JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 8 (2020), Issue 8 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Reviews

Corrigenda and Addenda

Original Paper

Impact of a Commercial Artificial Intelligence–Driven Patient Self-Assessment Solution on Waiting Times at General Internal Medicine Outpatient Departments: Retrospective Study

Yukinori Harada[1,2], MD; Taro Shimizu[1], MD, MPH, MBA, PhD

[1]Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, Mibu, Japan

[2]Department of General Internal Medicine, Nagano Chuo Hospital, Nagano, Japan

Corresponding Author:
Taro Shimizu, MD, MPH, MBA, PhD
Department of Diagnostic and Generalist Medicine
Dokkyo Medical University
Kitakobayashi 880
Mibu, 321-0293
Japan
Phone: 81 282861111
Fax: 81 282872502
Email: shimizutaro7@gmail.com

Abstract

Background: Patient waiting time at outpatient departments is directly related to patient satisfaction and quality of care, particularly in patients visiting the general internal medicine outpatient departments for the first time. Moreover, reducing wait time from arrival in the clinic to the initiation of an examination is key to reducing patients' anxiety. The use of automated medical history–taking systems in general internal medicine outpatient departments is a promising strategy to reduce waiting times. Recently, Ubie Inc in Japan developed AI Monshin, an artificial intelligence–based, automated medical history–taking system for general internal medicine outpatient departments.

Objective: We hypothesized that replacing the use of handwritten self-administered questionnaires with the use of AI Monshin would reduce waiting times in general internal medicine outpatient departments. Therefore, we conducted this study to examine whether the use of AI Monshin reduced patient waiting times.

Methods: We retrospectively analyzed the waiting times of patients visiting the general internal medicine outpatient department at a Japanese community hospital without an appointment from April 2017 to April 2020. AI Monshin was implemented in April 2019. We compared the median waiting time before and after implementation by conducting an interrupted time-series analysis of the median waiting time per month. We also conducted supplementary analyses to explain the main results.

Results: We analyzed 21,615 visits. The median waiting time after AI Monshin implementation (74.4 minutes, IQR 57.1) was not significantly different from that before AI Monshin implementation (74.3 minutes, IQR 63.7) (P=.12). In the interrupted time-series analysis, the underlying linear time trend (–0.4 minutes per month; P=.06; 95% CI –0.9 to 0.02), level change (40.6 minutes; P=.09; 95% CI –5.8 to 87.0), and slope change (–1.1 minutes per month; P=.16; 95% CI –2.7 to 0.4) were not statistically significant. In a supplemental analysis of data from 9054 of 21,615 visits (41.9%), the median examination time after AI Monshin implementation (6.0 minutes, IQR 5.2) was slightly but significantly longer than that before AI Monshin implementation (5.7 minutes, IQR 5.0) (P=.003).

Conclusions: The implementation of an artificial intelligence–based, automated medical history–taking system did not reduce waiting time for patients visiting the general internal medicine outpatient department without an appointment, and there was a slight increase in the examination time after implementation; however, the system may have enhanced the quality of care by supporting the optimization of staff assignments.

(JMIR Med Inform 2020;8(8):e21056) doi:10.2196/21056

KEYWORDS

artificial intelligence; automated medical history taking system; eHealth; interrupted time-series analysis; waiting time

Introduction

Background

The waiting time at outpatient departments is directly related to patient satisfaction [1]. Patient distrust regarding medical services increases with longer waiting time, specifically in patients visiting for the first time [1]. Compared to those in other departments, long waiting times in the general internal medicine outpatient departments are, particularly, an issue [2]. Low patient satisfaction may lead to poor patient safety from misunderstandings between patients and medical staff and from medical staff handling patient complaints about waiting time leaving less time for other duties such as medical care. Therefore, reducing waiting time for new patients in general internal medicine outpatient departments may play a vital role in maintaining and improving the quality of medical care. Moreover, reducing the waiting time from arrival in the clinic to the initiation of the medical examination appears to be particularly associated with a reduction of patient anxiety [3].

Clinical documentation is time consuming, taking approximately 34% of physician working time in the outpatient department setting [4]. Moreover, physicians can reduce their clinical evaluation time if summaries of patient histories have already been prepared prior to the examination [5]. Such summaries can be prepared by nurses using self-administered questionnaires provided to patients in the waiting room and completed by hand; this is already widely used in hospitals across Japan. This system, however, has several limitations. First, although a handwritten self-administered questionnaire is a patient-friendly and easy method for medical personnel to collect data, it takes a long time to transfer the detailed information correctly into electronic files. Second, some patients may fill the forms only partially [6], contributing to considerable missing data. Third, the quality of information depends on the skills of the nurse in collecting information. Finally, this system leaves nurses with less time to attend to other professional duties, including engaging in direct patient care [6].

Automated medical history–taking devices appear to be a promising solution for reducing the time spent on transferring handwritten data into digitized form. Automated medical history taking itself has a long history since it was introduced in the late 1960s [7,8]. Until recently, automated medical history taking was used outside of clinics or hospitals and took a long time to complete [9,10], but it has now been implemented in hospital and clinic waiting rooms through computing systems [11,12] and takes only 5 to 10 minutes to complete [11-14]. Its usability and acceptance by patients have been on the rise, and most patients (including older adults) can use automated medical history–taking devices without assistance [11-15]. Automated medical history taking is expected to assist physicians in developing differential diagnoses and to improve on accuracy of diagnoses, though this has not been the case previously [16-18]. Overall, computer-generated patient history recorded by an automated medical history–taking device was reported to be of higher quality, more comprehensive, better organized, and of greater relevance than patient information obtained through traditional methods of medical history taking [19]. Moreover, it was reported to be popular among patients, enabling better communication with physicians, helping to enhance the quality of patient care and making the patients more comfortable in answering sensitive questions [20].

However, there is a paucity of data on the efficacy of automated medical history taking in reducing waiting times. A previous study [21] reported that 45%-60% of physicians believed automated medical history taking could be time saving and efficient because fewer questions need be asked of the patient, less writing is necessary, the automated medical history taking provides a good basis for more detailed questioning, the history is more complete, and patients are forced to think about their problems beforehand [21]. Although not statistically analyzed, some physicians reported average time gained using automated medical history taking was 5 minutes (ranging from none to more than 15 minutes) [21].

Hypotheses and Study Goal

Recently, an artificial intelligence (AI)–based automated medical history–taking device, AI Monshin, was developed by Ubie Inc in Japan [22]. AI Monshin is not only an automated medical history–taking system but also a clinical decision support system trained to suggest differential diagnoses based on AI machine learning. Based on the positive outcomes of automated medical history–taking devices [21], we hypothesized that replacing the use of handwritten self-administered questionnaires with a new system using AI Monshin would reduce waiting time in a community hospital general internal medicine outpatient department.

Methods

Study Design

We conducted a retrospective observational study using data from outpatients who visited the Department of General Internal Medicine at the Nagano Chuo Hospital. The Nagano Chuo Hospital is a medium-sized, secondary community general hospital in Nagano City, Japan and has 332 inpatient beds. The Nagano Chuo Hospital Research Ethics Committee approved the study (serial number: Nagano Chuo Byoin 20-3). The requirement to obtain written informed consent from patients was waived because of the retrospective nature of the study.

Patient Population

We included patients who had visited the general internal medicine outpatient department in Nagano Chuo Hospital without an appointment between 8 AM and noon on ordinary weekdays (Monday to Friday, excluding hospital holidays) from April 1, 2017 to April 16, 2020. We implemented AI Monshin in the outpatient department on April 17, 2019.

AI Monshin Tool Presentation

AI Monshin is a software that converts data entered by patients on tablet terminals into technical terms and displays it in the electronic medical record [22]. While in the waiting room, patients enter their age, sex, and symptoms (details can be entered as free text) on a tablet. Consequently, the AI software chooses approximately 20 questions that are tailored to the patient from a pool of 3500 questions. Questions are displayed on the tablet one by one, and patients answer the questions by choosing from the items displayed. The questions are optimized according to previous answers to provide the most relevant list of potential differential diagnoses. It takes approximately 3 minutes to complete the questions [22]. Entered data are summarized and translated into compatible medical text automatically in the patient's electronic medical record. The top 10 possible differential diagnoses based on history generated by the AI software can be used to assist the physician during patient evaluation.

Intervention

Different patient flows were applied before and after the introduction of AI Monshin. Before AI Monshin implementation, patients—upon arrival in the clinic—wrote their symptoms, past medical history, family history, social history, and medication history by hand using self-administered questionnaire forms. Upon completion of the form, patients would be interviewed by a nurse, who would check their vital signs, triage the patient, and transfer the patient's information into the electronic medical record system. A doctor would then examine the patient. During and after the examination, the doctor could also edit the patient's medical records using unstructured free text clinical notes.

After AI Monshin implementation, when patients checked in, patients were asked to enter their medical information using the tablet; 5 tablets were introduced. While 3 nurses were engaged in pre-examination interviews prior to the implementation, after the implementation a clerk staff member was hired to assist patients when using the tablets, and one of these nurses was allocated to engaging in nursing work. Clerk staff assisted those who could not use the tablet. After completing the questions on the tablet, nurses would check the vital signs of the patient and triage. Patient data were automatically summarized and translated into compatible medical text in the electronic medical record. The doctor was able to edit the text during and after clinical examination. From the patients' perspective, the difference between before and after AI Monshin implementation were experienced in the waiting and examination rooms. After the implementation of AI Monshin, the patients were only required to fill the electronic form. Patients did not need to wait to be interviewed by a nurse, which usually was the rate-limiting step in outpatient flow prior to AI Monshin implementation. Moreover, patients could see their summary on the monitor in the examination room and could use the displayed information when communicating with doctors. The patient flow after examination was the same before and after AI Monshin implementation.

Data Collection, Outcomes, and Definitions

We retrospectively collected data, including age, sex, the time of arrival in the hospital, the time of entry into the examination room, and the first registered time of the doctor's data entry in the patient record for each individual visit. The primary outcome measure was median waiting time per patient. We collected data on waiting time both before and after AI Monshin implementation. The secondary outcome measure was the median waiting time per month. We defined the waiting time as the time between arriving in the hospital and the first recorded time of the doctor's data entry in the patient record since the time of entry into the examination room was not recorded in all patients.

Statistical Analyses

We compared the differences in median waiting time before and after AI Monshin implementation using the Wilcoxon rank-sum test. Moreover, we conducted a single-group interrupted time-series analysis [23-25] to evaluate changes in median waiting time per month before and after AI Monshin implementation. We set April 2019 as the start point of implementation. In these analyses, we excluded data with the first recorded time of doctor's data entry earlier than the time of patient's arrival in the hospital. Statistical tests were two-tailed, and a *P* value<.05 was considered statistically significant. We conducted all statistical analyses using R (version 3.6.3; The R Foundation for Statistical Computing).

Results

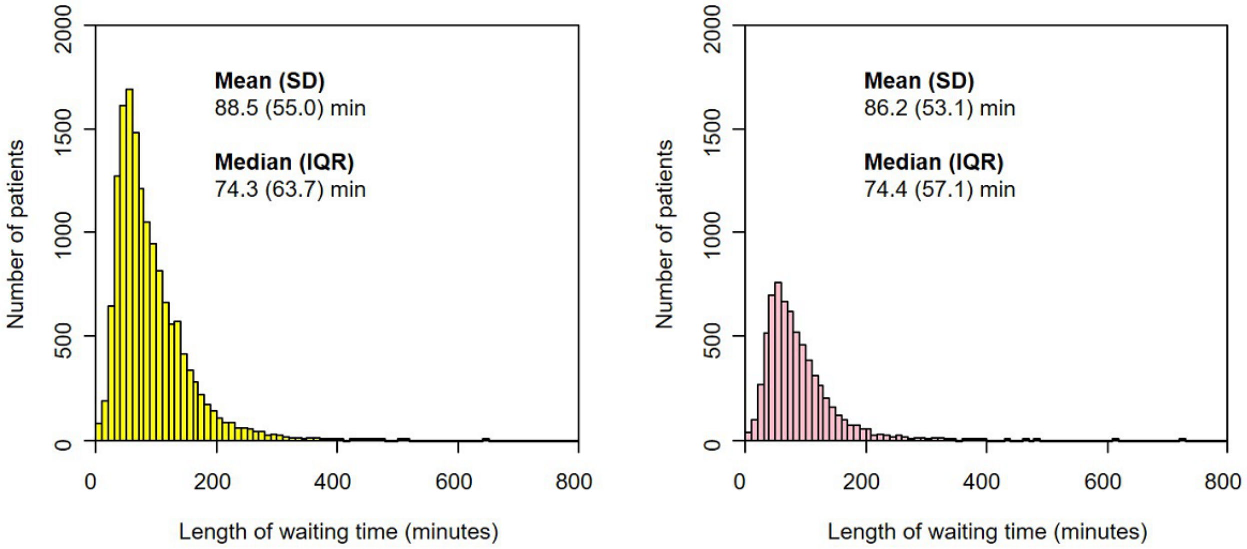
Population and Primary Outcome

From 21,723 eligible patient visits, we excluded 108 (0.5%) because the physicians' recorded data entry time was earlier than the patient's arrival time (this occurred for patients who did not follow the usual reception process, such as patients who were hospital staff or patients who visited the general internal medicine outpatient department after other departments on the same day). Hence, we included data from 21,615 patients in the study—15,000 patient visits before and 6615 patient visits after the implementation of AI Monshin. Patients who visited preimplementation were significantly older than those who visited postimplementation (age: mean 58.7 versus 56.8 years; *P*<.001). The proportions of men and women and the distribution of arrival times were not significantly different between the two groups (Table 1). Figure 1 shows the distribution of waiting time in the pre (left) and postimplementation (right) groups. Both groups showed the same distribution pattern with an extremely positive skew. The median waiting time was not significantly different between the groups (74.4 minutes versus 74.3 minutes, *P*=.12).

Table 1. Characteristics before and after AI Monshin implementation.

Characteristic	Preimplementation (n=15,000)	Postimplementation (n=6615)	*P* value
Age (years), mean (SD)	58.7 (19.6)	56.8 (20.2)	<.001
Gender, n (%)			.15
Men	6801 (45.3)	2930 (44.3)	
Women	8199 (54.7)	3685 (55.7)	
Arrival time, n (%)			.84
8 AM-9 AM	4369 (29.1)	1891 (28.6)	
9 AM-10 AM	4317 (28.8)	1906 (28.8)	
10 AM-11 AM	3489 (23.3)	1566 (23.7)	
11 AM-noon	2825 (18.8)	1252 (18.9)	

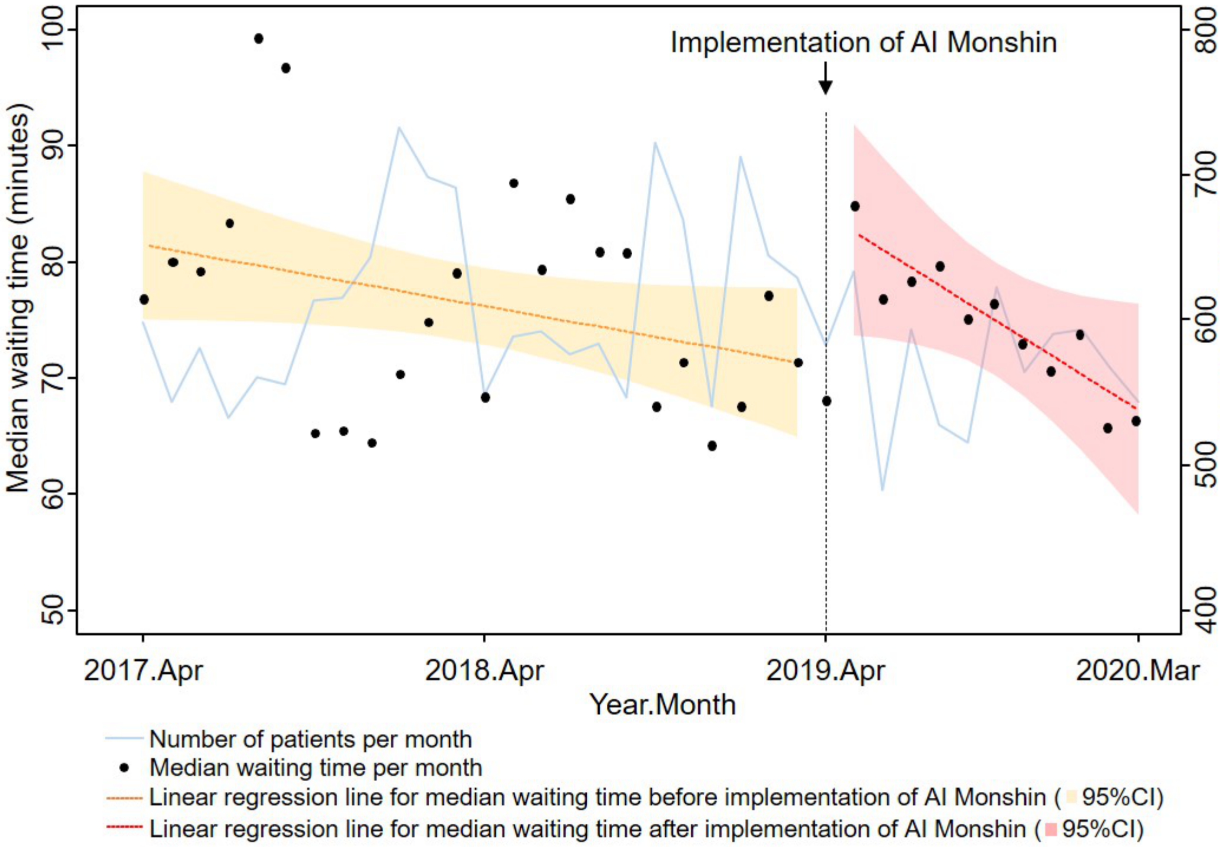
Figure 1. Distribution of waiting time before (left) and after (right) AI Monshin implementation.



Interrupted Time-Series Analysis

Figure 2 shows the trends in the number of patients and median waiting time by month from April 2017 to April 2020. The drops in waiting time in February 2020 and March 2020 (the last two dots in Figure 2) could have been partially influenced by the efforts to mitigate the risk of the spread of coronavirus disease 2019 in the waiting room. In the interrupted time-series analysis, the underlying linear time trend was –0.4 minutes per month (*P*=.06, 95% CI –0.9 to 0.02), the level change at April 2019 was 40.6 minutes (*P*=.09, 95% CI –5.8 to 87.0), and the slope change starting in April 2019 was –1.1 minutes per month (*P*=.16, 95% CI –2.7 to 0.4).

Figure 2. The trend in median waiting time and number of patients per month from April 2017 to April 2020.



Supplemental Analysis

We added supplemental analyses for data from 9054 of 21,615 patient visits (41.9%) for whom the time of entry into the examination room was recorded in addition to the doctor's first recorded data entry. We calculated the assumed examination time as the time between patient entry into the examination room and the first recorded time of doctor’s data entry in the patient record, in 2491 of 6615 (37.7%) and 6563 of 15,000 (43.8%) patient visits before and after AI Monshin implementation (*P*<.001), respectively. The median assumed examination time after AI Monshin implementation (6.0 minutes, IQR 5.2) was significantly longer compared to the median assumed examination time before AI Monshin implementation (5.7 minutes, IQR 5.0; *P*=.003).

Discussion

Principal Results

To the best of our knowledge, this study is the first to evaluate the associated change from an AI-based automated medical history–taking system with patient waiting times at a general internal medicine outpatient department using an extensive data set of approximately 21,500 visits. Our results showed that the median waiting times before and after AI Monshin implementation were not significantly different from one another (*P*=.12). Moreover, the interrupted time-series analysis also showed no significant change in median waiting time (level change: 40.6 minutes, *P*=.09, 95% CI –5.8 to 87.0; slope change: –1.1 minutes per month, *P*=.16, 95% CI –2.7 to 0.4). In addition, we observed a slight increase in the examination time (including writing the patient record), with statistical significance (*P*=.003), after implementing AI Monshin.

Limitations

This study had several limitations. First, there was the possibility of several confounding factors (such as staff skills, demographic changes, and case complexity) and other unmeasured confounding factors affecting the results. Therefore, we conducted time-series analysis in order to better interpret the results. Second, not all patients had data for examination start time. Thus, the waiting time in this study did not represent the actual waiting time in the waiting room. Moreover, the waiting time in this study may depend on when each doctor began entering data into the patient record; some doctors may prefer to enter data during patient examination, while others may prefer to enter data after the examination.

Interpretation and Comparison With Prior Works

In our study, the use of AI Monshin did not reduce waiting time, contrary to our hypothesis for the usefulness of implementation of automated medical history taking. This negative result appears to be due to the amount of time automated medical history taking required and the characteristics of patients visiting general internal medicine outpatient departments without an appointment. As previously mentioned, automated medical history taking saves up to 15 minutes of overall patient time in the clinic when used at home in advance to the visit [21]. This may be because doctors were able to spare enough time to grasp the complete information taken by automated medical history taking and prepare for the examination; however, in this study, automated medical history taking was used in the waiting room right before examination. In this situation, the doctors may not have been able to make use of the large amount of data taken by automated medical history taking in just a few minutes. In addition, the completeness of automated medical history taking could be paradoxically associated to more examination time in specific situations. Consequently, the implementation of automated medical history taking actually led to longer examination times. According to a previous study [18], physicians estimated that the use of an automated medical history–taking device has the potential to become time consuming in low-complexity cases, in which the medical history is easily taken. In the setting of the small- to medium-sized hospitals in Japan, case complexity is usually low for patients visiting general internal medicine outpatient departments without an appointment [26,27]. We conducted this study in a single center (small- to medium-sized hospital) in Japan. Therefore, there may have been a selection bias since most of cases were assumed to be low-complexity cases, though no stratification of data into the degree of complexity was performed. Hence, the increase of examination time after AI Monshin implementation in this study is consistent with the assumption. This could explain why AI Monshin implementation failed to reduce patient waiting time in our study.

Although the waiting time was not reduced in this study, AI Monshin implementation may have optimized the quality of care. Previous reports [21] revealed that while some physicians used the same amount of time before and after the implementation of automated medical history taking, they could perform a more complete evaluation of the patient with automated medical history taking. We could not judge whether these quality changes occurred in this study because we did not survey changes such as the quantity and quality of patient-physician communication, patient satisfaction, or the accuracy of diagnosis. However, we can hypothesize that the implementation of automated medical history taking has the potential to optimize staff assignment. Indeed, after AI Monshin implementation, one out of the three nurses was replaced with a medical clerk, and thus an additional nurse was available to attend to patients. This shift in resources could have enhanced the quality of care. Moreover, because approximately half of first-visit patients revisit the outpatient department [27], the comprehensive patient history taken by AI Monshin may enhance the quality of care for subsequent visits. Moreover, using an AI-based automated medical history–taking system may improve the quality and quantity of data records, which otherwise vary among physicians [19], ultimately resulting in enhancement in the quality of medical care.

Conclusions

The implementation of an AI-based automated medical history–taking system did not reduce the waiting time for patients visiting the general internal medicine outpatient department without an appointment. In addition, we noticed a slight increase in examination time after implementation. However, the implementation may have enhanced the quality of care by supporting the optimization of staff assignments. There may have been associations between case complexity and waiting time, examination time, and description time of patients. Therefore, we envision conducting further quantitative studies that take into account case complexity and that involve medical facilities of various sizes. Testing the effectiveness of automated medical history taking in reducing consultation time and explanation time between first versus second or subsequent visits is also a target issue for future study.

Acknowledgments

There was no funding to disclose for this study. The authors thank the project team at the Nagano Chuo Hospital for their efforts in their daily work.

Authors' Contributions

YH and TS designed the study. YH collected and analyzed the data. YH wrote the manuscript, and TS revised it. Both authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Bleustein C, Rothschild DB, Valen A, Valatis E, Schweitzer L, Jones R. Wait times, patient satisfaction scores, and the perception of care. Am J Manag Care 2014 May;20(5):393-400. [Medline: 25181568]
2. Thi Thao Nguyen S, Yamamoto E, Thi Ngoc Nguyen M, Bao Le H, Kariya T, Saw YM, et al. Waiting time in the outpatient clinic at a national hospital in Vietnam. Nagoya J Med Sci 2018 May;80(2):227-239 [FREE Full text] [doi: 10.18999/nagjms.80.2.227] [Medline: 29915440]
3. Swancutt D, Joel-Edgar S, Allen M, Thomas D, Brant H, Benger J, et al. Not all waits are equal: an exploratory investigation of emergency care patient pathways. BMC Health Serv Res 2017 Jun 24;17(1):436 [FREE Full text] [doi: 10.1186/s12913-017-2349-2] [Medline: 28646876]
4. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, et al. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. Ann Intern Med 2016 Sep 6. [doi: 10.7326/M16-0961] [Medline: 27595430]

5. Scott D, Hallett C, Fettiplace R. Data-to-text summarisation of patient records: using computer-generated summaries to access patient histories. Patient Educ Couns 2013 Aug;92(2):153-159. [doi: 10.1016/j.pec.2013.04.019] [Medline: 23746770]
6. Bachman J. Improving care with an automated patient history. Fam Pract Manag 2007;14(7):39-43 [FREE Full text] [Medline: 17696057]
7. Slack WV, Hicks GP, Reed CE, Van Cura LJ. A computer-based medical-history system. N Engl J Med 1966 Jan 27;274(4):194-198. [doi: 10.1056/NEJM196601272740406] [Medline: 5902618]
8. Zakim D. Development and significance of automated history-taking software for clinical medicine, clinical research and basic medical science. J Intern Med 2016 Sep;280(3):287-299 [FREE Full text] [doi: 10.1111/joim.12509] [Medline: 27071980]
9. Rockart JF, McLean ER, Hershberg PI, Bell GO. An automated medical history system. Experience of the Lahey Clinic Foundation with computer-processed medical histories. Arch Intern Med 1973 Sep;132(3):348-358. [doi: 10.1001/archinte.132.3.348] [Medline: 4593190]
10. Quaak MJ, Van der Voort PJ, Westerman RF, Hasman A, van Bemmel JH. Automation of the patient history--evaluation of ergonomic aspects. Int J Biomed Comput 1987 Nov;21(3-4):287-298. [doi: 10.1016/0020-7101(87)90094-8] [Medline: 3679586]
11. Herrick DB, Nakhasi A, Nelson B, Rice S, Abbott PA, Saber Tehrani AS, et al. Usability characteristics of self-administered computer-assisted interviewing in the emergency department: factors affecting ease of use, efficiency, and entry error. Appl Clin Inform 2013;4(2):276-292 [FREE Full text] [doi: 10.4338/ACI-2012-09-RA-0034] [Medline: 23874364]
12. Brahmandam S, Holland WC, Mangipudi SA, Braz VA, Medlin RP, Hunold KM, et al. Willingness and Ability of Older Adults in the Emergency Department to Provide Clinical Information Using a Tablet Computer. J Am Geriatr Soc 2016 Nov;64(11):2362-2367 [FREE Full text] [doi: 10.1111/jgs.14366] [Medline: 27804126]
13. Adang RP, Vismans FJ, Ambergen AW, Talmon JL, Hasman A, Flendrig JA. Evaluation of computerised questionnaires designed for patients referred for gastrointestinal endoscopy. Int J Biomed Comput 1991 Oct;29(1):31-44. [doi: 10.1016/0020-7101(91)90011-3] [Medline: 1959980]
14. Benaroia M, Elinson R, Zarnke K. Patient-directed intelligent and interactive computer medical history-gathering systems: a utility and feasibility study in the emergency department. Int J Med Inform 2007 Apr;76(4):283-288. [doi: 10.1016/j.ijmedinf.2006.01.006] [Medline: 16473548]
15. Smith PD, Grasmick M. Computer interviewing in a primary care office: the patients are ready. Stud Health Technol Inform 2004;107(Pt 2):1162-1165. [Medline: 15360995]
16. Quaak MJ, Westerman RF, Schouten JA, Hasman A, van Bemmel JH. Appraisal of computerized medical histories: comparisons between computerized and conventional records. Comput Biomed Res 1986 Dec;19(6):551-564. [doi: 10.1016/0010-4809(86)90029-7] [Medline: 3539503]
17. Quaak MJ, Westerman RF, van Bemmel JH. Comparisons between written and computerised patient histories. Br Med J (Clin Res Ed) 1987 Jul 18;295(6591):184-190 [FREE Full text] [doi: 10.1136/bmj.295.6591.184] [Medline: 3115371]
18. Schwitzguebel AJ, Jeckelmann C, Gavinio R, Levallois C, Benaïm C, Spechbach H. Differential Diagnosis Assessment in Ambulatory Care With an Automated Medical History-Taking Device: Pilot Randomized Controlled Trial. JMIR Med Inform 2019 Nov 04;7(4):e14044 [FREE Full text] [doi: 10.2196/14044] [Medline: 31682590]
19. Almario CV, Chey W, Kaung A, Whitman C, Fuller G, Reid M, et al. Computer-generated vs. physician-documented history of present illness (HPI): results of a blinded comparison. Am J Gastroenterol 2015 Jan;110(1):170-179 [FREE Full text] [doi: 10.1038/ajg.2014.356] [Medline: 25461620]
20. Arora S, Goldberg AD, Menchine M. Patient impression and satisfaction of a self-administered, automated medical history-taking device in the Emergency Department. West J Emerg Med 2014 Feb;15(1):35-40 [FREE Full text] [doi: 10.5811/westjem.2013.2.11498] [Medline: 24695871]
21. Rockart JF, McLean ER, Hershberg PI, Bell GO. An automated medical history system. Experience of the Lahey Clinic Foundation with computer-processed medical histories. Arch Intern Med 1973 Sep;132(3):348-358. [doi: 10.1001/archinte.132.3.348] [Medline: 4593190]
22. Highlighting Japan. 2019. URL: https://dwl.gov-online.go.jp/video/cao/dl/public_html/gov/book/hlj/20191201/book.pdf
23. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ 2015 Jun 09;350:h2750 [FREE Full text] [doi: 10.1136/bmj.h2750] [Medline: 26058820]
24. Lopez BJ, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol 2016 Jun 09 [FREE Full text] [doi: 10.1093/ije/dyw098] [Medline: 27283160]
25. Hoffman SJ, Poirier MJP, Rogers Van Katwyk S, Baral P, Sritharan L. Impact of the WHO Framework Convention on Tobacco Control on global cigarette consumption: quasi-experimental evaluations using interrupted time series analysis and in-sample forecast event modelling. BMJ 2019 Jun 19;365:l2287 [FREE Full text] [doi: 10.1136/bmj.l2287] [Medline: 31217191]
26. Takeshima T, Kumada M, Mise J, Ishikawa Y, Yoshizawa H, Nakamura T, et al. Reasons for encounter and diagnoses of new outpatients at a small community hospital in Japan: an observational study. Int J Gen Med 2014;7:259-269 [FREE Full text] [doi: 10.2147/IJGM.S62384] [Medline: 24940078]

27. Kajiwara N, Hayashi K, Misago M, Murakami S, Ueoka T. First-visit patients without a referral to the Department of Internal Medicine at a medium-sized acute care hospital in Japan: an observational study. Int J Gen Med 2017;10:335-345 [FREE Full text] [doi: 10.2147/IJGM.S146830] [Medline: 29042808]

Abbreviations

AI: artificial intelligence

Edited by C Lovis; submitted 04.06.20; peer-reviewed by F Palmieri, I Gabashvili, J Li; comments to author 07.07.20; revised version received 28.07.20; accepted 03.08.20; published 31.08.20.

Please cite as:
Harada Y, Shimizu T
Impact of a Commercial Artificial Intelligence–Driven Patient Self-Assessment Solution on Waiting Times at General Internal Medicine Outpatient Departments: Retrospective Study
JMIR Med Inform 2020;8(8):e21056
URL: http://medinform.jmir.org/2020/8/e21056/
doi:10.2196/21056
PMID:32865504

Original Paper

Improving Diagnostic Classification of Stillbirths and Neonatal Deaths Using ICD-PM (International Classification of Diseases for Perinatal Mortality) Codes: Validation Study

Hiu Mei Luk[1], MD, MMed; Emma Allanson[2], PhD, FRANZCOG; Wai-Kit Ming[3], MPH, MD, PhD; Wing Cheong Leung[1], MD, FRCOG

[1]Department of Obstetrics and Gynaecology, Kwong Wah Hospital, Hong Kong SAR, China (Hong Kong)
[2]Institute of Health Research, University of Notre Dame, Fremantle, Western Australia, Australia
[3]Department of Public Health and Preventive Medicine, School of Medicine, Jinan University, Guangzhou, China

Corresponding Author:
Wing Cheong Leung, MD, FRCOG
Department of Obstetrics and Gynaecology
Kwong Wah Hospital
Hong Kong SAR
China (Hong Kong)
Phone: 852 35175091
Email: leungwc@ha.org.hk

Abstract

Background: Stillbirths and neonatal deaths have long been imperfectly classified and recorded worldwide. In Hong Kong, the current code system is deficient (>90% cases with unknown causes) in providing the diagnoses of perinatal mortality cases.

Objective: The objective of this study was to apply the International Classification of Diseases for Perinatal Mortality (ICD-PM) system to existing perinatal death data. Further, the aim was to assess whether there was any change in the classifications of perinatal deaths compared with the existing classification system and identify any areas in which future interventions can be made.

Methods: We applied the ICD-PM (with International Statistical Classification of Diseases and Related Health Problems, 10th Revision) code system to existing perinatal death data in Kwong Wah Hospital, Hong Kong, to improve diagnostic classification. The study included stillbirths (after 24 weeks gestation) and neonatal deaths (from birth to 28 days). The retrospective data (5 years) from May 1, 2012, to April 30, 2017, were recoded by the principal investigator (HML) applying the ICD-PM, then validated by an overseas expert (EA) after she reviewed the detailed case summaries. The prospective application of ICD-PM from May 1, 2017, to April 30, 2019, was performed during the monthly multidisciplinary perinatal meetings and then also validated by EA for agreement.

Results: We analyzed the data of 34,920 deliveries, and 119 cases were included for analysis (92 stillbirths and 27 neonatal deaths). The overall agreement with EA of our codes using the ICD-PM was 93.2% (111/119); 92% (78/85) for the 5 years of retrospective codes and 97% (33/34) for the 2 years of prospective codes (*P*=.44). After the application of the ICD-PM, the overall proportion of unknown causes of perinatal mortality dropped from 34.5% (41/119) to 10.1% (12/119) of cases (*P*<.001).

Conclusions: Using the ICD-PM would lead to a better classification of perinatal deaths, reduce the proportion of unknown diagnoses, and clearly link the maternal conditions with these perinatal deaths.

(JMIR Med Inform 2020;8(8):e20071) doi:10.2196/20071

KEYWORDS

stillbirths; perinatal deaths; neonatal deaths; ICD-PM; ICD-10

Introduction

Background

More than 5 million perinatal deaths occur globally each year, and this largely silent epidemic significantly impacts and burdens families [1,2]. Despite this, perinatal deaths have long been frequently invisible and poorly recorded worldwide.

A large proportion of stillbirth and neonatal death cases take place in less developed countries [3]. Classification methods used in these countries can be obscure. The numbers of skilled medical personnel are often inadequate, which contributes to less than comprehensive recording of clinical data at the time of stillbirth and neonatal deaths [4,5]. In contrast, stillbirth affects proportionally fewer births in high-income countries, and data related to these deaths tend to be comprehensively recorded [6]. Many high-income countries have developed their own classification systems for perinatal deaths; for example, the United Kingdom uses the Codac system [7], Sweden has the Stockholm system [8], Australia and New Zealand use the Perinatal Society of Australia and New Zealand perinatal death classification system [9], and the Netherlands uses the Tulip classification system [10]. In the United States, the Stillbirth Collaborative Research Network developed the initial causes of fetal death system [11]. While we see an annual reduction in the rate in stillbirth globally [12], this trend differs widely among high- and low- income countries. An internationally recognized classification system of perinatal mortality would be invaluable so that the data could easily be compared between different countries to facilitate classification and research in driving programs to reduce the overall perinatal mortality.

Based on the *International Statistical Classification of Diseases and Related Health Problems, 10th Revision* (ICD-10), the International Classification of Diseases for Perinatal Mortality (ICD-PM) was released by the World Health Organization in August 2016 [13]. It is the first perinatal death classification system developed to be used worldwide. This system links the cause of perinatal death, using ICD-10 codes [14], separated by timing of death, with the maternal conditions that contribute to perinatal death.

Currently, in Hong Kong under the Hospital Authority, the coding system (for stillbirths only) used [15] is a direct one (ie, if the cause of the stillbirth could be identified, it would be stated directly such as congenital abnormalities, pregnancy-induced hypertension, cord accident, antepartum hemorrhage, maternal disease). The remaining cases would be placed in the categories unclassified, unexplained, and miscellaneous/uninvestigated, which essentially refer to unknown causes. The problem is that up to 92.8% (800/862) of stillbirths from 2012 to 2018 were classified under these 3 categories [15]. This phenomenon has significantly impaired the potential of using this perinatal database to further study and design programs to reduce perinatal mortality.

We hypothesized that by using a globally applicable classification system such as ICD-PM that recognizes stillbirths and neonatal deaths together with the contributing maternal conditions, coding of stillbirth and neonatal death could be significantly improved. Therefore, a validation study was performed to apply the ICD-PM coding system to our stillbirths and neonatal death cases.

Outcome

The primary outcome was the reduction in unknown causes for stillbirth and neonatal death after applying the ICD-PM coding system. The secondary outcome was the percentage agreement with expert (EA) in applying ICD-PM codes.

Methods

Recruitment

The ICD-PM system, the World Health Organization application of the ICD-10 to deaths during the perinatal period, was applied to existing perinatal death data from May 1, 2012, to April 30, 2019, in Kwong Wah Hospital, a regional public hospital with 34,920 deliveries during the 7-year study period.

Selection Criteria

Inclusion criteria:

- Stillbirth cases diagnosed after 24 completed weeks gestation
- Neonatal death cases within 28 days of birth

Exclusion criteria:

- Miscarriage at less than 24 completed weeks of gestation
- Termination of pregnancy due to fetal anomalies or maternal abnormal medical conditions

Ethics

Ethics approval for the study was granted by the Kowloon Central/Kowloon East Research Ethics Committee (KC/KE-19-0193/ER-1). As this clinical study did not involve active patient participation, no patient consent was required.

Diagnostic Classification

In our department, every case of stillbirth and neonatal death is discussed in our monthly perinatal meetings attended by consultants, specialists, trainees, and senior labor ward midwives and senior nurses from the departments of obstetrics and gynecology and pediatrics. A detailed case summary is presented by the resident trainee directly involved in the clinical management of that particular case. Diagnosis of the cause of stillbirth and neonatal death is based on the clinical findings and investigation results.

We performed routine investigations for stillbirths, including:

- Maternal: complete blood counts; liver and renal function tests; urate; clotting profile; thyroid function test; lupus anticoagulant; anticardiolipin antibodies; antinuclear antibodies +/– anti-ds DNA; rheumatoid factor; Kleihauer test; toxoplasmosis, rubella, cytomegalovirus, and herpes simplex virus tests; oral glucose tolerance test; hemoglobin A1c; high vaginal swab; midstream urine for culture; and others.
- Placenta: swabs for culture, histopathology, karyotyping (if the stillbirth showed dysmorphic features)

- Stillbirth: body surface swabs for culture, autopsy (with consent from parents)

Data Validation

Our validation study started in 2017 and was divided into two parts.

