of rheumatology, cardiology, psychiatry, asthma, alcoholism, pain assessment, gastrointestinal disease, diabetes, and allergies. In contrast, a study of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30 found small but statistically significant differences in scale mean scores (3 to 7 points on a 100-point scale) associated with mode of administration [29]. Various validated web-based questionnaires in oncology have been demonstrated to be reliable and effective tools for assessing PROMs in therapeutic clinical and research settings [30-33]. Currently in China, web-based versions of clinical research questionnaires using WeChat are rapidly growing in number, and various studies have validated the WeChat-based administration of health-related questionnaires [34-36]. To cover the large growing patient base in China, we expected web-based administration of the NFBSI-16 to be a reliable methodology to assess the impact of disease, treatment, and well-being status among patients with breast cancer. Additionally, it could be a more cost-effective and efficient method to apply in the growing number of patients in certain demographics.

The aim of this study was to analyze the reliability of a web-based NFBSI-16 questionnaire (Chinese language) for measuring disease- and treatment-related symptoms and concerns in breast cancer patients, comparing it with the validated paper-based version.

Methods

Study Design and Patient Enrolment

Patients were recruited from the Department of Breast Surgery of the First Affiliated Hospital of China Medical University, Shenyang, China, between October 2019 and January 2020. The inclusion criteria were female gender, full legal age, proven diagnosis of breast cancer, being under systemic anticancer treatment, ability to follow study instructions, sufficient literacy and fluency in Chinese to comprehend the questionnaires, ability and willingness to complete the study protocol, and signed declaration of consent. Potential participants were excluded if they could not provide informed consent or participated in other studies (burden of participation). Participants had an initial clinic visit at which eligibility was assessed. All eligible participants were randomly chosen from the hospital’s central registry and invited to volunteer for the study via face-to-face interview with a trained research clinician. Written informed consents were obtained. The study protocol was approved by the First Affiliated Hospital of China Medical University ethics committee.

The study was a randomized crossover design in which all participants completed both a standard paper questionnaire and a web-based version of the NFBSI-16 (Multimedia Appendix 1). Patients in group A were assigned to start with the paper-based version followed by the web-based version on their smartphone in the same session. Patients in group B completed the web-based version followed by the paper-based version. Participants were randomized immediately after enrolment to group A or B in a 1:1 ratio using a computer-generated randomization list with a specified seed and block size of 6, based on the mode of administration to be completed first. Between each session from paper-based to web-based and web-based to paper-based, participants were given a break of 15 minutes during which they were invited into a quick patient education seminar, which was also a routine activity in our...
department as a distractor task to lower the potential carryover effect. All participants were provided with written instructions for completion of the paper- or web-based questionnaires prior to their questionnaires being administered. After completing both versions of the NFBSI-16 questionnaire, participants were invited to state their preference for either the paper- or web-based NFBSI-16 questionnaire.

**Questionnaire**

The NFBSI-16 contains items from the original FBSI and FACIT measures selected by patients and clinicians according to their priority concern [15], which presented as a more direct tool to reflect the effectiveness of treatments for advanced breast cancer. The NFBSI-16 comprises 16 items with 4 dimensions for ease of use and scoring: Disease-Related Symptoms Subscale–Physical (DRS-P), Disease-Related Symptoms Subscale–Emotional (DRS-E), Treatment Side Effects Subscale (TSE), and Function and Well-Being Subscale (FWB). Therefore, clinicians and researchers can individually view and assess subscale scores when concerned about a particular class of symptoms. The questionnaires were self-completed, and careful attention was paid to the design and layout of the web-based version. In order to reduce the risk of errors in posing, interpretation, recording, and coding responses and potential interrater variability, the theory-based guidelines for self-administered questionnaire design were followed by the authors (Multimedia Appendix 1) [37]. The web-based user interface and paper for the paper-based questionnaires were free from all other information such as logos, slogans, advertisement, etc. The instructions for completing the web-based and paper-based questionnaires were included at the beginning of the web-based interface and header of the paper, respectively. In brief, while participating in the web-based assessment, patients had to scan a redesignated Quick Response code using their smartphone. This action automatically took them to a web-based test, and the user had to select the intensity or severity of the 16 items. After completing the 16th question, the interface turned into a blank screen indicating the test was over. On the other hand, the paper-based questionnaire test was conducted using white paper and pencil. The text was printed using clear 12-point font.

**Testing of the Instrument**

During pretesting and pilot testing, 3 colleagues specializing in oncology and 3 nonexperts evaluated the web-based questionnaire’s usability, accessibility, and clarity of the user interface. This testing was only conducted on the functionality of the web-based questionnaire since the format, structure, and sequence of items in the web-based questionnaire were the same as in the validated paper-based questionnaire.

**Computation of Subscale Scores**

Data from the paper questionnaires were entered manually into an electronic patient management system by the authors, and data from the web-based questionnaires were automatically captured after the participant completed the online questionnaire and downloaded to the electronic patient management system. All data was anonymized. We assessed the completeness of the data on a per-item basis and questionnaire basis. The total scores were obtained by taking the mean score across completed items and multiplying by 16, the number of items (following official guidelines) [15]. All subscale totals ranged from 0 to 4, with a score of 0 representing that the patient agrees with the item “not at all” and 4 representing “very much”. Subscale scores and total scores were computed for each participant and each mode of administration separately. Comparative analyses of individual items, subscales, and total score were the primary goal of the study.

**Statistical Analysis**

All statistical analyses were conducted using SPSS Statistics, version 25 (IBM Corp). Frequency analysis was performed to determine the descriptive sociodemographic characteristics of patients. Referring to ISPOR ePRO Good Research Practices Task Force recommendations [21], we conducted the evaluation of measurement equivalence. Reliability, internal consistency, disparity of responses, and the rate of consistency between paper- and web-based responses were assessed. Reliability was calculated for the 16 individual items as well as for scores of the 4 subscales (DRS-P, DRS-E, TSE, and FWB) and the NFBSI-16 total score in accordance with the NFBSI-16 guidelines [15]. The primary study outcome was to assess the reliability of single items and total score of the web-based questionnaire. The Wilcoxon test was used to identify possible statistically significant differences in the test of parallel forms reliability, both between the single items and the scores due to the ordinal nature of the data. The secondary outcome measure was to assess the consistency and agreement of the web-based questionnaire with the paper-based questionnaire. The mean values of the paper- and web-based measures were calculated, consistency analyses were performed by calculation of the Spearman rank correlation coefficient (Spearman ρ), and agreement rates for each item were assessed using rank correlation (Kendall τ) for each scale. As a second measure of test-retest reliability, we calculated the Lin concordance correlation coefficient (CCC) [38]. Finally, all answers to the “preference” questionnaire were compared between the web-based and the paper version of the NFBSI-16 using χ² tests. In all analyses, P<.05 (2-tailed) was considered indicative of statistically significant differences (α=.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-16 items</th>
<th>Paper-based patient score</th>
<th>Web-based patient score</th>
<th>P value</th>
<th>Δ [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 (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>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.
Figure 2. Box plot comparison of paper-based and web-based distribution of total scores.

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

<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>Δ [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</td>
<td>(9.59)</td>
<td>35.64</td>
<td>(9.58)</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>(30–42)</td>
<td>34</td>
<td>(10–42)</td>
</tr>
<tr>
<td>Disease-Related Symptoms Subscale—Emotional</td>
<td>32.00</td>
<td>(16.54)</td>
<td>32.05</td>
<td>(16.84)</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>(16–32)</td>
<td>32</td>
<td>(16–32)</td>
</tr>
<tr>
<td>Treatment Side Effects Subscale</td>
<td>39.20</td>
<td>(13.39)</td>
<td>39.20</td>
<td>(12.86)</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>(28–48)</td>
<td>36</td>
<td>(32–48)</td>
</tr>
<tr>
<td>Function and Well-Being Subscale</td>
<td>43.37</td>
<td>(13.37)</td>
<td>43.62</td>
<td>(13.72)</td>
</tr>
<tr>
<td></td>
<td>42.67</td>
<td>(32–53.33)</td>
<td>42.67</td>
<td>(32–53.33)</td>
</tr>
</tbody>
</table>


Test of Internal Consistency

Table 4 shows the Spearman ρ correlation values between the individual items from the paper- and web-based questionnaires. All 16 items demonstrated a high correlation (>0.8) between paper- and web-based items. Individual item internal consistency test was performed by Kendall τ analysis between the two versions. In all items, the rank correlation was high as the Kendall τ coefficients ranged between 0.787 and 0.877 and were all statistically significant. With each data point reflecting an individual patient’s total NFBSI-16 score, Figure 3 depicts a positive correlation between total paper-based and web-based scores. Overall, CCC agreement between paper-based and web-based questionnaires’ item scores were all comparably high at 0.94 (fair: 0.21–0.40; moderate: 0.41–0.60; substantial: 0.61–0.80; almost perfect: 0.81–1.00), as represented in Table 5.
Table 4. Correlation between test-retest in individual items and subscale (Spearman ρ and Kendall τ analysis).