Retrospective Validation Study (Five Years)

The recoding of the retrospective stillbirth and neonatal death data from May 1, 2012, to April 30, 2017, was performed by the principal investigator (HML) using the ICD-10 [14] to get a specific ICD-10 code that was then converted to the ICD-PM categories [13].

Our ICD-10 and ICD-PM codes together with the detailed case summaries for each stillbirth and neonatal death case were then forwarded to an overseas expert (EA; via emails without patient identity) for validation of our codes based on the detailed case summaries, further discussion, and verification. EA has extensive experience using the ICD-PM [1-4]. The proportion of ICD-PM codes EA disagreed with was noted.

Prospective Validation Study (Two Years)

The prospective application of ICD-10 and then ICD-PM was performed from May 1, 2017, to April 30, 2019. The coding for each stillbirth and neonatal death case was validated during our monthly perinatal meetings.

Our ICD-PM codes together with the detailed case summaries in this prospective case series were also forwarded to the overseas expert (EA) to be checked for agreement.

Statistical Analysis

All frequency data were analyzed by summary statistics. SPSS Statistics version 25.0 (IBM Corporation) was used for the analysis. The Pearson chi-square test was used where appropriate, such as to determine whether there was a significant difference between the frequency of unknown causes of death classified in our original Hospital Authority coding system and the newly applied ICD-PM coding system. *P*<.05 was considered statistically significant.

Results

We analyzed data for 34,920 deliveries, and 119 cases were included for analysis (92 stillbirths and 27 neonatal deaths):

- Retrospective ICD-PM codes (5 years; May 1, 2012, to April 30, 2017)—total 85 cases
- Prospective ICD-PM codes (2 years; May 1, 2017, to April 30, 2019)—total 34 cases

All stillbirth cases had the full set of routine investigations described under Methods. However, only 25.2% (30/119) of cases had a postmortem examination.

EA verified every single one of the 119 stillbirth and neonatal death cases during the study period. The overall agreement rate of our codes using ICD-10 and then ICD-PM was 93.2% (111/119 cases) with EA: 92% (78/85 cases) for the 5 years of retrospective cases and 97% (33/34 cases) for the 2 years of prospective cases (*P*=.44). It was interesting and educational to look at how EA disagreed with our codes (Table 1).

Table 2 illustrates the application of the ICD-PM for perinatal death and maternal condition in stillbirth and neonatal death cases, respectively. In the ICD-PM, there are 6 groups of antepartum causes for stillbirth (A1 to A6), 7 groups of intrapartum causes for stillbirth (I1 to I7), 11 groups of causes for neonatal death (N1 to N11), and 5 groups of maternal conditions (M1 to M5) to be associated with stillbirth or neonatal death.

The most common causes for antepartum stillbirths were A3 (antepartum hypoxia, 24/91, 26%), followed by A5 (disorders related to fetal growth, 17/91, 19%), and A1 (congenital malformations, deformations, and chromosomal abnormalities, 10/91, 11%). In this case series, there was only one intrapartum stillbirth (I1). The most common corresponding maternal conditions were M1 (complications of placenta, cord, and membranes, 39/91, 43%), followed by M4 (maternal medical and surgical conditions, 21/91, 23%), and M2 (maternal complications of pregnancy, 12/91, 13%). The most common associations were A3;M1 (20/91, 22%), A6;M5 (12/91, 13%), and A6;M4 (9/91, 10%).

On the other hand, the most common causes for neonatal deaths were N9 (low birth weight and prematurity, 9/27, 33%), followed by N8 (neonatal conditions, 5/27, 19%), N6 (infection, 4/27, 15%), and N1 (congenital malformations, deformations and chromosomal abnormalities, also 4/27, 15%). The most common corresponding maternal conditions were M1 (complications of placenta, cord, and membranes, 10/27, 37%), followed by M2 (maternal complications of pregnancy, 6/27, 22%) and M4 (maternal medical and surgical conditions, 5/27, 19%). The most common associations were N9;M1 (3/27, 11%), N9;M4 (3/27, 11%), and N6; M1 (3/27, 11%).

Table 1. The 8 cases in which the overseas expert (EA) disagreed with our codes.

Case number	Our original codes	Our ICD-PM[a] codes	EA's codes	Comment
1	Unknown	A6; M1	A6; M5	EA: wonder if the placental pathology showing chorioamnionitis (M1) is related to the stillbirth in the absence of other evidence
9	Preeclampsia with placenta abruptio	A3; M2	A3; M4	EA: would classify this as M4 (PET[b]) as the fetus died before the mother, likely as a result of the PET, rather than the maternal death (M2) being the cause of the fetal death
11	Twin-twin transfusion syndrome (TTTS[c]), post-mortem–hypoplastic adrenals	A1, M1	A1, M5	EA: Do you think the hypoplastic adrenals (A1) was the cause of death? Or hypoxia as a result of the TTTS (M1) with the hypoplastic adrenals being a secondary issue?
19	Unknown	A6; M1	A6; M5	EA: As long as you are certain of the chorioamnionitis (M1) again, or is this a postmortem change (M5) between fetal death and delivery?
23	Unknown	A5; M4	A6; M4	EA: The birthweight is surely to be expected with the delay between fetal death and delivery, and there is no evidence of IUGR[d] (A5); keep M4 (GDM[e]).
31	Unknown	A6; M1	A6; M5	EA: Are you confident of the chorioamnionitis (M1) as the cause of death?
71	Fetal syndromal abnormality	A1; M3	A1; M5	EA: The cesarean delivery was done after the fetal death. Cesarean delivery as a cause is more when there are complications (M3) from the cesarean delivery that lead to the fetal death
89	Cord accident, drug addict	A3; M4	A3; M1	EA: Cord accident (A3; M1) as the cause of fetal death rather than mother is drug addict (M4)

[a] ICD-PM: International Classification of Disease for Perinatal Mortality.

[b] PET: pre-eclampsia.

[c] TTTS: twin-twin transfusion syndrome.

[d] IUGR: intrauterine growth restriction.

[e] GDM: gestational diabetes mellitus.

Table 2. ICD-PM codes for stillbirths (antepartum [A] and intrapartum [I]) and neonatal [N] deaths with maternal conditions (n=119).

Perinatal cause of death	Maternal condition					
	M1: complications of placenta, cord, and membrane	M2: maternal complications of pregnancy	M3: other complications of labor and delivery	M4: maternal medical and surgical conditions	M5: no maternal conditions	Total (%)
Antenatal death (A)						
A1: congenital malformations, deformations, and chromosomal abnormalities	2	2	0	2	4	10 (11.0)
A2: infection	6	0	0	0	1	7 (7.7)
A3: antepartum hypoxia	20	1	0	3	0	24 (26.4)
A4: other specified antepartum disorder	1	0	0	1	0	2 (2.2)
A5: disorders related to fetal growth	6	5	0	6	0	17 (18.7)
A6: fetal death of unspecified cause	4	4	2	9	12	31 (34.0)
Total (%)	39 (42.9)	12 (13.2)	2 (2.2)	21 (23.0)	17 (18.7)	91 (100.0)
Intrapartum death (I)						
I1: congenital malformations, deformations, and chromosomal abnormalities	0	0	1	0	0	1 (100.0)
I2: birth trauma	0	0	0	0	0	0 (0.0)
I3: acute intrapartum event	0	0	0	0	0	0 (0.0)
I4: infection	0	0	0	0	0	0 (0.0)
I5: other specified intrapartum disorder	0	0	0	0	0	0 (0.0)
I6: disorders related to fetal growth	0	0	0	0	0	0 (0.0)
I7: intrapartum death of unspecified cause	0	0	0	0	0	0 (0.0)
Total (%)	0	0	1 (100.0)	0	0	1 (100.0)
Neonatal death (N)						
N1: congenital malformations, deformations, and chromosomal abnormalities	1	1	2	0	0	4 (14.8)
N2: disorders related to fetal growth	0	0	0	1	0	1 (3.7)
N3: birth trauma	0	0	0	0	0	0 (0.0)
N4: complications of intrapartum events	1	0	0	0	0	1 (3.7)
N5: convulsions and disorders of cerebral status	0	0	0	0	0	0 (0.0)
N6: infection	3	1	0	0	0	4 (14.8)
N7: respiratory and cardiovascular disorders	0	2	1	0	0	3 (11.1)
N8: neonatal conditions	2	1	0	1	1	5 (18.5)
N9: low birth weight and prematurity	3	1	1	3	1	9 (33.4)
N10: miscellaneous	0	0	0	0	0	0 (0.0)
N11: neonatal death of unspecified cause	0	0	0	0	0	0 (0.0)
Total (%)	10 (37.1)	6 (22.2)	4 (14.8)	5 (18.5)	2 (7.4)	27 (100.0)

Before the application of the ICD-PM coding system, 34.5% (41/119) of stillbirths and neonatal deaths were classified as having unknown causes. After the application of the ICD-PM system, the cases with unknown causes of perinatal death dropped to 10.1% (12/119) cases (*P*<.001). In our study, all neonatal deaths had specific causes identified.

After retrospectively applying ICD-10 and ICD-PM codes in the cases with unknown causes (30/85, 35%) of stillbirth from our original classification on the 85 cases from 2012 to 2017 (retrospective validation study), we can further reduce the unknown causes (ie, A6;M5) to 8% (7/85) of cases (*P*<.001). Table 3 showed how we could change the 23 unknown (30 minus 7) to known causes using ICD-PM.

Table 3. In the retrospective 5-year validation study (May 1, 2012, to April 30, 2017), 23/30 unknown causes of stillbirth can be assigned a diagnosis category after applying ICD-PM code.

Case number	Original codes	Remarks	ICD-PM[a] codes
3	Unknown	Funisitis, GDM[b]	A2; M1
8	Unknown	Maternal hypothyroidism	A6; M4
14	Unknown	History of deep vein thrombosis, placental histopathology: fetal thrombotic vasculopathy and placenta infarction	A3; M1
23	Unknown	480 g at 28+ weeks, GDM	A6; M4
28	Unknown	375 g at 24+ weeks, placenta: thrombotic vasculopathy	A5; M1
32	Unknown	Placenta: focal fetal thrombotic vasculopathy	A6; M1
39	Unknown	GDM	A6; M4
48	Unknown	GDM	A6; M4
50	Unknown	2520 g at 39 weeks; placenta: thrombotic vasculopathy	A5; M1
55	Unknown	400 g at 25 weeks; severe oligohydramnios	A5; M2
57	Unknown	Preeclampsia, DM[c] in pregnancy	A6; M4
58	Unknown	Polyhydramnios	A6; M2
59	Unknown	250 g at 25 weeks; twisted cord	A5; M1
60	Unknown	255 g at 24 weeks; GDM	A5; M4
61	Unknown	330 g at 24 weeks; GDM	A5; M4
62	Unknown	Breech presentation	A6; M3
65	Unknown	Twin pregnancy	A6; M2
66	Unknown	Twin pregnancy	A6; M2
72	Unknown	Placenta: chorioamnionitis (Escherichia coli); fetal thrombotic vasculopathy	A2; M1
73	Unknown	Placenta: fetal thrombotic vasculopathy	A6; M1
74	Unknown	Oligohydramnios	A6; M2
80	Unknown	Maternal infection with fever; placenta: focal intervillous thrombus	A2; M1
85	Unknown	Placenta: minor infarcted villi	A6; M1

[a] ICD-PM: International Classification of Diseases for Perinatal Mortality.

[b] GDM: gestational diabetes mellitus.

[c] DM: diabetes mellitus.

Discussion

Principal Findings

Perinatal deaths remain problematic worldwide. Before the advent of the ICD-PM, there were several independent systems in high-income countries and few systems in low- and middle-income countries, causing significant disparity in perinatal death data recorded among countries [16]. The classification system used in Hong Kong Hospital Authority obstetrics units is no better, with 92.8% of cases having unknown causes [15]. It seems that we were somehow reluctant to give a specific diagnosis under the original coding system unless the cause was crystal clear. Hence detailed and accurate information that can be retrieved from our original classification of perinatal mortality cases was scarce. The ICD-PM is the first system that addresses the issue internationally. ICD-10 and ICD-PM are user-friendly as shown by the high level of agreement (93.2% overall) between our codes with that by the international expert (EA). The agreement with EA of our codes using ICD-PM was even higher, reaching 97.1% for the 2 years of prospective cases compared with 91.8% for the 5 years of retrospective cases, although not statistically significant (*P*=.44). This further supports our observation that the final consensus code made during perinatal meetings with multidisciplinary discussion on the perinatal death cases gives the most accurate answer, although our principal investigator (HML) had already done a good job with the 5-year retrospective codes.

The ICD-PM is designed to be used for classifying stillbirths and neonatal deaths at three levels. First, it identifies the time of death as antepartum (before the onset of labor), intrapartum (during labor but before delivery), or neonatal (up to day 7 of postnatal life, can be extended to late neonatal deaths, so as within 28 days of life, like in our classification). Second, it is multilayered such that the depth of classification can reflect the locally available intensity of investigation. In our Hospital Authority obstetrics units, the investigations for stillbirths and neonatal deaths are quite in-depth, as described under Methods, but this was not be reflected by the original classification system. Third, the ICD-PM links the contributing maternal condition, if any, with perinatal death. This is very important because most of these contributing maternal conditions would not have shown up in our original coding system. Ultimately, the ICD-PM classification at all three levels allows easy identification of where a program intervention should be targeted in order to improve the perinatal outcomes.

After the application of the ICD-PM system, the ratio of unknown causes of perinatal mortality dropped from 34.5% to 10.1%, which was statistically significant (*P*<.001). A total of 77% of stillbirth cases initially classified as unknown were now assigned a diagnostic category after applying the ICD-PM (Table 3). In our study, all neonatal deaths had specific causes identified. In other words, the ICD-PM is more useful for coding stillbirths than neonatal deaths in our locality.

The main benefit of using ICD-10 and ICD-PM codes was to have a better understanding of the perinatal deaths in terms of the timing of death, the depth of investigations, and any contributing maternal conditions. We also consider the ICD-PM code to be user-friendly for changing an existing local perinatal death classification to one that is global and can be compared with international data. Using the ICD-PM code can lead to better classification of perinatal deaths, reduce the proportion of unknown diagnoses, and clearly link the maternal conditions with these perinatal deaths. There are some countries such as the United Kingdom [3], South Africa [3,16,17], India [18], and Colombia [19] using the ICD-PM to interpret the perinatal mortality data which showed that the ICD-PM classification was feasible and enabling the characterization of perinatal mortality. This is the first validation study demonstrating the application of the ICD-PM coding system in Hong Kong.

Limitations

However, there are limitations to our study. The study was focused on one hospital only, so the sample size was small. The improvement in coding might be related to the enthusiastic principal investigator (HML) as the coder in the retrospective validation study; the code in the prospective validation study was done during our monthly perinatal meetings with multidisciplinary input from various stakeholders, which is our usual practice before and after the study. Whether there is an overall Hawthorne effect (improving performance of coding during the study period) could be seen by continuous monitoring of the diagnostic classification of stillbirths and neonatal deaths after the ICD-PM is formally used for coding from 2020 onward.

Further study can be performed in all Hong Kong Hospital Authority obstetrics units using the ICD-10 and ICD-PM so as to draw an overall picture of perinatal mortality across the territory.

Conclusions

The ICD-PM is a user-friendly system that can enhance the understanding of data [5]. Using the ICD-PM coding system could lead to a more comprehensive classification of perinatal deaths, reduce the proportion of unknown causes as well as providing better linkage to maternal conditions in these perinatal deaths. The ICD-PM classifications are more extensive in covering diagnostic categories, with more specific details. Implementing this new coding system in Hong Kong Hospital Authority obstetrics units will be of great help in improving clinical practice and reducing perinatal mortality in the long run.

Authors' Contributions

All authors had full access to the data, contributed to the study, approved the final version for publication, and take responsibility for its accuracy and integrity. HML and WCL were responsible for concept or design of the study. HML acquired the data and drafted the article. All authors contributed to the analysis or interpretation of data. WCL, EA, and WKM performed critical revision for important intellectual content.

Conflicts of Interest

None declared.

References

1. Allanson ER, Tunçalp O, Gardosi J, Pattinson RC, Francis A, Vogel JP, et al. Optimising the International Classification of Diseases to identify the maternal condition in the case of perinatal death. BJOG 2016 Nov;123(12):2037-2046 [FREE Full text] [doi: 10.1111/1471-0528.14246] [Medline: 27527550]
2. Allanson ER, Tunçalp O, Gardosi J, Pattinson RC, Vogel JP, Erwich J, et al. Giving a voice to millions: developing the WHO application of ICD-10 to deaths during the perinatal period: ICD-PM. BJOG 2016 Nov;123(12):1896-1899 [FREE Full text] [doi: 10.1111/1471-0528.14243] [Medline: 27526957]
3. Allanson ER, Tunçalp O, Gardosi J, Pattinson RC, Francis A, Vogel JP, et al. The WHO application of ICD-10 to deaths during the perinatal period (ICD-PM): results from pilot database testing in South Africa and United Kingdom. BJOG 2016 Nov;123(12):2019-2028 [FREE Full text] [doi: 10.1111/1471-0528.14244] [Medline: 27527122]
4. Allanson ER, Vogel JP, Tunçalp O, Gardosi J, Pattinson RC, Francis A, et al. Application of ICD-PM to preterm-related neonatal deaths in South Africa and United Kingdom. BJOG 2016 Nov;123(12):2029-2036 [FREE Full text] [doi: 10.1111/1471-0528.14245] [Medline: 27527390]

5. Aminu M, van den Broek N. Stillbirth in low- and middle-income countries: addressing the 'silent epidemic'. Int Health 2019 Jul 01;11(4):237-239 [FREE Full text] [doi: 10.1093/inthealth/ihz015] [Medline: 31081893]
6. Flenady V, Wojcieszek AM, Middleton P, Ellwood D, Erwich JJ, Coory M, Lancet Ending Preventable Stillbirths study group, Lancet Stillbirths In High-Income Countries Investigator Group. Stillbirths: recall to action in high-income countries. Lancet 2016 Feb 13;387(10019):691-702. [doi: 10.1016/S0140-6736(15)01020-X] [Medline: 26794070]
7. Frøen JF, Pinar H, Flenady V, Bahrin S, Charles A, Chauke L, et al. Causes of death and associated conditions (Codac): a utilitarian approach to the classification of perinatal deaths. BMC Pregnancy Childbirth 2009 Jun 10;9:22 [FREE Full text] [doi: 10.1186/1471-2393-9-22] [Medline: 19515228]
8. Varli IH, Petersson K, Bottinga R, Bremme K, Hofsjö A, Holm M, et al. The Stockholm classification of stillbirth. Acta Obstet Gynecol Scand 2008;87(11):1202-1212. [doi: 10.1080/00016340802460271] [Medline: 18951207]
9. Lu JR, McCowan L. A comparison of the Perinatal Society of Australia and New Zealand-Perinatal Death Classification system and relevant condition at death stillbirth classification systems. Aust N Z J Obstet Gynaecol 2009 Oct;49(5):467-471. [doi: 10.1111/j.1479-828X.2009.01066.x] [Medline: 19780727]
10. Korteweg FJ, Gordijn SJ, Timmer A, Erwich JJHM, Bergman KA, Bouman K, et al. The Tulip classification of perinatal mortality: introduction and multidisciplinary inter-rater agreement. BJOG 2006 Apr;113(4):393-401 [FREE Full text] [doi: 10.1111/j.1471-0528.2006.00881.x] [Medline: 16553651]
11. Boyd TK, Wright CA, Odendaal HJ, Elliott AJ, Sens MA, Folkerth RD, PASS Network. The stillbirth classification system for the safe passage study. Pediatr Dev Pathol 2017;20(2):120-132. [doi: 10.1177/1093526616686251] [Medline: 28326963]
12. International Stillbirth Alliance Collaborative for Improving Classification of Perinatal Deaths, Flenady V, Wojcieszek AM, Ellwood D, Leisher SH, Erwich JJHM, et al. Classification of causes and associated conditions for stillbirths and neonatal deaths. Semin Fetal Neonatal Med 2017 Jun;22(3):176-185. [doi: 10.1016/j.siny.2017.02.009] [Medline: 28285990]
13. The WHO application of ICD-10 to deaths during the perinatal period: ICD-PM. Geneva: World Health Organization; 2016. URL: https://apps.who.int/iris/bitstream/handle/10665/249515/9789241549752-eng.pdf;jsessionid=A18FDCCE417B827672FB54E387A8CC63?sequence=1 [accessed 2020-07-02]
14. International Statistical Classification of Diseases and Related Health Problems, 10th Revision ICD-10 Version: 2019. URL: https://icd.who.int/browse10/2019/en [accessed 2020-07-02]
15. The Hong Kong Hospital Authority Annual Obstetric Reports 2012-2018. URL: https://www.ekg.org.hk/html/gateway/obs-gyn/aor2018.pdf [accessed 2020-07-02]
16. Lavin T, Allanson ER, Nedkoff L, Preen DB, Pattinson RC. Applying the international classification of diseases to perinatal mortality data, South Africa. Bull World Health Organ 2018 Dec 01;96(12):806-816 [FREE Full text] [doi: 10.2471/BLT.17.206631] [Medline: 30505028]
17. Aminu M, Mathai M, van den Broek N. Application of the ICD-PM classification system to stillbirth in four sub-Saharan African countries. PLoS One 2019;14(5):e0215864 [FREE Full text] [doi: 10.1371/journal.pone.0215864] [Medline: 31071111]
18. Sharma B, Siwatch S, Kakkar N, Suri V, Raina A, Aggarwal N. Classifying stillbirths in a tertiary care hospital of India: International Classification of Disease-Perinatal Mortality (ICD-PM) versus cause of death-associated condition (CODAC) system. J Obstet Gynaecol 2020 Apr 29:1-5. [doi: 10.1080/01443615.2020.1736016] [Medline: 32347769]
19. Salazar-Barrientos M, Zuleta-Tobón JJ. [Application of the International Classification of Diseases for Perinatal Mortality (ICD-PM) to vital statistics records for the purpose of classifying perinatal deaths in Antioquia, Colombia]. Rev Colomb Obstet Ginecol 2019 Dec;70(4):228-242. [doi: 10.18597/rcog.3406] [Medline: 32142238]

Abbreviations

ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision
ICD-PM: International Classification of Diseases for Perinatal Mortality

Edited by G Eysenbach; submitted 11.05.20; peer-reviewed by V Della Mea; comments to author 04.06.20; revised version received 08.06.20; accepted 25.06.20; published 03.08.20.

Please cite as:
Luk HM, Allanson E, Ming WK, Leung WC
Improving Diagnostic Classification of Stillbirths and Neonatal Deaths Using ICD-PM (International Classification of Diseases for Perinatal Mortality) Codes: Validation Study
JMIR Med Inform 2020;8(8):e20071
URL: https://medinform.jmir.org/2020/8/e20071
doi:10.2196/20071
PMID:32744510

Review

Utilization Barriers and Medical Outcomes Commensurate With the Use of Telehealth Among Older Adults: Systematic Review

Clemens Kruse[1*], MSIT, MHA, MBA, PhD; Joanna Fohn[1*], BS, MHA; Nakia Wilson[1*], BA, MHA; Evangelina Nunez Patlan[1*], BS, MHA; Stephanie Zipp[1*], BS, MHA; Michael Mileski[1*], DC, MHA, MPH, MSHEd

School of Health Administration, Texas State University, San Marcos, TX, United States

[*]all authors contributed equally

Corresponding Author:
Clemens Kruse, MSIT, MHA, MBA, PhD
School of Health Administration
Texas State University
601 University Dr
Encino Hall 250
San Marcos, TX, 78666
United States
Phone: 1 210 355 4742
Email: scottkruse@txstate.edu

Abstract

Background: Rising telehealth capabilities and improving access to older adults can aid in improving health outcomes and quality of life indicators. Telehealth is not being used ubiquitously at present.

Objective: This review aimed to identify the barriers that prevent ubiquitous use of telehealth and the ways in which telehealth improves health outcomes and quality of life indicators for older adults.

Methods: This systematic review was conducted and reported in accordance with the Kruse protocol and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Reviewers queried the following four research databases: Cumulative Index of Nursing and Allied Health Literature (CINAHL), PubMed (MEDLINE), Web of Science, and Embase (Science Direct). Reviewers analyzed 57 articles, performed a narrative analysis to identify themes, and identified barriers and reports of health outcomes and quality of life indicators found in the literature.

Results: Reviewers analyzed 57 studies across the following five interventions of telehealth: eHealth, mobile health (mHealth), telemonitoring, telecare (phone), and telehealth video calls, with a Cohen κ of 0.75. Reviewers identified 14 themes for barriers. The most common of which were technical literacy (25/144 occurrences, 17%), lack of desire (19/144 occurrences, 13%), and cost (11/144 occurrences, 8%). Reviewers identified 13 medical outcomes associated with telehealth interventions. The most common of which were decrease in psychological stress (21/118 occurrences, 18%), increase in autonomy (18/118 occurrences, 15%), and increase in cognitive ability (11/118 occurrences, 9%). Some articles did not report medical outcomes (18/57, 32%) and some did not report barriers (19/57, 33%).

Conclusions: The literature suggests that the elimination of barriers could increase the prevalence of telehealth use by older adults. By increasing use of telehealth, proximity to care is no longer an issue for access, and thereby care can reach populations with chronic conditions and mobility restrictions. Future research should be conducted on methods for personalizing telehealth in older adults before implementation.

Trial Registration: PROSPERO CRD42020182162; https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020182162.

International Registered Report Identifier (IRRID): RR2-10.2196/15490

(JMIR Med Inform 2020;8(8):e20359) doi:10.2196/20359

KEYWORDS

telehealth; telemedicine; older adults; barriers; health outcomes

Introduction

Background

A demographic shift has been evident globally since 2015. Specifically, the aging population has been growing at a rapid rate and has been predicted to reach 22% by the year 2050 [1]. In fact, the World Health Organization (WHO) estimates that during 2020, adults aged 60 years or older will outnumber children aged 5 years or younger [1]. The United States Census Bureau published a graphic on March 13, 2018, depicting the population pyramid from 1960 and comparing it with the 2060 prediction [2]. The graphic demonstrated the gradual change of the US population pyramid to a pillar shape [2]. This graphic is key to understanding the demands on the health care system in the area of geriatric, long-term, and end-of-life care, because it highlights the larger number of older adults living longer lives. By 2030, 60 million people in the "baby boomer" generation (born between 1946 and 1964) will have reached 65 years of age or older and will be eligible for age-related state entitlements in most countries [3,4]. This demographic shift is an impending issue facing health care, as geriatric, long-term, and end-of-life care will experience a surge in demand. Health care organizations and their providers must find ways to effectively allocate resources and provide the right care at the right time and at the right place [5].

Telemedicine has the potential to increase access among elderly people and relieve the stress regarding care for the unusually large number of elderly people. The WHO defines telemedicine as "healing from a distance." More specifically, it is healing through the use of information and communication technologies "to improve patient outcomes by increasing access to care and medical information" [6]. The WHO also does not differentiate between the terms telemedicine and telehealth.

There has not been much work on the use of telehealth based on age; however, we know that a technology gap or digital divide exists. It is established by tiers of race, age, and economic disparities [7]. In the United States, for instance, the elder-care entitlement Medicare imposes restrictions on the use of telehealth for the primary population [8]. The Coronavirus Aid, Relieve, and Economic (CARES) Act provides a regulatory waiver to extend reimbursements to telemedicine, but this is only a relief act and not permanent legislation [9]. Previous reviews have investigated facilitators and barriers to the adoption of telehealth, the use of eHealth and mobile health (mHealth) tools in health promotion and primary prevention among older adults, and patient satisfaction with telehealth interventions [10-12]. A narrative analysis on mHealth solutions for the aging population used a generational analysis that included culture and trust of other people and a distrust of technology [13]. This work noted an increase in the use of technology for health purposes and an increase in the use of the internet for health purposes. It also noted concerns of security and privacy and technical troubleshooting. A review from 6 years ago spanned 10 years, analyzed 14 articles, and focused on older adults over 65 years old [10]. The most recent review on a topic most like this work was published 5 years ago, spanned 10 years, analyzed 45 articles, and focused on older adults aged over 50 years [11].

With an aging population, telehealth services are becoming more common to aid in independent living and health management [14]. An example of telehealth is virtual home health care, where health care providers provide guidance in specific procedures while the patients are in the comfort of their home. Telehealth programs can improve access to health care and have a positive effect on patients' medical outcomes, especially for the treatment of chronic illnesses in vulnerable populations, such as elderly people [15]. Utilizing age-friendly technology could improve the care providers give to older adults through telehealth services and improve the usability of telehealth for older adults [16]. It is essential to first understand the barriers that affect the usability of telehealth services among older adults in order to find opportunities for improving health outcomes. Barriers to using telehealth can affect the accessibility of health services to older adults. When it comes to technology, older adults are often stereotyped as laggards in technology adoption [7]. However, owing to rising telehealth capabilities, improvement of access, especially to older adults, can aid in improving health outcomes [15]. Understanding the perspectives of older adults is important when evaluating telehealth barriers because older adults generally develop different perspectives compared with those of other age demographics [16]. Other studies on this topic have focused on conditions like depression, heart failure, and falls [17-19]. However, no review has looked at medical outcomes, including indicators of quality of life, that come as a benefit of using telehealth and the barriers that exist to the use of telehealth internationally. This review intends to examine these issues and what has changed in telehealth for older adults in the last 5 years.

Objectives

The purpose of this systematic review was to evaluate the current literature to help identify and understand health-related quality of life enhancers and general health outcomes that are commensurate with and barriers to the use of telehealth services by older adults. Health outcomes, including quality of life enhancers, provide the "so what" to the use of telehealth modalities. Recognizing barriers can help develop solutions for broadening the use of telehealth services in older adults. During the COVID-19 crisis, providers and patients alike were thrust into the world of telehealth. An overview of the benefits and barriers would be helpful to those deciding whether to continue the use of telehealth modalities.

Methods

Protocol and Registration

This review used the Kruse protocol published in 2019 and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [20,21]. The review was registered with PROSPERO on May 2, 2020 (ID: CRD42020182162). In accordance with the rules at PROSPERO, the registration was completed before analysis began.

Eligibility Criteria

Studies were eligible for this review if participants were older adults (older than 50 years), if the intervention was some form of telehealth (including mHealth, eHealth, and all forms of telehealth), if the authors reported either barriers to the use of telehealth or health outcomes, and if the article was published in a research journal in the English language in the last 5 years. Adults older than 50 years were chosen out of trial and error. When we initially wrote the methods for this study, we chose a more universal definition of older adults as those over 65 years of age. Once we started filtering articles for analysis, we noticed a large number of articles that were being eliminated, despite the high level of quality of these studies. If we had stuck with age over 65 years as our screening criteria, we would have eliminated more than half of the group of articles for analysis. As a result, we chose age over 50 years, which is supported by other reviews in this field [11]. This is a limitation we list later.

Information Sources

The following four databases were queried: Cumulative Index of Nursing and Allied Health Literature (CINAHL), PubMed (MEDLINE), Web of Science (WoS), and Embase (Science Direct). Additionally, a specific journal search was conducted in the journal of choice for publication (Journal of Medical Internet Research). Databases were filtered for the last 5 years. Database searches occurred between February 2 and 14, 2020. A period of 5 years was chosen because it has been that long since the last review was published on a similar topic. We expect to find advances in technology and advances in adoption by elderly people because younger people who use technology regularly have advanced into the observation group of over 50 years old. We hope to find fewer barriers.

Search

Reviewers carefully analyzed the MEDLINE Medical Subject Headings (MeSH) for key terms related to telehealth and elderly people. Based on the established hierarchy of indexed terms at MeSH and a series of experimental searches, the final search terms were "Telehealth AND 'older adults.'" This combination of terms yielded the maximum number of results in all four databases. Reviewers used available filters to eliminate other reviews and focus on academic or peer-reviewed journals over the last 5 years.

Study Selection

Reviewers followed the Kruse protocol, which entails a series of three consensus meetings. The results of the first consensus meeting identified the studies for analysis. After filtering the results of the four databases to meet the eligibility criteria, all reviewers screened the abstracts of the results to ensure that articles were germane to the topic, they were actually studies (not protocols), and they contained tangible results to enable analysis toward the review's objectives. The first consensus meeting discussed whether to keep articles for analysis. The reasons for rejection included opinion article (not a study), protocol (no results), concept or design paper (no results), review, no use of telehealth, and no reporting of either outcomes or barriers. A kappa statistic was calculated from the results of this meeting [20]. Before consensus meeting number two, the group leader assigned workload to ensure that each article was analyzed by at least two reviewers. Reviewers independently analyzed articles using a piloted form. Reviewers collected several standard items used for summary, such as PICOS (Participants, Intervention, Comparison [to the control group], Outcome, Study design), and analysis, such as forms of telehealth interventions, barriers to the use of telehealth by older adults, and the medical outcomes observed in older adults using telehealth solutions [20]. After making a list of observations, reviewers attempted to make sense of the observations using a narrative analysis [22].

Data Collection Process

The group leader divided analysis workload to ensure all articles were reviewed by at least two reviewers. Reviewers independently analyzed articles using a standardized Excel spreadsheet as a piloted form for data extraction.

Data Items

The piloted form collected data, including participants, intervention, study design, results compared to a control group (where applicable), medical outcomes, sample size, bias within studies, effect size, country of origin, statistics used, barriers to the use of telehealth, and quality assessment from the John Hopkins Nursing Evidence-Based Practice (JHNEBP) rating scale, as well as general observations about the article that would help in interpretation of the results [23]. These data items were independently collected and discussed in the second consensus meeting.

Risk of Bias Within and Across Studies

General observations of bias were made about each study, such as selection bias. These observations were independently collected and discussed in the second consensus meeting. The JHNEBP rating scale was used to assess the risk and quality of each study analyzed. Within the JHNEBP rating scale, level I indicates experimental studies, randomized controlled trials (RCTs), or meta analyses of RCTs; level II indicates quasiexperimental studies; level III indicates nonexperimental studies, qualitative studies, or meta-syntheses; level IV indicates opinions of nationally recognized experts based on research evidence or expert consensus panels (systematic reviews or clinical practice guidelines); and level V indicates opinions of individual experts based on nonresearch evidence. There are three levels of quality of evidence, which are listed as A (high quality), B (good quality), and C (low quality or major flaws). Each of these levels define the following four thresholds: research, summative reviews, organizational opinion, and expert opinion. For instance, in level A, studies have consistent results with sufficient sample size, adequate control, and definitive conclusions. In level C, studies have little evidence with inconsistent results and insufficient sample size, and conclusions cannot be drawn. To limit the inherent bias and limitations commensurate with low-quality studies, the ratings from the JHNEBP rating scale serve as screening criteria. Articles with evidence ratings below level IV were not accepted. Quality of evidence ratings below level B were highly suspect.

Summary Measures and Additional Analysis

The review analyzed both qualitative and quantitative methods, so the summary measures sought were not consistent. The preferred summary statistic was the risk ratio, but other summary statistics were also sufficient. The summary statistics were independently collected and discussed in the second consensus meeting.

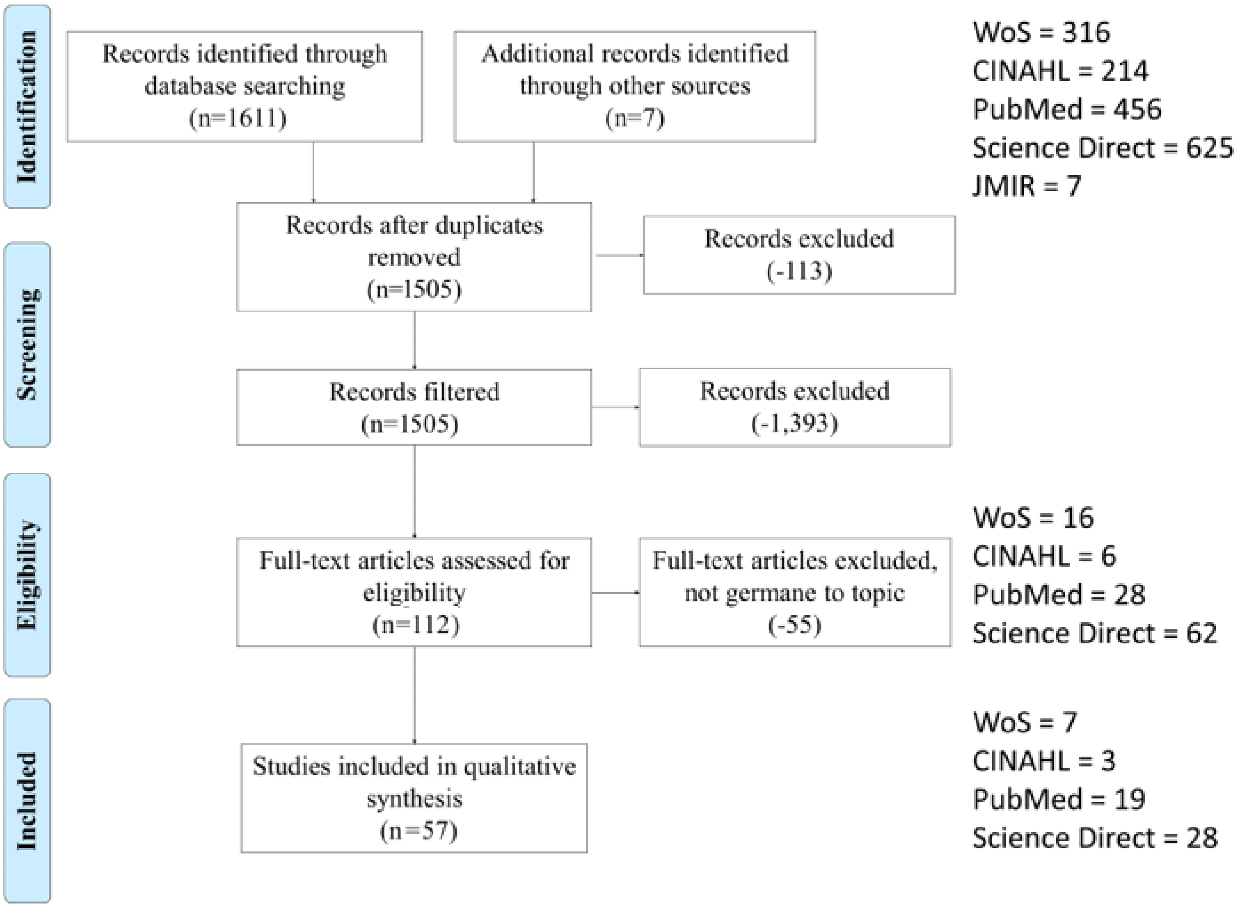
A narrative analysis summarized themes for barriers, interventions, and medical outcomes. They were reported in summary statistics in affinity matrices. These themes were independently collected and discussed in the third consensus meeting. After themes were identified, interactions between themes were observed using a spreadsheet.

Results

Study Selection

Figure 1 illustrates the study selection process. A kappa statistic was calculated to measure the reliability of article selection between reviewers. The κ value was 0.75, representing moderate agreement [24,25].

Figure 1. Study selection process.


Identification
Screening
Eligibility
Included
Records identified through database searching (n=1611)
Additional records identified through other sources (n=7)
Records after duplicates removed (n=1505)
Records excluded (-113)
Records filtered (n=1505)
Records excluded (-1,393)
Full-text articles assessed for eligibility (n=112)
Full-text articles excluded, not germane to topic (-55)
Studies included in qualitative synthesis (n=57)
WoS = 316
CINAHL = 214
PubMed = 456
Science Direct = 625
JMIR = 7
WoS = 16
CINAHL = 6
PubMed = 28
Science Direct = 62
WoS = 7
CINAHL = 3
PubMed = 19
Science Direct = 28

Study Characteristics

Table 1 lists the ancillary data extracted from the studies analyzed in reverse chronological order as follows: 2020 [26], 2019 [5,26-34], 2018 [4,15,16,35-46], 2017 [14,47-56], 2016 [19,57-63], and 2015 [64-76].

Table 1. PICOS characteristics.

Authors, year	Participants	Intervention	Comparator	Medical outcomes reported	Study design
Hamilton et al, 2020 [26]	765 older adults; ≥55 years; Medicare/Medicaid beneficiaries; English 76% (581), Spanish 20% (153), and others 4% (31); low income	Telemonitoring Remote patient monitoring (RPM): blood-pressure cuffs, pulse oximeters, and body weight scales Telehealth Intervention Programs for Seniors, RPM, extensive social wraparound services, care coordination, and intergenerational socialization aimed at improving health care options to assist low-income high health risk older adults who live in subsidized congregate housing or attend local community centers for older adults. A survey instrument was collected each week.	None	Hospital visits and readmissions	Observational study
Theis et al, 2019 [5]	551 older adults, ≥60 years, 51.3% male and 48.7% female 441 participants (80%) already retired, 109 (19.8%) still working	eHealth	None	Satisfaction: 64% (353) of older adults were satisfied with the health information they received, 34% (187) were neutral, and 2% (11) were dissatisfied	Analytical observational study
Wildenbos et al, 2019 [27]	13 older adults, ≥50 years, primarily Dutch speaking Additional inclusion criteria for App 2: heart failure (HF) patient and chronic obstructive pulmonary disease	mHealth[a] Investigated these interaction issues in two different case studies; an app for older adults facilitating their hospital appointment attendance (App 1) and a self-monitoring app for chronically ill older patients	None	Cognitive impairment was reported but not compared with a control.	Case study
Jakobsson et al, 2019 [28]	9 older adults, 65-85 years, cognitive impairment of different origin (eg, stroke, dementia, and mild cognitive impairment)	Telehealth, smartphone, computer, and landline	None	Cognitive impairment was reported but not compared with a control.	Qualitative study
Karlsen et al, 2019 [29]	18 older adults, ≥60 years, living in their own homes and having recently received telecare service (within the last 0-3 months), received home care services, Norwegian speaking, no limitations considering disease or chronic conditions	Telemonitoring Personal alarm (16), light sensors (3), stove alarm (4), GPS tracking (3), medication reminders (8), bed sensors (1), door sensor (2), video surveillance (2)	None	Safety, satisfaction, security, independence, responsibility, mindfulness of failty	Qualitative study

Authors, year	Participants	Intervention	Comparator	Medical outcomes reported	Study design
Coley et al, 2019 [30]	341 (quantitative) and 46 (qualitative) older adults; ≥65 years; Finland, France, and Netherlands; response rate 79% (Finland: 81%, France: 72%, Netherlands: 87%, *P*=.04); 48% (164) male; 51% (174) university-level education	eHealth Participants were randomized to either an interactive internet platform designed to encourage goal setting and lifestyle changes with the remote support of a lifestyle coach or a control platform with basic health information but no interactive features or coach support. Owing to the nature of the intervention, complete double blinding was not possible, but masking was attempted by informing participants that they would be randomized to one of two internet platforms (without further details on the content).	Control	Not reported	Cross-sectional mixed-methods randomized controlled trial (RCT)
Giesbrecht & Miller, 2019 [31]	18 older adults, ≥50 years, resided in the community, self-propelled using both hands at least 1 hour per day inside and outside their home, English speaking	eHealth The treatment group incorporated two in-person training sessions with a trainer and 4 weeks of monitored home training using a computer tablet (mHealth) wheelchair skills program. The control group did not receive skills training, as is typical practice with this population.	Control	Skill capacity and safety	RCT
Brodbeck et al, 2019 [32]	110 older adults, >50 years, 79% (87) female	eHealth Internet-based self-help intervention for prolonged grief symptoms after spousal bereavement or separation/divorce	Control	Grief, depression, psychological distress, embitterment, loneliness, and life satisfaction	Mixed methods, RCT
Mosley et al, 2019 [33]	112 older adults, ≥60 years, 58% (65) female, English speaking	eHealth Etymotic home hearing test compared with traditional manual audiometry	Control	Not reported	Quasiexperimental study
Jensen et al, 2019 [34]	20 older adults, hip fracture	eHealth “My Hip Fracture Journey” on iPad (provided) education through pictographs, video clips, illustrated exercises, and written information. This was used to augment home visits and subsequent interviews.	None	Autonomy and self-care	Qualitative study
Rasche et al, 2018 [15]	576 older adults, ≥60 years, 48.7% (280) female, German speaking	eHealth The national survey queried the use of health apps and their perceived usefulness.	Control	Not reported	Quasiexperimental study
Portz et al, 2018 [35]	30 older adults, ≥60 years, location at the University of Colorado Hospital and the University Hospital Cleveland Medical Center in Cleveland (Ohio), 60% (18) female, 63% (19) black people	mHealth The HF app was developed to allow patients to track their symptoms of HF. Thirty older adults completed an acceptability survey after using the mobile app. The survey used Likert items and open-ended feedback questions.	None	Awareness of the condition and self-care	Quantitative acceptability survey analysis

Authors, year	Participants	Intervention	Comparator	Medical outcomes reported	Study design
Castro et al, 2018 [36]	501 older adults, ≥65 years, Medicare population	eHealth Participants were matched into geographically based small groups with an assigned health coach, and they began the program at the same time. Group members were connected to each other through a private online social forum where they could post comments and questions, engage in health coach–moderated discussions, and provide social support to one another. Using internet-enabled devices (laptop, tablet, or smartphone), program participants were able to asynchronously complete weekly interactive curriculum lessons, reflections, and goal-setting activities in relation to the weekly topic.	Pretest	Weight: participants lost an average of 13-14 pounds (8%) HbA_{1c}: 0.14% absolute decrease at 6 months and 12 months (*P*<.001) Cholesterol: mean reduction of -12.92 mg/dL (*P*<.001).	Single-arm pretest/posttest design
Joe et al, 2018 [37]	43 older adults, 70% (30) female	eHealth A focus-group method was used to brainstorm designs for telehealth for older adults.	None	Not reported	Qualitative analysis study (focus groups)
Dham et al, 2018 [4]	134 older adults, 60% (80) female	Telemedicine Telepsychiatry assessments	None	Not reported	Mixed-methods cross-sectional cohort study with retrospective chart review and prospective feedback survey
Paige et al, 2018 [16]	384 older adults, 74.3% (285) female, 57.7% (222) Caucasian people, 42.3% (162) black people	eHealth eHealth awareness and eHealth literacy scale	None	Not reported	Qualitative measurement invariance study
Cajita et al, 2018 [38]	10 older adults, ≥65 years, history of HF, spoke English, difficulty with mobile technology	mHealth	None	Not reported	Descriptive exploratory study
Harte et al, 2018 [39]	22 older adults, >65 years, difficulty using smartphones	mHealth Training on a smartphone-based fall detection and prevention system	None	Not reported	Usability and learnability case study
Gordon & Hornbrook, 2018 [40]	2602 older adults, >65 years, 54% (1,405) female, 79% (2,056) Caucasian people	eHealth Online forms, online tracking systems, and patient portal	None	Not reported	Mixed-methods cross-sectional study
Bao et al, 2018 [41]	12 older adults, ≥65 years, 75% (8) female	eHealth Online training	Pretest	Sensory organization test, mini balance evaluation system test, five times sit to stand test, and no statistical significance in other clinical outcomes	Pretest posttest true experiment
Egede et al, 2018 [42]	241 older adults, >63 years, 98% (236) male, 60% (144) Caucasian people, veterans having major depressive disorder	Telemedicine Telepsychotherapy	Control	Baseline depression severity, generalized anxiety disorder, alcohol misuse, cannabis misuse, and cannabis dependence	RCT

Authors, year	Participants	Intervention	Comparator	Medical outcomes reported	Study design
Platts-Mills et al, 2018 [43]	75 older adults, <50 years, musculoskeletal pain	Telecare Telephone call and protocol-guided follow up	Control	Pain	Randomized controlled pilot study
Lopez-Villegas et al, 2018 [44]	50 older adults, >65 years, 48% (24) women, seen in the cardiology clinic, using a pacemaker	Telemonitoring Pacemakers	Control	EQ-5D VAS[b] (health-related quality of life)	RCT
Dugas et al, 2018 [45]	27 older adults, >60 years	mHealth DiaSocial for glucose control, exercise, nutrition, and medication adherence	None	Glucose management and HbA_{1c}	Pilot study
Nalder et al, 2018 [46]	8 older adults, >55 years, type 2 diabetes	eHealth Three internet-based platforms: 1. Chronic disease management 2. Real-world strategy training 3. Learning the ropes	None	HbA_{1c}, independence, emotional support, and motivation to self-manage	Qualitative pilot program
Buck et al, 2017 [47]	12 older adults, >60 years, 42% (5) female	eHealth PSHA, a web-based tablet-delivered intervention developed internally, which encourages the participant to record daily medication intake, weight, and time spent with a brief exercise program using an aerobic stepper. The tablet records daily information, and the participant watches a short heart health educational video.	None	Documentation for nutrition and eating and instructional video exposure	Proof-of-concept trial Qualitative semistructured interviews after the study protocol
Ware et al, 2017 [14]	15 older adults, ≥50 years, 73% (11) female	eHealth	None	Not reported	Two focus groups and pragmatic thematic analysis
Chang et al, 2017 [48]	18 older adults, >65 years, diabetes	Telehealth Diabetes management	None	Self-management and independence	Qualitative research design and 1-1 semistructured interviews
Cajita et al, 2017 [49]	129 older adults, >65 years, 73.6% (95) male, 56.6% (73) Caucasian people	mHealth Simple linear regression was used to test the relationship between the main study variables (eHealth literacy, social influence, perceived financial cost, perceived ease of use, and perceived usefulness) and intention to use mHealth.	None	Not reported	Cross-sectional correlational study
LaMonica et al, 2017 [50]	221 older adults, ≥50 years, 57.7% (128) female	eHealth Memory aids and mental acuity exercises	None	Memory	Qualitative study
Bahar-Fuchs et al, 2017 [51]	45 older adults; >65 years; mild cognitive impairment (n=9), mood-related neuropsychiatric symptoms (n=11), or both (n=25)	eHealth Tailored and adaptive computer cognitive training in older adults at risk for dementia	Control	Memory, global cognition, learning, and mood	RCT
Nahm et al, 2017 [52]	866 older adults, >50 years, bone health issues	eHealth Bone Power program	Control	Osteoporosis knowledge, self-efficacy/outcome expectations, and exercise behaviors	Two-arm RCT

Authors, year	Participants	Intervention	Comparator	Medical outcomes reported	Study design
Knaevelsrud et al, 2017 [53]	47 older adults, >50 years, 64.9% (31) female, posttraumatic stress disorder symptoms, German speaking	eHealth Internet-based therapist-guided intervention	Control	Comfort (from not being able to see the therapist), satisfaction, motivation, feeling of being understood	RCT
Reijnders et al, 2017 [54]	376 older adults, >50 years, 67.5% female	eHealth Cognitive functioning	Control	Feelings of stability, memory functioning, and locus of control	RCT
Hamblin et al, 2017 [56]	60 older adults, >85 years	Telemonitoring	None	Autonomy, awareness of danger areas like gardens or staircases, and safety	Mixed-methods study
Mageroski et al, 2016 [55]	25 older adults, >50 years	Telemonitoring Remote sensors in homes of older adults	None	Not reported	Mixed-methods study
Wang et al, 2016 [57]	29 older adults, >65 years, 71% (21) female	Telemonitoring Wearables, mobile devices, trackers, and in-home telemonitoring	None	Not reported	Cross-sectional study
Gordon & Hornbrook, 2016 [58]	231,082 older adults for database arm, 2602 older adults for survey arm	eHealth	None	Not reported	Mixed methods, database, and survey study
Williams et al, 2016 [59]	7 older adults, >60 years, dementia	eHealth	None	Not reported	Pilot study
Evans et al, 2016 [19]	41 older adults, >55 years, 57.1% (23) female, English speaking	mHealth Remote monitoring, wrist wearable, and wireless tablet	Control	Documentation for weight and blood pressure	Single-arm quasi-experimental study
Muller et al, 2016 [60]	43 older adults, ≥55 years, mobile phone use, no regular exercise	mHealth SMS and Physical Activity for Health Study	Control	Exercise, mood, fitness, health, mindfulness of the importance of exercise, and guilt	Two-arm parallel RCT
Quinn et al, 2016 [61]	118 older adults, >50 years, 66% (78) female, diabetes	mHealth Mobile diabetes intervention study	Control	HbA_{1c}	True experiment
Royackers et al, 2016 [62]	8 older adults, caring for loved ones in their last days	eHealth Point of care technology through eShift (home-based palliative care)	None	Comfort, independence, and autonomy	Qualitative pilot study
Duh et al, 2016 [63]	45 older adults, >60 years	Telecare CareMe	None	Not reported	Qualitative participatory design
Depatie & Bigbee, 2015 [64]	30 older adults, ≥60 years, 80% (24) female	mHealth Mobile health technology for older adults in rural communities	None	Not reported	Mixed-methods study
Moore et al, 2015 [65]	26 older adults, >55 years, 77% (20) male	eHealth Internet-based hearing health care for older adults	None	Not reported	Training study
Currie et al, 2015 [66]	168 older adults, ≥60 years, living in rural areas, long-term chronic pain	eHealth	None	Pain	Mixed-methods study

Authors, year	Participants	Intervention	Comparator	Medical outcomes reported	Study design
Grant et al, 2015 [67]	762 older adults, >60 years, 67% (511) female, 90% (686) Caucasian people	Telemonitoring LivingWell@Home, sensors (motion, bed, and humidity), emergency response systems, and biometric monitors (heart rate, blood pressure, weight, pulse oximetry, and blood glucose)	Control	Satisfaction, autonomy, and independence	RCT
Brenes et al, 2015 [68]	141 older adults, ≥60 years, 81% (114) female, living in rural areas, diagnosis of generalized anxiety disorder (GAD)	Telecare Telephone-delivered cognitive behavior therapy and telephone-delivered nondirective supportive therapy	Control	Worry, GAD, depression, and anxiety	RCT
Corbett et al, 2015 [69]	2192 older adults, ≥60 years	eHealth Online cognitive training package	Control	Reasoning, verbal learning, and instrumental activities of daily living	RCT
Mavandadi et al, 2015 [70]	1018 older adults, ≥65 years, 83.2% (847) female, community-dwelling, low-income, mental health symptoms	Telecare SUSTAIN care management system (assessment, monitoring, care management, and brief therapies)	Control	Depressive symptoms, anxiety symptoms, and MH functioning	RCT
Egede et al, 2015 [71]	90 older adults, ≥58 years, 98% (88) male, diagnosis of diabetes	Telemedicine Telepsychotherapy	Control	Geriatric depression scale, Beck depression inventory, and Diagnostic and Statistical Manual, version 4 symptoms	RCT
Chang et al, 2015 [72]	192 older adults, >60 years, 81% (156) female, cardiology diagnoses	Telemonitoring Remote cardiology management	None	Cardiac arrhythmias detected and paroxysmal atrial fibrillation detected	Pilot study
Boulos et al, 2015 [73]	27 older adults, 31 caregivers, 43 healthcare professionals	eHealth LiveWell Parkinson intervention and learning modules	None	Communication of the condition with the provider	Pilot study
Dino & deGuzman, 2015 [74]	82 older adults, demographics not reported	Telemedicine, mHealth, and eHealth	None	Not reported	Structured equation modeling
Czaja et al, 2015 [75]	24 older adults, >60 years, 71% (17) female, 94% (23) Hispanic people, diagnosis of hypertension	Telemonitoring Telehealth system that monitors blood pressure and body weight	Control	Self-management, health, and independence	Randomized feasibility study
Choi et al, 2015 [76]	42 older adults, ≥60 years, 81% (34) female, low-income, homebound, score of 15 or above on the 24-item Hamilton Rating Scale for Depression	Telecare Video tele-problem-solving therapy (PST) to in-person PST and telephone care calls	None	Depressive symptoms, understanding of depression, and social interaction	Qualitative

[a] mHealth: mobile health.

[b] EQ-5D VAS: European health-related quality of life utility with a visual analogue scale.

Risk of Bias Within Studies

At the study level, reviewers recorded observations of bias. The most frequently observed form of bias was selection bias (asking for volunteers for a research study involving technology will result in volunteers who already gravitate toward technology), which occurred in 7 out of 57 (13%) articles analyzed [15,26,30-32,37,39]. There were six instances of convenience samples from a local population [34,49-52,64]. Both examples of bias limit the external validity of the results.

Results of Individual Studies

Themes that resulted from the narrative analysis are listed in Table 2. Repetition can be observed in a frame of a theme owing to multiple observations from the same article for that theme. Translations from observations to themes for interventions, medical outcomes, and barriers are listed in Multimedia Appendix 1, Multimedia Appendix 2, and Multimedia Appendix 3, respectively. These appendices illustrate the logical inference reviewers made for each theme. For instance, one article listed remote patient monitoring for blood pressure, pulse oximeter, and body weight scales. These were categorized under telemonitoring [26]. The same article listed a decrease in hospital visits and a decrease in readmissions. These were categorized under an increase in hospital metrics. Additional data collected (bias, statistics, country of origin, and quality assessments) are displayed in Multimedia Appendix 4. In consensus meeting number two, we identified general observations, as depicted in the tables [20].

Table 2. Summary of the analysis of each article.

Authors, year	Intervention	Medical outcome theme	Theme of barriers
Hamilton et al, 2020 [26]	Telemonitoring	Increase in hospital metrics	Not reported
Theis et al, 2019 [5]	eHealth	Increase in satisfaction	Medical literacy Trust of the internet Ownership of technology[a]
Wildenbos et al, 2019 [27]	mHealth[b]	Increase in cognitive ability	Visual acuity[a] Mental acuity Technical literacy
Jakobsson et al, 2019 [28]	mHealth eHealth Telecare (phone)	Increase in cognitive ability	Social implications Privacy and security[a] Technical literacy Lack of desire Ownership of technology Lack of technical support
Karlsen et al, 2019 [29]	Telemonitoring	Increase in safety or security Increase in health-related quality of life Increase in safety or security[a] Increase in autonomy Increase in mindfulness of the condition	Mental acuity Visual acuity Social implications
Coley et al, 2019 [30]	eHealth	Not reported	Trust of the internet
Giesbrecht & Miller, 2019 [31]	eHealth	Increase in cognitive ability Increase in safety or security	Not reported
Brodbeck et al, 2019 [32]	eHealth	Decrease in psychological distress[a] Increase in health-related quality of life	Not reported
Mosley et al, 2019 [33]	eHealth	Not reported	Cost
Jensen et al, 2019 [34]	eHealth	Increase in autonomy[a]	Privacy and security Ownership of technology Lack of desire Lack of technical support Technical literacy
Rasche et al, 2018 [15]	eHealth	Not reported	Trust of the internet Technical literacy[a] Privacy and security
Portz et al, 2018 [35]	mHealth	Increase in mindfulness of the condition Increase in autonomy	Technical literacy Medical literacy
Castro Sweet et al, 2018 [36]	eHealth	Decrease in medical conditions surrounding diabetes[a]	Not reported
Joe et al, 2018 [37]	eHealth	Not reported	Visual acuity[a] Hand-eye coordination Technical literacy Lack of desire
Dham et al, 2018 [4]	Telehealth video call	Increase in satisfaction	Visual acuity Auditory acuity
Paige et al, 2018 [16]	eHealth	Not reported	Technical literacy Trust of the internet

Authors, year	Intervention	Medical outcome theme	Theme of barriers
Cajita et al, 2018 [38]	mHealth	Not reported	Medical literacy Mental acuity Lack of desire Technical literacy Ownership of technology Cost
Harte et al, 2018 [39]	mHealth	Not reported	Technical literacy
Gordon & Hornbrook, 2018 [40]	eHealth	Not reported	Cost Technical literacy
Bao et al, 2018 [41]	eHealth	Increase in cognitive ability Increase in activity or coordination[a]	Not reported
Egede et al, 2018 [42]	Telehealth video call	Decrease in psychological distress[a] Decrease in medical conditions surrounding pain[a]	Not reported
Platts-Mills et al, 2018 [43]	Telecare (phone)	Decrease in medical conditions surrounding pain	Not reported
Lopez-Villegas et al, 2018 [44]	Telemonitoring	Increase in health-related quality of life	Not reported
Dugas et al, 2018 [45]	mHealth	Decrease in medical conditions surrounding diabetes[a]	Not reported
Nalder et al, 2018 [46]	eHealth	Decrease in medical conditions surrounding diabetes Increase in autonomy Decrease in psychological distress Increase in autonomy	Technical literacy
Buck et al, 2017 [47]	eHealth	Increase in documentation to give the provider Increase in mindfulness of the condition	Technical literacy
Ware et al, 2017 [14]	eHealth	Not reported	Trust of the internet Medical literacy Technical literacy Social implications Lack of technical support Privacy and security
Chang et al, 2017 [48]	mHealth	Increase in autonomy[a]	Cost
Cajita et al, 2017 [49]	mHealth	Not reported	Medical literacy Lack of desire Cost Technical literacy Social implications
LaMonica et al, 2017 [50]	eHealth	Increase in cognitive ability	Auditory acuity Cost Auditory acuity
Bahar-Fuchs et al, 2017 [51]	eHealth	Increase in cognitive ability[a] Decrease in psychological distress	Not reported
Nahm et al, 2017 [52]	eHealth	Increase in mindfulness of the condition Increase in autonomy Increase in activity or coordination	Not reported

Authors, year	Intervention	Medical outcome theme	Theme of barriers
Knaevelsrud et al, 2017 [53]	eHealth	Increase in safety or security Increase in satisfaction Increase in autonomy Increase in health-related quality of life	Not reported
Reijnders et al, 2017 [54]	eHealth	Increase in activity or coordination Increase in cognitive ability Increase in autonomy	Not reported Privacy and security
Hamblin et al, 2017 [56]	Telemonitoring	Increase in autonomy Increase in mindfulness of the condition Increase in safety or security	Technical literacy Medical literacy Social implications[a]
Mageroski et al, 2016 [55]	Telemonitoring	Not reported	Cost
Wang et al, 2016 [57]	Telemonitoring	Not reported	Lack of desire
Gordon & Hornbrook, 2016 [58]	eHealth	Not reported	Ownership of technology Lack of technical support Cost Technical literacy Hand-eye coordination Trust of the internet Social implications Lack of desire
Williams et al, 2016 [59]	eHealth	Not reported	Technical literacy Lack of technical support Mental acuity Visual acuity Hand-eye coordination
Evans et al, 2016 [19]	mHealth	Increase in documentation to give the provider	Lack of desire Technical literacy Lack of desire Ownership of technology
Muller et al, 2016 [60]	mHealth	Increase in activity or coordination[a] Decrease in psychological distress Decrease in medical conditions surrounding diabetes Increase in mindfulness of the condition Increase in guilt	Lack of desire
Quinn et al, 2016 [61]	mHealth	Decrease in medical conditions surrounding diabetes	Visual acuity Auditory acuity
Royackers et al, 2016 [62]	eHealth	Increase in safety or security Increase in autonomy[a]	Not reported
Duh et al, 2016 [63]	Telecare (phone)	Not reported	Mental acuity Lack of desire Lack of technical support Technical literacy Cost

Authors, year	Intervention	Medical outcome theme	Theme of barriers
Depatie & Bigbee, 2015 [64]	mHealth	Not reported	Cost Lack of desire Social implications Technical literacy Lack of technical support Privacy and security
Moore et al, 2015 [65]	eHealth	Not reported	Technical literacy Computer anxiety Lack of technical support
Currie et al, 2015 [66]	eHealth	Decrease in medical conditions surrounding pain	Social implications
Grant et al, 2015 [67]	Telemonitoring	Increase in health-related quality of life Increase in autonomy[a]	Lack of desire Cost Privacy and security
Brenes et al, 2015 [68]	Telecare (phone)	Decrease in psychological distress[a]	Not reported
Corbett et al, 2015 [69]	eHealth	Increase in cognitive ability[a] Increase in health-related quality of life	Not reported
Mavandadi et al, 2015 [70]	Telecare (phone)	Decrease in psychological distress[a]	Not reported
Egede et al, 2015 [71]	Telehealth video call	Decrease in psychological distress[a]	Not reported
Chang et al, 2015 [72]	Telemonitoring	Increase in mindfulness of the condition[a]	Not reported
Boulos et al, 2015 [73]	eHealth	Increase in documentation to give the provider	Medical literacy Lack of technical support Mental acuity Hand-eye coordination Visual acuity
Dino & deGuzman, 2015 [74]	mHealth eHealth Telemonitoring	Not reported	Lack of desire Lack of technical support
Czaja et al, 2015 [75]	Telemonitoring	Increase in autonomy[a] Decrease in medical conditions surrounding diabetes	Technical literacy
Choi et al, 2015 [76]	Telehealth video call	Decrease in psychological distress Increase in mindfulness of the condition Increase in autonomy	Ownership of technology Lack of desire

[a] Multiple uses of this theme in the same article. See Multimedia Appendix 1 for a complete list of individual observations and their translation to themes.

[b] mHealth: mobile health.

Risk of Bias Across Studies and Quality Assessments

Table 3 summarizes the quality indicators identified by the JHNEBP tool [15]. The most frequent strength rating was III, followed by I, II, and IV. The most frequent evidence rating was A, followed by B and C. No strengths below IV were encountered. A full list of quality assessments is presented in Multimedia Appendix 4. Articles that did not meet the minimum standards of quality were not included in the analysis. This decision was made to limit the bias inherent to nondata-driven opinions or conclusions that do not logically follow the data.

Table 3. Summary of quality indicators.

Quality indicator	Value (N=57), n (%)
Strength of evidence	
I (experimental study, RCT[a], or meta-analysis of RCT)	18 (32%)
II (quasiexperimental study)	10 (17%)
III (nonexperimental, qualitative, or meta-synthesis study)	28 (49%)
IV (opinion)	1 (2%)
Quality of evidence	
A (high quality)	33 (58%)
B (good quality)	23 (40%)
C (low quality or major flaws)	1 (2%)

[a]RCT: randomized controlled trial.

Additional Analysis

The results of consensus meeting three identified the themes that corresponded with telehealth interventions, barriers to the use of telehealth, and medical outcomes. These are summarized in Tables 4-6.

Interventions of Telehealth

Five themes for interventions were identified. Two studies used multiple telehealth interventions. Table 4 lists the interventions with the associated references, number of occurrences, and probability of occurrence in the review. The most common intervention was eHealth (computer driven), followed by mHealth (smart device driven), telemonitoring (remote sensors), telecare (phone), and telehealth video call.

Table 4. Affinity matrix for telehealth interventions.

Intervention	References	Number of occurrences (N=62)	Probability of occurrence
eHealth	[5,14-16,28,30-34,36,37,40,41,46,47,50-54,58,59,62,65,66,69,73,74]	29	47%
mHealth[a]	[19,27,28,35,38,39,45,48,49,60,61,64,74]	13	21%
Telemonitoring	[26,29,44,55-57,67,72,74,75]	10	16%
Telecare (phone)	[28,43,63,68,70]	5	8%
Video call	[4,29,42,71,76]	5	8%

[a]mHealth: mobile health.

Medical Outcomes and Health-Related Quality of Life Enhancers

Thirteen themes and one observation that did not correspond with a theme for medical outcomes and quality of life factors were reported. Table 5 lists the outcomes with their associated references, number of occurrences, and probability of occurrence in this review. The most common theme for medical outcomes associated with telehealth interventions was *decrease in psychological distress* (decrease in anxiety symptoms, decrease in depressive symptoms, decrease in embitterment, decrease in grief, decrease in worry, decrease in loneliness, increase in emotional support, and increase in mood), with 21 of 118 (18%) occurrences [32,42,46,51,60,68,70,71,76]. The theme associated with quality of life factors was listed as an *increase in autonomy* (increase in locus of control, increase in autonomy, increase in responsibility, increase in motivation to self-manage, and increase in independence), with 18 of 118 (15%) occurrences [29,34,35,46,48,52-54,56,62,67,75,76]. One theme was associated with an *increase in cognitive ability* (increase in skill ability, increase in sensory organization, increase in memory, increase in cognitive activity, and increase in reasoning), with 11 of 118 (9%) occurrences [19,20,23,32,41,42,60]. Another theme was associated with a *decrease in symptoms surrounding diabetes* (decrease in HbA$_{1c}$, decrease in cholesterol, increase in glucose management, and increase in diabetes health), with 9 of 118 (8%) occurrences [28,36,37,51,52,66]. Another theme was associated with an *increase in mindfulness of the condition* (increase in medical events detected, increase in education exposure, and more awareness of danger areas for falls like outside or stairwells), with 8 of 118 (7%) occurrences [21,27,38,43,47,51,63,67]. The next theme summarized observations of an *increase in the sense of safety, security, or comfort*, with 7 of 118 (6%) occurrences [21,23,44,47,53]. The last set of themes comprised 25% of the observations, and they were an *increase in health-related quality of life* (increase in life satisfaction and increase in the feeling of being understood); *increase in activity or coordination* (increase in mobility, increase in activity, increase in exercise, decrease in weight, decrease in BMI, increase in balance evaluation, and increase in the feeling of stability); *decrease in medical conditions surrounding pain* (decrease in alcohol abuse, decrease in cannabis misuse, decrease in cannabis dependence, and decrease in pain); *increase in documentation to give to the provider* (documentation and communication with the provider); *increase in satisfaction* (satisfaction with the health care system); and *increase in hospital metrics* (decrease in readmissions and decrease in hospital visits). The last observation was the only negative outcome. One participant noted that the SMS text messages she received as part of an exercise RCT *increased her level of guilt* because she was not exercising.

Table 5. Affinity matrix for medical outcomes and quality of life factors observed by older adults using telehealth.

Theme or observation	References	Number of occurrences (N=118)	Probability of occurrence
Decrease in psychological distress	[32,42,46,51,60,68,70,71,76]	21	19%
Increase in autonomy	[29,34,35,46,48,52-54,56,62,67,75,76]	18	16%
Not reported	[14-16,30,33,37-40,49,55,57-59,63-65,74]	18	16%
Increase in cognitive ability	[27,28,31,41,50,51,54,69]	11	10%
Decrease in medical conditions surrounding diabetes	[36,45,46,60,61,75]	9	8%
Increase in mindfulness of the condition	[29,35,47,52,56,60,72,77]	8	7%
Increase in safety or security	[29,31,53,56,62]	7	6%
Increase in health-related quality of life	[29,32,44,53,67,69]	6	5%
Increase in activity or coordination	[41,52,54,60]	6	5%
Decrease in medical conditions surrounding pain	[42,43,66]	5	4%
Increase in documentation to give the provider	[19,47,73]	3	3%
Increase in satisfaction	[4,5,53]	3	3%
Increase in hospital metrics	[26]	2	2%
Increase in guilt	[60]	1	1%

Barriers

Fourteen themes and one observation that did not fit into a theme for barriers were observed. Table 6 lists the barriers with their associated references, number of occurrences, and probability of occurrence in this review. The barrier that was reported most often was *technical literacy* (I do not understand technology, I cannot navigate menus, I do not know how, etc) [14-16,19,27,28,34,35,37-40,46,47,49,56,58,59,63-65,75]. The theme noted the second most often was *lack of desire* (laziness, I do not want to, I am too busy, etc) [19,28,34,37,38,49,57,58,60,63,64,67,74,76]. Another theme was *cost* (too expensive, we live off a fixed income, etc) [33,38,40,48-50,55,58,63,64,67]. The theme *lack of technical support* included the following: my friends or family are not able to help me, I do not understand the interface, etc [14,28,34,58,63-65,73,74]. The theme *visual acuity* included the following: fonts or icons are too small, color contrast, etc [4,27,29,37,59,61,73]. The next observation was a surprise to our reviewing team; the theme was *social implications of using a telemonitoringdevice* (I do not want to bother a first responder, I do not want a stranger coming to my house, I do not want anyone coming to my house late at night, I had a bad experience the last time I used the telemonitoring device, I do not want my neighbor to overhear me using this thing, I do not have my own email, I do not understand social media, etc) [14,28,29,49,56,58,64,66]. The next theme was *ownership of technology* (no phone, no computer, no internet access, etc) [5,19,28,34,38,58,76]. The last set of themes and observations comprised less than 25% of the observations, and they were *privacy and security concerns*, *medical literacy* (I do not understand terminology, I do not understand test results, etc), *trust of the internet*, *mental acuity* (computers confuse me, the interface is too complex, I cannot focus for very long, how did I get to this page? etc), *hand-eye coordination* (particularly with those who have Parkinson disease, but not exclusively), *auditory acuity*, and *computer anxiety*.

Table 6. Affinity matrix for barriers to the use of telehealth by older adults.

Themes of barriers	References	Number of occurrences (N=144)	Probability of occurrence
Technical literacy	[14-16,19,27,28,34,35,37-40,46,47,49,56,58,59,63-65,75]	25	17%
Not reported	[26,31,32,36,41-45,51-54,62,68-72]	19	13%
Lack of desire	[19,28,34,37,38,49,57,58,60,63,64,67,74,76]	15	10%
Cost	[33,38,40,48-50,55,58,63,64,67]	11	8%
Lack of technical support	[14,28,34,58,63-65,73,74]	10	7%
Visual acuity	[4,27,29,37,59,61,73]	10	7%
Social implications	[14,28,29,49,56,58,64,66]	9	6%
Ownership of technology	[5,19,28,34,38,58,76]	8	6%
Privacy and security	[5,19,28,34,38,58,76]	8	6%
Medical literacy	[5,14,15,35,38,49,56,73]	8	6%
Trust of the internet	[5,14-16,30,58]	6	4%
Mental acuity	[27,29,38,59,63,73]	6	4%
Hand-eye coordination	[37,58,59,73]	4	3%
Auditory acuity	[4,50,61]	4	3%
Computer anxiety	[65]	1	1%

Interactions Between Observations

There were several interactions worth discussing. We analyzed the interactions between interventions and barriers. Ten instances of eHealth interventions were mentioned with *technical literacy* [14-16,30,33,37,40,58,59,65,74]. Eight instances of eHealth interventions were mentioned with *lack of technical support* [14,28,34,58,59,65,73,74]. There were eight instances of mHealth interventions associated with *technical literacy* [19,27,28,35,38,39,49,64], but these were hardly mentioned at all with *lack of technical support* [28,74]. The interventions of mHealth were also associated with the barrier of *lack of desire*. This occurred six times in the literature [28,38,39,60,64,74]. Contrary to literature on the digital divide, eHealth and mHealth were only marginally associated with *ownership of technology*, which occurred four [5,28,34,58] and three times [19,28,38], respectively. Commensurate with literature on generational trends, both eHealth and mHealth were associated with *privacy and security concerns*, which occurred four [14,15,34,35] and two times [28,64], respectively. Both eHealth and mHealth were associated with the barrier *medical literacy*, which occurred four [5,14,15,73] and three times [35,38,39], respectively. Surprisingly, eHealth was associated with *hand-eye coordination*, but mHealth was not [37,58,60,73]. Finally, eHealth was associated with *lack of trust of the internet*, which occurred six times in the literature [5,14-16,30,58].

We also analyzed the interactions between interventions and medical outcomes. eHealth interventions were associated with an increase in cognitive ability. This interaction occurred seven times in the literature [28,31,41,50,51,54,69].

Results Summary

This review identified 13 themes and one lone observation of medical outcomes incident with the adoption of five types of telehealth approaches. This review also identified 14 themes and one observation of barriers to the adoption of telehealth.

Discussion

Common Barriers to Telehealth

In this review, we were able to identify the common barriers associated with older adults utilizing telehealth. The most frequent barriers were lack of desire, cost, lack of technical support, visual acuity, social implications of use, ownership of technology, privacy and security, medical literacy, trust of the internet, mental acuity, hand-eye coordination, auditory acuity, and computer anxiety. Each of these barrier areas could present hurdles for elderly people dealing with telehealth and reasons to not use it. Lack of technical literacy is a large area of concern, as many elderly people have issues using computers to check email or smartphones to make telephone calls [13]. Because this is new to this population, they are also being held back from acceptance by a simple lack of wanting to do it [28,34,37-39,57,58,60,63,64,67,74,76]. It seems to be an easy thing to add to one's daily tasks, but when one has lived largely without the use of these technologies, it can become an arduous task to "sell" the benefits of the sudden use of new technology and learning how to use new technology. They have the attitude "as it was not needed before, why bother to learn it now?" This can prove to be an uphill battle for providers who are attempting to utilize new technologies in different ways.

The cost of technology is also quite prohibitive, as computers, smartphones, and other devices cost hundreds to thousands of dollars. Those living on fixed incomes are cash strapped and may not be able to afford to purchase or use such new technologies. Not owning such technologies presents its own concerns for the provision of care. Besides cost, there are concerns in this population regarding the ability to actually utilize the modality of telehealth efficiently. Issues with visual acuity [4,27,29,37,59,61,73], mental acuity [27,29,38,59,63,73], hand-eye coordination [37,58,59,73], and auditory acuity [4,50,61] are all relevant concerns for elderly people. Many people, as they become older, experience decreases in the efficiencies of the operations of many body systems, including their senses. Many develop disease processes that can affect their mental status, vision, and hearing, and any or all of these could easily lead to problems with being able to use technology, let alone having a clear understanding of what they need to be doing with the device or even how to interact with it.

The elderly population also has relevant concerns with trust and technology, as they are one of the prime targets for abuse from their use of technology according to popular media [13,78]. This is where lack of technical support for the use of technology can become a very relevant area of concern. There is no affordable and adequate source of “technical support” to simply learn how to use devices [14,28,59,73]. This lack of knowledge and available education can be a very problematic barrier for the use of the modality of telehealth. Furthermore, problems surrounding trust of the internet [5,14-16,30,58], concerns of privacy and security [5,19,28,34,38,58,76], and even computer anxiety [65] can figure into the use of technology. As there are concerns with privacy and security, telehealth could easily cause patients to succumb to some level of anxiety. Not understanding the modality of telehealth or how to use it can add to the level of this anxiety at an exponential rate.

Another consideration with the use of telehealth is that it requires a certain level of user knowledge. The utilization of medical applications requires the user to have some knowledge of medical terms, procedures, etc [5,14,15,35,38,49,56,73]. This is often not the case, as this population was raised without the internet or medical knowledge. Medical knowledge came from physicians during their younger years, and only recently, the approach has changed to the utilization of internet web searches to garner knowledge about symptoms and diagnoses. This is an entirely new world for the elderly population and a relevant barrier to the use of these applications overall. Overcoming this knowledge gap could prove to be an insurmountable task or one that requires any telehealth use to be kept to an absolute minimum for knowledge or know-how on the part of the user.