<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>HI7 (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><strong>Disease-Related Symptoms Subscale – Physical (DRS-P)</strong></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><strong>Disease-Related Symptoms Subscale – Emotional (DRS-E)</strong></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><strong>Treatment Side Effects Subscale (TSE)</strong></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><strong>Function and Well-Being Subscale (FWB)</strong></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><strong>Total score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score</td>
<td>0.94</td>
<td>0.93-0.95</td>
</tr>
</tbody>
</table>

$R_c^a$: 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</td>
<td>98</td>
<td>177</td>
<td>-79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>questionnaire</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preferred web-based</td>
<td>256</td>
<td>177</td>
<td>79</td>
<td>70.50 (1)</td>
<td>.001b</td>
</tr>
<tr>
<td>Total</td>
<td>354</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 177.0.
bStatistically significant difference.

Estimation of the Carryover Effect

To assess the carryover effect, we let $s_A$ denote the sum (total scores from web-based items plus the total scores from paper-based items for each respondent) from group A and let $s_B$ denote the sum from group B. We estimate the carryover effect in both groups (A and B) using the Wilcoxon test on the sum values $s_A$ and $s_B$, and at a level of significance of 5%, the
possible carryover effect is not significantly different between the different sequences ($P=0.84$).

**Discussion**

**Principal Results**

Overall, reliability was considered to be excellent for the web-based version as measured with the Wilcoxon signed rank test and CCC. Additionally, Spearman $\rho$ correlation and Kendall $\tau$ 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. 2013225585).

**Conflicts of Interest**

None declared.
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

Please cite as:
URL: https://medinform.jmir.org/2021/3/e18269
doi:10.2196/18269
PMID:33650978

©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.
Machine Learning Approach to Predict the Probability of Recurrence of Renal Cell Carcinoma After Surgery: Prediction Model Development Study

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 naive 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.

Keywords
renal cell carcinoma; recurrence; machine learning; naive Bayes; algorithm; cancer; surgery; web-based; database; prediction; probability; carcinoma; kidney; model; development

Introduction

Renal cell carcinoma (RCC) accounts for 90% of malignant tumors in the kidney and is twice as common in men as in women [1]. Kidney cancer, therefore, generally refers to RCC. It is the sixth most frequently diagnosed cancer in men and the 10th most frequently diagnosed cancer in women worldwide [2]. According to the cancer statistics from the National Cancer
Center, the number of new kidney cancer cases in Korea in 2017 was 5299, accounting for approximately 2.3% of the total of 232,255 cancer cases. Further, the incidence of kidney cancer per 100,000 people has been increasing since 1999 [3]. RCC is one of the most lethal types of malignant tumors in urology, with approximately 20% to 30% of patients with RCC suffering from metastatic diseases, and more than 40% of patients eventually die of the disease [4-6]. The main treatment for RCC is radical nephrectomy; for small tumors, partial nephrectomy is performed to preserve kidney function [7].

RCC can be completely cured through full surgical resection if there is no evidence of preoperative metastatic disease. However, it has a high recurrence rate of 20% to 30% [8,9], and approximately 50% of recurrences occur within 2 years [8,10]. RCC recurrence is generally classified as early recurrence or late recurrence based on the 5-year threshold [11]. Most recurrences occur during the early recurrence period (within 5 years) [11,12], whereas approximately 10% occur during the late recurrence period (after 5 years) [11,13].

RCC is generally resistant to radiation and chemotherapy, making treatment of its recurrence difficult [4]. Therefore, it is necessary to predict the probability of RCC recurrence so that risk factors can be managed in advance. The Memorial Sloan Kettering Cancer Center (MSKCC) in the United States developed a nomogram that predicts the probability of recurrence within 5 years using the symptoms and histology of 601 patients with kidney cancer who received surgical treatment in 2001 [14]. Additionally, in 2005, a nomogram was developed to predict the recurrence probability within 5 years using the pathological stage, Fuhrman nuclear grade, tumor size, necrosis, vascular invasion, and clinical presentation variables of 701 patients with kidney cancer [15]. Previous studies have used small-scale RCC cohorts from single institutions, and the data have included censored data, where the values of the observations were only partially known. If censored data are included, they can be applied in the Cox proportional hazards model, a standard statistical technique for modeling censored data, but they are difficult to apply to other machine learning (ML) techniques [16].