Common Outcomes Associated With Telehealth Interventions

The research supports strong medical outcomes incident to the use of telehealth as follows: *decreased psychological distress* [32,42,46,51,60,68,70,71,76], *increased autonomy* [29,34,35,46,48,52-54,56,62,67,75,76], *increased cognitive ability* [27,28,31,41,50,51,54,69], and many others. This review supports an *increased quality of life* for those who adopt telehealth [29,32,44,53,67,69]. The use of telehealth can lead to less psychological distress, as users know that they have a way of communicating their medical concerns to their providers in a much easier and faster way. This could eventually enable better health due to better management, thus allowing for fewer associated medical conditions for those patients who use telehealth for assistance in the management of their care.

The observation of greater documentation for providers demonstrated that the use of telehealth is not all about the patient. It is just as much about practitioners providing care. The use of telehealth allows for much faster accessibility to documentation to provide care or even real-time information about the patient to allow for immediate diagnosis or intervention, based on information being gathered by the used technology. This can make the provision of care easier and much more efficient for the field, which is already seeing more patients than it can comfortably manage.

Interactions Among Outcomes, Barriers, and Types of Interventions

eHealth interventions were the most frequently observed interventions in the literature, and these interventions were most frequently associated with the barriers of technical literacy and lack of technical support. This observation is interesting because general technical support, whether from friends, neighbors, family, or caregivers, or professionally acquired technical support is a control for the barrier of technical literacy. The interaction between eHealth and *technical literacy* is interesting as well. This could signal that older adults are more adept at mobile technology than computer technology for application of telehealth. This supposition is supported by the literature because many older adults are turning to mobile technology to communicate with children and grandchildren [13]. The interaction between mHealth and *lack of desire* is noteworthy. This seems to indicate that older adults are willing to interact with mobile technology to communicate with children and grandchildren, but they are not as willing to use it for telehealth interventions.

Study Quality and Literature Bias

The assessment of the quality of the articles studied is worthy of discussion. The majority (27/57, 49%) of the articles analyzed were level III (nonexperimental, qualitative, or meta-synthesis studies). The reviewers would have preferred to analyze only the highest level (level I) (experimental study or RCT), but only 10 (17%) such studies were available. Fortunately, 98% (56/57) of the articles were rated as quality level A (high quality) or B (good quality). The importance of this rating cannot be understated. If the findings from this review were from low-quality articles, the results would not be as strong. By analyzing high-quality articles with strong levels of evidence, readers can be more assured of the results. Research articles with strong study designs and sufficiently large samples are generally accepted in the scientific field for their veracity.

Limitations

The authors identified the low number of articles analyzed as a limitation of this systematic review. If the authors conduct another systematic review on the same topic, they would like to have a larger analysis pool. This could be achieved by broadening the years of study in the selection or by reducing the threshold of quality. However, the additional years of study would only repeat the results from previously published reviews of a similar topic, and lowering the threshold of quality would introduce articles with dubious results.

Although not intentional, the authors realized that selection bias may be present in this article. To combat selection bias, the authors worked to minimize its effects by ensuring each article was reviewed by at least two authors. The authors held consensus meetings after each screening to provide feedback and reach total agreement on the inclusion and exclusion of articles for the analysis.

Another source of bias that could have affected this article is publication bias. To control for publication bias, the authors searched the Boolean search string in Google Scholar. This action was intended to identify articles from lesser-known journals that may not have appeared in MEDLINE or CINAHL.

Another limitation is our inclusion of people aged 50 years or above in the study of older adults. Most studies categorize older adults as those aged 65 years or above. The elderly population currently spans baby boomers and the silent generation. The youngest members of the former group are still working and are most likely using technology fluently. It is possible that our generalizations do not apply to all members of the elderly demographic.

Future Research

Health care systems can utilize knowledge of these barriers to develop solutions for broadening the use of telehealth among older adults. A multidisciplinary approach and culture of collaboration between administrative leadership and providers may be the most effective and immediate manner of implementing solutions to breach these barriers and strengthen the reach of health care services. However, some barriers may be out of the scope of impact, and policy makers should consider supporting the efforts. Future research should be conducted on methods for personalizing telehealth in older adults before implementation.

Conclusion

Providing sufficient health care access to the rapidly growing aging population has been an imminent issue, and telehealth is a useful tool that can provide a solution. While health care systems increase their telehealth efforts to improve access to health care services among vulnerable populations, such as older adults, some health care organizations do not consider the technological, educational, financial, and behavioral barriers before implementing telehealth solutions. It is imperative that health care systems use a multidisciplinary approach and collaborate with health care providers, community partners, and policy makers to address these barriers of utilizing telehealth among older adults and to successfully implement telehealth solutions. This systematic review provides some understanding of older adults’ perspectives and experiences with the barriers of implementing telehealth services.

Authors' Contributions

Conceptualization: CSK, NW, JF, ENP, and SZ; methodology: CSK, NW, JF, ENP, and SZ; formal analysis: CSK, NW, JF, ENP, SZ, and MM; investigation: CSK, NW, JF ENP, and SZ; writing—original draft preparation: NW, JF, ENP, and SZ; writing—review and editing: JF; visualization: ENP; supervision: NW; project administration: SZ; final draft: CSK and MM.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed observations on interventions and corresponding themes.
[DOCX File , 29 KB - medinform_v8i8e20359_app1.docx]

Multimedia Appendix 2
Detailed observations on medical outcomes and corresponding themes.
[DOCX File , 42 KB - medinform_v8i8e20359_app2.docx]

Multimedia Appendix 3
Detailed observations on barriers and corresponding themes.
[DOCX File , 37 KB - medinform_v8i8e20359_app3.docx]

Multimedia Appendix 4
Bias, country of origin, statistics, and quality assessments.
[DOCX File , 51 KB - medinform_v8i8e20359_app4.docx]

References

1. Ageing and health. World Health Organization. 2018 Feb 05. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2020-04-02]

2. From Pyramid to Pillar: A Century of Change, Population of the U.S. US Census Bureau. 2018 Mar 13. URL: https://www.census.gov/library/visualizations/2018/comm/century-of-change.html [accessed 2020-04-02]
3. Colby S, Ortman J. The Baby Boom Cohort in the United States: 2012-2060. US Census Bureau. 2014 May 01. URL: https://www.census.gov/library/publications/2014/demo/p25-1141.html [accessed 2020-04-02]
4. Dham P, Gupta N, Alexander J, Black W, Rajji T, Skinner E. Community based telepsychiatry service for older adults residing in a rural and remote region- utilization pattern and satisfaction among stakeholders. BMC Psychiatry 2018 Sep 27;18(1):316 [FREE Full text] [doi: 10.1186/s12888-018-1896-3] [Medline: 30261845]
5. Theis S, Schäfer D, Bröhl C, Schäfer K, Rasche P, Wille M, et al. Predicting technology usage by health information need of older adults: Implications for eHealth technology. Work 2019;62(3):443-457. [doi: 10.3233/WOR-192878] [Medline: 30909259]
6. Ryu S. Telemedicine: Opportunities and Developments in Member States: Report on the Second Global Survey on eHealth 2009 (Global Observatory for eHealth Series, Volume 2). Healthc Inform Res 2012;18(2):153. [doi: 10.4258/hir.2012.18.2.153]
7. Yoon H, Jang Y, Vaughan PW, Garcia M. Older Adults' Internet Use for Health Information: Digital Divide by Race/Ethnicity and Socioeconomic Status. J Appl Gerontol 2020 Jan;39(1):105-110. [doi: 10.1177/0733464818770772] [Medline: 29661052]
8. Dorsey ER, Topol EJ. Telemedicine 2020 and the next decade. The Lancet 2020 Mar;395(10227):859. [doi: 10.1016/s0140-6736(20)30424-4]
9. Trump Administration Issues Second Round of Sweeping Changes to Support U.S. Healthcare System During COVID-19 Pandemic. Centers for Medicare & Medicaid Services. 2020 Apr 30. URL: https://www.cms.gov/newsroom/press-releases/trump-administration-issues-second-round-sweeping-changes-support-us-healthcare-system-during-covid [accessed 2020-04-30]
10. Foster MV, Sethares KA. Facilitators and barriers to the adoption of telehealth in older adults: an integrative review. Comput Inform Nurs 2014 Nov;32(11):523-33; quiz 534. [doi: 10.1097/CIN.0000000000000105] [Medline: 25251862]
11. Kampmeijer R, Pavlova M, Tambor M, Golinowska S, Groot W. The use of e-health and m-health tools in health promotion and primary prevention among older adults: a systematic literature review. BMC Health Serv Res 2016 Sep 05;16 Suppl 5:290 [FREE Full text] [doi: 10.1186/s12913-016-1522-3] [Medline: 27608677]
12. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ Open 2017 Aug 03;7(8):e016242 [FREE Full text] [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]
13. Kruse CS, Mileski M, Moreno J. Mobile health solutions for the aging population: A systematic narrative analysis. J Telemed Telecare 2017 May;23(4):439-451. [doi: 10.1177/1357633X16649790] [Medline: 27255207]
14. Ware P, Bartlett SJ, Paré G, Symeonidis I, Tannenbaum C, Bartlett G, et al. Using eHealth Technologies: Interests, Preferences, and Concerns of Older Adults. Interact J Med Res 2017 Mar 23;6(1):e3 [FREE Full text] [doi: 10.2196/ijmr.4447] [Medline: 28336506]
15. Rasche P, Wille M, Bröhl C, Theis S, Schäfer K, Knobe M, et al. Prevalence of Health App Use Among Older Adults in Germany: National Survey. JMIR Mhealth Uhealth 2018 Jan 23;6(1):e26 [FREE Full text] [doi: 10.2196/mhealth.8619] [Medline: 29362211]
16. Paige SR, Miller MD, Krieger JL, Stellefson M, Cheong J. Electronic Health Literacy Across the Lifespan: Measurement Invariance Study. J Med Internet Res 2018 Jul 09;20(7):e10434 [FREE Full text] [doi: 10.2196/10434] [Medline: 29986848]
17. Choi N. Telehealth Depression Treatments for Older Adults. ClinicalTrials.gov. 2015 Nov 09. URL: https://clinicaltrials.gov/ct2/show/NCT02600754 [accessed 2020-04-02]
18. Scogin F, Lichstein K, DiNapoli EA, Woosley J, Thomas SJ, LaRocca MA, et al. Effects of Integrated Telehealth-Delivered Cognitive-Behavioral Therapy for Depression and Insomnia in Rural Older Adults. J Psychother Integr 2018 Sep;28(3):292-309 [FREE Full text] [doi: 10.1037/int0000121] [Medline: 30930607]
19. Evans J, Papadopoulos A, Silvers CT, Charness N, Boot WR, Schlachta-Fairchild L, et al. Remote Health Monitoring for Older Adults and Those with Heart Failure: Adherence and System Usability. Telemed J E Health 2016 Jun;22(6):480-488 [FREE Full text] [doi: 10.1089/tmj.2015.0140] [Medline: 26540369]
20. Kruse CS. Writing a Systematic Review for Publication in a Health-Related Degree Program. JMIR Res Protoc 2019 Oct 14;8(10):e15490 [FREE Full text] [doi: 10.2196/15490] [Medline: 31527018]
21. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009 Jul 21;339:b2535. [doi: 10.1136/bmj.b2535] [Medline: 19621072]
22. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
23. Dang D, Dearholt S. Johns Hopkins Nursing Evidence-based Practice: Model and Guidelines. Indianapolis, IN: SIGMA Theta Tau International; 2018.
24. Light RJ. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin 1971;76(5):365-377. [doi: 10.1037/h0031643]
25. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]

26. Hamilton T, Johnson L, Quinn BT, Coppola J, Sachs D, Migliaccio J, et al. Telehealth Intervention Programs for Seniors: An Observational Study of a Community-Embedded Health Monitoring Initiative. Telemed J E Health 2020 Apr;26(4):438-445. [doi: 10.1089/tmj.2018.0248] [Medline: 30994409]
27. Wildenbos GA, Jaspers MW, Schijven MP, Dusseljee-Peute LW. Mobile health for older adult patients: Using an aging barriers framework to classify usability problems. Int J Med Inform 2019 Apr;124:68-77. [doi: 10.1016/j.ijmedinf.2019.01.006] [Medline: 30784429]
28. Jakobsson E, Nygård L, Kottorp A, Malinowsky C. Experiences from using eHealth in contact with health care among older adults with cognitive impairment. Scand J Caring Sci 2019 Jun;33(2):380-389. [doi: 10.1111/scs.12634] [Medline: 30628736]
29. Karlsen C, Moe CE, Haraldstad K, Thygesen E. Caring by telecare? A hermeneutic study of experiences among older adults and their family caregivers. J Clin Nurs 2019 Apr;28(7-8):1300-1313. [doi: 10.1111/jocn.14744] [Medline: 30552788]
30. Coley N, Rosenberg A, van Middelaar T, Soulier A, Barbera M, Guillemont J, MIND-AD, HATICE groups. Older Adults' Reasons for Participating in an eHealth Prevention Trial: A Cross-Country, Mixed-Methods Comparison. J Am Med Dir Assoc 2019 Jul;20(7):843-849.e5. [doi: 10.1016/j.jamda.2018.10.019] [Medline: 30541689]
31. Giesbrecht EM, Miller WC. Effect of an mHealth Wheelchair Skills Training Program for Older Adults: A Feasibility Randomized Controlled Trial. Arch Phys Med Rehabil 2019 Nov;100(11):2159-2166. [doi: 10.1016/j.apmr.2019.06.010] [Medline: 31336101]
32. Brodbeck J, Berger T, Biesold N, Rockstroh F, Znoj HJ. Evaluation of a guided internet-based self-help intervention for older adults after spousal bereavement or separation/divorce: A randomised controlled trial. J Affect Disord 2019 Jun 01;252:440-449. [doi: 10.1016/j.jad.2019.04.008] [Medline: 31003114]
33. Mosley CL, Langley LM, Davis A, McMahon CM, Tremblay KL. Reliability of the Home Hearing Test: Implications for Public Health. J Am Acad Audiol 2019 Mar;30(3):208-216 [FREE Full text] [doi: 10.3766/jaaa.17092] [Medline: 30461396]
34. Jensen CM, Overgaard S, Wiil UK, Clemensen J. Can Tele-Health Support Self-Care and Empowerment? A Qualitative Study of Hip Fracture Patients' Experiences With Testing an "App". SAGE Open Nursing 2019 Feb 21;5:237796081982575 [FREE Full text] [doi: 10.1177/2377960819825752]
35. Portz JD, Vehovec A, Dolansky MA, Levin JB, Bull S, Boxer R. The Development and Acceptability of a Mobile Application for Tracking Symptoms of Heart Failure Among Older Adults. Telemed J E Health 2018 Feb;24(2):161-165 [FREE Full text] [doi: 10.1089/tmj.2017.0036] [Medline: 28696832]
36. Castro Sweet CM, Chiguluri V, Gumpina R, Abbott P, Madero EN, Payne M, et al. Outcomes of a Digital Health Program With Human Coaching for Diabetes Risk Reduction in a Medicare Population. J Aging Health 2018 Jun;30(5):692-710 [FREE Full text] [doi: 10.1177/0898264316688791] [Medline: 28553807]
37. Joe J, Hall A, Chi N, Thompson H, Demiris G. IT-based wellness tools for older adults: Design concepts and feedback. Inform Health Soc Care 2018 Mar;43(2):142-158. [doi: 10.1080/17538157.2017.1290637] [Medline: 28350186]
38. Cajita MI, Hodgson NA, Lam KW, Yoo S, Han H. Facilitators of and Barriers to mHealth Adoption in Older Adults With Heart Failure. Comput Inform Nurs 2018 Aug;36(8):376-382 [FREE Full text] [doi: 10.1097/CIN.0000000000000442] [Medline: 29742549]
39. Harte R, Hall T, Glynn L, Rodríguez-Molinero A, Scharf T, Quinlan LR, et al. Enhancing Home Health Mobile Phone App Usability Through General Smartphone Training: Usability and Learnability Case Study. JMIR Hum Factors 2018 Apr 26;5(2):e18 [FREE Full text] [doi: 10.2196/humanfactors.7718] [Medline: 29699969]
40. Gordon NP, Hornbrook MC. Older adults' readiness to engage with eHealth patient education and self-care resources: a cross-sectional survey. BMC Health Serv Res 2018 Mar 27;18(1):220 [FREE Full text] [doi: 10.1186/s12913-018-2986-0] [Medline: 29587721]
41. Bao T, Carender WJ, Kinnaird C, Barone VJ, Peethambaran G, Whitney SL, et al. Effects of long-term balance training with vibrotactile sensory augmentation among community-dwelling healthy older adults: a randomized preliminary study. J Neuroeng Rehabil 2018 Jan 18;15(1):5 [FREE Full text] [doi: 10.1186/s12984-017-0339-6] [Medline: 29347946]
42. Egede LE, Walker RJ, Payne EH, Knapp RG, Acierno R, Frueh BC. Effect of psychotherapy for depression via home telehealth on glycemic control in adults with type 2 diabetes: Subgroup analysis of a randomized clinical trial. J Telemed Telecare 2018 Oct;24(9):596-602. [doi: 10.1177/1357633X17730419] [Medline: 28945160]
43. Platts-Mills TF, Hollowell AG, Burke GF, Zimmerman S, Dayaa JA, Quigley BR, et al. Randomized controlled pilot study of an educational video plus telecare for the early outpatient management of musculoskeletal pain among older emergency department patients. Trials 2018 Jan 05;19(1):10 [FREE Full text] [doi: 10.1186/s13063-017-2403-8] [Medline: 29304831]
44. Lopez-Villegas A, Catalan-Matamoros D, Lopez-Liria R, Enebakk T, Thunhaug H, Lappegård KT. Health-related quality of life on tele-monitoring for users with pacemakers 6 months after implant: the NORDLAND study, a randomized trial. BMC Geriatr 2018 Sep 21;18(1):223 [FREE Full text] [doi: 10.1186/s12877-018-0911-3] [Medline: 30241511]
45. Dugas M, Crowley K, Gao GG, Xu T, Agarwal R, Kruglanski AW, et al. Individual differences in regulatory mode moderate the effectiveness of a pilot mHealth trial for diabetes management among older veterans. PLoS One 2018;13(3):e0192807 [FREE Full text] [doi: 10.1371/journal.pone.0192807] [Medline: 29513683]

46. Nalder E, Marziali E, Dawson DR, Murphy K. Delivering cognitive behavioural interventions in an internet-based healthcare delivery environment. British Journal of Occupational Therapy 2018 Mar 23;81(10):591-600. [doi: 10.1177/0308022618760786]
47. Buck H, Pinter A, Poole E, Boehmer J, Foy A, Black S, et al. Evaluating the older adult experience of a web-based, tablet-delivered heart failure self-care program using gerontechnology principles. Geriatr Nurs 2017;38(6):537-541. [doi: 10.1016/j.gerinurse.2017.04.001] [Medline: 28554497]
48. Chang C, Lee T, Mills ME. Experience of Home Telehealth Technology in Older Patients With Diabetes. Comput Inform Nurs 2017 Oct;35(10):530-537. [doi: 10.1097/CIN.0000000000000341] [Medline: 28291156]
49. Cajita MI, Hodgson NA, Budhathoki C, Han H. Intention to Use mHealth in Older Adults With Heart Failure. J Cardiovasc Nurs 2017;32(6):E1-E7 [FREE Full text] [doi: 10.1097/JCN.0000000000000401] [Medline: 28248747]
50. LaMonica HM, English A, Hickie IB, Ip J, Ireland C, West S, et al. Examining Internet and eHealth Practices and Preferences: Survey Study of Australian Older Adults With Subjective Memory Complaints, Mild Cognitive Impairment, or Dementia. J Med Internet Res 2017 Oct 25;19(10):e358 [FREE Full text] [doi: 10.2196/jmir.7981] [Medline: 29070481]
51. Bahar-Fuchs A, Webb S, Bartsch L, Clare L, Rebok G, Cherbuin N, et al. Tailored and Adaptive Computerized Cognitive Training in Older Adults at Risk for Dementia: A Randomized Controlled Trial. J Alzheimers Dis 2017;60(3):889-911. [doi: 10.3233/JAD-170404] [Medline: 28922158]
52. Nahm E, Resnick B, Brown C, Zhu S, Magaziner J, Bellantoni M, et al. The Effects of an Online Theory-Based Bone Health Program for Older Adults. J Appl Gerontol 2017 Sep;36(9):1117-1144 [FREE Full text] [doi: 10.1177/0733464815617284] [Medline: 26675352]
53. Knaevelsrud C, Böttche M, Pietrzak RH, Freyberger HJ, Kuwert P. Efficacy and Feasibility of a Therapist-Guided Internet-Based Intervention for Older Persons with Childhood Traumatization: A Randomized Controlled Trial. Am J Geriatr Psychiatry 2017 Aug;25(8):878-888. [doi: 10.1016/j.jagp.2017.02.024] [Medline: 28365000]
54. Reijnders JS, Geusgens CA, Ponds RW, van Boxtel MP. "Keep your brain fit!" Effectiveness of a psychoeducational intervention on cognitive functioning in healthy adults: A randomised controlled trial. Neuropsychol Rehabil 2017 Jun;27(4):455-471. [doi: 10.1080/09602011.2015.1090458] [Medline: 26414279]
55. Mageroski A, Alsadoon A, Prasad P, Pham L, Elchouemi A. Impact of wireless communications technologies on elder people healthcare: Smart home in Australia. 2016 Presented at: 13th International Joint Conference on Compouter Science and Software Engineering (JCSSE); 2016; Khon Kaen, Thailand p. 1-6. [doi: 10.1109/jcsse.2016.7748862]
56. Hamblin K, Yeandle S, Fry G. Researching telecare: the importance of context. Journal of Enabling Technologies 2017 Sep 18;11(3):75-84. [doi: 10.1108/jet-04-2017-0016]
57. Wang J, Carroll D, Peck M, Myneni S, Gong Y. Mobile and Wearable Technology Needs for Aging in Place: Perspectives from Older Adults and Their Caregivers and Providers. Stud Health Technol Inform 2016;225:486-490. [Medline: 27332248]
58. Gordon NP, Hornbrook MC. Differences in Access to and Preferences for Using Patient Portals and Other eHealth Technologies Based on Race, Ethnicity, and Age: A Database and Survey Study of Seniors in a Large Health Plan. J Med Internet Res 2016 Mar 04;18(3):e50 [FREE Full text] [doi: 10.2196/jmir.5105] [Medline: 26944212]
59. Williams K, Pennathur P, Bossen A, Gloeckner A. Adapting Telemonitoring Technology Use for Older Adults: A Pilot Study. Res Gerontol Nurs 2016;9(1):17-23 [FREE Full text] [doi: 10.3928/19404921-20150522-01] [Medline: 26020575]
60. Müller AM, Khoo S, Morris T. Text Messaging for Exercise Promotion in Older Adults From an Upper-Middle-Income Country: Randomized Controlled Trial. J Med Internet Res 2016 Jan 07;18(1):e5 [FREE Full text] [doi: 10.2196/jmir.5235] [Medline: 26742999]
61. Quinn CC, Shardell MD, Terrin ML, Barr EA, Park D, Shaikh F, et al. Mobile Diabetes Intervention for Glycemic Control in 45- to 64-Year-Old Persons With Type 2 Diabetes. J Appl Gerontol 2016 Feb;35(2):227-243. [doi: 10.1177/0733464814542611] [Medline: 25098253]
62. Royackers A, Regan S, Donelle L. The eShift model of care: informal caregivers' experience of a new model of home-based palliative care. Progress in Palliative Care 2016 Apr 13;24(2):84-92. [doi: 10.1179/1743291x15y.0000000006]
63. Duh ES, Guna J, Pogačnik M, Sodnik J. Applications of Paper and Interactive Prototypes in Designing Telecare Services for Older Adults. J Med Syst 2016 Apr;40(4):92. [doi: 10.1007/s10916-016-0463-z] [Medline: 26860915]
64. Depatie A, Bigbee JL. Rural Older Adult Readiness to Adopt Mobile Health Technology: A Descriptive Study. OJRNHC 2015 May 29;15(1):150-184. [doi: 10.14574/ojrnhc.v15i1.346]
65. Moore AN, Rothpletz AM, Preminger JE. The Effect of Chronological Age on the Acceptance of Internet-Based Hearing Health Care. Am J Audiol 2015 Sep;24(3):280-283. [doi: 10.1044/2015_AJA-14-0082] [Medline: 26649530]
66. Currie M, Philip LJ, Roberts A. Attitudes towards the use and acceptance of eHealth technologies: a case study of older adults living with chronic pain and implications for rural healthcare. BMC Health Serv Res 2015 Apr 16;15:162 [FREE Full text] [doi: 10.1186/s12913-015-0825-0] [Medline: 25888988]
67. Grant LA, Rockwood T, Stennes L. Client Satisfaction with Telehealth in Assisted Living and Homecare. Telemed J E Health 2015 Dec;21(12):987-991. [doi: 10.1089/tmj.2014.0218] [Medline: 26126079]
68. Brenes GA, Danhauer SC, Lyles MF, Hogan PE, Miller ME. Telephone-Delivered Cognitive Behavioral Therapy and Telephone-Delivered Nondirective Supportive Therapy for Rural Older Adults With Generalized Anxiety Disorder: A

Randomized Clinical Trial. JAMA Psychiatry 2015 Oct;72(10):1012-1020 [FREE Full text] [doi: 10.1001/jamapsychiatry.2015.1154] [Medline: 26244854]

69. Corbett A, Owen A, Hampshire A, Grahn J, Stenton R, Dajani S, et al. The Effect of an Online Cognitive Training Package in Healthy Older Adults: An Online Randomized Controlled Trial. J Am Med Dir Assoc 2015 Nov 01;16(11):990-997. [doi: 10.1016/j.jamda.2015.06.014] [Medline: 26543007]
70. Mavandadi S, Benson A, DiFilippo S, Streim JE, Oslin D. A Telephone-Based Program to Provide Symptom Monitoring Alone vs Symptom Monitoring Plus Care Management for Late-Life Depression and Anxiety: A Randomized Clinical Trial. JAMA Psychiatry 2015 Dec;72(12):1211-1218. [doi: 10.1001/jamapsychiatry.2015.2157] [Medline: 26558530]
71. Egede LE, Acierno R, Knapp RG, Lejuez C, Hernandez-Tejada M, Payne EH, et al. Psychotherapy for depression in older veterans via telemedicine: a randomised, open-label, non-inferiority trial. Lancet Psychiatry 2015 Aug;2(8):693-701. [doi: 10.1016/S2215-0366(15)00122-4] [Medline: 26249300]
72. Chang W, Hou CJ, Wei S, Tsai J, Chen Y, Chen M, et al. Utilization and Clinical Feasibility of a Handheld Remote Electrocardiography Recording Device in Cardiac Arrhythmias and Atrial Fibrillation: A Pilot Study. International Journal of Gerontology 2015 Dec;9(4):206-210. [doi: 10.1016/j.ijge.2015.06.002]
73. Boulos M, Ifeachor E, Zhao P, Escudero J, Carroll C, Costa P, et al. LiveWell-Promoting Healthy Living and Wellbeing for Parkinson Patients through Social Network and ICT Training: Lessons Learnt and Best Practices. International Journal of Healthcare Information Systems and Informatics 2015;10(3):24-41. [doi: 10.4018/IJHISI.2015070102]
74. Diño MJ, de Guzman AB. Using Partial Least Squares (PLS) in Predicting Behavioral Intention for Telehealth Use among Filipino Elderly. Educational Gerontology 2014 Jun 25;41(1):53-68. [doi: 10.1080/03601277.2014.917236]
75. Czaja SJ, Lee CC, Arana N, Nair SN, Sharit J. Use of a telehealth system by older adults with hypertension. J Telemed Telecare 2014 Jun;20(4):184-191 [FREE Full text] [doi: 10.1177/1357633X14533889] [Medline: 24803275]
76. Choi NG, Wilson NL, Sirrianni L, Marinucci ML, Hegel MT. Acceptance of home-based telehealth problem-solving therapy for depressed, low-income homebound older adults: qualitative interviews with the participants and aging-service case managers. Gerontologist 2014 Aug;54(4):704-713 [FREE Full text] [doi: 10.1093/geront/gnt083] [Medline: 23929664]
77. Chi N, Demiris G. A systematic review of telehealth tools and interventions to support family caregivers. J Telemed Telecare 2015 Jan;21(1):37-44 [FREE Full text] [doi: 10.1177/1357633X14562734] [Medline: 25475220]
78. Knowles B, Hanson V. Older Adults' Deployment of 'Distrust'. ACM Trans. Comput.-Hum. Interact 2018 Sep 15;25(4):1-25. [doi: 10.1145/3196490]

Abbreviations

JHNEBP: John Hopkins Nursing Evidence-Based Practice
MeSH: Medical Subject Headings
mHealth: mobile health
RCT: randomized controlled trial
WHO: World Health Organization

Edited by J Hefner, G Eysenbach; submitted 17.05.20; peer-reviewed by S MacEwan, J Portz; comments to author 08.06.20; revised version received 11.06.20; accepted 06.07.20; published 12.08.20.

Please cite as:
Kruse C, Fohn J, Wilson N, Nunez Patlan E, Zipp S, Mileski M
Utilization Barriers and Medical Outcomes Commensurate With the Use of Telehealth Among Older Adults: Systematic Review
JMIR Med Inform 2020;8(8):e20359
URL: http://medinform.jmir.org/2020/8/e20359/
doi:10.2196/20359
PMID:32784177

Review

Effects of Computerized Decision Support Systems on Practitioner Performance and Patient Outcomes: Systematic Review

Clemens Scott Kruse[1*], MHA, MBA, MSIT, PhD; Nolan Ehrbar[1*], BHA

School of Health Administration, Texas State University, San Marcos, TX, United States

[*]all authors contributed equally

Corresponding Author:
Clemens Scott Kruse, MHA, MBA, MSIT, PhD
School of Health Administration
Texas State University
601 University Dr
San Marcos, TX,
United States
Phone: 1 512 245 4462
Fax: 1 512 245 8712
Email: scottkruse@txstate.edu

Abstract

Background: Computerized decision support systems (CDSSs) are software programs that support the decision making of practitioners and other staff. Other reviews have analyzed the relationship between CDSSs, practitioner performance, and patient outcomes. These reviews reported positive practitioner performance in over half the articles analyzed, but very little information was found for patient outcomes.

Objective: The purpose of this review was to analyze the relationship between CDSSs, practitioner performance, and patient medical outcomes. PubMed, CINAHL, Embase, Web of Science, and Cochrane databases were queried.

Methods: Articles were chosen based on year published (last 10 years), high quality, peer-reviewed sources, and discussion of the relationship between the use of CDSS as an intervention and links to practitioner performance or patient outcomes. Reviewers used an Excel spreadsheet (Microsoft Corporation) to collect information on the relationship between CDSSs and practitioner performance or patient outcomes. Reviewers also collected observations of participants, intervention, comparison with control group, outcomes, and study design (PICOS) along with those showing implicit bias. Articles were analyzed by multiple reviewers following the Kruse protocol for systematic reviews. Data were organized into multiple tables for analysis and reporting.

Results: Themes were identified for both practitioner performance (n=38) and medical outcomes (n=36). A total of 66% (25/38) of articles had occurrences of positive practitioner performance, 13% (5/38) found no difference in practitioner performance, and 21% (8/38) did not report or discuss practitioner performance. Zero articles reported negative practitioner performance. A total of 61% (22/36) of articles had occurrences of positive patient medical outcomes, 8% (3/36) found no statistically significant difference in medical outcomes between intervention and control groups, and 31% (11/36) did not report or discuss medical outcomes. Zero articles found negative patient medical outcomes attributed to using CDSSs.

Conclusions: Results of this review are commensurate with previous reviews with similar objectives, but unlike these reviews we found a high level of reporting of positive effects on patient medical outcomes.

(JMIR Med Inform 2020;8(8):e17283) doi:10.2196/17283

KEYWORDS

CDSS; performance; outcomes

Introduction

Rationale

Computerized decision support systems (CDSSs) are software programs that support the decision making of patients, practitioners, and staff with knowledge and person-specific information. CDSSs present several tools and alerts to enhance the decision-making process within the clinical workflow [1]. Knowledge-based CDSSs were the earliest classes of CDSSs using a data repository to draw conclusions. Knowledge-based systems use traditional computing methods giving programmed results. Non–knowledge-based CDSSs are the most common forms used today. These systems use artificial intelligence (AI) assistance to augment clinical decisions made at the point of care. AI-supported CDSSs use patient data to analyze relationships between symptoms, treatments, and patient outcomes to make clinical decisions. These patient data are usually derived from electronic health records (EHRs): digital forms of patient records that include patient information such as personal contact information, patient's medical history, allergies, test results, and treatment plan [2]. Artificial intelligence, software, or algorithms able to perform tasks that normally require human intelligence are integrated into CDSS processes. Data mining, a process usually assisted by AI, is often used by CDSSs to identify new data patterns from large data sets (like patient EHRs) [3]. The conclusions reached by AI used for data mining can be used by both non–knowledge-based CDSSs and knowledge-based CDSSs [3]. CDSSs are integrated into technologies such as computerized physician order entry (CPOE) [4] tools and electronic medical record (EMR)/EHR databases and use a wide variety of drug, patient, and treatment data and more to make clinical decisions that provide the best recommendations for treatment. CDSS utility varies widely, drawing conclusions about different ailments, disorders, and syndromes. Prospects for this technology may employ patient preferences or financial capabilities.

In prior studies, CDSSs have been shown to improve practitioner performance, but the effects on patient outcomes were inconsistent and required further study. A review conducted in 1998 evaluated studies for the previous 5 years and found a benefit to physician performance in 66% of studies analyzed (n=65), but only 14 of those analyzed discussed outcomes, so no conclusions were made [5]. The review was repeated in 2005 with a larger sample (n=100) and found a positive impact on physician performance in 64% of studies analyzed, but like the 1998 review, effects on patient outcomes were insufficient to make generalizations [6]. In 2010, a research protocol was registered to repeat the review, but no publication followed. In 2011, the review was repeated with a similar size of articles analyzed (n=91) and identified a positive effect of CDSSs on practitioner performance for 57% of articles analyzed; however, consistent with previous reviews, no conclusions could be made concerning patient outcomes [7].

Since the last publication on this topic in 2011, CDSSs have seen significant industry growth, becoming more accessible, cost-effective, and reliable and possessing greater computational power [8]. In addition to hardware improvements, the inclusion of software such as artificial intelligence (AI) programs is growing rapidly in CDSSs, but as of yet these improvements have not been systematically reviewed to determine any impacts they might have on patient outcomes and practitioner performance.

Objective

The purpose of this systematic review is to conduct a similar review to those from 1998 and 2005 to analyze the association between CDSSs, practitioner performance, and patient outcomes. The methods used in the 2010 manuscript were never published, and those used in the 2011 review were significantly different than those in 1998 and 2005. The taxonomy of CDSSs has changed greatly since 1998, so search terms used 23 years ago will not be relevant today. CDSS employment is rapidly growing, especially with increased access to CDSS AI-supported software. Because the effects are understudied, our goal is to review the effectiveness of CDSS technologies, their employment, and their overall utility.

Methods

Protocol Registration and Eligibility Criteria

This review was not registered. The methods followed a technique of sharing workload from the Assessment of Multiple Systematic Reviews (AMSTAR) [9]. The format of the review uses the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [10]. Conceptualization of the overall review, including standardized data extraction tools, follows the Kruse protocol for writing systematic reviews in a health-related program [11]. Articles were eligible for inclusion if they were published in the English language within the last 10 years, had full text available, and reported on the elements of the objective statement: measures of effectiveness of CDSSs on practitioner performance or patient outcomes. A 10-year window was justified because we wanted the research to be current, and this exceeds the window of the 1998 and 2005 reviews, which used only 5 years. At first, we limited the search to studies in peer-reviewed journals, but because our sample was too small, we expanded the search to include grey literature. However, we limited our choices to use only those that had results.

Information Sources

Five common research databases were queried: PubMed (the web-based components of MEDLINE, life science journals, and online books), CINAHL, Embase, Web of Science, and Cochrane (reviews, controlled trials, methodologies, and health technology assessments). Searches were conducted from January 29 to January 31, 2020. Databases were chosen at the recommendation of the National Institutes of Health, which recommends at least three databases: PubMed, Embase, and Cochrane [12]. This practice also follows established practice in published systematic reviews [11].

Search and Study Selection

Searches in each database were identical: ("Clinical decision support systems" OR "computerized provider order entry" OR "diagnosis, computer assisted" OR "drug therapy, computer-assisted" OR "expert systems") AND ("patient reported outcomes" OR "practitioner performance"). Embase and Web of Science do not allow Boolean searches, so an advanced search was used. Articles were eligible for inclusion if they were published in the last 10 years and discussed both CDSSs and either practitioner performance or patient-reported outcomes. We excluded reviews. In CINAHL, we excluded MEDLINE to avoid duplication with the results from PubMed.

The search strings for the 1998 and 2005 reviews were not available, but the search string for the 2011 study was available: (literature review[tiab] OR critical appraisal[tiab] OR meta analysis[pt] OR systematic review[tw] OR medline[tw]) AND (medical order entry systems[mh] OR medical order entry system*[tiab] OR computerized order entry[tiab] OR computerized prescriber order entry[tiab] OR computerized provider order entry[tiab] OR computerized physician order entry[tiab] OR electronic order entry[tiab] OR electronic prescribing[mh] OR electronic prescribing[tiab] OR cpoe[tiab] OR drug therapy, computer assisted[mh] OR computer assisted drug therapy[tiab] OR decision support systems, clinical[mh] OR decision support system*[tiab] OR reminder system*[tiab] OR decision making, computer assisted[mh] OR computer assisted decision making [tiab] OR diagnosis, computer assisted[mh] OR computer assisted diagnosis[tiab] OR therapy, computer assisted[mh] OR computer assisted therapy[tiab] OR expert systems[mh] OR expert system*[tiab]). It is important to note the limited terms used for CDSSs also included lesser known terms indexed by PubMed's Medical Subject Headings: clinical decision support; clinical decision supports; decision support, clinical; support, clinical decision; supports, clinical decision; decision support, clinical; and decision support systems, clinical. Searching for CPOE also included order entry systems, medical; medication alert systems; alert system, medication; medication alert system; system, medication alert; alert systems, medication; computerized physician order entry system; CPOE; computerized provider order entry; and computerized physician order entry. Searching for diagnosis, computer assisted also included the following: computer-assisted diagnosis; computer assisted diagnosis; computer-assisted diagnoses; and diagnoses, computer assisted. Searching for drug therapy included the following: drug therapy, computer assisted; therapy, computer-assisted drug; computer-assisted drug therapies; drug therapies, computer-assisted; therapies, computer-assisted drug; therapy, computer assisted drug; computer-assisted drug therapy; computer assisted drug therapy; protocol drug therapy, computer-assisted; and protocol drug therapy, computer assisted. A search of expert systems also included expert system; system, expert; and systems, expert.

Abstracts were independently screened by each reviewer, and a consensus meeting was called to discuss disagreement. A kappa score was calculated to provide a measure of agreement between reviewers.

Data Collection and Data Items

A standardized Excel spreadsheet (Microsoft Corporation) was used as a data extraction tool, in accordance with the Kruse protocol [11]. This tool acted as a template for reviewers to collect study design, participants, sample size, intervention, observed bias, and effect size, where applicable. A literature matrix was created to list and organize all articles, extract data between multiple reviewers, and discuss observations in consensus meetings. Three consensus meetings were held for reviewers to discuss disagreement and share observations. This practice created a synergy effect and ensured everyone progressed with a like mind.

Risk of Bias in Individual Studies

Reviewers noted any observation of bias. We used the Johns Hopkins Nursing Evidence-Based Practice (JHNEBP) tool as a quality assessment of studies analyzed. Other forms of bias were noted as well, which are described in risk of bias across studies.

Synthesis of Results

The Excel spreadsheet was used to synthesize our observations and data collected. The spreadsheet enabled a narrative analysis which identified themes, as is the practice in multiple disciplines. We did not combine results of studies because this was not a meta-analysis.

Risk of Bias Across Studies

Additional forms of bias other than selection bias were noted on the spreadsheet such as localized studies or surveillance bias.

Additional Analysis

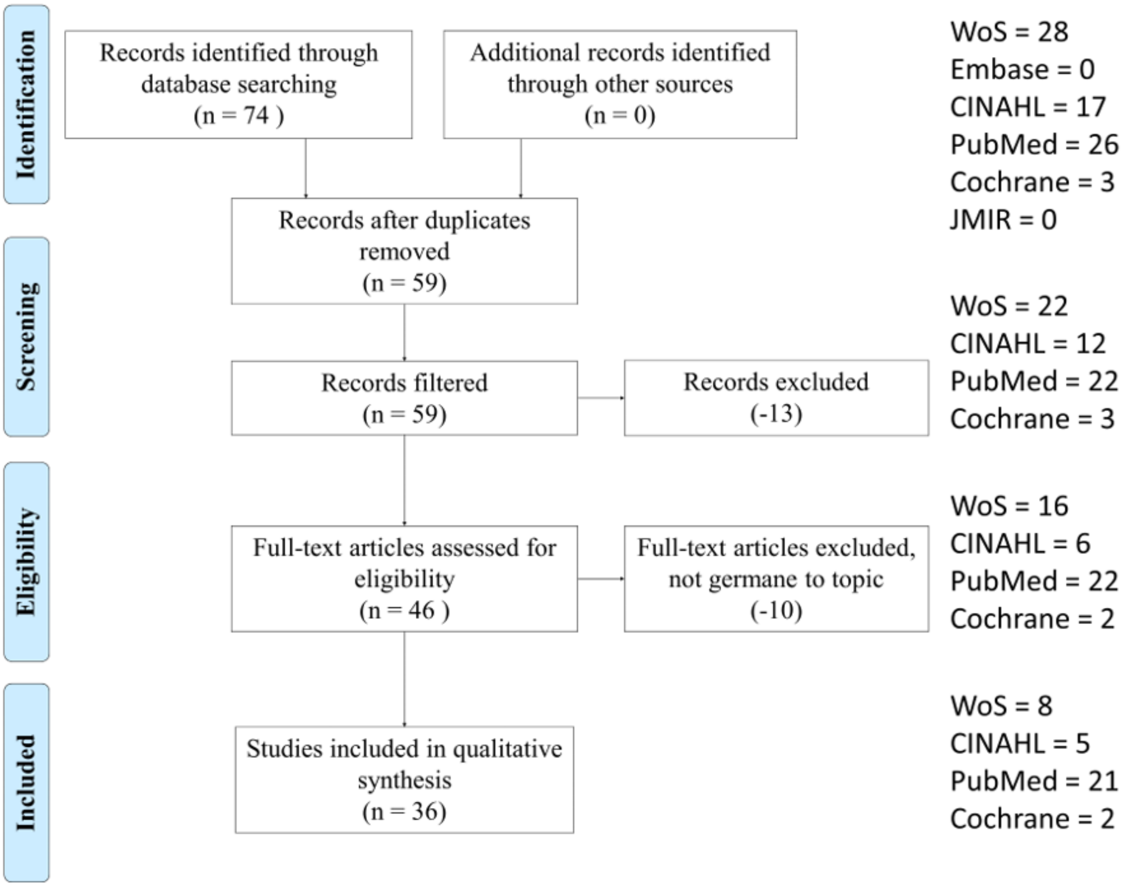
Reviewers read each article two times [11]. During the second reading, reviewers made independent notes of major themes related to the objective, using the Excel data extraction tool. After a third consensus meeting debriefing the observations and themes, detailed notes were formulated about health policy implications of telemedicine. Frequency of occurrence of each of the major common themes was captured in affinity matrices for further analysis. Data and calculations are available upon request.

Results

Study Selection and Study Characteristics

The study selection process is illustrated in Figure 1. The 74 results from the search string in five databases were placed into an Excel spreadsheet and shared among reviewers for selection and analysis. Filters were applied in each database to capture only the last 10 years (January 30, 2011, to January 30, 2020). Reviewers independently removed duplicates and screened abstracts. A statistic of agreement, kappa, was calculated. The kappa score produced was .98, showing almost complete agreement on all reviewed articles [13,14]. The remaining 36 results were read in full for relevance. Observations for the 36 articles that remained were placed in an Excel spreadsheet for independent data analysis.

Figure 1. Article selection process with selection criteria.


Identification
Records identified through database searching (n = 74)
Additional records identified through other sources (n = 0)
WoS = 28
Embase = 0
CINAHL = 17
PubMed = 26
Cochrane = 3
JMIR = 0
Screening
Records after duplicates removed (n = 59)
Records filtered (n = 59)
Records excluded (-13)
WoS = 22
CINAHL = 12
PubMed = 22
Cochrane = 3
Eligibility
Full-text articles assessed for eligibility (n = 46)
Full-text articles excluded, not germane to topic (-10)
WoS = 16
CINAHL = 6
PubMed = 22
Cochrane = 2
Included
Studies included in qualitative synthesis (n = 36)
WoS = 8
CINAHL = 5
PubMed = 21
Cochrane = 2

Reviewers collected standard patient/participants, intervention, comparison, outcome, study design (PICOS) observations plus indications of either practitioner performance or patient medical outcomes (Multimedia Appendix 1). Bias was also noted. Following the Kruse protocol, observations were distilled into themes for further analysis. Three consensus meetings were used to discuss disagreement. A summary of all observations is listed in Table 1. Articles are listed in reverse chronological order. The details extracted were year of publication, authors, title, study design, participants, sample size, intervention, bias, and observations about barriers or facilitators to the adoption of telemedicine.

Risk of Bias Within Studies

Bias was not observed in all studies analyzed. A full review of the bias observed is provided in Multimedia Appendix 1. The JHNEBP tool found no quality measure below Level IV or C.

Results of Individual Studies

General observations and thematic analysis are listed in Table 1. Articles are listed in reverse chronological order. A table of PICOS is provided in Multimedia Appendix 1.

Table 1. Summary of analysis.

Authors	Efficiency (practitioner performance)	Efficiency themes	Effectiveness (medical outcomes)	Effectiveness themes
Grout et al [15]	Practitioner performance not discussed	Not reported or discussed	Self-reporting by adolescents increased (doubled) by 19.3 percentage points	Improved screening
Connelly et al [16]	Number of prescriptions written for migraines increased significantly; average length of time per use of tool was 3.3 minutes	More accurate prescribing	Medical outcomes not reported or discussed	Not reported or discussed
Salz et al [17]	Facilitated a more comprehensive visit	Improved care plans	Provided a way for patients to chronicle other physicians who had been involved in medical decisions enabling doctors to communicate	Improved feedback
Kirby et al [18]	Increased referral awareness by providers for patients with severe aortic stenosis (which is a known quality issue); increase in referral rate from 72% to 98%	Increased awareness	Medical outcomes not reported or discussed	Not reported or discussed
Dolan and Veazie [19]	Practitioner performance not statistically different	No difference reported	Medical outcomes not reported or discussed	Not reported or discussed
Jackson and De Cruz [20]	Practitioner performance not discussed	Not reported or discussed	Improvement in relapse duration, medication adherence, cost, and number of clinic visits	Improved symptoms
Caballero-Ruiz et al [21]	Face-to-face visits with patients reduced by 89% time devoted by clinicians to patient evaluation was reduced by 27%; automatic detection of 100% of patients who needed insulin therapy	Improved performance	Diet prescriptions provided without clinician intervention; patients were very pleased with the tool	Improved disease management
Raj et al [22]	System did not improve pain intensity, therefore no significant differences in dose of opiates compared with control; had no effect on practitioner performance	No difference reported	Did not improve or worsen pain management	No difference reported
Mooney et al [23]	Enabled providers to follow up based on feedback from patients	Better follow-up with patients	Intervention group demonstrated fewer severe and moderate symptoms	Improved symptoms
Baypinar et al [24]	Correct prescribing increased from 54% to 91% (*P*<.01) for folic acid and 11% to 40% (*P*<.001) for vitamin D, and stopped orders increased from 3% to 14% (*P*<.002)	More accurate prescribing	Medical outcomes not reported or discussed	Not reported or discussed
Zini et al [25]	Practitioners improved prevention, diagnosis, and treatment	Improved care plans	Medical outcomes not reported or discussed	Not reported or discussed
Muro et al [26]	Practitioner performance not discussed	Not reported or discussed	Improved symptoms; decreased adverse events	Improved symptoms
Kistler et al [27]	Practitioner performance not discussed	Not reported or discussed	More patients agreed to screening in the intervention group than the control	Improved screening
Lawes and Grissinger [28]	Practitioners performed worse when CDSS[a] was not available or when incorrect data were entered for weight	More accurate prescribing	Adverse drug events no doubt occurred because of error, but no outcomes were discussed	Not reported or discussed
Kouladjian et al [29]	Average time to complete task to recognize sedative and anticholinergic medicines in practice was 7:20 (SD 1:45) minutes	Improved performance	Medical outcomes not reported or discussed	Not reported or discussed
Norton et al [30]	Surgeons rated the tool very useful or moderately useful (25%), neutral (47%), or moderately useless or not useful (28%)	Improved care plans	Medical outcomes not reported or discussed	Not reported or discussed
Pombo et al [31]	Resolving missing data	Improved documentation	Resolving missing data in daily diary improved the feedback loop to the pain manager	Improved feedback
Cox and Pieper [32]	Practitioner performance not discussed	Not reported or discussed	Treatment in the doxazosin arm was stopped early due to a 1.25-fold increase in the incidence of CVD[b] and a 2-fold increase in the incidence of heart failure compared with the diuretic arm	Improved efficacy

Authors	Efficiency (practitioner performance)	Efficiency themes	Effectiveness (medical outcomes)	Effectiveness themes
Schneider et al [33]	Once CDSS scored significantly more exams as appropriate; better interface of one CDSS versus the other influenced provider willingness to use the CDS system	Improved screening; improved buy-in of CDSSs	Medical outcomes not reported or discussed	Not reported or discussed
Zhu and Cimino [34]	Accuracy improved: reduced inaccuracy	Improved accuracy and performance	Improved patient safety	Improved safety
Utidjian et al [35]	Proportions of doses administered declined during the baseline seasons (from 72% to 62%) with partial recovery to 68% during the intervention season; palivizumab-focused group improved by 19.2 percentage points in the intervention season compared with the prior baseline season (*P*<.001), while the comprehensive intervention group only improved 5.5 percentage points (*P*=.29); difference in change between study groups was significant (*P*=.05)	More accurate prescribing	A quality improvement initiative supported by CDS and workflow tools integrated in the EHR[c] improved recognition of eligibility and may have increased palivizumab administration rates; palivizumab-focused group performed significantly better than a comprehensive intervention	Improved disease management
Semler et al [36]	No statistically significant difference in performance (also low use of tool)	No difference reported	No statistically significant difference: mortality 14% versus 15%, ICU[d]-free days 17 versus 19, vasopressor-free days 22.2 versus 22.6	No difference reported
Peiris et al [37]	Patients more likely to receive screening with CDSS (63% vs 53%); no improvements in prescription of recommended medications at the end of the study	Improved screening	Improved cardiovascular disease risk management; no difference in prescription rates	Improved disease management
Chow et al [38]	Only one-quarter of patients received antibiotics despite recommendations of CDSSs	More accurate prescribing	Patients aged <65 years had greater mortality benefit (OR[e] 0.45, 95% CI 0.20-1.00; *P*=.05) than patients >65 years (OR 1.28, 95% CI 0.91-1.82; *P*=.16); no effect was observed on incidence of *Clostridium difficile* (OR 1.02, 95% CI 0.34-3.01) and multidrug-resistant organism (OR 1.06, 95% CI 0.42-2.71) infections; no increase in infection-related readmission (OR 1.16, 95% CI 0.48-2.79) was found in survivors; receipt of CDSS-recommended antibiotics reduced mortality risk in patients aged ≤65 years and did not increase risk in older patients	No difference reported
Wilson et al [39]	Practitioner performance not discussed	Not reported or discussed	Improved self-efficacy and decreased fecal aversion	Improved efficacy
Loeb et al [40]	Training greatly improved documentation	Improved documentation	Medical outcomes not reported or discussed	Not reported or discussed
Mishuris et al [41]	Practitioner performance not discussed	Not reported or discussed	Patients who visited clinics missing at least one of the CDSS functions were more likely to have controlled blood pressure (86% vs 82%; OR 1.3, 95% CI 1.1-1.5) and more likely to not have adverse drug event visits (99.9% vs 99.8%; OR 3.0, 95% CI 1.3-7.3)	Improved symptoms
Dexheimer et al [42]	No difference in time to disposition decision; no change in hospital admission rate; no difference in ED[f] length of stay	No difference reported	CDSS supported communication between patient and provider	Improved feedback
Heisler et al [43]	Practitioner performance not discussed	Not reported or discussed	Decrease in diabetes distress, but no difference in other outcomes	Improved symptoms
Eckman et al [44]	Decisions are based on >0.1 QALYs[g]; tool identified the 50% who would benefit from this threshold	Improved performance	Significant gain in quality-adjusted life expectancy	Improved mortality
Zaslansky et al [45]	Audit, feedback, and benchmarking provided to practitioners to identify when their practice is not in line with data	Improved benchmarking	Provides real-time feedback on PROs[h]	Improved feedback

Authors	Efficiency (practitioner performance)	Efficiency themes	Effectiveness (medical outcomes)	Effectiveness themes
Lobach et al [46]	No treatment-related differences between groups	No difference reported	Among patients <18 years, those in the email group had fewer low severity (7.6 vs 10.6/100 enrollees; *P*<.001) and total ED encounters (18.3 vs 23.5/100 enrollees; *P*<.001) and lower ED ($63 vs $89, *P*=.002) and total medical costs ($1736 vs $2207, *P*=.009); patients who were ≥18 years in the latter group had greater outpatient medical costs	Improved symptoms
Barlow and Krassas [47]	Annual cycle of care plans increased by 12%	Improved care plans	Patients better able to meet targets for microalbumin; glycemic control well managed	Improved symptoms
Robbins et al [48]	A total of 90% of providers involved with the RCT[i] supported adopting the intervention	Improved buy-in of CDSSs	Increased CD4+ lymphocyte count and reduced suboptimal follow-up appointment	Improved symptoms
Chen et al [49]	New CDSS identified 70 records needing reassessment of triglyceride level	Improved screening	Medical outcomes not discussed	Not reported or discussed
Seow et al [50]	A total of 87% of respondents strongly agreed or somewhat agreed that the “ESAS[j] was important to complete because it helped the health care team to know what symptoms [they] were having and how severe they were”	Improved screening	A total of 79% of respondents rated that their “pain and other symptoms have been controlled to a comfortable level” always or most of the time compared with 8% of respondents who rated this as rarely or never occurring	Improved symptoms

[a] CDSS: computerized decision support system.

[b] CVD: cardiovascular disease.

[c] EHR: electronic health record.

[d] ICU: intensive care unit.

[e] OR: odds ratio.

[f] ED: emergency department.

[g] QALY: quality-adjusted life year.

[h] PRO: patient-reported outcome.

[i] RCT: randomized controlled trial.

[j] ESAS: Edmonton Symptom Assessment System.

Risk of Bias Across Studies

Multimedia Appendix 1 provides a table of PICOS and bias. Outcomes are reported in Table 1. Bias was similar across articles reviewed: most research took place in one facility, organization, or state, which is a form of selection bias and limits the broad application of results. A sample taken from a limited geographic area is inherently limited in its ability to generalize results to the general population unless steps have been taken to ensure the sample is representative of the population.

Additional Analysis

Twelve themes were identified for practitioner performance, two of which were no difference and not discussed. These themes are listed in Table 2 in order of occurrence first for positive effect followed by no difference and not discussed.

Table 2. Summary of themes identified for practitioner performance (n=38).

Efficiency themes	Occurences	Incidence, n (%)
More accurate prescribing	16,24,28,35,38	5 (13)
Improved screening	33,37,49,50	4 (11)
Improved performance	21,29,34,44	4 (11)
Improved care plans	17,25,30,47	4 (11)
Improved documentation	31,40	2 (5)
Improved buy-in of CDSSs[a]	33,48	2 (5)
Increased awareness	18	1 (3)
Better follow-up with patients	23	1 (3)
Improved accuracy	34	1 (3)
Improved benchmarking	45	1 (3)
No difference reported	19,22,36,42,46	5 (13)
Not reported or discussed	15,20,26,27,32,39,41,43	8 (21)

[a]CDSS: computerized decision support system.

As illustrated, 66% (25/38) of the occurrences of themes identified 10 positive indicators of practitioner performance [16-18,21,23-25,28-31,33-35,37,38,40,44,45,47-50]. Practitioner performance was reported as more accurate prescribing, improved screening of patients, improved overall performance, increased awareness of patient conditions, improved follow-up due to better communication with patients, improved accuracy of diagnosis, improved documentation, improved benchmarking, improved care plans, and improved buy-in of CDSSs. A total of 21% (8/38) of articles did not discuss practitioner performance [15,20,26,27,32,39,41,43].

Practitioners using CDSSs experienced more accurate prescribing [16,24,28,35,38], improved screening [33,37,49,50], improved overall performance [21,29,34,44], improved care plans [17,25,30,47], improved documentation [31,40], overall improved buy-in for CDSSs [33,48], increased awareness of needs of patients [18], improved follow-up with patients due to enhanced communication channels enabled by the application [23], improved accuracy of diagnosis [34], and improved benchmarking [45].

Nine themes were identified for patient medical outcomes, two of which were no difference and not discussed. These themes are listed in Table 3 by order of greatest occurrence for positive effect followed by no difference and not discussed.

Table 3. Summary of themes identified for patient medical outcomes (n=36).

Effectiveness themes	Occurences	Incidence, n (%)
Improved symptoms	20,23,26,41,43,46-48,50	9 (25)
Improved feedback	17,31,42,45	4 (11)
Improved disease management	21,35,37	3 (8)
Improved efficacy	32,39	2 (6)
Improved screening	15,27	2 (6)
Improved safety	34	1 (3)
Improved mortality	44	1 (3)
No difference reported	22,36,38	3 (8)
Not reported or discussed	16,18,19,24,25,28-30,33,40,49	11 (31)

As illustrated, 61% (22/36) of occurrences of themes identified 7 positive patient medical outcomes as a result of using CDSSs [15,17,20,21,23,26,27,31,32,34,35,37,39,41-48,50]. Patients experienced improved symptoms [20,23,26,41,43,46-48,50], improved feedback from provider [17,31,42,45], improved disease management [21,35,37], improved efficacy of treatment [32,39], improved screening [15,27], and improved safety [34], and one study even reported improved mortality [44]. Although 11 articles did not discuss patient medical outcomes [16,18,19,24,25,28-30,33,40,49], only 3 reported no statistically significant difference in outcomes between control and intervention groups [22,36,38].

Discussion

Summary of Evidence

Our review methodology enabled a meticulous evaluation of the efficiency and effectiveness of CDSSs for practitioner performance and medical outcomes. A summary of the findings from the review are listed in Table 1. Of the 36 articles analyzed that reported efficiency or effectiveness, 25 reported positive performance and 22 reported positive outcomes; 9 did not report practitioner performance and 11 did not report patient medical outcomes.

Commensurate with previous reviews on this topic [6,7], a majority of articles analyzed reported improvement in practitioner performance [16-18,21,23-25,28-31,33-35,37,38,40,44,45,47-50], but contrary to the previous reviews, our review found articles that reported patient outcomes, and a majority were positive outcomes [15,17,20,21,23,26,27,31,32,34,35,37,39,41-48,50]. Although 9 articles did not discuss practitioner performance [15,20,26,27,32,39,41,43], only 5 articles reported no difference in productivity [19,22,36,42,46].

The decision of whether to adopt a CDSS is one of complexity and change management. Providers and administrators need to discuss the advantages and disadvantages. The organization’s infrastructure must support the application, providers must be trained on how to implement it, and administrators must ensure that budget and organizational dynamics can afford acquisition and implementation. The literature is clear in the efficacy of CDSSs, and this should assist organizations in gaining user acceptance. Providers should carefully integrate CDSSs into their processes and clinical practice guidelines to ensure they are an asset more than a hindrance. They should be used to augment patient care rather than coming between patients and providers.

It is interesting that previous reviews did not find results of medical outcomes. This could have been a limitation in search strategy. It could also be due to the maturation of CDSSs in general. At the time the other reviews were conducted, it may have just been too soon for reviews to see the positive results in medical outcomes.

Because CDSSs present providers with knowledge-based information at the point of care, they augment decision making. Timely tools are available to providers through CDSSs that may not otherwise be available at the point of care. AI-supported recommendations provided by CDSSs analyze symptoms, possible treatments, clinical practice guidelines, and patient outcomes [1,2]. These capabilities are most likely the catalyst for improved practitioner performance and patient outcomes.

There does not appear to be one CDSS panacea for all practices, specialties, or templates. The literature is mixed on which products are best of breed systems. Clearly, additional research should continue to be conducted in this valuable area of medical practice. While other industries have fully embraced the digitized environment, health care in general has been slow to adopt, which is understandable when health is at stake. Based on the results of this review compared with similar ones in the past, CDSSs are diffusing across the health care industry as the systems improve. Further research into CDSSs should look to improve productivity and standardize their integration into clinical practice guidelines.

Another interesting note is that alert fatigue was not raised in any of the studies analyzed. Alert fatigue is a known phenomenon and worthy of note [51]. It is attributed to medical error in the areas of pharmacy and physician ordering systems, which are common attributes in CDSSs [52]. Even in clinical trials, alert fatigue is known to be persistent over time [53]. It is interesting that it was not noted, and if it was not noted, it was not controlled for in the studies analyzed.

Limitations

The small group of articles for analysis was a limitation. Only 36 articles met the selection criteria. A larger group for analysis would strengthen the external validity of the results because we could be better assured that our group is representative of the population. The effects of selection bias were reduced using multiple reviewers to screen and analyze articles [9]. Only two reviewers screened abstracts and analyzed articles for themes. One additional reviewer might have increased the number of observations. Publication bias was reduced through the inclusion of grey literature that included more than just peer-reviewed material; however, these articles were discarded if they did not include results. We considered only articles published in the English language. It is possible that additional observations could have been gained by expanding the search to other languages. This review is also limited by the techniques used in the trials analyzed, and statistics and effect sizes could not be combined due to the wide range used in the articles. We analyzed both qualitative and quantitative methods, and effect size is only viable for the latter. Sample sizes were widely different between studies analyzed, ranging from 6 to 900 million. Such a wide disparity makes consolidation of results difficult. We also did not analyze or compare the heuristics and algorithms used by CDSSs within the studies. To compensate for a limitation from a similar review in 2005, we expanded our analysis beyond randomized controlled trials to pre-post and other designs [6].

Conclusion

Overall , the research generally supports the efficiency of CDSS technologies for practitioner performance [16-18,21,23-25,28-31,33-35,37,38,40,44,45,47-50] and effectiveness in patient medical outcomes [15,17,20,21,23,26,27,31,32,34,35,37,39,41-48,50]; however, a further in-depth review of their effectiveness, in particular for aspects such as the avoidance of alert fatigue and extension of CDSS utility, is important. Decision-support tools extend beyond the practitioner to the patient, and some tools are not software-based but based on patient-reported data [46]. The implementation of CDSSs can mutually benefit the practitioner and patient, and they show great promise for health care in the future.

Authors' Contributions

NE drafted the initial introduction and discussion and analyzed the articles. CSK designed the review, served as senior editor, analyzed the articles, and wrote the final version of all sections. Special thanks to my assistant, Leah Frye, who helped organize the references.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Participants, intervention, comparison, outcome, study design table.
[DOCX File , 33 KB - medinform_v8i8e17283_app1.docx]

References

1. Clinical decision support. URL: https://www.healthit.gov/topic/safety/clinical-decision-support [accessed 2020-08-05]
2. Electronic health records. Centers for Medicare and Medicaid Services. URL: https://www.cms.gov/Medicare/E-Health/EHealthRecords/index.html [accessed 2020-08-05]
3. Han J, Kamber M. Data Mining: Concepts and Techniques. 3rd edition. New York: Elsevier; 2012.
4. Dixon B, Zafar M. Inpatient computerized provider order entry. Rockville: Agency for Health Care Research and Quality; 2009. URL: https://digital.ahrq.gov/sites/default/files/docs/page/09-0031-EF_cpoe.pdf [accessed 2020-08-05]
5. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA 1998 Oct 21;280(15):1339-1346. [doi: 10.1001/jama.280.15.1339] [Medline: 9794315]
6. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
7. Jaspers MW, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc 2011 May 01;18(3):327-334 [FREE Full text] [doi: 10.1136/amiajnl-2011-000094] [Medline: 21422100]
8. Sim I, Gorman P, Greenes RA, Haynes RB, Kaplan B, Lehmann H, et al. Clinical decision support systems for the practice of evidence-based medicine. J Am Med Inform Assoc 2001;8(6):527-534 [FREE Full text] [Medline: 11687560]
9. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017 Dec 21;358:j4008 [FREE Full text] [doi: 10.1136/bmj.j4008] [Medline: 28935701]
10. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009;151(4):264-269 [FREE Full text] [Medline: 19622551]
11. Kruse CS. Writing a systematic review for publication in a health-related degree program. JMIR Res Protoc 2019 Oct 14;8(10):e15490 [FREE Full text] [doi: 10.2196/15490] [Medline: 31527018]
12. Systematic review service. National Institutes of Health. URL: https://www.nihlibrary.nih.gov/services/systematic-review-service [accessed 2020-08-05]
13. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
14. Light RJ. Measures of response agreement for qualitative data: some generalizations and alternatives. Psychol Bull 1971;76(5):365-377. [doi: 10.1037/h0031643]
15. Grout RW, Cheng ER, Aalsma MC, Downs SM. Let them speak for themselves: improving adolescent self-report rate on pre-visit screening. Acad Pediatr 2019 Jul;19(5):581-588. [doi: 10.1016/j.acap.2019.04.010] [Medline: 31029741]
16. Connelly M, Bickel J. Primary care access to an online decision support tool is associated with improvements in some aspects of pediatric migraine care. Acad Pediatr 2019 Dec 04. [doi: 10.1016/j.acap.2019.11.017] [Medline: 31809810]
17. Salz T, Schnall RB, McCabe MS, Oeffinger KC, Corcoran S, Vickers AJ, et al. Incorporating multiple perspectives into the development of an electronic survivorship platform for head and neck cancer. JCO Clin Cancer Inform 2018 Dec;2:1-5 [FREE Full text] [doi: 10.1200/CCI.17.00105] [Medline: 30652547]
18. Kirby AM, Kruger B, Jain R, O Hair DP, Granger BB. Using clinical decision support to improve referral rates in severe symptomatic aortic stenosis: a quality improvement initiative. Comput Inform Nurs 2018 Nov;36(11):525-529. [doi: 10.1097/CIN.0000000000000471] [Medline: 30134257]
19. Dolan JG, Veazie PJ. The feasibility of sophisticated multicriteria support for clinical decisions. Med Decis Making 2018 May;38(4):465-475 [FREE Full text] [doi: 10.1177/0272989X17736769] [Medline: 29083251]
20. Jackson BD, Con D, De Cruz P. Design considerations for an eHealth decision support tool in inflammatory bowel disease self-management. Intern Med J 2018 Jun;48(6):674-681. [doi: 10.1111/imj.13677] [Medline: 29136332]

21. Caballero-Ruiz E, García-Sáez G, Rigla M, Villaplana M, Pons B, Hernando ME. A web-based clinical decision support system for gestational diabetes: automatic diet prescription and detection of insulin needs. Int J Med Inform 2017 Dec;102:35-49. [doi: 10.1016/j.ijmedinf.2017.02.014] [Medline: 28495347]
22. Raj SX, Brunelli C, Klepstad P, Kaasa S. COMBAT study—computer-based assessment and treatment: a clinical trial evaluating impact of a computerized clinical decision support tool on pain in cancer patients. Scand J Pain 2017 Oct;17:99-106. [doi: 10.1016/j.sjpain.2017.07.016] [Medline: 28850380]
23. Mooney KH, Beck SL, Wong B, Dunson W, Wujcik D, Whisenant M, et al. Automated home monitoring and management of patient-reported symptoms during chemotherapy: results of the symptom care at home RCT. Cancer Med 2017 Mar;6(3):537-546 [FREE Full text] [doi: 10.1002/cam4.1002] [Medline: 28135050]
24. Baypinar F, Kingma HJ, van der Hoeven RTM, Becker ML. Physicians' compliance with a clinical decision support system alerting during the prescribing process. J Med Syst 2017 Jun;41(6):1-6. [doi: 10.1007/s10916-017-0717-4] [Medline: 28480481]
25. Zini EM, Lanzola G, Bossi P, Quaglini S. An environment for guideline-based decision support systems for outpatients monitoring. Methods Inf Med 2017 Aug 11;56(4):283-293. [doi: 10.3414/ME16-01-0142] [Medline: 28726971]
26. Muro N, Larburu N, Bouaud J, Seroussi B. Weighting experience-based decision support on the basis of clinical outcomes' assessment. Stud Health Technol Inform 2017;244:33-37. [Medline: 29039372]
27. Kistler CE, Golin C, Morris C, Dalton AF, Harris RP, Dolor R, et al. Design of a randomized clinical trial of a colorectal cancer screening decision aid to promote appropriate screening in community-dwelling older adults. Clin Trials 2017 Dec;14(6):648-658 [FREE Full text] [doi: 10.1177/1740774517725289] [Medline: 29025270]
28. Lawes S, Grissinger M. Medication errors attributed to health information technology. Pa Patient Saf Advis 2017 Mar;14(1):1-8.
29. Kouladjian L, Gnjidic D, Chen TF, Hilmer SN. Development, validation and evaluation of an electronic pharmacological tool: the Drug Burden Index Calculator©. Res Social Adm Pharm 2016;12(6):865-875. [doi: 10.1016/j.sapharm.2015.11.002] [Medline: 26655397]
30. Norton WE, Hosokawa PW, Henderson WG, Volckmann ET, Pell J, Tomeh MG, et al. Acceptability of the decision support for safer surgery tool. Am J Surg 2015 Jun;209(6):977-984 [FREE Full text] [doi: 10.1016/j.amjsurg.2014.06.037] [Medline: 25457241]
31. Pombo N, Rebelo P, Araújo P, Viana J. Combining data imputation and statistics to design a clinical decision support system for post-operative pain monitoring. Procedia Comput Sci 2015;64:1018-1025. [doi: 10.1016/j.procs.2015.08.621]
32. Cox JL, Pieper K. Harnessing the power of real-life data. Eur Heart J Suppl 2015 Jul 10;17(suppl D):D9-D14. [doi: 10.1093/eurheartj/suv036]
33. Schneider E, Zelenka S, Grooff P, Alexa D, Bullen J, Obuchowski NA. Radiology order decision support: examination-indication appropriateness assessed using 2 electronic systems. J Am Coll Radiol 2015 Apr;12(4):349-357. [doi: 10.1016/j.jacr.2014.12.005] [Medline: 25842015]
34. Zhu X, Cimino JJ. Clinicians' evaluation of computer-assisted medication summarization of electronic medical records. Comput Biol Med 2015 Apr;59:221-231 [FREE Full text] [doi: 10.1016/j.compbiomed.2013.12.006] [Medline: 24393492]
35. Utidjian LH, Hogan A, Michel J, Localio AR, Karavite D, Song L, et al. Clinical decision support and palivizumab: a means to protect from respiratory syncytial virus. Appl Clin Inform 2015;6(4):769-784 [FREE Full text] [doi: 10.4338/ACI-2015-08-RA-0096] [Medline: 26767069]
36. Semler MW, Weavind L, Hooper MH, Rice TW, Gowda SS, Nadas A, et al. An electronic tool for the evaluation and treatment of sepsis in the ICU: a randomized controlled trial. Crit Care Med 2015 Aug;43(8):1595-1602 [FREE Full text] [doi: 10.1097/CCM.0000000000001020] [Medline: 25867906]
37. Peiris D, Usherwood T, Panaretto K, Harris M, Hunt J, Redfern J, et al. Effect of a computer-guided, quality improvement program for cardiovascular disease risk management in primary health care: the treatment of cardiovascular risk using electronic decision support cluster-randomized trial. Circ Cardiovasc Qual Outcomes 2015 Jan;8(1):87-95 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.114.001235] [Medline: 25587090]
38. Chow ALP, Lye DC, Arah OA. Mortality benefits of antibiotic computerised decision support system: modifying effects of age. Sci Rep 2015 Nov 30;5:17346 [FREE Full text] [doi: 10.1038/srep17346] [Medline: 26617195]
39. Wilson CJ, Flight IH, Turnbull D, Gregory T, Cole SR, Young GP, et al. A randomised controlled trial of personalised decision support delivered via the internet for bowel cancer screening with a faecal occult blood test: the effects of tailoring of messages according to social cognitive variables on participation. BMC Med Inform Decis Mak 2015 Apr 09;15(1):25 [FREE Full text] [doi: 10.1186/s12911-015-0147-5] [Medline: 25886492]
40. Loeb D, Sieja A, Corral J, Zehnder NG, Guiton G, Nease DE. Evaluation of the role of training in the implementation of a depression screening and treatment protocol in 2 academic outpatient internal medicine clinics utilizing the electronic medical record. Am J Med Qual 2015;30(4):359-366 [FREE Full text] [doi: 10.1177/1062860614532681] [Medline: 24829154]
41. Mishuris RG, Linder JA, Bates DW, Bitton A. Using electronic health record clinical decision support is associated with improved quality of care. Am J Manag Care 2014 Oct 01;20(10):e445-e452 [FREE Full text] [Medline: 25414982]

42. Dexheimer JW, Abramo TJ, Arnold DH, Johnson K, Shyr Y, Ye F, et al. Implementation and evaluation of an integrated computerized asthma management system in a pediatric emergency department: a randomized clinical trial. Int J Med Inform 2014 Nov;83(11):805-813 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.07.008] [Medline: 25174321]
43. Heisler M, Choi H, Palmisano G, Mase R, Richardson C, Fagerlin A, et al. Comparison of community health worker-led diabetes medication decision-making support for low-income Latino and African American adults with diabetes using e-health tools versus print materials: a randomized, controlled trial. Ann Intern Med 2014 Nov 18;161(10 Suppl):S13-S22 [FREE Full text] [doi: 10.7326/M13-3012] [Medline: 25402398]
44. Eckman MH, Wise RE, Speer B, Sullivan M, Walker N, Lip GYH, et al. Integrating real-time clinical information to provide estimates of net clinical benefit of antithrombotic therapy for patients with atrial fibrillation. Circ Cardiovasc Qual Outcomes 2014 Sep;7(5):680-686 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.114.001163] [Medline: 25205788]
45. Zaslansky R, Rothaug J, Chapman RC, Backström R, Brill S, Engel C, et al. PAIN OUT: an international acute pain registry supporting clinicians in decision making and in quality improvement activities. J Eval Clin Pract 2014 Dec;20(6):1090-1098. [doi: 10.1111/jep.12205] [Medline: 24986116]
46. Lobach DF, Kawamoto K, Anstrom KJ, Silvey GM, Willis JM, Johnson FS, et al. A randomized trial of population-based clinical decision support to manage health and resource use for Medicaid beneficiaries. J Med Syst 2013 Feb;37(1):1-10. [doi: 10.1007/s10916-012-9922-3] [Medline: 23321963]
47. Barlow J, Krassas G. Improving management of type 2 diabetes: findings of the Type2Care clinical audit. Aust Fam Physician 2013;42(1-2):57-60 [FREE Full text] [Medline: 23529464]
48. Robbins GK, Lester W, Johnson KL, Chang Y, Estey G, Surrao D, et al. Efficacy of a clinical decision-support system in an HIV practice: a randomized trial. Ann Intern Med 2012 Dec 04;157(11):757-766 [FREE Full text] [doi: 10.7326/0003-4819-157-11-201212040-00003] [Medline: 23208165]
49. Chen CC, Chen K, Hsu C, Li YJ. Developing guideline-based decision support systems using protégé and jess. Comput Methods Programs Biomed 2011 Jun;102(3):288-294. [doi: 10.1016/j.cmpb.2010.05.010] [Medline: 20594609]
50. Seow H, King S, Green E, Pereira J, Sawka C. Perspectives of patients on the utility of electronic patient-reported outcomes on cancer care. J Clin Oncol 2011 Nov 01;29(31):4213-4214. [doi: 10.1200/JCO.2011.37.9750] [Medline: 21931027]
51. Cash J. Alert fatigue. Am J Health Sys Pharm 2009;66(23):2098-2101. [doi: 10.2146/ajhp090181]
52. Kesselheim AS, Cresswell K, Phansalkar S, Bates DW, Sheikh A. Clinical decision support systems could be modified to reduce 'alert fatigue' while still minimizing the risk of litigation. Health Affairs 2011 Dec;30(12):2310-2317. [doi: 10.1377/hlthaff.2010.1111]
53. Embi PJ, Leonard AC. Evaluating alert fatigue over time to EHR-based clinical trial alerts: findings from a randomized controlled study. J Am Med Informat Assoc 2012 Jun 01;19(e1):e145-e148. [doi: 10.1136/amiajnl-2011-000743]

Abbreviations

AI: artificial intelligence
AMSTAR: Assessment of Multiple Systematic Reviews
CDSS: computerized decision support system
CPOE: computerized physician order entry
EHR: electronic health record
EMR: electronic medical record
JHNEBP: Johns Hopkins Nursing Evidence-Based Practice
PICOS: patient/participants, intervention, comparison, outcome, study design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by G Eysenbach; submitted 02.12.19; peer-reviewed by M Mileski, N Larburu, N Muro, ZSY Wong, P Martin, S Lundell, SA Steindal; comments to author 24.12.19; revised version received 08.04.20; accepted 27.07.20; published 11.08.20.

Please cite as:
Kruse CS, Ehrbar N
Effects of Computerized Decision Support Systems on Practitioner Performance and Patient Outcomes: Systematic Review
JMIR Med Inform 2020;8(8):e17283
URL: http://medinform.jmir.org/2020/8/e17283/
doi:10.2196/17283
PMID:32780714

Original Paper

Diagnostic Model for In-Hospital Bleeding in Patients with Acute ST-Segment Elevation Myocardial Infarction: Algorithm Development and Validation

Yong Li[1], MSc

[1]Emergency and Critical Care Center, Beijing Anzhen Hospital, Capital Medical University, Beijing, China

Corresponding Author:
Yong Li, MSc
Emergency and Critical Care Center
Beijing Anzhen Hospital
Capital Medical University
No. 405, Building No. 5
Madian Nancun, Xicheng District
Beijing, 100088
China
Phone: 86 13910227262
Email: liyongdoctor@sina.com

Abstract

Background: Bleeding complications in patients with acute ST-segment elevation myocardial infarction (STEMI) have been associated with increased risk of subsequent adverse consequences.