In this study, we used a multicenter, large-scale RCC cohort collected from eight tertiary care hospitals in Korea; we removed censored data and used only the fully observed data. ML focuses on building new predictive models by performing extensive searches on multiple models and parameters and then performing validation [17]. The objective of this study was to develop an algorithm that could predict the recurrence probability of RCC after surgery within 5 and 10 years by applying eight representative ML algorithms to a large-scale Korean RCC cohort. Using the developed algorithm, clinicians can manage postoperative patient outcomes and establish personalized treatment strategies.

Methods

Study Population

The data used in this study were obtained from a large-scale cohort of Korean patients with RCC assembled from the KOrean Renal Cell Carcinoma (KORCC) web-based database. It consisted of 206 variables, including demographic information such as age, height, and weight, as well as pathological information, including clinical stage, pathological stage, Fuhrman nuclear grade, and survival period [18]. The study protocol was approved by the institutional review board of the Catholic University of Korea (IRB No. KC20ZIDI0966). The data of 6849 patients who participated in the KORCC study group as of July 1, 2015, were collected from eight tertiary hospitals.

Variable Selection and Data Cleansing

The t test for continuous variables and the chi-square test for categorical variables were used to explore variables that significantly affect recurrence. In both tests, variables with missing values were removed to ensure that the data used were complete and without missing values. At a significance level of $P=.05$, we first extracted 31 variables showing significant differences between the recurring and nonrecurring groups. Of the 31 variables extracted, 10 variables that had significant effects on recurrence in actual clinical trials were finally extracted based on the expert advice of a urologist. The final 10 selected variables were gender, age, BMI, smoking, pathological tumor stage, histological type, necrosis, lymphovascular invasion, capsular invasion, and Fuhrman nuclear grade.

Several studies reported that age $\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>226 (10.04)</td>
<td>2025 (89.96)</td>
</tr>
<tr>
<td>After</td>
<td>2025 (50.00)</td>
<td>2025 (50.00)</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 ≥pT2 was approximately 60.4% (168/278) in the recurrence group and 15.2% (386/2536) in the nonrecurrence group. Approximately 77.7% (216/278) of the patients in the recurrence group and 44.8% (1135/2536) of those in the nonrecurrence group had Fuhrman nuclear grades ≥3; thus, the recurrence group had higher Fuhrman nuclear grades. The distribution of each category of pathological variables is shown in Table 2.
<table>
<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/m²), mean (SD)</td>
<td>23.6 (3.2)</td>
<td>24.7 (3.3)</td>
</tr>
<tr>
<td><strong>Gender, n (%)</strong></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><strong>Smoking, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsmoker</td>
<td>2026 (79.9)</td>
<td>2026 (79.9)</td>
</tr>
<tr>
<td>Current smoker</td>
<td>207 (74.5)</td>
<td>2026 (79.9)</td>
</tr>
<tr>
<td><strong>Pathological tumor stage, n (%)</strong></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><strong>Histologic type, n (%)</strong></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><strong>Necrosis, n (%)</strong></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><strong>Lymphovascular invasion, n (%)</strong></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><strong>Capsular invasion, n (%)</strong></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><strong>Fuhrman nuclear grade, n (%)</strong></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></td>
<td>5-year</td>
<td>10-year</td>
<td>5-year</td>
<td>10-year</td>
</tr>
<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></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(100, 100)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.556</td>
<td>0.519</td>
<td>0.905</td>
<td>0.898</td>
</tr>
<tr>
<td>(10, 10)</td>
<td>0.467</td>
<td>0.426</td>
<td>0.947</td>
<td>0.928</td>
</tr>
<tr>
<td>(50, 50)</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></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(10, 100)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.822</td>
<td>0.731</td>
<td>0.850</td>
<td>0.828</td>
</tr>
<tr>
<td>Random forest (number of trees)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5, 5)&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>(10, 10)</td>
<td>0.511</td>
<td>0.423</td>
<td>0.866</td>
<td>0.861</td>
</tr>
<tr>
<td>(50, 50)</td>
<td>0.511</td>
<td>0.442</td>
<td>0.875</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></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(50, 200)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.733</td>
<td>0.692</td>
<td>0.815</td>
<td>0.810</td>
</tr>
<tr>
<td>(10, 10)</td>
<td>0.600</td>
<td>0.577</td>
<td>0.895</td>
<td>0.845</td>
</tr>
<tr>
<td>(50, 50)</td>
<td>0.733</td>
<td>0.673</td>
<td>0.815</td>
<td>0.824</td>
</tr>
<tr>
<td>(100, 100)</td>
<td>0.711</td>
<td>0.692</td>
<td>0.835</td>
<td>0.802</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></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(50, 100)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.688</td>
<td>0.635</td>
<td>0.819</td>
<td>0.826</td>
</tr>
<tr>
<td>(10, 10)</td>
<td>0.756</td>
<td>0.596</td>
<td>0.667</td>
<td>0.849</td>
</tr>
<tr>
<td>(50, 50)</td>
<td>0.688</td>
<td>0.615</td>
<td>0.819</td>
<td>0.826</td>
</tr>
<tr>
<td>(100, 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: naïve 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: KOrean Renal Cell Carcinoma
ML: machine learning
MSKCC: Memorial Sloan Kettering Cancer Center
NB: naive Bayes
RCC: renal cell carcinoma
SMOTE: synthetic minority oversampling technique
SVM: support vector machine
Predictive Modeling of 30-Day Emergency Hospital Transport of German Patients Using a Personal Emergency Response: Retrospective Study and Comparison with the United States