Objective: The objective of our study was to develop and externally validate a diagnostic model of in-hospital bleeding.

Methods: We performed multivariate logistic regression of a cohort for hospitalized patients with acute STEMI in the emergency department of a university hospital. Participants: The model development data set was obtained from 4262 hospitalized patients with acute STEMI from January 2002 to December 2013. A set of 6015 hospitalized patients with acute STEMI from January 2014 to August 2019 were used for external validation. We used logistic regression analysis to analyze the risk factors of in-hospital bleeding in the development data set. We developed a diagnostic model of in-hospital bleeding and constructed a nomogram. We assessed the predictive performance of the diagnostic model in the validation data sets by examining measures of discrimination, calibration, and decision curve analysis (DCA).

Results: In-hospital bleeding occurred in 112 of 4262 participants (2.6%) in the development data set. The strongest predictors of in-hospital bleeding were advanced age and high Killip classification. Logistic regression analysis showed differences between the groups with and without in-hospital bleeding in age (odds ratio [OR] 1.047, 95% CI 1.029-1.066; *P*<.001), Killip III (OR 3.265, 95% CI 2.008-5.31; *P*<.001), and Killip IV (OR 5.133, 95% CI 3.196-8.242; *P*<.001). We developed a diagnostic model of in-hospital bleeding. The area under the receiver operating characteristic curve (AUC) was 0.777 (SD 0.021, 95% CI 0.73576-0.81823). We constructed a nomogram based on age and Killip classification. In-hospital bleeding occurred in 117 of 6015 participants (1.9%) in the validation data set. The AUC was 0.7234 (SD 0.0252, 95% CI 0.67392-0.77289).

Conclusions: We developed and externally validated a diagnostic model of in-hospital bleeding in patients with acute STEMI. The discrimination, calibration, and DCA of the model were found to be satisfactory.

Trial Registration: ChiCTR.org ChiCTR1900027578; http://www.chictr.org.cn/showprojen.aspx?proj=45926

(JMIR Med Inform 2020;8(8):e20974) doi:10.2196/20974

KEYWORDS

coronary disease; ST-segment elevation myocardial infarction; hemorrhage; nomogram

Introduction

Hemorrhagic complications occur in nearly 8.5% of patients with acute ST-segment elevation myocardial infarction (STEMI) during hospitalization [1,2]. Bleeding events were associated with an increased risk of adverse outcomes in patients with STEMI [3-7]. Prevention of bleeding may represent an achievable step. Mehran et al [8] developed a model to predict bleeding in patients with acute coronary syndromes; however, the model has not been validated. Alexander et al [9] developed a model to predict in-hospital major bleeding in acute myocardial infarction, but their models were only internally validated. Moa Simonsson et al [6] developed a model to predict in-hospital major bleeding in acute myocardial infarction, and the internal and temporal validity of the model was assessed. The aim of our study was to develop and externally validate a diagnostic model of in-hospital bleeding in patients with acute STEMI.

Methods

Statement of Ethics and Data Availability

The Ethics Committee of Beijing Anzhen Hospital Capital Medical University approved the study (approval no. 2019044X, November 18, 2019). We registered this study with the WHO International Clinical Trials Registry Platform (ICTRP) (ChiCTR.org ChiCTR1900027578, November 19, 2019).

This was a retrospective analysis, and informed consent was waived by the Ethics Committee of Beijing Anzhen Hospital Capital Medical University. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional or national research committees and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study was not conducted with animals. All data generated or analyzed during this study are included in the published paper and in Multimedia Appendix 1.

Participant Selection

We used a Type 2b predictive model study, which is covered by a TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement [9]. The data were nonrandomly divided into two groups according to time: one group was used to develop a prediction model, and the other group was used for validation [9]. A Type 2b study is considered to be an external verification study [9].

The derivation cohort was 4262 hospitalized patients with acute STEMI from January 2002 to December 2013 in Beijing Anzhen Hospital, Capital Medical University. The validation cohort was 6015 hospitalized patients with acute STEMI from January 2014 to August 2019 in Beijing Anzhen Hospital, Capital Medical University. The participants were consecutively hospitalized patients with STEMI aged older than 18 years. We established the diagnosis of acute myocardial infarction (AMI) and STEMI based on the fourth universal definition of myocardial infarction [10].

Outcomes

The outcome of interest was all-cause in-hospital bleeding not related to coronary artery bypass graft surgery or catheterization during hospitalization, as defined according to the Bleeding Academic Research Consortium criteria 2, 3, and 5 [4]. The presence or absence of in-hospital bleeding was decided blinded to the predictor variables and based on the medical record.

We selected 13 predictors according to clinical relevance and the results of baseline descriptive statistics. The potential candidate variables were age, sex, Killip classification, atrioventricular (AV) block, atrial fibrillation (AF), underwent percutaneous coronary intervention (PCI) during hospitalization, and medical history such as hypertension, diabetes, myocardial infarction, PCI, coronary artery bypass graft (CABG), cerebrovascular disease, and chronic kidney disease (CKD). All these variables were determined based on the patients' medical records. AF was defined as all types of AF during hospitalization. AV block was defined as all types of AV block during hospitalization.

Our numbers of samples and events exceeded the minima required for all approaches; each candidate variable included at least 10 events for model derivation and at least 100 events for validation studies [9].

We excluded patients who lacked information on the key predictors of age and Killip classification. The reason for exclusion of all patients was lack of Killip classification.

We maintained all continuous data as continuous and retained the original scale. Based on the significant variables generated by univariate logistic regression, we constructed a multivariate logistic regression model using the backward variable selection method. We used the Akanke information criterion (AIC) and Bayesian information criterion (BIC) to select predictors. These criteria considered the model fitting and penalized the estimated number of parameters, which was equivalent to using α=.157 [9].

We assessed the predictive performance of the diagnostic model in the validation data set by examining measures of discrimination, calibration, and decision curve analysis (DCA) [9,11].

Discrimination was defined as the ability of the diagnostic model to differentiate between patients with and without in-hospital bleeding. This measure was quantified by calculating the area under the receiver operating characteristic (ROC) curve (AUC) [9].

Calibration referred to how closely the predicted in-hospital bleeding agrees with the observed in-hospital bleeding [9]. The Brier score is an aggregate measure of disagreement between the observed outcome and a prediction based on the average squared error difference.

We used DCA to describe and compare the clinical effects of the diagnostic model [9].

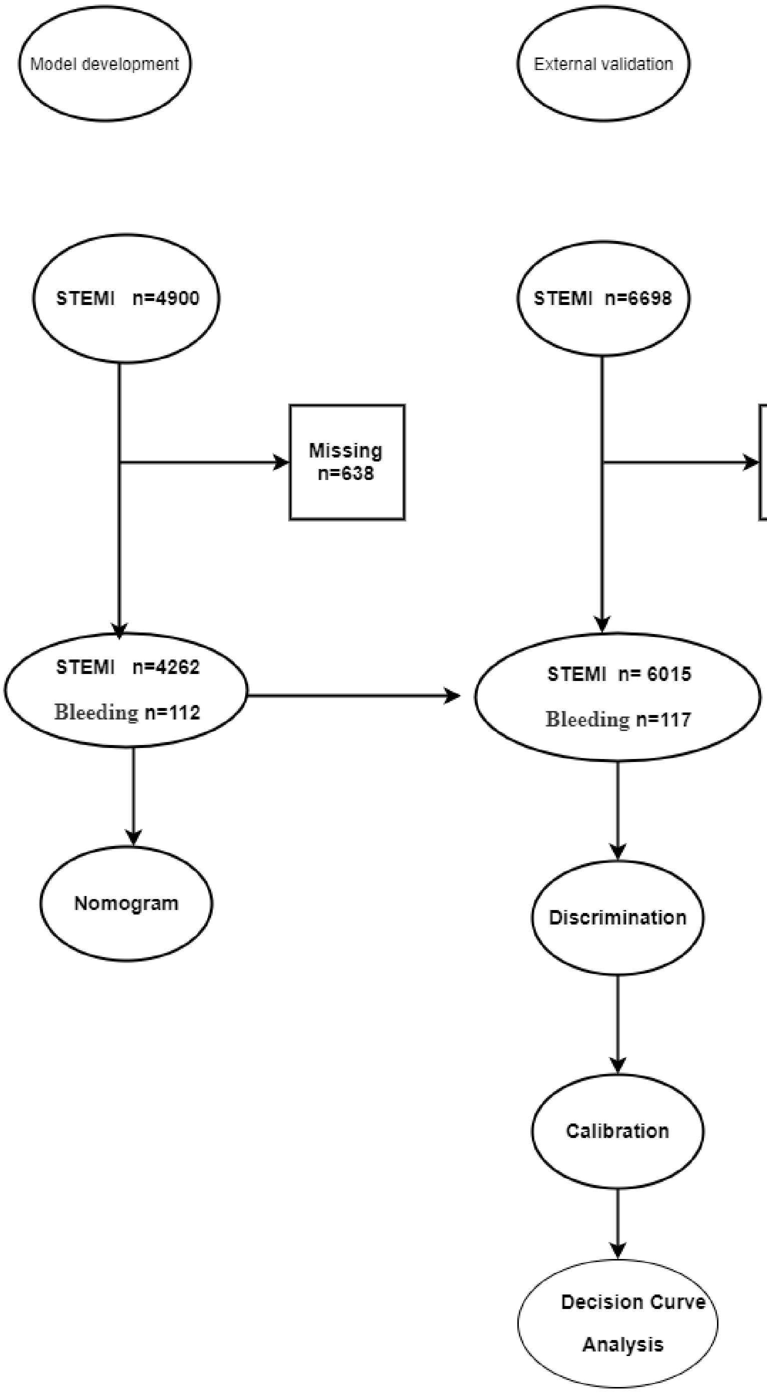
We performed statistical analyses with STATA version 15.1 (StataCorp), R version 4.0.0 (R Project), and the RMS package developed by Harrell et al [12].

Results

The study was approved by the ethics committee on November 18, 2019. Data collection started on November 26, 2019. As of submission of the manuscript, 10,277 people had been recruited for the study.

A flow diagram of the study is presented in Figure 1.

Figure 1. Flow diagram of the study. STEMI: ST-segment elevation myocardial infarction.



In the development data set, 112 of 4262 hospitalized patients (2.6%) experienced in-hospital bleeding. The patients' baseline characteristics are shown in Table 1. Nine variables (age, sex, Killip classification, AVB, AF, history of CABG, history of diabetes, history of CKD, and underwent PCI during hospitalization) were significantly different in the two groups of patients (α=.157).

Table 1. Demographic and clinical characteristics of patients with and without in-hospital bleeding in the development data set (N=4262).

Characteristic	Total (N=4262)	In-hospital bleeding (n=112)	No bleeding (n=4150)	*P* value
Age (years, range 21-99), mean (SD)	60 (13)	70 (10)	60 (13)	<.001
Male sex, n (%)	3248 (76.2)	73 (65.2)	3175 (76.5)	.006
Medical history, n (%)				
Hypertension	2372 (55.7)	63 (56.3)	2309 (55.6)	.90
Diabetes	1246 (29.2)	42 (37.5)	1204 (29.0)	.053
Myocardial infarction	426 (10.0)	13 (11.6)	413 (10.0)	.57
PCI[a]	228 (5.3)	7 (6.3)	221 (5.3)	.67
CABG[b]	28 (0.7)	2 (1.8)	26 (0.6)	.15
CKD[c]	95 (2.2)	7 (6.3)	88 (2.1)	.006
HCD[d]	338 (7.9)	11 (9.8)	327 (7.9)	.45
Killip classification, n (%)				
I	769 (18)	13 (11.6)	756 (18.2)	.08
II	2565 (60.2)	31 (27.7)	2534 (61.1)	<.001
III	533 (12.5)	32 (28.6)	501 (12.1)	<.001
IV	395 (9.3)	36 (32.1)	359 (8.7)	<.001
AF[e]	243 (5.7)	15 (13.4)	228 (5.5)	.001
AVB[f]	197 (4.6)	13 (11.6)	184 (4.4)	.001
Underwent PCI	3103 (72.8)	50 (44.6)	3053 (73.6)	<.001

[a] PCI: percutaneous coronary intervention.

[b] CABG: coronary artery bypass graft.

[c] CKD: chronic kidney disease.

[d] HCD: history of cerebrovascular disease.

[e] AF: atrial fibrillation.

[f] AVB: atrioventricular block.

After application of the backward variable selection method, AIC, and BIC, age remained a significant independent predictor of in-hospital bleeding; Killip classification remained a rank variable of in-hospital bleeding. These results are shown in Table 2 and Table 3.

Table 2. Predictors of in-hospital bleeding obtained from multivariable logistic regression models (odds ratio) in the development data set.

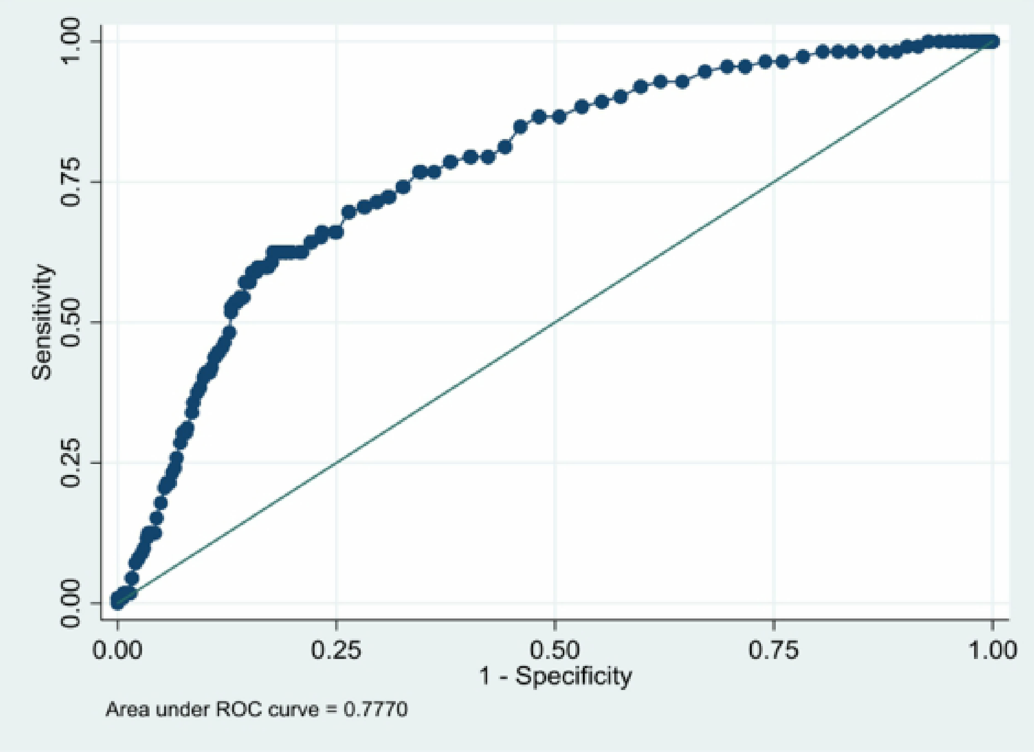
In-hospital bleeding	Odds ratio	Standard error	Z	Pr>| Z |	95% CI
Age	1.047443	0.0095986	5.06	<.001	1.028798-1.066426
Killip III	3.265072	0.8100203	4.77	<.001	2.007804-5.309632
Killip IV	5.132613	1.240357	6.77	<.001	3.196212-8.242169
Constant	0.0007621	0.0004685	–11.68	<.001	0.0002285-0.0025424

Table 3. Predictor of in-hospital bleeding obtained from multivariable logistic regression models (coefficients) in the development data set.

In-hospital bleeding	Coefficient	Standard error	Z	Pr>| Z |	95% CI
Age	0.0463523	0.0091638	5.06	<.001	0.0283915 to 0.0643131
Killip III	1.183282	0.2480865	4.77	<.001	0.6970414 to 1.669523
Killip IV	1.635615	0.2416619	6.77	<.001	1.161966 to 2.109263
Constant	–7.179377	0.6146614	–11.68	<.001	–8.384092 to –5.974663

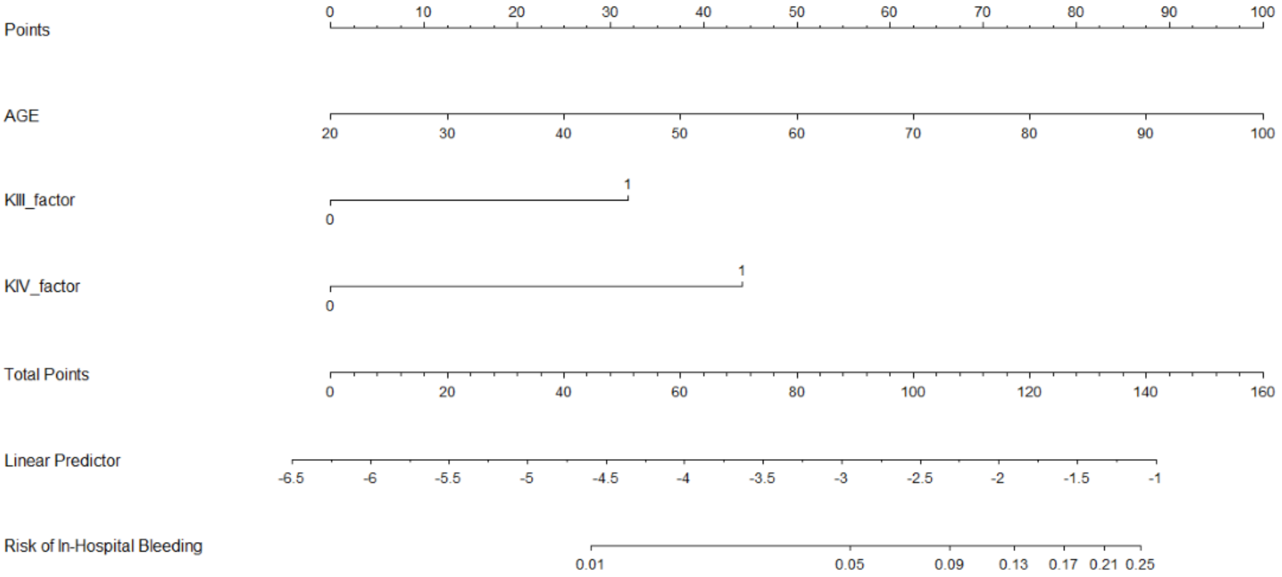
According to the above risk factors, we can calculate the predicted probability of in-hospital bleeding using the formula P= 1/(1 + exp(–(–7.179377 + 0.0463523 × AGE(years) + 1.183282 × KIII + 1.635615 × KIV))), where KIII is Killip III (0 = No, 1 = Yes) and KIV is Killip IV (0 = No, 1 = Yes). The ROC curve was drawn (Figure 2). The AUC was 0.777 (SD 0.021, 95% CI 0.73576-0.81823).

Figure 2. ROC curve for the identification of patients with in-hospital bleeding in the development dataset. ROC: receiver operating characteristic.



We constructed the nomogram (Figure 3) using the development database based on an independent prognostic marker (age) and a rank variable (Killip classification). To use the nomogram, the patient's age is found on the AGE axis, and a straight line is then drawn upward to the Points axis to determine how many points toward progression the patient receives for their age. The steps are repeated for the other axes, with a straight line drawn upward each time toward the points axis. The points received for each predictor are summed, and the sum is found on the total points axis. A straight line is drawn down to the Risk of In-Hospital Bleeding axis to find the patient's probability of in-hospital bleeding.

Figure 3. Nomogram for predicting in-hospital bleeding in patients with acute ST-segment elevation myocardial infarction. AGE: age (years); KIII-factor: Killip III; KIV-factor: Killip IV.



A total of 117 of 6015 hospitalized patients in the validation data set (1.9%) suffered in-hospital bleeding. The baseline characteristics of the patients are shown in Table 4. We can calculate the predicted probability of in-hospital bleeding using the formula P= 1/(1 + exp(–(–7.179377 + .0463523 × age (years) + 1.183282 × KIII + 1.635615 × KIV))), where KIII is Killip III (0 = No, 1 = Yes) and KIV is Killip IV (0 = No, 1 = Yes).

We drew the ROC curve (Figure 4). The AUC was 0.7234 (SD 0.0252, 95% CI 0.67392-0.77289).

We drew a calibration plot (Figure 5) with the distribution of the predicted probabilities for individuals with and without in-hospital bleeding in the validation data set. The Hosmer-Lemeshow χ^2_{10} value was 10.64, Pr>χ^2 was 0.3859>.05, and the Brier score was .0188 (<.25).

Figure 6 shows the DCA of the validation data set.

Table 4. Demographic and clinical characteristics of patients with and without in-hospital bleeding in the validation data set (N=6015).

Characteristic	Total (N=6015)	In-hospital bleeding (n=117)	No bleeding (n=5898)	*P* value
Age (years, range 21-92), mean (SD)	59 (12)	64 (12)	58 (12)	<.001
Male sex, n (%)	4894 (81.4)	86 (73.5)	4808 (81.5)	.03
Medical history, n (%)				
Hypertension	3427 (57.0)	65 (55.6)	3362 (57)	.75
Diabetes	1822 (30.3)	40 (34.2)	1782 (30.2)	.36
Myocardial infarction	433 (7.2)	14 (12)	419 (7.1)	.047
PCI[a]	575 (9.6)	18 (15.4)	557 (9.4)	.03
CABG[b]	51 (0.8)	3 (2.6)	48 (0.8)	.05
CKD[c]	145 (2.4)	4 (3.4)	141 (2.4)	.48
HCD[d]	421 (7.0)	12 (10.3)	409 (6.9)	.17
Killip classification, n (%)				
I	4234 (70.4)	45 (38.5)	4189 (71.0)	<.001
II	1188 (19.7)	31 (26.5)	1157 (19.6)	.07
III	266 (4.4)	11 (9.4)	255 (4.3)	.01
IV	330 (5.5)	30 (25.6)	300 (5.1)	<.001
AF[e], n (%)	275 (4.6)	12 (10.3)	263 (4.5)	.004
AVB[f], n (%)	119 (2.0)	3 (2.6)	116 (2.0)	.65
Underwent PCI n (%)	4564 (75.9)	70 (59.8)	4494 (76.2)	<.001

[a] PCI: percutaneous coronary intervention.

[b] CABG: coronary artery bypass grafting.

[c] CKD: chronic kidney disease.

[d] HCD: cerebrovascular disease.

[e] AF: atrial fibrillation.

[f] AVB: atrioventricular block.

Figure 4. ROC curve for the identification of patients with in-hospital bleeding in the validation data set. ROC: receiver operating characteristic.

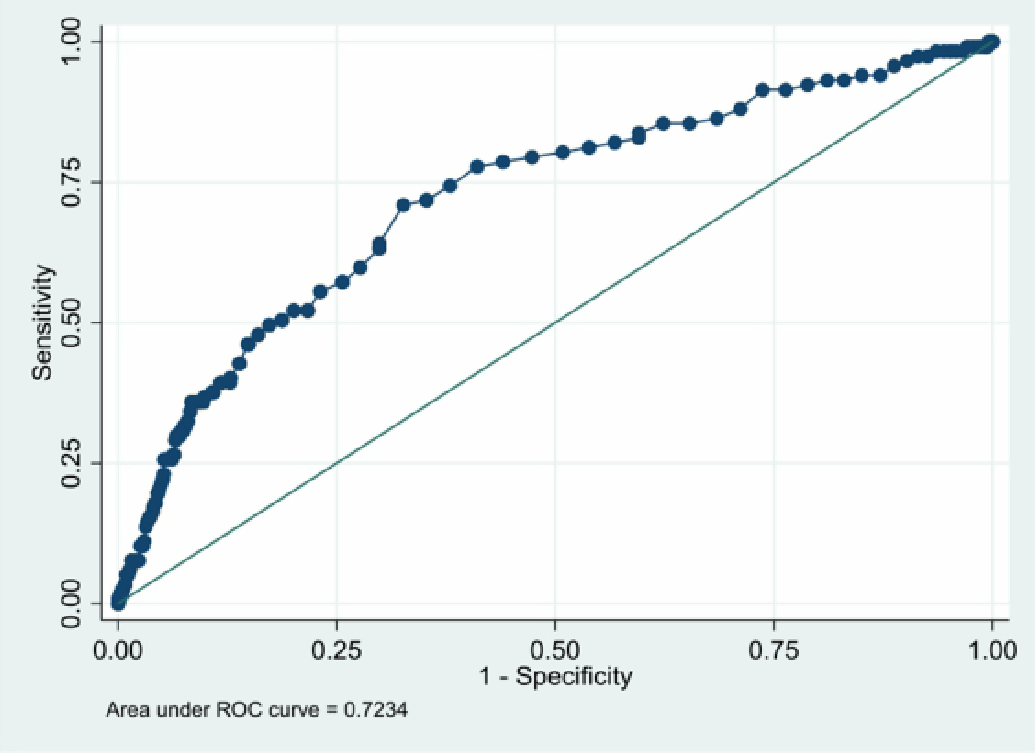


Figure 5. Calibration plot with distribution of the predicted probabilities for individuals with and without in-hospital bleeding in the validation data set.

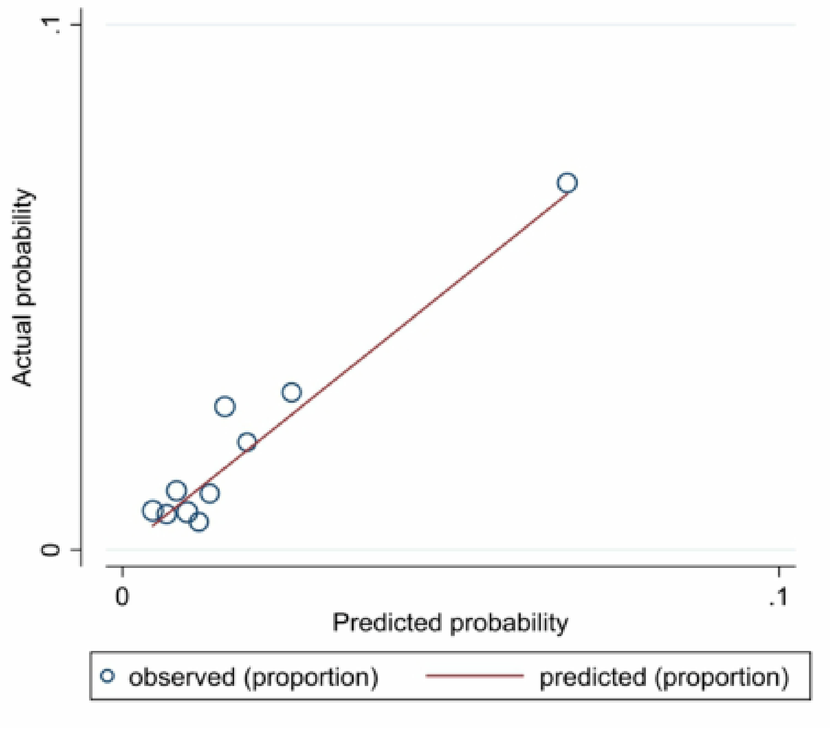
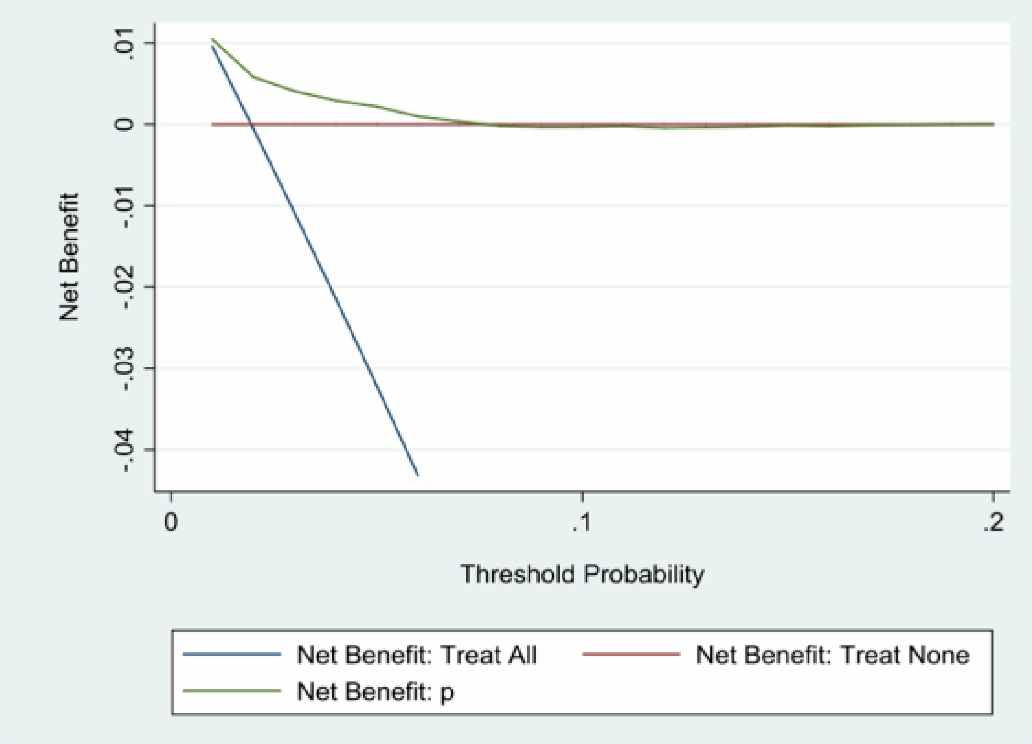


Figure 6. Decision curve analysis of the validation data set.



Discussion

Principal Findings

In our study, advanced age and high Killip classification were associated with increased risk of in-hospital bleeding in patients with acute STEMI. The formula or nomogram could be used to predict in-hospital bleeding. Specific strategies should be used to reduce the risk of in-hospital bleeding, such as ensuring the appropriate dose of antithrombotic drugs.

The predictive performance of the diagnostic model in the validation data set was assessed by examining measures of discrimination, calibration, and DCA. The AUC was 0.7234 (SD 0.0252, 95% CI 0.67392-0.77289) in the validation data set. The Hosmer-Lemeshow χ^2_{10} value was10.64, Pr>χ^2 was 0.3859>.05. and the Brier score was <.25. The discrimination, calibration, and DCA results were satisfactory.

A high Killip classification has been associated with increased risk of bleeding [3,7,13]. In our study, patients with Killip class IV were at 5.1 times higher risk of in-hospital bleeding than patients with Killip classes I to III. Insufficient tissue perfusion adversely affected the coagulation system and platelet function [13]. Insufficient tissue perfusion may cause gastritis or ulceration and increase the possibility of gastrointestinal bleeding [13].

Advanced age has been reported to be an independent risk factor of bleeding [3,13-17]. Age may change the balance between the risks and benefits of treatment strategies [18]. The cause of the higher risk of bleeding in older people may be multifactorial, including decreased kidney function and increased sensitivity to anticoagulants [19]. It has been speculated that the presence of local vascular changes is an explanation for the increased incidence of bleeding complications in older patients [20]. Stomach protection is recommended for older patients [21].

Moscucci et al [20] observed that older age, female sex, history of bleeding, and renal insufficiency were independent predictors of major bleeding among 8151 patients with STEMI, 7440 patients with non–ST-segment elevation myocardial infarction (NSTEMI), and 8454 patients with unstable angina registered in the Global Acute Coronary Events Registry (GRACE). Spencer et al [22] found that major bleeding occurred in 2.8% of 40,087 patients with AMI enrolled in the GRACE. These patients were older, more severely ill, and more likely to undergo invasive procedures. Subherwal et al [23] used 71,277 patients to derive and 17,857 patients to validate a model to stratify the risk of major bleeding in patients with NSTEMI. This was a form of internal validation, as their development and validation cohorts were created randomly rather than nonrandomly [9]. Nikolsky et al [19] found 7 independent predictors of major bleeding after PCI using the femoral approach, and the AUC was 0.62 in the validation data set.

Roxana Mehran et al [8] used 17,421 patients to derive a model that identifies 6 independent baseline predictors to predict bleeding in patients with acute coronary syndromes; however, this model has not been validated. KP Alexander et al [24] used 72,313 patients to develop and 17,960 patients to validate a model to predict in-hospital major bleeding during myocardial infarction care. This was also a form of internal validation because their cohorts were randomly created [9]. Moa Simonsson et al [6] used 97,597 patients to develop a model to predict in-hospital major bleeding in acute myocardial infarction. The internal and temporal validity of the model were assessed; the temporal validity of the score was assessed using internal-external cross-validation [6].

Our diagnostic model of in-hospital bleeding builds upon these studies in several ways. Our model was externally validated. It provides an absolute value rather than a relative value. It includes only two baseline factors, namely age and Killip classification. It can be easily calculated at patient presentation. It can remain discriminatory irrespective of which treatment was used (eg, invasive care or antithrombotic drugs), thereby improving its effectiveness in clinical decision-making. It was developed using unselected real-world populations, including patients who underwent initial invasive strategies and revascularization as well as patients who were conservatively treated without catheterization. Algorithms that can help physicians evaluate diagnoses should be simple and easy to apply, and they should use clinical data that is routinely provided by the hospital. The nomogram we constructed for in-hospital bleeding captures most of the diagnostic information provided by the complete logistic regression model and is easy to use.

Limitations

The present analysis has a few limitations. This was a single-center study. Some patients were selected >10 years ago; therefore, their treatment may not represent current standards and techniques. We did not include bleeding related to catheterization. The use of antithrombotic drugs and previous bleeding history were not obtained in this study; therefore, we could not determine the impact of anticoagulation or previous bleeding history on bleeding risk. Finally, the C statistics of the in-hospital bleeding model in the study were modest (0.777 in the derivation cohort and 0.7234 in the validation cohort).

Conclusion

We developed and externally validated a diagnostic model of in-hospital bleeding in patients with acute STEMI.

Acknowledgments

This research received no external funding.

Authors' Contributions

YL contributed to the generation of the study data, analyzed and interpreted the study data, drafted the manuscript, and revised the manuscript. YL is responsible for the overall content as guarantor. The author read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary materials. The data are the demographic and clinical characteristics of hospitalized patients with acute STEMI. AGE: age; ALLAF: atrial fibrillation; AVB: atrioventricular block; BLOOD: all-cause bleeding; CABG: history of coronary artery bypass graft; CKD: history of chronic kidney disease; DM: history of diabetes; HBP: history of hypertension; HCD: history of cerebrovascular disease; HPCI: history of percutaneous coronary intervention; KI: Killip I; KII: Killip II; KIII: Killip III; KIV: Killip IV; OMI: history of myocardial infarction; PCI: underwent PCI during hospitalization; S: sex.