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.](https://medinform.jmir.org/2021/3/e25121)

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 (e.g., 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 (i.e., that shared a pseudonym with another subscriber) were also excluded. After applying inclusion and exclusion criteria, data from 5805 subscribers remained for analysis.

Data Processing

The retrospective data set included a table consisting of subscriber data with a single row for each subscriber and a case data table with each row representing a single case. The tables were processed in the statistical programming software R (R Foundation for Statistical Computing).

The case data were characterized in terms of case types, case reasons, and case outcomes (Table 1). The case data for each subscriber were then aggregated by determining the frequency and recency of each of the case types, reasons, and outcomes. The frequency represents the number of a particular case that the subscriber has experienced, while the recency represents the time that has passed since the subscriber has experienced a particular case. Up to 2 years of historic case data were used to derive these features. The frequency and recency features of case data cases were then combined with subscriber demographics, support network, and self-reported medical conditions and medications from the subscriber data table. Tables were merged based on the pseudonymized subscriber IDs.

Table 1. Case types, reasons, and outcomes for which frequency and recency features are derived for input into the predictive model. Examples are given per category.

<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>

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 (i.e., 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—the training and test sets were mutually exclusive with regard to users. The US test set comprised 109,966 PERS users. Since the German data set was smaller, multiple prediction dates were considered by splitting the time range January 1, 2012, through January 1, 2018, into 73 equally spaced 30-day windows. This resulted in a training set containing 96,273 (79.5%) prediction dates and a test set with 24,847 (20.5%) prediction dates.

PERS users were on average aged 84.0 years in the German training set. Average age was slightly, but statistically significantly, younger in the test set (83.6 years). The average age in the US test set was statistically significantly lower at 81.2 years compared with the German test set. About two-thirds (2997/4644, 64.5%) of German PERS users were female, with no statistically significant difference between training and test sets. However, the US test set showed a significantly higher proportion of females (88,433/109,966, 80.4%).

In the German training set, more than half of users (2598/4644, 55.9%) were on the service 2 years or less, 23.2% (1079/4644) of users were 2 to 4 years on the service, and 20.8% (967/4644) of users were more than 4 years on the service. These percentages were not statistically significantly different in the test set. In the US test set, 44.4% (48,922/109,966) of users were less than 2 years on the service, which was significantly lower than in the Germany test set (630/1161, 54.3%). A similar percentage of US PERS users were 2 to 4 years on the service (26,193/109,966, 23.8%, vs 264/1161, 22.7%), while more users were 4 or more years on the service (34,851/109,966, 31.7%, vs 267/1161, 23.0%).