[ZIP File (Zip Archive), 31 KB - medinform_v8i8e20974_app1.zip]

References

1. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, American Heart Association Council on EpidemiologyPrevention Statistics CommitteeStroke Statistics Subcommittee. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. Circulation 2019 Mar 05;139(10):e56-e528. [doi: 10.1161/CIR.0000000000000659] [Medline: 30700139]
2. Masoudi FA, Ponirakis A, de Lemos JA, Jollis JG, Kremers M, Messenger JC, et al. Executive Summary: Trends in U.S. Cardiovascular Care: 2016 Report From 4 ACC National Cardiovascular Data Registries. J Am Coll Cardiol 2017 Mar 21;69(11):1424-1426 [FREE Full text] [doi: 10.1016/j.jacc.2016.12.004] [Medline: 28025066]
3. Albeiruti R, Chaudhary F, Alqahtani F, Kupec J, Balla S, Alkhouli M. Incidence, Predictors, and Outcomes of Gastrointestinal Bleeding in Patients Admitted With ST-Elevation Myocardial Infarction. Am J Cardiol 2019 Aug 01;124(3):343-348. [doi: 10.1016/j.amjcard.2019.05.008] [Medline: 31182211]
4. Mehran R, Rao SV, Bhatt DL, Gibson CM, Caixeta A, Eikelboom J, et al. Standardized bleeding definitions for cardiovascular clinical trials: a consensus report from the Bleeding Academic Research Consortium. Circulation 2011 Jun 14;123(23):2736-2747. [doi: 10.1161/CIRCULATIONAHA.110.009449] [Medline: 21670242]
5. Cornara S, Somaschini A, De Servi S, Crimi G, Ferlini M, Baldo A, et al. Prognostic Impact of in-Hospital-Bleeding in Patients With ST-Elevation Myocardial Infarction Treated by Primary Percutaneous Coronary Intervention. Am J Cardiol 2017 Nov 15;120(10):1734-1741. [doi: 10.1016/j.amjcard.2017.07.076] [Medline: 28865893]
6. Simonsson M, Winell H, Olsson H, Szummer K, Alfredsson J, Hall M, et al. Development and Validation of a Novel Risk Score for In-Hospital Major Bleeding in Acute Myocardial Infarction:-The SWEDEHEART Score. J Am Heart Assoc 2019 Mar 05;8(5):e012157 [FREE Full text] [doi: 10.1161/JAHA.119.012157] [Medline: 30803289]

7. Sadjadieh G, Engstrøm T, Høfsten DE, Helqvist S, Køber L, Pedersen F, et al. Bleeding Events After ST-segment Elevation Myocardial Infarction in Patients Randomized to an All-comer Clinical Trial Compared With Unselected Patients. Am J Cardiol 2018 Oct 15;122(8):1287-1296. [doi: 10.1016/j.amjcard.2018.07.008] [Medline: 30115422]
8. Mehran R, Pocock SJ, Nikolsky E, Clayton T, Dangas GD, Kirtane AJ, et al. A risk score to predict bleeding in patients with acute coronary syndromes. J Am Coll Cardiol 2010 Jun 08;55(23):2556-2566 [FREE Full text] [doi: 10.1016/j.jacc.2009.09.076] [Medline: 20513595]
9. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015 Jan 06;162(1):W1-73. [doi: 10.7326/M14-0698] [Medline: 25560730]
10. Thygesen K, Alpert JS, Jaffe AS, Chaitman BR, Bax JJ, Morrow DA, Executive Group on behalf of the Joint European Society of Cardiology (ESC)/American College of Cardiology (ACC)/American Heart Association (AHA)/World Heart Federation (WHF) Task Force for the Universal Definition of Myocardial Infarction. Fourth Universal Definition of Myocardial Infarction (2018). J Am Coll Cardiol 2018 Oct 30;72(18):2231-2264 [FREE Full text] [doi: 10.1016/j.jacc.2018.08.1038] [Medline: 30153967]
11. Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. BMJ 2012 Jun 21;344:e4181 [FREE Full text] [doi: 10.1136/bmj.e4181] [Medline: 22723603]
12. Harrell FE, Margolis PA, Gove S, Mason KE, Mulholland EK, Lehmann D, et al. Development of a clinical prediction model for an ordinal outcome: the World Health Organization Multicentre Study of Clinical Signs and Etiological agents of Pneumonia, Sepsis and Meningitis in Young Infants. WHO/ARI Young Infant Multicentre Study Group. Stat Med 1998 Apr 30;17(8):909-944. [doi: 10.1002/(sici)1097-0258(19980430)17:8<909::aid-sim753>3.0.co;2-o] [Medline: 9595619]
13. Matić DM, Ašanin MR, Stanković SD, Mrdović IB, Marinković JM, Kočev NI, et al. Incidence, predictors and prognostic implications of bleeding complicating primary percutaneous coronary intervention. Vojnosanit Pregl 2015 Jul;72(7):589-595. [doi: 10.2298/vsp140223064m] [Medline: 26364451]
14. Ko SQ, Valsdottir LR, Strom JB, Cheng Y, Hirayama A, Liu P, et al. Meta-Analysis of Bleeding Risk Prediction Scores in Patients After Percutaneous Coronary Intervention on Dual Antiplatelet Therapy. Am J Cardiol 2018 Dec 01;122(11):1843-1852. [doi: 10.1016/j.amjcard.2018.08.025] [Medline: 30309627]
15. Jeger RV, Pfisterer M, Vogt DR, Galatius S, Abildgaard U, Naber C, et al. Competing risks of major bleeding and thrombotic events with prasugrel-based dual antiplatelet therapy after stent implantation - An observational analysis from BASKET-PROVE II. PLoS One 2019;14(1):e0210821 [FREE Full text] [doi: 10.1371/journal.pone.0210821] [Medline: 30645635]
16. Luo P, Lin X, Lin C, Luo J, Hu H, Ting P, et al. Risk factors for upper gastrointestinal bleeding among aspirin users: An old issue with new findings from a population-based cohort study. J Formos Med Assoc 2019 May;118(5):939-944 [FREE Full text] [doi: 10.1016/j.jfma.2018.10.007] [Medline: 30366771]
17. Lenti MV, Pasina L, Cococcia S, Cortesi L, Miceli E, Caccia Dominioni C, REPOSI Investigators. Mortality rate and risk factors for gastrointestinal bleeding in elderly patients. Eur J Intern Med 2019 Mar;61:54-61. [doi: 10.1016/j.ejim.2018.11.003] [Medline: 30522789]
18. Roe MT, Goodman SG, Ohman EM, Stevens SR, Hochman JS, Gottlieb S, et al. Elderly patients with acute coronary syndromes managed without revascularization: insights into the safety of long-term dual antiplatelet therapy with reduced-dose prasugrel versus standard-dose clopidogrel. Circulation 2013 Aug 20;128(8):823-833. [doi: 10.1161/CIRCULATIONAHA.113.002303] [Medline: 23852610]
19. Nikolsky E, Mehran R, Dangas G, Fahy M, Na Y, Pocock SJ, et al. Development and validation of a prognostic risk score for major bleeding in patients undergoing percutaneous coronary intervention via the femoral approach. Eur Heart J 2007 Aug;28(16):1936-1945. [doi: 10.1093/eurheartj/ehm194] [Medline: 17575270]
20. Moscucci M, Fox KAA, Cannon CP, Klein W, López-Sendón J, Montalescot G, et al. Predictors of major bleeding in acute coronary syndromes: the Global Registry of Acute Coronary Events (GRACE). Eur Heart J 2003 Oct;24(20):1815-1823. [doi: 10.1016/s0195-668x(03)00485-8] [Medline: 14563340]
21. Ibanez B, James S, Agewall S, Antunes MJ, Bucciarelli-Ducci C, Bueno H, ESC Scientific Document Group. 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: The Task Force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC). Eur Heart J 2018 Jan 07;39(2):119-177. [doi: 10.1093/eurheartj/ehx393] [Medline: 28886621]
22. Spencer FA, Moscucci M, Granger CB, Gore JM, Goldberg RJ, Steg PG, GRACE Investigators. Does comorbidity account for the excess mortality in patients with major bleeding in acute myocardial infarction? Circulation 2007 Dec 11;116(24):2793-2801. [doi: 10.1161/CIRCULATIONAHA.107.694273] [Medline: 18025530]
23. Subherwal S, Bach RG, Chen AY, Gage BF, Rao SV, Newby LK, et al. Baseline risk of major bleeding in non-ST-segment-elevation myocardial infarction: the CRUSADE (Can Rapid risk stratification of Unstable angina patients Suppress ADverse outcomes with Early implementation of the ACC/AHA Guidelines) Bleeding Score. Circulation 2009 Apr 14;119(14):1873-1882 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.108.828541] [Medline: 19332461]

24. Mathews R, Peterson ED, Chen AY, Wang TY, Chin CT, Fonarow GC, et al. In-hospital major bleeding during ST-elevation and non-ST-elevation myocardial infarction care: derivation and validation of a model from the ACTION Registry®-GWTG™. Am J Cardiol 2011 Apr 15;107(8):1136-1143. [doi: 10.1016/j.amjcard.2010.12.009] [Medline: 21324428]

Abbreviations

AF: atrial fibrillation
AIC: Akanke information criterion
AMI: acute myocardial infarction
AUC: area under the receiver operating characteristic curve
AV: atrioventricular
BIC: Bayesian information criterion
CABG: coronary artery bypass graft
CKD: chronic kidney disease
DCA: decision curve analysis
HCD: history of cerebrovascular disease
GRACE: Global Registry of Acute Coronary Events
MI: myocardial infarction
NSTEMI: non–ST-segment elevation myocardial infarction
PCI: percutaneous coronary intervention
ROC: receiver operating characteristic
STEMI: ST-segment elevation myocardial infarction
TIMI: thrombolysis in myocardial infarction
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Edited by G Eysenbach; submitted 02.06.20; peer-reviewed by DM Matić, N Onyeakusi; comments to author 29.06.20; revised version received 06.07.20; accepted 14.07.20; published 14.08.20.

Please cite as:
Li Y
Diagnostic Model for In-Hospital Bleeding in Patients with Acute ST-Segment Elevation Myocardial Infarction: Algorithm Development and Validation
JMIR Med Inform 2020;8(8):e20974
URL: http://medinform.jmir.org/2020/8/e20974/
doi:10.2196/20974
PMID:32795995

Original Paper

Model-Based Algorithms for Detecting Peripheral Artery Disease Using Administrative Data From an Electronic Health Record Data System: Algorithm Development Study

Elizabeth Hope Weissler[1], MD; Steven J Lippmann[2], PhD; Michelle M Smerek[2], BSc; Rachael A Ward[3], MD, MPH; Aman Kansal[3], MD; Adam Brock[3], MD; Robert C Sullivan[3], MD; Chandler Long[1], MD; Manesh R Patel[2,3,4], MD; Melissa A Greiner[2], MSc; N Chantelle Hardy[2], MPH; Lesley H Curtis[2,4], PhD; W Schuyler Jones[2,3,4], MD

[1]Division of Vascular and Endovascular Surgery, Duke University School of Medicine, Durham, NC, United States

[2]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC, United States

[3]Department of Medicine, Duke University School of Medicine, Durham, NC, United States

[4]Duke Clinical Research Institute, Durham, NC, United States

Corresponding Author:
W Schuyler Jones, MD
Department of Medicine
Duke University School of Medicine
DUMC Box 3330
Durham, NC, 27710
United States
Phone: 1 919 668 8917
Email: schuyler.jones@duke.edu

Abstract

Background: Peripheral artery disease (PAD) affects 8 to 10 million Americans, who face significantly elevated risks of both mortality and major limb events such as amputation. Unfortunately, PAD is relatively underdiagnosed, undertreated, and underresearched, leading to wide variations in treatment patterns and outcomes. Efforts to improve PAD care and outcomes have been hampered by persistent difficulties identifying patients with PAD for clinical and investigatory purposes.

Objective: The aim of this study is to develop and validate a model-based algorithm to detect patients with peripheral artery disease (PAD) using data from an electronic health record (EHR) system.

Methods: An initial query of the EHR in a large health system identified all patients with PAD-related diagnosis codes for any encounter during the study period. Clinical adjudication of PAD diagnosis was performed by chart review on a random subgroup. A binary logistic regression to predict PAD was built and validated using a least absolute shrinkage and selection operator (LASSO) approach in the adjudicated patients. The algorithm was then applied to the nonsampled records to further evaluate its performance.

Results: The initial EHR data query using 406 diagnostic codes yielded 15,406 patients. Overall, 2500 patients were randomly selected for ground truth PAD status adjudication. In the end, 108 code flags remained after removing rarely- and never-used codes. We entered these code flags plus administrative encounter, imaging, procedure, and specialist flags into a LASSO model. The area under the curve for this model was 0.862.

Conclusions: The algorithm we constructed has two main advantages over other approaches to the identification of patients with PAD. First, it was derived from a broad population of patients with many different PAD manifestations and treatment pathways across a large health system. Second, our model does not rely on clinical notes and can be applied in situations in which only administrative billing data (eg, large administrative data sets) are available. A combination of diagnosis codes and administrative flags can accurately identify patients with PAD in large cohorts.

(JMIR Med Inform 2020;8(8):e18542) doi:10.2196/18542

KEYWORDS

peripheral artery disease; patient selection; electronic health records; cardiology; health data

Introduction

Lower extremity peripheral artery disease (PAD) is a prevalent chronic vascular condition that is estimated to affect over 200 million patients globally [1]. Although most patients with PAD are asymptomatic, more severe disease is associated with negative health and quality of life effects including claudication (leg pain caused by insufficient blood flow), ischemia (blood flow insufficient to meet the extremity's metabolic demands), and tissue loss from small wounds that worsen without adequate blood for healing. Severe ischemia with enlarging or infected wounds can require amputation [1,2]. Given the morbidity and mortality burden of PAD, investigation of novel therapies and implementation efforts is an ongoing necessity.

Improvement in the quality of PAD treatment and research requires correct and efficient identification of patients who truly have the disease. Although prospective studies can confirm patients' diagnoses through multiple methods, studies that rely on the review of electronic health records (EHRs) or billing claims are limited to preexisting data. Computable phenotypes based on billing codes are sufficient to identify affected patients for many conditions, but for others, current diagnosis codes do not adequately differentiate the condition of interest from other related conditions [3,4]. PAD detection algorithms using administrative code sets, such as combinations of International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis codes or Current Procedural Terminology (CPT) procedure codes, have been shown to be relatively inaccurate compared to diagnostic "gold standards" such as the ankle-brachial index, especially when applied beyond patients seen in a vascular laboratory or clinic setting [5-8].

Our initial attempt to identify patients with PAD within a large US academic health system using billing diagnosis codes had a very low positive predictive value. In this paper, we describe the two-staged "learning" approach that we adopted by first determining the PAD status for a random sample of the initially selected patients; training and validating a model using that patient set; and then scoring the remaining patients from the initial patient query to identify charts with a high likelihood of PAD to review for model validation and PAD cohort inclusion. The goal of this research was to develop and validate a model-based algorithm to accurately detect patients with peripheral artery disease using diagnostic billing codes and administrative information available in the EHRs data system.

Methods

Data Source and Study Population

The study population was selected using a query to Duke Enterprise Data Unified Content Explorer (DEDUCE). DEDUCE interfaces with and supports queries of the EHR data repository for all patients seen within the Duke University Health System (DUHS), an integrated health system that includes 3 hospitals and a large number of outpatient clinical offices in the Raleigh-Durham region of North Carolina. To be eligible for inclusion in this study, patients needed to have had at least one clinical encounter at DUHS resulting in one or more PAD-related diagnosis codes between January 1, 2015, and March 31, 2016. This study period was chosen in part because it included the period during which the ICD-9-CM to ICD-10-CM (10th Revision) transition occurred, thereby facilitating incorporation of codes from both systems into our algorithm. Encounter-level EHRs were obtained for all clinical encounters during the study period, including hospital admissions, emergency department visits, and outpatient clinic visits. This research was approved by the Duke University Institutional Review Board (protocol ID number Pro00075637).

Selection of Diagnosis Codes, Procedure Codes, and Other Administrative Data Flags

Our initial list of diagnosis codes related to lower extremity PAD (including peripheral vascular disease, atherosclerosis, diabetes with peripheral circulatory disorders, lower extremity ulcers, arterial thromboembolism, and gangrene) contained 31 ICD-9-CM and 375 ICD-10-CM diagnosis codes (Multimedia Appendix 1). The ICD-9-CM codes used in this study were drawn from cohort eligibility criteria or outcome definitions from prior studies of PAD [9,10], as well as from clinician review of the ICD-9-CM classification system. ICD-10-CM codes were forward- and backward-mapped from ICD-9-CM codes using General Equivalence Mappings (GEMs) and were screened by the clinical team to eliminate spurious mappings [11]. The mapped corresponding ICD-9 and ICD-10 codes were included as separate flags. However, there were two codes (ICD-9-CM 443.9 and ICD-10-CM I73.9 for "Peripheral vascular disease, unspecified") that were grouped into a single flag because the terminology for this code did not change with the ICD-10-CM transition. In addition, nearly half of all patients in the study population had one or both of those codes (443.9 or I73.9) present during the study period. There were 247 PAD-related codes that were not detected for any patients, and an additional 50 codes that were used for only 1 or 2 patients; these codes were removed from the analysis, leaving 108 diagnosis code flags including the combined flag for 443.9/I73.9.

Additionally, 4 indicator variables were created to increase the likelihood that a PAD-related diagnosis code indicated true PAD, rather than an encounter devoting to "ruling out" PAD. Two were procedure code-based: one for having any revascularization procedure and another for having any diagnostic imaging code associated with an encounter with a PAD-related diagnosis code. We selected ICD-9-PCS, ICD-10-PCS, and CPT procedure codes based on prior literature and clinical expertise (Multimedia Appendix 2) [12,13]. Revascularization procedures included codes for atherectomy, angioplasty, dilation, bypass, replacement, or supplementation procedures related to the lower extremity arteries. Diagnostic studies included noninvasive hemodynamic studies, ultrasound, magnetic resonance imaging, computed tomography angiography, and catheter-based angiography.

Finally, we also derived two indicator variables based on other administrative information contained within the EHR. One was a flag for having two or more encounters with a PAD-related diagnosis code within the study period. The other was a flag for encounters associated with PAD codes in which the primary physician was listed as "Cardiology," "Vascular Surgery," "Cardiovascular Medicine," "Interventional Radiology," "Podiatry," or "Wound Care" (the most common provider types who frequently care for patients with PAD).

Chart Abstraction Process, Model Development, and Validation

Chart abstraction was necessary for the larger PAD outcomes study that this project was a part of because there are potential confounders of the associations between patient characteristics and clinical outcomes that must be obtained through review of clinical data. It was impractical to abstract data from and confirm the very large number of potential patients with PAD identified from the initial billing diagnosis codes. Instead, we took a two-staged "learning" approach to abstraction by first reviewing charts from a random subgroup of patients and then using this PAD-adjudicated subgroup to model which of the diagnosis and administrative flags were most predictive of true PAD diagnosis. We then used the probabilities generated from this model to decide which of the remaining patients' records to abstract.

Chart review was performed in accordance with a written manual to standardize abstraction. There were 6 medical abstractors in total, and each reviewer was trained to complete the forms completely. When discordant information, inconclusive data, or uncertainty remained after initial review, the file was marked and the senior author (WSJ) reviewed the file and made a final determination. In the first stage of abstraction, we reviewed charts to adjudicate PAD status for a random sample of 2500 patients from the original cohort (Figure 1). PAD was confirmed using either ankle-brachial index (ABI), history of prior revascularization, or lower extremity amputation for an indication of symptomatic PAD. ABI 0.9 or ABI 1.4 in either limb was diagnostic of PAD, and toe pressures were used if lower extremity vessels were noncompressible. Revascularization procedures performed between January 1, 2010 and the index visit date during the study period within DUHS were considered to be prior revascularizations.

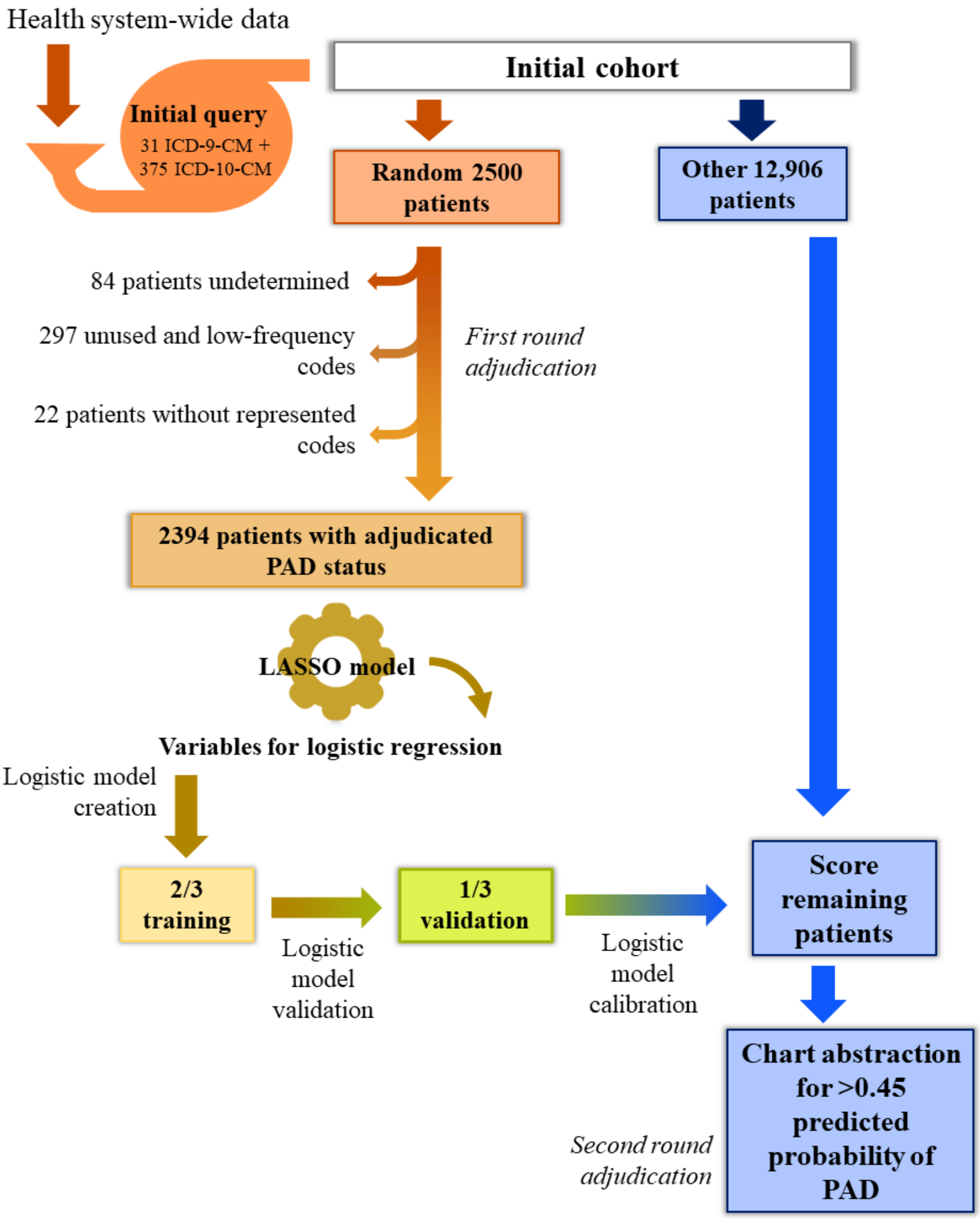
To avoid overfitting the prediction model, we used the least absolute shrinkage and selection operator (LASSO) approach to reduce the number of variables [14]. The starting point of the LASSO model included the binary flags for each diagnostic code, as well as for revascularization procedures, diagnostic imaging, specialist provider, and having ≥2 PAD encounters. Using the chart abstraction PAD status determination as the "ground truth," we fit the LASSO logistic regression model with all adjudicated patients. The LASSO model was performed using the SAS (SAS Institute Inc) HPGENSELECT procedure using the Bayesian information criterion as the selection criterion, and setting the stop criterion to "none." The LASSO-reduced variable list was stored for use in the next stage of analysis.

Adjudicated patients were then randomly partitioned 2:1 into training and validation sets. Using only the training set, we fit a logistic regression model with the LASSO-reduced variable list, stored the model coefficients, estimated the C statistic, and produced a receiver operating characteristic (ROC) curve. We then applied the stored model coefficients to score the patients in the validation set and reestimated the C statistic. To assess model calibration, we divided both the training and validation sets into deciles of predicted probability and plotted the proportion of adjudicated true PAD within each decile.

After confirming that the model was performing similarly in both the training and validation sets, we recombined the sets and fit a final logistic model using the LASSO-reduced variable list and all adjudicated patients to obtain the final trained coefficient estimates. Using the predicted probabilities from the final trained logistic model, we evaluated potential discrimination thresholds to classify true presence of PAD.

We then scored the remaining, unadjudicated patients from the original data query using the final model coefficients. Patients with a predicted probability of ≥45% of truly having PAD were then included in the second round of chart abstraction. This threshold was chosen to favor sensitivity and was based on examination of both the ROC curve and the distribution of the predicted probabilities among these patients. To further validate the model performance, we also evaluated the concordance between the predicted PAD probability and actual PAD presence for each level of predicted probability.

Figure 1. Overview of the chart abstraction and analysis process. LASSO: least absolute shrinkage and selection operator; PAD: peripheral artery disease.


Health system-wide data
Initial query
31 ICD-9-CM +
375 ICD-10-CM
Initial cohort
Random 2500 patients
Other 12,906 patients
84 patients undetermined
297 unused and low-frequency codes
22 patients without represented codes
First round adjudication
2394 patients with adjudicated PAD status
LASSO model
Variables for logistic regression
Logistic model creation
2/3 training
Logistic model validation
1/3 validation
Logistic model calibration
Score remaining patients
Chart abstraction for >0.45 predicted probability of PAD
Second round adjudication

Results

Overview

In the initial data pull from the DUHS EHRs data repository, we identified 15,406 patients who had ≥1 clinical encounter within the health system during the period from January 1, 2015, through March 31, 2016, that was coded with one of the 406 PAD-related ICD-9-CM and ICD-10-CM diagnosis codes. Of the 2500 patients who were randomly selected for the first round of chart abstraction, 2416 had a definitive “yes” or “no” decision adjudicated. The remaining 84 patients were considered “undetermined” due to insufficient evidence in the charts, and were removed from the cohort.

Initial Code Inclusion and Exclusion Decisions

We began the analysis with 406 PAD-related ICD diagnosis codes (31 ICD-9-CM, 375 ICD-10-CM). Of those 406 codes, there were 247 codes that were not assigned during the study period to any of the 2416 clinically adjudicated patients. Additionally, 35 codes were assigned to only 1 patient, and 15 codes were assigned to only 2 patients during the study period. One of the most common diagnoses was “Peripheral vascular disease, unspecified,” which is 443.9 in ICD-9-CM and I73.9 in ICD-10-CM; 1190 (49.7%) of all patients had either one or both of those codes present during the study period. We grouped these two codes into a single flag because of their prevalence and because the terminology for this code did not change with the ICD-10-CM transition. Prior to LASSO modeling, we removed 297 unused and very low frequency (1-2 uses) diagnosis codes and combined the flag for 443.9/I73.9, leaving 108 diagnosis code flags to be included in the LASSO model.

In addition, 22 patients who no longer had any of the retained diagnosis codes were removed, leaving an analysis cohort of 2394 patients with adjudicated yes/no PAD status (Figure 1). Among these 2394 patients, only 780 (32.6%) were adjudicated as having confirmed PAD (Table 1; baseline characteristics by training versus validation roles available in Multimedia Appendix 3).

Table 1. Baseline characteristics of initially-adjudicated patients by confirmed peripheral artery disease status.

Demographics	Patients without confirmed peripheral artery disease (n=1614)	Patients with confirmed peripheral artery disease (n=780)	*P* value
Age (years), mean (SD)	66.9 (15.0)	69.8 (10.9)	.001
Gender (male), n (%)	824 (51.1)	464 (59.5)	<.001
Race, n (%)			<.001
White	1161 (71.9)	464 (59.5)	N/A[a]
Black/African American	380 (23.5)	252 (32.3)	N/A
Other	73 (4.5)	64 (8.2)	N/A

[a] N/A: not applicable.

Model Construction and Evaluation

We first assessed multicollinearity by fitting a linear regression model and evaluating the variance inflation factor for each of the 108 retained diagnosis flags and the 4 other indicator variables. Most of the variance inflation factor values were below 1.5, and the maximum VIF was 2.85, indicating that the variables in the model were sufficiently noncollinear to proceed, using a rule of thumb of <3.

We then entered the 108 diagnosis code flags and the 4 administrative flags for revascularization, diagnostic testing, specialist service, and ≥2 PAD-related encounters into the LASSO logistic regression prediction model. This yielded 15 flags for inclusion, including all 4 administrative flags and 11 diagnosis code flags.

Using the 15 LASSO-selected variables, we fit another logistic regression model to the adjudicated training set (2/3 partition, n=1604). Odds ratios and 95% confidence intervals from this training model are presented in Table 2.

In the training set, the C statistic was 0.8618 (95% CI 0.8427-0.8810). We then applied the model coefficients derived from the training set to score the observations in the adjudicated validation set (1/3 partition, n=790). In the validation set, the C statistic was 0.8618 (95% CI 0.8352-0.8884). Figure 2 displays the ROC curves for both the training and validation sets. Additionally, we ranked both the training and validation sets into deciles of predicted probability and plotted the relationship between the mean predicted probability in each decile to the observed prevalence of confirmed PAD in that decile (Figure 2). Overall, it appeared that the model derived from the training set fit the validation data equally well.

Finally, we refit the model using all 2394 PAD-adjudicated patients to obtain the final odds ratios, which are displayed in Table 2. The C statistic for the area under the ROC curve for this final model was 0.8618 (95% CI 0.8463-0.8774). We then generated a classification table to assess the impact of potential thresholds of predicted probability on the discrimination measures. At a threshold of predicted probability ≥0.45, the estimated sensitivity was 75.3% and the estimated specificity was 81.7%, with an estimated positive predictive value of 66.5% and negative predictive value of 87.2%.

We then applied the final model coefficients to score the remaining 12,801 patients from the original data pull. Of these patients, 4753 (37.1%) had a predicted PAD probability of ≥0.45 (Figure 3). PAD status was definitively adjudicated in 4493 patients and 260 patients were assigned an “Undetermined” status. Of the 4493 patients, 2981 (66.3%) were confirmed to have PAD. Figure 4 illustrates the proportion of patients who had confirmed PAD at each level of predicted probability.

Table 2. Odds ratio estimates and 95% confidence intervals from the training set and the final model using the variables selected in the least absolute shrinkage and selection operator (LASSO) model.

Diagnosis code or flag type	ICD[a] version	ICD code description or study definition	Training set (n=1604), odds ratio (95% CI)	Final model (n=2394), odds ratio (95% CI)
250.70	9	Diabetes with peripheral circulatory disorders, type II or unspecified type, not stated as uncontrolled	1.81 (0.78-4.24)	1.62 (0.78-3.39)
440.20	9	Atherosclerosis of native arteries of the extremities, unspecified	2.04 (0.79-5.29)	3.62 (1.60-8.19)
440.21	9	Atherosclerosis of native arteries of the extremities with intermittent claudication	6.28 (2.84-13.92)	5.81 (3.07-10.97)
440.23	9	Atherosclerosis of native arteries of the extremities with ulceration	24.18 (4.48-130.45)	13.73 (3.22-58.64)
440.9	9	Generalized and unspecified atherosclerosis	0.73 (0.36-1.49)	0.71 (0.39-1.30)
444.22	9	Arterial embolism and thrombosis of lower extremity	2.14 (0.67-6.78)	4.37 (1.67-11.48)
707.10	9	Ulcer of lower limb, unspecified	0.29 (0.12-0.70)	0.41 (0.21-0.80)
785.4	9	Gangrene	2.59 (0.89-7.48)	2.71 (1.11-6.59)
I702.13	10	Atherosclerosis of native arteries of extremities with intermittent claudication, bilateral legs	103.23 (11.78-904.75)	20.55 (5.97-70.80)
I96	10	Gangrene, not elsewhere classified	2.61 (0.87-7.86)	2.15 (0.84-5.51)
443.9 or I739	9 and 10	Peripheral vascular disease, unspecified	13.28 (9.53-18.50)	14.22 (10.77-18.77)
Specialist	N/A[b]	Any PAD-related specialist during study period	1.64 (1.22-2.19)	1.66 (1.31-2.10)
Revascularization	N/A	Any revascularization procedure during study period	3.38 (1.60-7.14)	2.37 (1.32-4.26)
Diagnostic imaging	N/A	Any PAD-related diagnostic imaging test during study period	1.08 (0.71-1.65)	0.99 (0.70-1.40)
≥2 encounters	N/A	2 PAD-related encounters during study period	1.70 (1.26-2.28)	1.86 (1.46-2.36)

[a] ICD: International Classificiation of Diseases.

[b] N/A: not applicable.

Figure 2. Comparison of results from training and validation sets. Left panel: receiver operating characteristic curves for training and validation sets. Right panel: Comparison of deciles of predicted probabilities in training set versus validation set. PAD: peripheral artery disease.

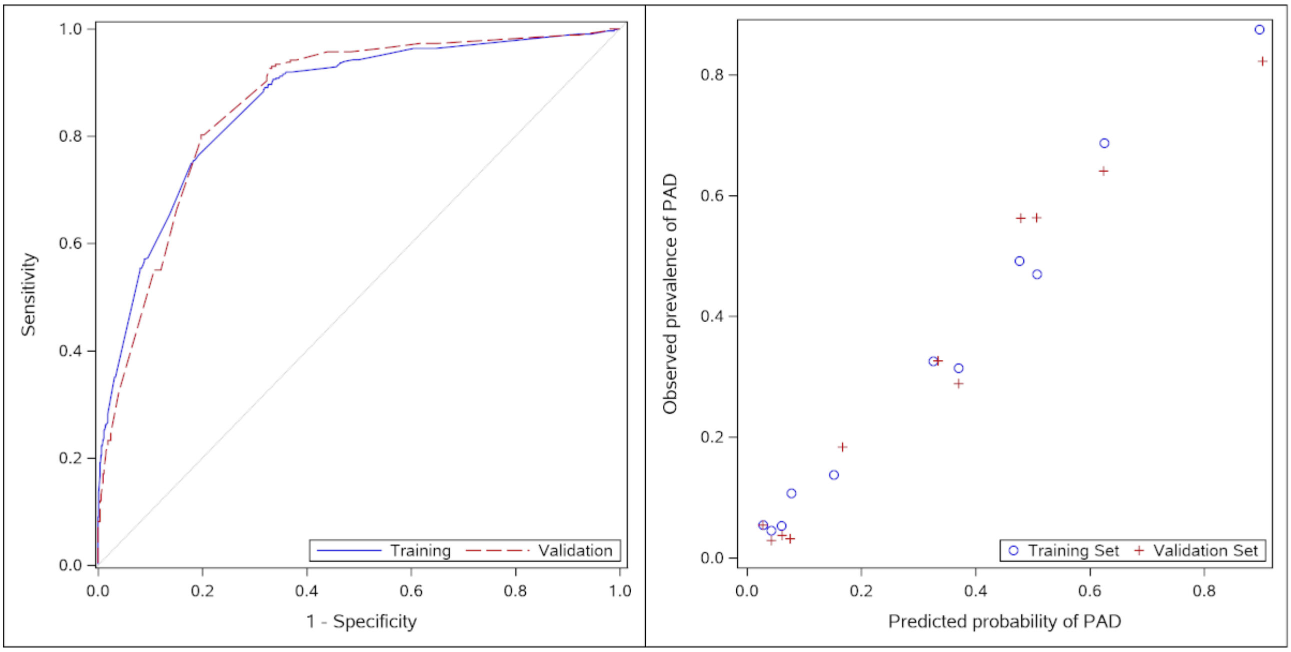


Figure 3. Histogram of predicted PAD probabilities in the remaining unadjudicated patients using the final model coefficients. The distribution of predicted PAD probabilities contributed to a chosen probability threshold of 0.45 for second round chart adjudication. PAD: peripheral artery disease.

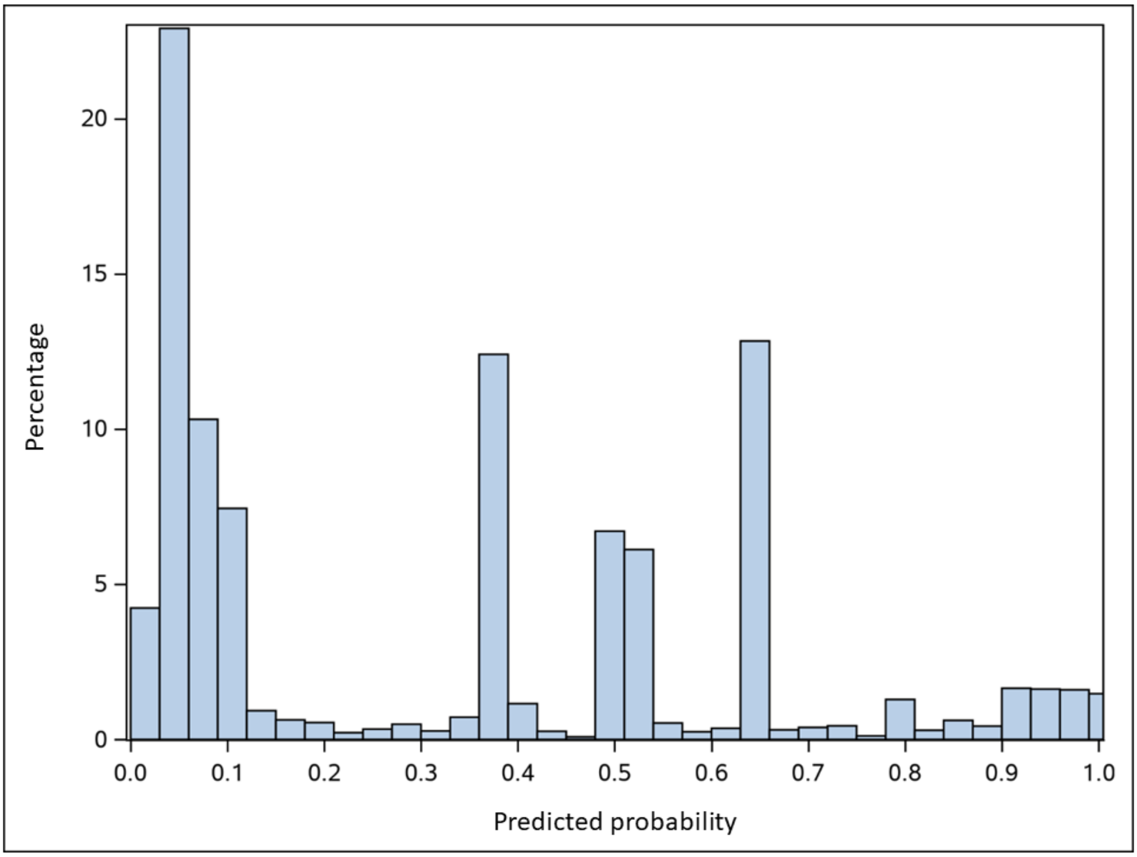
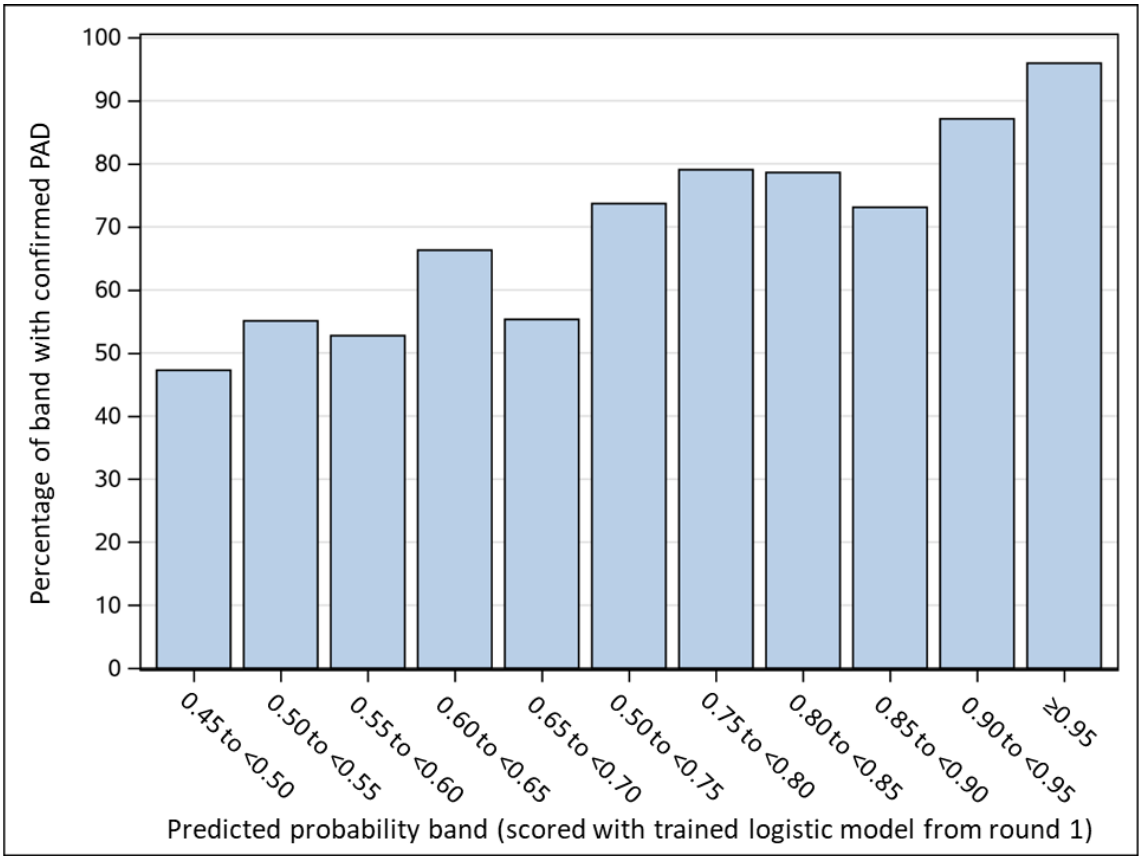


Figure 4. Model performance. Proportion of patients who were selected for the second round of abstraction who were confirmed to have PAD via abstraction, by bands of the predicted probabilities obtained with the trained logistic model. PAD: peripheral artery disease.



Discussion

Principal Results

We created a model-based algorithm for detecting PAD because diagnosis codes were an unacceptably low-yield way to find patients with PAD in our health system's EHRs. Out of 2394 randomly selected patients from our initial query of 15,406 patients with PAD codes, only 780 actually had PAD, a positive predictive value of 32.6%. Faced with the prohibitively labor-intensive process of chart extraction and adjudication for thousands of patients in our initial EHR-identified cohort with low probabilities for true PAD (despite the presence of PAD codes), we attempted to improve our yield using the LASSO approach for selecting administrative codes and flags most predictive of PAD. Our final model included 10 individual ICD-9/10 CM codes, one combined ICD-9+10 CM code, and flags for visits with PAD-related specialists, prior revascularizations, PAD-related diagnostic imaging, and ≥2 PAD-related encounters during the study period. This model had a C statistic of 0.8618. When we applied the full model to the remaining 12,801 patients and abstracted PAD status from those predicted to have a 45% or greater chance of PAD, we found that our yield of true PAD diagnoses tracked with the underlying predicted probability of PAD, as seen in Figure 4. That is, roughly 45% of the patients predicted to have a 45% probability of PAD actually had PAD and approximately 95% of the patients predicted to have a 95% probability of PAD actually had PAD.