In the training set for the Germany predictive model, 94.3% (4397/4644) of users had at least one self-reported medical condition, with 35.7% (1657/4644) reporting 5 or more conditions. There were no statistically significant differences between the number of self-reported conditions in the training and test set for the Germany predictive model. In contrast, 77.3% (85,056/109,966) of users in the test set for the US predictive model self-reported one or more medical conditions, which was statistically significantly lower than in the test set for the Germany predictive model.

The prevalence of the dependent variable “emergency hospital transport in the next 30 days” was 1.6% (1506/96,273) in the training set and 1.4% (350/24,847) in the test set for the Germany predictive model. The latter was statistically significantly lower than the prevalence of the dependent variable in the test set of the US predictive model (2455/109,966, 2.2%).
Table 2. Subscriber characteristics and prevalence of the dependent variable in the training and test sets for the Germany predictive model compared with the previously published results of the US predictive model in the test set. *P* values are reported for differences between German test and training sets, and between US and German test sets.

<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>General</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction dates</td>
<td>Jan 1, 2012, to Jan 1, 2018</td>
<td>—</td>
</tr>
<tr>
<td># of unique PERS(^a) 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></td>
<td></td>
</tr>
<tr>
<td>0-2</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></td>
<td></td>
</tr>
<tr>
<td>None</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>

\(^a\)PERS: personal emergency response system.

**Predictive Model Evaluation**

The performance of the Germany predictive model on the test set is detailed in Table 3 for various prediction score thresholds. AUC was 0.749 (95% CI 0.721-0.777) for emergency hospital transport in 30 days. This was slightly but not statistically significantly lower than the AUC for the US predictive model (0.778 [95% CI 0.769-0.788]), as 95% CIs were overlapping.

Positive predictive values for the Germany predictive model were low due to the low prevalence of 30-day emergency hospital transport in 30 days. This was 1.4% (350/24,847) in the test set (Table 1). By increasing the prediction score threshold, PPV increased but at the expense of decreased sensitivity. At a prediction score threshold corresponding to the 90th percentile, the Germany predictive model identified 40.3% (95% CI 35.1%-45.4%) of the subscribers who had emergency transport in the 30 days following the prediction date (sensitivity); however, only 5.7% (95% CI 4.9%-6.4%) of flagged subscribers had emergency transport in the following 30 days (PPV) at this threshold. At thresholds corresponding to the 95th and 99th percentiles, the sensitivity dropped to 26.9% (95% CI 22.3%-31.1%) and 10.6% (95% CI 7.4%-14.0%), respectively, while the PPV increased to 7.5% (95% CI 6.3%-8.8%) and 15.0% (95% CI 10.7%-19.3%), respectively. When the threshold was set at the 99th percentile, the PPV was 10.7 times higher than the prevalence of 1.4%. This lift of the prediction model was similar for the US prediction model, namely 11.9.

The US predictive model demonstrated similar sensitivity and specificity values for the different thresholds. However, PPV was significantly higher across all thresholds compared with the Germany predictive model due to the higher prevalence of the target variable in the US data set.
Table 3. Performance of the Germany and US predictive models on the corresponding test sets, evaluated by positive predictive value, sensitivity, and specificity using the 90th, 95th, and 99th percentiles as a threshold and area under receiver operator characteristic curve.

<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>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>5.7 (4.9-6.4)b</td>
<td>9.4 (8.9-9.8)</td>
</tr>
<tr>
<td>95%</td>
<td>7.5 (6.3-8.8)b</td>
<td>13.6 (12.9-14.3)</td>
</tr>
<tr>
<td>99%</td>
<td>15.0 (10.7-19.3)b</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>c</td>
<td>0.749 (0.721-0.777)</td>
<td>0.778 (0.769-0.788)</td>
</tr>
</tbody>
</table>

aPPV: positive predictive value.
bNonoverlapping 95% CI between Germany and US predictive models.
cAUC: area under the receiver operator characteristic curve.

The predictive model produced for each subscriber a probability from 0% to 100% indicating the risk of having 30-day emergency hospital transport. The actually observed percentage of subscribers with 30-day emergency hospital transport and average predicted probabilities are presented in Figure 4 to indicate calibration across deciles of risk for both models. Each decile consists of 10% of the test set sorted by predicted probability. Probabilities increased from 0.4% in the lowest risk decile to 5.7% in the highest risk decile observed in the Germany test set and from 0.2% to 9.7% for the US test set. Both models were well calibrated with $R^2=.9935$ and $R^2=.9992$ for the Germany and US predictive models, respectively.
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>
</tr>
</thead>
<tbody>
<tr>
<td>Case data-based predictors</td>
<td>48</td>
<td>72.9</td>
</tr>
<tr>
<td>US</td>
<td>62</td>
<td>87.7</td>
</tr>
<tr>
<td>Self-reported medical conditions</td>
<td>43</td>
<td>9.6</td>
</tr>
<tr>
<td>US</td>
<td>44</td>
<td>3.7</td>
</tr>
<tr>
<td>Other predictors</td>
<td>7</td>
<td>17.5</td>
</tr>
<tr>
<td>US</td>
<td>15</td>
<td>8.7</td>
</tr>
<tr>
<td>Total</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>US</td>
<td>121</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 running 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.
Regional Resource Assessment During the COVID-19 Pandemic in Italy: Modeling Study

Pietro H Guzzi¹*, PhD; Giuseppe Tradigo²*, PhD; Pierangelo Veltri¹, PhD

¹Department of Surgical and Medical Sciences, University of Catanzaro, CZ, Italy
²Ecampus University, Novedrate, Italy
*these authors contributed equally

Corresponding Author:
Pietro H Guzzi, PhD
Department of Surgical and Medical Sciences
University of Catanzaro
Catanzaro
CZ, Italy
Phone: 39 09613694148
Email: hguzzi@unicz.it

Abstract

Background: COVID-19 has been declared a worldwide emergency and a pandemic by the World Health Organization. It started in China in December 2019, and it rapidly spread throughout Italy, which was the most affected country after China. The pandemic affected all countries with similarly negative effects on the population and health care structures.

Objective: The evolution of the COVID-19 infections and the way such a phenomenon can be characterized in terms of resources and planning has to be considered. One of the most critical resources has been intensive care units (ICUs) with respect to the infection trend and critical hospitalization.

Methods: We propose a model to estimate the needed number of places in ICUs during the most acute phase of the infection. We also define a scalable geographic model to plan emergency and future management of patients with COVID-19 by planning their reallocation in health structures of other regions.

Results: We applied and assessed the prediction method both at the national and regional levels. ICU bed prediction was tested with respect to real data provided by the Italian government. We showed that our model is able to predict, with a reliable error in terms of resource complexity, estimation parameters used in health care structures. In addition, the proposed method is scalable at different geographic levels. This is relevant for pandemics such as COVID-19, which has shown different case incidences even among northern and southern Italian regions.

Conclusions: Our contribution can be useful for decision makers to plan resources to guarantee patient management, but it can also be considered as a reference model for potential upcoming waves of COVID-19 and similar emergency situations.

(JMIR Med Inform 2021;9(3):e18933) doi:10.2196/18933

KEYWORDS
COVID-19; data analysis; ICU; management; intensive care unit; pandemic; outbreak; infectious disease; resource; planning

Introduction

COVID-19 is a disease that was reported in Wuhan, China in December 2019 [1]. It has been stressing health structures and governments worldwide due to the difficulties in containing its diffusion [2-4]. The virus appears as a flu, but it attacks the pulmonary apparatus, and on average, 6% of symptomatic patients require hospitalization in intensive care units (ICUs) due to severe respiratory syndromes. The rapid diffusion of the virus caused a high number of infections. Only a fraction of patients who are infected need hospitalization in ICUs, a relatively high number of which do not survive the virus. In Italy, in almost 2 months during the first peak, there have been more than 100,000 known infections and nearly 35,000 COVID-19–related deaths [5-7].

The virus also spread in other countries across the world with different modalities and aggressiveness. During its first wave, from March to May 2020, Italy had the second highest number
of patients who were infected. All countries reported difficulties in answering to the high number of requests for ICU beds and reliable detection of the real infection numbers. In countries where the virus spread later with respect to Italy, such as the United States, France, Spain, and other European countries, it diffused with similar trends but with different absolute numbers, as reported by the World Health Organization [3,8-13]. Nevertheless, the impact on health structure resources has been similar. In fact, the number of infections detected is strictly related to the number of swab tests performed in the population. Tradigo et al [14] assessed the real number of people who are infected with respect to the known ones. In contrast, the number of infections is much higher than the known ones and includes patients who are asymptomatic (i.e., those who got the infection but who did not manifest any symptoms). The number of ICU beds is always related to the number of real infections.

Regions such as Lombardia in the north of Italy have been strongly affected by COVID-19, and it seems that this is due to a large number of patients who were asymptomatic already in the middle of January 2020 and late adoption of containment measures such as limitation of circulation and delay in applying rules such as smart working [15-17].

We focus on the problem of rapidly estimating resources during the exponential phase of the COVID-19 emergency, in particular, being aware of the differences among regions in terms of health structure resources such as ICU beds. We show how the proposed model scales at the regional level and how it can help decision makers plan expansions of resources near saturation or reroute patients to neighboring regions.

The prediction of COVID-19 diffusion is a relevant problem, and it has been discussed in other papers [18-20]. Some papers do not consider the regional level and local differences that are relevant in some countries such as Italy. For instance, Li et al [21] developed a model starting from Hubei Province data, and they used the model for prediction in other countries such as Italy and Korea. Other papers did not consider the prediction of ICU resources. Conversely, our study is scalable on a regional level, and it is able to predict ICU needs. In a previous study [22-24], we developed a preliminary model for resource planning; here, we present an extension of the model with a particular focus on the assessment of the predictions.

The model presented here has been assessed comparing the simulated and predicted resource values with the measured ones. The rapid diffusion trend in high-income countries’ populations and in cities with high density stressed the health structure in many countries. Indeed, a small portion of patients with COVID-19 require ICU admission. The exponential diffusion in terms of an increased number of infections per day required a larger number of ICU beds than the ones available. We report our model as being scalable at both the regional and subregional levels. We claim that it can be used in different countries and in future contexts where virus diffusion will require well-planned health resource management [13,25].

The paper is structured as follows. The Methods section reports the proposed assessed model and the Italian infection data. The Results section reports the application of the model on three sample regions out of 20, Lombardia (north), Toscana (central), and Sicilia (south). The Discussion section reports on the limitations of this study and comparisons with other work. However, our model is general enough to be successfully applied to other pandemic situations in other contexts.

### Methods

#### ICU Situation in Italy

Italy was affected by COVID-19 by the end of January 2020, starting from northern regions such as Lombardia and Veneto. By the end of February, the increasing trend of infection numbers per day obliged the governments at the regional level—and at the national level starting on March 10—to introduce containment measures.

For example, on March 26, 2020, in Italy, we had 24,747 total reported COVID-19 infection cases, of which 20,603 had the disease, 1809 had died, and 2335 had recovered from it. Regarding patients who were infected, 9268 were treated in their homes since they did not have severe illness, 9663 were hospitalized, and 1672 were admitted to ICUs. The trend continued increasing until April 19, 2020, which has been the peak of COVID-19 infections in Italy.

In reaction to the exponential growth of patients who are infected that require hospitalization, one possible measure adopted by many countries has been to build emergency hospitals dedicated to patients with COVID. In Italy, one strategy consisted in improving existing structures by extending the number of ICU resources and beds, and using dedicated health structures. For example, one study [26] focuses on accelerating the process of acquiring and furnishing hospitals with assisted breathing devices.

Italy has approximately 5200 ICU beds in total, which have been dimensioned by design to be equal to 80% of their average occupancy at any given time. In addition, they are allocated at a regional level proportional to the local population and are usually managed locally. Table 1 reports the ICU bed distribution among regions associated with the demography. The COVID-19 pandemic called these choices into question, thus introducing the necessity of emergency units in cities where the virus rapidly diffused and where existing resources were limited.
Table 1. Distribution of ICU beds in each Italian region ordered by regional population. The number of beds could increase in the future due to government investments for the emergency.

<table>
<thead>
<tr>
<th>Region</th>
<th>ICU^a 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 $y = b e^{ax}$, 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].
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

Edited by G Eysenbach; submitted 27.03.20; peer-reviewed by B Smith, M Hamasha; comments to author 04.05.20; revised version received 05.06.20; accepted 17.01.21; published 09.03.21.

Please cite as:

Guzzi PH, Tradigo G, Veltri P

Regional Resource Assessment During the COVID-19 Pandemic in Italy: Modeling Study

JMIR Med Inform 2021;9(3):e18933

URL: https://medinform.jmir.org/2021/3/e18933

doi: 10.2196/18933

PMID: 33629957

©Pietro H Guzzi, Giuseppe Tradigo, Pierangelo Veltri. 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.