We are currently using the cohort derived from our code- and administrative data–based model to analyze patient, provider, and health system factors associated with PAD care and outcomes in our health system. We set a 45% threshold for manual chart abstraction both because of the underlying characteristics of the model and because we wanted to derive a cohort that broadly reflected all patients with PAD in our health system. Depending on a researcher's goals, the threshold for inclusion or chart abstraction can be adjusted accordingly to favor sensitivity (using lower probability threshold) or specificity (using higher probability threshold) as needed, allowing for more efficient cohort construction. For instance, if the model were applied to a larger population with the intent to find patients for a PAD-related interventional study, a researcher might choose to increase the threshold probability to obtain a more specific though less broadly representative cohort with less manual effort.

Comparison With Prior Work

The use of diagnostic codes in administrative data sets is an appealing method of identifying patients with PAD, but it can be challenging. Although the use of PAD-associated procedure codes generally is sensitive and specific for the subgroup of patients with PAD undergoing a given procedure, diagnosis codes alone have poor predictive value [6,15]. We believe our model combining administrative data with diagnosis codes offers two main advantages.

First, the training population we used to build the model is representative of all patients with PAD at our institution, regardless of what location they received care in, what care they received, and who provided the care. This has not been true of prior similar efforts, which have used preexisting groups of patients with known PAD status from which to construct their models. For instance, Fan et al [7] designed and tested an administrative code–based algorithm in a population of 22,723 Mayo Clinic patients with PAD codes who underwent ABI testing. Their model, which included diagnosis codes, imaging procedures, and toe amputation, yielded an area under the curve of 0.912 in a test subset of the initial vascular lab population. However, when tested in a community sample, the sensitivity dropped from 85.5% to 68%. Hong et al [8] pooled patients from two prior prospective trials that had collected ABIs to create a cohort of patients with known PAD status from whom to construct various models combining diagnosis and procedure codes. They tested their models' abilities to find the patients already known to have PAD within administrative data sets, reaching a maximum sensitivity of 34.7%. Bekwelem et al [16] used a similar approach to discriminate between patients with and without critical limb ischemia (a more specific and severe kind of PAD) in a preadjudicated database and reported a maximum sensitivity of 92% by using either diagnosis or procedure codes. They then applied their model to unadjudicated health system data, but never confirmed their findings. We believe that training our model in a cohort containing diverse representations of PAD is a significant benefit for its applicability in multiple circumstances.

Second, we believe that another strength of the algorithm is its use of structured data. Though there have been some reports of natural language processing for PAD cohort identification [16,17], free text is not always available, either in adequate amounts to train an algorithm or at all for a given study population of interest. Examples of free text–limited circumstances include feasibility analyses for future studies, the construction of cohorts for further investigation, and research carried out entirely in an administrative context. Natural language processing approaches also require more time, expertise, and computing resources.

Limitations

Our approach does have some limitations. To maximize sensitivity, our initial DEDUCE query included a large number of PAD diagnosis codes, some of which were not used for any patient in the DEDUCE cohort or were used for only 1 or 2 patients. This may have driven down our initial positive PAD yield rate. Furthermore, we chose to treat each ICD-CM code as an individual flag because the mapping between ICD-9 and ICD-10 was not entirely concordant except in the instance of 443.9 and I73.9 (unspecified peripheral vascular disease). ICD-10 codes for PAD often specify the disease state followed by a specific affected anatomic location. Rather than treat these codes independently, we could have combined all of the ICD-10 codes with similar disease processes across multiple anatomic locations. This may have increased the likelihood for some ICD codes on the margin to make it into the model. In addition, we did not have access to revascularization procedures prior to 2010, which may have minimally decreased the ability of our model to find patients with PAD. The final and most significant limitation of our approach is that, thus far, we have validated it only internally, and are therefore unsure of how it will perform in different EHRs and health systems. As we look toward deploying this model as part of collaborative research with other institutions, we will need to remain vigilant for signs of model performance degradation. Furthermore, deployment in other health systems will require some level of chart adjudication for validation, the necessary amount of which will be determined on the basis of the threshold chosen, intended cohort use, and initial performance in the new health system.

Conclusions

We selected all patients from an entire health system with PAD-related diagnosis codes between January 1, 2015, and March 31, 2016. Using a random subset of patients, we constructed a code- and administrative data–based model including 10 individual ICD-9/10 CM flags, one combined ICD-9+10 CM flag, and flags for visits with PAD-related specialists, prior revascularizations, PAD-related diagnostic imaging, and ≥2 PAD-related encounters during the study period. This model had a C statistic of 0.8618. Use of only nonselective PAD diagnosis codes to identify patients for research purposes is unacceptably nonspecific for many studies and should not be done without supplementary methods of cohort confirmation.

Acknowledgments

This work was supported by the Doris Duke Charitable Foundation (Grant 2016091). EHW receives support from a R38 HL143612/HL/NHLBI NIH HHS/United States.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ICD-9-CM and ICD-10-CM diagnosis codes included in initial data pull: 31 ICD-9-CM codes, 375 ICD-10-CM codes.
[DOCX File , 21 KB - medinform_v8i8e18542_app1.docx]

Multimedia Appendix 2
Procedure codes for lower extremity revascularization (ICD-9-CM, ICD-10-PCS, and CPT procedure codes) and procedure codes for diagnostic imaging.
[DOCX File , 40 KB - medinform_v8i8e18542_app2.docx]

Multimedia Appendix 3
Table of baseline demographics by training versus validation roles.
[DOCX File , 15 KB - medinform_v8i8e18542_app3.docx]

References

1. Fowkes FGR, Aboyans V, Fowkes FJI, McDermott MM, Sampson UKA, Criqui MH. Peripheral artery disease: epidemiology and global perspectives. Nat Rev Cardiol 2017 Mar;14(3):156-170. [doi: 10.1038/nrcardio.2016.179] [Medline: 27853158]
2. Hiramoto JS, Teraa M, de Borst GJ, Conte MS. Interventions for lower extremity peripheral artery disease. Nat Rev Cardiol 2018 Jun;15(6):332-350. [doi: 10.1038/s41569-018-0005-0] [Medline: 29679023]
3. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-Treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken) 2010 Aug;62(8):1120-1127 [FREE Full text] [doi: 10.1002/acr.20184] [Medline: 20235204]
4. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc 2012 Jun;19(e1):e162-e169 [FREE Full text] [doi: 10.1136/amiajnl-2011-000583] [Medline: 22374935]
5. Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. J Am Med Inform Assoc 2010;17(5):568-574 [FREE Full text] [doi: 10.1136/jamia.2010.004366] [Medline: 20819866]
6. Lasota AN, Overvad K, Eriksen HH, Tjønneland A, Schmidt EB, Grønholdt MM. Validity of Peripheral Arterial Disease Diagnoses in the Danish National Patient Registry. Eur J Vasc Endovasc Surg 2017 May;53(5):679-685 [FREE Full text] [doi: 10.1016/j.ejvs.2016.12.031] [Medline: 28187995]
7. Fan J, Arruda-Olson AM, Leibson CL, Smith C, Liu G, Bailey KR, et al. Billing code algorithms to identify cases of peripheral artery disease from administrative data. J Am Med Inform Assoc 2013 Dec;20(e2):e349-e354 [FREE Full text] [doi: 10.1136/amiajnl-2013-001827] [Medline: 24166724]
8. Hong Y, Sebastianski M, Makowsky M, Tsuyuki R, McMurtry MS. Administrative data are not sensitive for the detection of peripheral artery disease in the community. Vasc Med 2016 Aug;21(4):331-336. [doi: 10.1177/1358863X16631041] [Medline: 27114456]

9. Vemulapalli S, Greiner MA, Jones WS, Patel MR, Hernandez AF, Curtis LH. Peripheral arterial testing before lower extremity amputation among Medicare beneficiaries, 2000 to 2010. Circ Cardiovasc Qual Outcomes 2014 Jan;7(1):142-150. [doi: 10.1161/CIRCOUTCOMES.113.000376] [Medline: 24425703]
10. Jones WS, Mi X, Qualls LG, Vemulapalli S, Peterson ED, Patel MR, et al. Trends in settings for peripheral vascular intervention and the effect of changes in the outpatient prospective payment system. J Am Coll Cardiol 2015 Mar 10;65(9):920-927 [FREE Full text] [doi: 10.1016/j.jacc.2014.12.048] [Medline: 25744009]
11. 2016 ICD-10-CM and GEMs. Centers for Medicare and Medicaid Services. 2015 Oct 08. URL: https://www.cms.gov/Medicare/Coding/ICD10/2016-ICD-10-CM-and-GEMs [accessed 2020-02-17]
12. Goodney PP, Beck AW, Nagle J, Welch HG, Zwolak RM. National trends in lower extremity bypass surgery, endovascular interventions, and major amputations. J Vasc Surg 2009 Jul;50(1):54-60 [FREE Full text] [doi: 10.1016/j.jvs.2009.01.035] [Medline: 19481407]
13. Jaff MR, Cahill KE, Yu AP, Birnbaum HG, Engelhart LM. Clinical outcomes and medical care costs among medicare beneficiaries receiving therapy for peripheral arterial disease. Ann Vasc Surg 2010 Jul;24(5):577-587. [doi: 10.1016/j.avsg.2010.03.015] [Medline: 20579582]
14. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 2018 Dec 05;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
15. Mell MW, Pettinger M, Proulx-Burns L, Heckbert SR, Allison MA, Criqui MH, WHI PVD Writing Workgroup. Evaluation of Medicare claims data to ascertain peripheral vascular events in the Women's Health Initiative. J Vasc Surg 2014 Jul;60(1):98-105 [FREE Full text] [doi: 10.1016/j.jvs.2014.01.056] [Medline: 24636641]
16. Savova GK, Fan J, Ye Z, Murphy SP, Zheng J, Chute CG, et al. Discovering peripheral arterial disease cases from radiology notes using natural language processing. AMIA Annu Symp Proc 2010 Nov 13;2010:722-726 [FREE Full text] [Medline: 21347073]
17. Afzal N, Sohn S, Abram S, Scott CG, Chaudhry R, Liu H, et al. Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. J Vasc Surg 2017 Jun;65(6):1753-1761 [FREE Full text] [doi: 10.1016/j.jvs.2016.11.031] [Medline: 28189359]

Abbreviations

ABI: ankle-brachial index
CPT: Current Procedural Technology
DEDUCE: Duke Enterprise Data Unified Content Explorer
DUHS: Duke University Health System
EHR: electronic health record
ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification
ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification
LASSO: least absolute shrinkage and selection operator
PAD: peripheral artery disease
ROC: receiver operating characteristic

Edited by C Lovis; submitted 03.03.20; peer-reviewed by A Arruda-Olson, IY Jang; comments to author 07.06.20; revised version received 21.06.20; accepted 28.06.20; published 19.08.20.

Please cite as:
Weissler EH, Lippmann SJ, Smerek MM, Ward RA, Kansal A, Brock A, Sullivan RC, Long C, Patel MR, Greiner MA, Hardy NC, Curtis LH, Jones WS
Model-Based Algorithms for Detecting Peripheral Artery Disease Using Administrative Data From an Electronic Health Record Data System: Algorithm Development Study
JMIR Med Inform 2020;8(8):e18542
URL: http://medinform.jmir.org/2020/8/e18542/
doi:10.2196/18542
PMID:32663152

Original Paper

Analysis of Benzodiazepine Prescription Practices in Elderly Appalachians with Dementia via the Appalachian Informatics Platform: Longitudinal Study

Niharika Bhardwaj[1], MBBS, MS; Alfred A Cecchetti[1], MSc, MSIS, PhD; Usha Murughiyan[1], MBBS; Shirley Neitch[2], MD, FACP, AGSF

[1]Department of Clinical and Translational Science, Joan C Edwards School of Medicine, Marshall University, Huntington, WV, United States
[2]Department of Internal Medicine, Joan C Edwards School of Medicine, Marshall University, Huntington, WV, United States

Corresponding Author:
Niharika Bhardwaj, MBBS, MS
Department of Clinical and Translational Science
Joan C Edwards School of Medicine
Marshall University
1600 Medical Center Drive
Suite 265
Huntington, WV, 25701
United States
Phone: 1 304 691 5397
Email: bhardwaj1@marshall.edu

Abstract

Background: Caring for the growing dementia population with complex health care needs in West Virginia has been challenging due to its large, sizably rural-dwelling geriatric population and limited resource availability.

Objective: This paper aims to illustrate the application of an informatics platform to drive dementia research and quality care through a preliminary study of benzodiazepine (BZD) prescription patterns and its effects on health care use by geriatric patients.

Methods: The Maier Institute Data Mart, which contains clinical and billing data on patients aged 65 years and older (N=98,970) seen within our clinics and hospital, was created. Relevant variables were analyzed to identify BZD prescription patterns and calculate related charges and emergency department (ED) use.

Results: Nearly one-third (4346/13,910, 31.24%) of patients with dementia received at least one BZD prescription, 20% more than those without dementia. More women than men received at least one BZD prescription. On average, patients with dementia and at least one BZD prescription sustained higher charges and visited the ED more often than those without one.

Conclusions: The Appalachian Informatics Platform has the potential to enhance dementia care and research through a deeper understanding of dementia, data enrichment, risk identification, and care gap analysis.

(JMIR Med Inform 2020;8(8):e18389) doi:10.2196/18389

KEYWORDS

dementia; Alzheimer disease; benzodiazepines; Appalachia; geriatrics; informatics platform; interactive visualization; eHealth; clinical data

Introduction

Dementia is the fifth leading cause of death among people older than 65 in the United States [1]. The prevalence of dementia has been escalating, especially in West Virginia, a state with one of the highest percentages of older adults in its population [2]. Not only that, but more than half (52.5%) of these older adults also reside in rural areas [3]. As of early 2019, an estimated 38,000 people with Alzheimer's disease (AD) were living in West Virginia, and this number is expected to increase to 44,000 by 2025 [4]. Although age is the greatest risk factor for AD, comorbidities such as stroke, cardiovascular disease, smoking, high cholesterol, obesity, poor nutrition, physical inactivity, and diabetes also contribute to the disease burden in AD [5]. According to the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System, West Virginia has been ranked among the worst of the 50 states and

District of Columbia in the prevalence of smoking, diabetes, hypertension, and obesity [6]. Associated contributory factors to dementia, such as excessive prescription of medications (eg, benzodiazepines [BZD]), poor rates of health screening, and high illiteracy, have also been found to be highly prevalent in West Virginia and the rest of the Appalachia [7-9]. Thus, caring for patients with dementia is challenging, especially in Appalachia, because of the complexities that arise due to the increased burden of aforementioned comorbidities and contributory factors [10]. Moreover, 42% of the Appalachian population resides in remote and rural settings, limiting access to health care [11,12].

Technological advancements over the past few years have only added to the challenges by leading to the production of massive amounts of data, known as big data, originating from a wide variety of disparate sources, such as electronic health records (EHRs), specialized registries, smart home health devices, genomic data, etc [13-16]. In order to transform this siloed data into actionable knowledge to further dementia research and care, it is vital to connect them and create a longitudinal record across the care continuum. This can be achieved through the application of numerous current and emerging big data approaches for data storage, management, analytics, and mining [17]. These techniques offer benefits such as data quality, data structure, data accessibility, quality improvement, population management and health, early detection of disease, improved decision making, and cost reduction. However, they pose some challenges concerning security, infrastructure, ethics, and scientific evidence and theory that still need to be overcome [18-26].

The Appalachian Clinical and Translational Science Institute (ACTSI) at Marshall University Joan C Edwards School of Medicine (MU/JCESOM) has recently established the Maier Institute for Excellence in Therapeutics for Elders with Dementia, which aims to enhance patient care and advance research in AD and other dementias in the Appalachian elderly population. The focus of Maier Institute is on strategic approaches that will identify existing gaps and improve the quality of care for patients with dementia. The Maier Institute is dedicated to ensuring that every person with dementia receives optimal treatment through our discovery of new knowledge and dissemination of information regarding appropriate therapeutics.

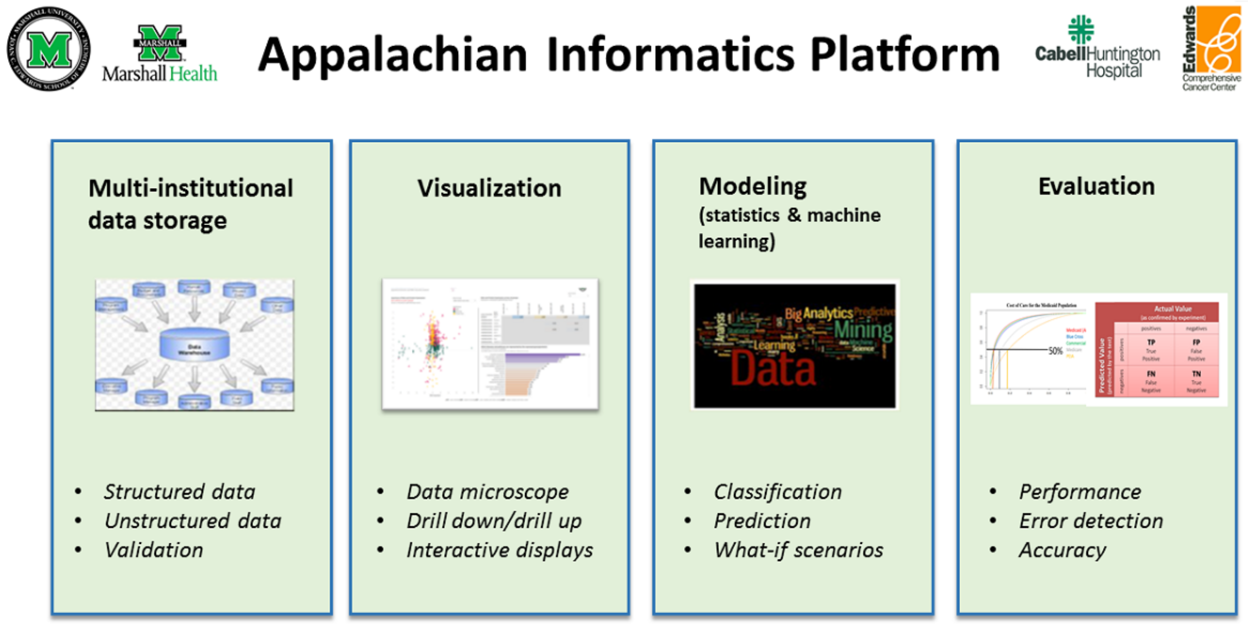
This paper describes one of the Maier Institute's approaches to improving therapeutics for this very vulnerable population using the Appalachian Informatics Platform, which will be very valuable going forward in our pursuit of the mission of Maier Institute. As a foundation for all future studies, the ACTSI Maier Institute Data Mart was built using the Clinical Data Warehouse (CDW), which will serve as a source of consolidated information across the care continuum for our geriatric population. Using data from this data mart, a pilot exploratory study was conducted.

Given the current climate of crisis in the use of controlled and addictive drugs and the potential for abuse of BZDs, we studied the prescribing patterns of the BZD class of medications within our patient population. While very useful when properly prescribed, BZDs can be harmful to elderly persons otherwise. Thus, the goal of this paper was to provide a better understanding of their use, which is critical to good clinical care, through the use of the Appalachian Informatics Platform, thereby demonstrating the value of this platform in driving dementia research.

Methods

The ACTSI's Division of Clinical Informatics has a functional Appalachian Informatics Platform (Figure 1) that is composed of 4 major components to be described in detail in a future paper. The CDW, containing over 9 years of billing and clinical data, forms an integral part of the Appalachian Informatics Platform, which contains, in addition to the CDW, embedded data analytics (modeling and evaluation) and interactive visualization tools (eg, Tableau [Tableau Software Inc] and Power BI [Microsoft Corp]). The information contained within the CDW consists of internal structured EHR data (eg, vitals, medication, procedure, diagnosis, etc), non-EHR survey data, and unstructured (text) information received from Marshall Health practice plan, Cabell Huntington Hospital (CHH), and MU/JCESOM’s Edwards Comprehensive Cancer Center. The source data are ingested daily incrementally through SQL Server Integration Services (Microsoft Corp). These data are tested and validated via standard extract, transform, load testing, which includes but is not limited to comparing data in production systems against source data, source to target data and count testing, metadata testing, and data quality testing (accepting default values, reporting invalid data, etc). Any missing values are recoded as unknown and outliers are corrected, if possible, or recoded as unknown. The CDW serves as a secure source of quality data for research studies and development and evaluation of machine learning algorithms. It also stores the results from the resulting machine learning model following its deployment. The visual analytical tools enable initial exploratory data analysis and interactive presentation of data as well as model information for further analysis and review. After an exploratory visual analysis, detailed statistical analysis is performed using a statistics application (eg, SPSS [IBM Corp], Stata [StataCorp], R [R Foundation for Statistical Computing]).

Figure 1. Components of the Appalachian Informatics Platform.



ACTSI's Division of Clinical Informatics accessed the ACTSI CDW to develop a data mart, called the ACTSI Maier Institute Data Mart. The ACTSI Maier Institute Data Mart comprises information about persons 65 years of age and older who have been seen at the geriatric clinic or were admitted to the primary hospital used by the school's medical practice. It is a regional data mart specially designed to capture the unique needs of the vulnerable population in Appalachia. It supports the inclusion of socioeconomic determinants of health that greatly affect the population in this area and is supplemented by a questionnaire to gather clinical information missing from the data warehouse that is critical to dementia research. For this study, we extracted the following data for each patient aged 65 years and older between 2010 and 2019: (1) use of BZD, (2) number of BZD prescriptions, (3) dementia status, (4) visits to the CHH Emergency Department (ED), (5) source of admission and discharge disposition for inpatient admissions if available, (6) patient ZIP code, and (7) charges incurred per visit for any services received through CHH or Marshall Health physicians.

Regarding use of BZD, if a patient received one or more prescriptions of BZD or reported taking a BZD as home medication at or after the age of 65 years, they were classified as a BZD user. Otherwise, they were classified as a nonuser. A list of generic drug names used to categorize a drug as BZD are listed in Multimedia Appendix 1.

To determine the number of BZD prescriptions, the number of BZD prescriptions ordered was included, but the number of refills was not taken into account. As long as the prescription number in the EHR stayed the same, it was counted as a single prescription.

Dementia status was determined by diagnosis code. Patients that had a diagnosis code (International Classification of Diseases, 9th Revision or International Classification of Diseases, 10th Revision code) for AD at any point between 2010 and 2019 were classified as having AD, those with diagnosis codes for dementia apart from AD were classified as having other dementia, and those without any dementia diagnosis codes were classified as having no dementia. The codes are listed in Multimedia Appendix 1.

Using this extracted data, we developed an interactive dashboard that was then used for initial exploratory analysis to help outline the BZD-prescribing patterns in this population.

Results

Between the years of 2010 and 2019, there were 98,952 patients aged 65 and older who received any service from Marshall University physicians in CHH or the ambulatory geriatric clinic. Over the span of those 10 years, $4.29 billion in total charges were accrued, with an average charge per patient of over $43,000. The mean number of ED visits per patient was 2.64. The geriatric population was predominantly female (54,887/98,952, 55.47%) with the prevalence of dementia reaching 14.06% (13,910/98,952) (see Figure 2).

Approximately 31.24% (4346/13,910) of patients with dementia received one or more BZD prescriptions compared with 11.22% (9540/85,042) of those without dementia. A slightly higher percentage of patients specifically diagnosed with AD (761/2251, 33.81%) were found to have at least one BZD prescription. Further, fewer men than women received at least one BZD prescription (4830/44,055, 10.96% vs 9056/54,887, 16.50%) (data not shown).

The percentage of elderly patients with any type of dementia receiving one or more BZD prescriptions declined appreciably (by about 10 percentage points in AD and by about 7 percentage points in other dementia from 2010 to 2019). A slight downward trend was also seen in patients with no dementia (see Figure 3).

Figure 2. Tableau dashboard displaying key information on the geriatric population, including heat map, dementia prevalence, benzodiazepine use, and trends in charges and visits by dementia status. BZD: benzodiazepine; ED: emergency department.

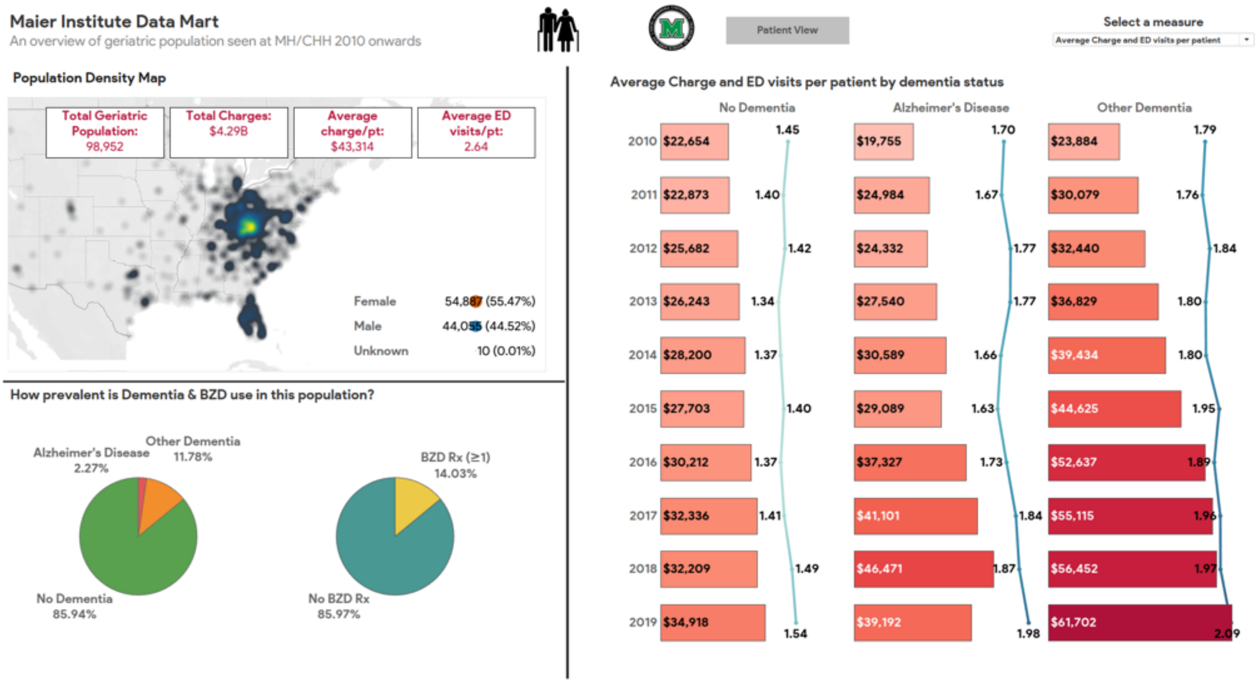
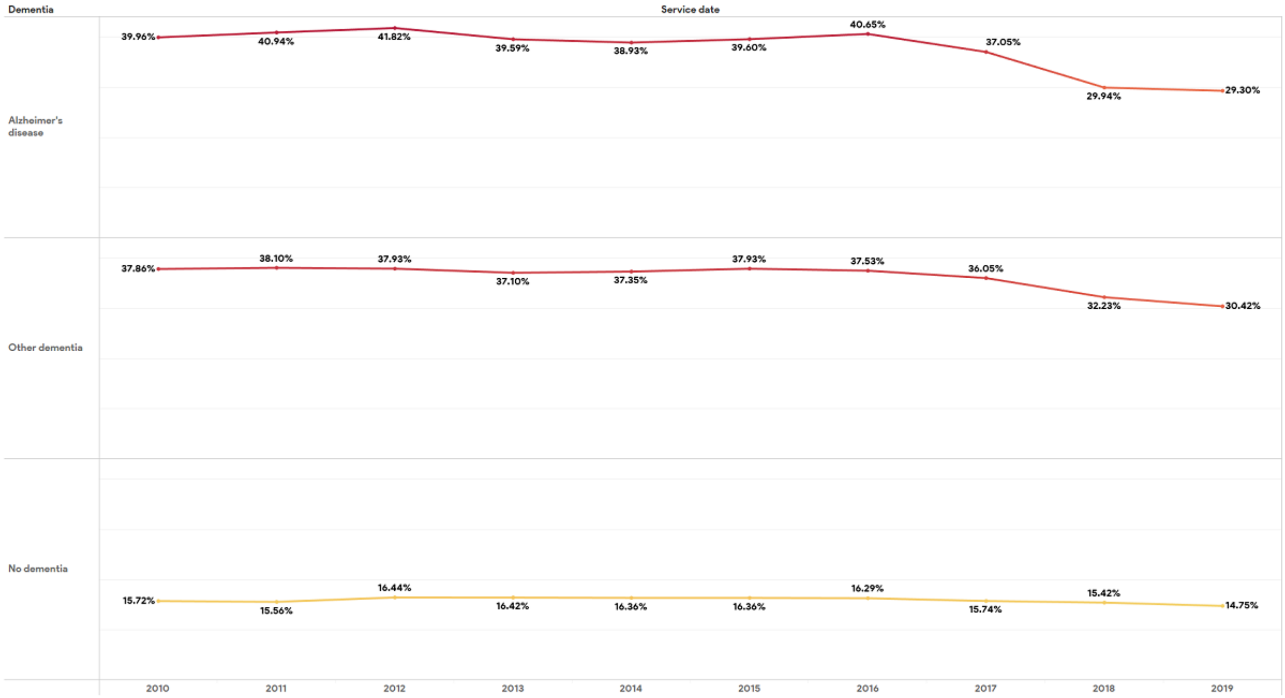


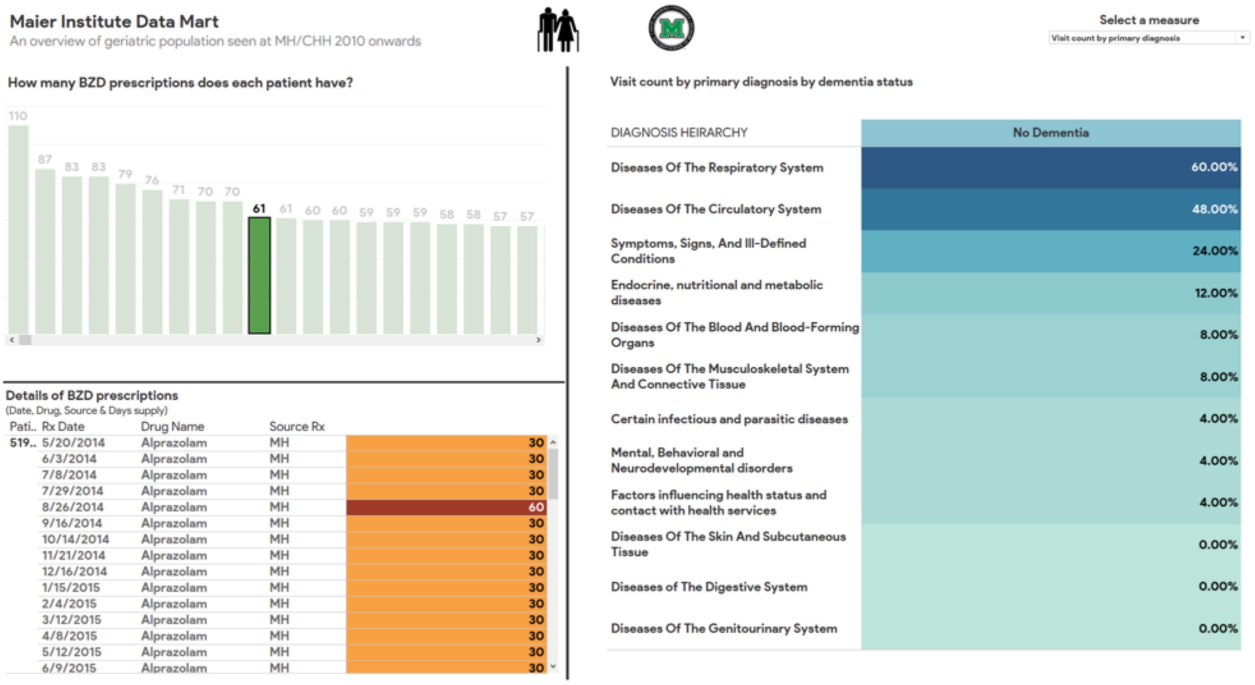
Figure 3. Tableau dashboard showing the trend in the percentage of patients with at least 1 benzodiazepine prescription between 2010 and 2019 by dementia status.



We also found that patients with other dementia and AD, on average, incurred charges 3.3 times (approximately $109,000) and 2.3 times (approximately $76,000) those incurred by patients without dementia (approximately $33,000), respectively. The patients with any type of dementia, on average, also visited the ED 83% more compared with those without dementia (3.82 vs 2.08 visits). The average charges and number of ED visits were even higher (33%-42% increase in average charges; 34%-54% increase in average ED visits) for patients with at least one BZD prescription compared with those without a BZD prescription for patients with or without dementia.

A patient-centered view also enabled patient-level analysis of patients' BZD prescription history and use of health care services over the years (see Figure 4).

Figure 4. Tableau dashboard showing a patient-level view for detailed analysis (the patient IDs have been deidentified). BZD: benzodiazepine.



Maier Institute Data Mart

An overview of geriatric population seen at MH/CHH 2010 onwards

Select a measure

Visit count by primary diagnosis

How many BZD prescriptions does each patient have?

Details of BZD prescriptions

(Date, Drug, Source & Days supply)

Pati..	Rx Date	Drug Name	Source Rx	
519..	5/20/2014	Alprazolam	MH	30
	6/3/2014	Alprazolam	MH	30
	7/8/2014	Alprazolam	MH	30
	7/29/2014	Alprazolam	MH	30
	8/26/2014	Alprazolam	MH	60
	9/16/2014	Alprazolam	MH	30
	10/14/2014	Alprazolam	MH	30
	11/21/2014	Alprazolam	MH	30
	12/16/2014	Alprazolam	MH	30
	1/15/2015	Alprazolam	MH	30
	2/4/2015	Alprazolam	MH	30
	3/12/2015	Alprazolam	MH	30
	4/8/2015	Alprazolam	MH	30
	5/12/2015	Alprazolam	MH	30
	6/9/2015	Alprazolam	MH	30

Visit count by primary diagnosis by dementia status

DIAGNOSIS HEIRARCHY	No Dementia
Diseases Of The Respiratory System	60.00%
Diseases Of The Circulatory System	48.00%
Symptoms, Signs, And Ill-Defined Conditions	24.00%
Endocrine, nutritional and metabolic diseases	12.00%
Diseases Of The Blood And Blood-Forming Organs	8.00%
Diseases Of The Musculoskeletal System And Connective Tissue	8.00%
Certain infectious and parasitic diseases	4.00%
Mental, Behavioral and Neurodevelopmental disorders	4.00%
Factors influencing health status and contact with health services	4.00%
Diseases Of The Skin And Subcutaneous Tissue	0.00%
Diseases of The Digestive System	0.00%
Diseases Of The Genitourinary System	0.00%

Discussion

Principal Findings and Comparison With Prior Work

The first project undertaken using the newly established ACTSI Maier Institute Data Mart explored the use of BZDs by elderly persons in our practices. Several interesting trends and patterns were noted, such as the higher prevalence of BZD use in patients with dementia and female geriatric patients and the higher mean ED visits and mean charges in patients with dementia plus at least one BZD prescription.

BZD use (receiving one or more BZD prescriptions) was found to be almost thrice as prevalent in elderly patients with dementia diagnoses compared with those without a dementia diagnosis (4346/13,910, 31.24% vs 9540/85,042, 11.22%). This is much higher than the estimate by a systematic review of past studies on BZD use in patients with AD, which estimated that 10% to 20% of these patients receive a BZD at least once during the course of the disease [27]. However, since most of the studies included in the review occurred more than 6 years ago and were heterogeneous regarding patient populations and disease stages of AD, the estimate may not accurately reflect the true prevalence of BZD use in patients with AD. Further, even though the percentage of elderly patients and patients within dementia status subgroups that received one or more BZD prescription dropped overall, the average number of prescriptions per patient rose. However, since we did not account for the number of refills per prescription and quantity of supply, it is hard to assess the implication of this finding.

Past studies have found BZD use to be more prevalent in women than in men [28,29]. This is consistent with our finding of a larger percentage of women receiving a BZD prescription compared with men.

Another study found patients with AD, in general, had more ED visits and were more likely to have a BZD-related adverse drug event, but had similar mean charges for ED visits when compared with patients without AD [30]. We found that on average, even in our population, patients with AD and other dementia visited the ED more often, but they also had higher mean charges compared with patients without dementia. Additionally, in patients receiving at least one BZD prescription over their lifetime, these numbers were even higher. This difference in charges could be because their study focused only on ED and inpatient charges, while our study included outpatient charges as well. A detailed investigation is needed to determine whether BZD use contributed to the increase in charges and ED visits.

A detailed analysis of the data is underway to better understand our initial findings, but this paper demonstrates the value of searching the data through the Appalachian Informatics Platform and exploring said data interactively with Tableau. Generally, health care providers that serve rural and indigent populations in Appalachia do not have the resources to gather and analyze quality data to understand the unique needs of the patients with dementia in this region, which as a result, remain largely unknown. In this paper, we have exemplified that these health care providers can, despite limited resources, develop and use inexpensive data warehousing and visualization tools with a small clinical informatics team to explore their data to obtain a comprehensive and near real-time picture of the current state of dementia care. This will help them identify and address care gaps, driving dementia research and quality care in the geriatric Appalachian population.

This study adds to the limited knowledge of dementia care in Appalachia through this effort. The authors hope to improve the understanding of the effect of the geographical, environmental, cultural, and socioeconomic factors on dementia care in Appalachia and in rural populations affected by similar factors through future studies.

Limitations

We relied only on the presence of diagnosis codes specific to dementia in the billing systems to identify whether a patient had dementia. Thus, it is possible that some of the patients had dementia that was not documented in the billing system or that a dementia diagnosis was documented but later found to have been erroneous. Further, we determined BZD use based on whether a patient received a prescription for BZD or reported a BZD as a home medication. This may not indicate the actual use of BZD, since we do not know whether the patient actually filled the prescription and took the drug. Also, patients may have received BZD and other services outside our clinic or hospital. This may have resulted in an underestimation of BZD use, ED visits, and total charges.

Future Directions: Incorporation of a Clinical Questionnaire

During this initial analysis of the ACTSI Maier Institute Data Mart, some patterns and trends emerged that warrant more detailed analysis. In the second phase of our use of the data mart, we plan to explore the clinical features of dementia. Much is known about risk factors for dementia, yet how the known risks act and interact in individual cases and whether there are other factors that contribute to the development of dementia remain unknown. Specifically, long-term clinical care of large numbers of patients with dementia by one author (SN) has raised the question of whether there could be risks specific to any of the ethnic/demographic groups in the Appalachian region of the United States or to persons with a history of exposure to environmental aspects of Appalachia.

Since the Maier Institute Data Mart is designed to serve as a repository of searchable information about these questions, we are developing a data collection instrument to gather in-depth clinical information with a more detailed patient background. This will help advance the clinical care of patients who seek evaluation and ongoing care at MU/JCESOM's cognitive assessment clinic, known as the Susan Edwards Drake Memory Clinic. The Susan Edwards Drake Memory Clinic Questionnaire (SEDQ) will consist of 3 sections—demographics, health history, and dementia evaluation—with the resulting data integrated into the data warehouse.

In addition to commonly needed general information, which would be sought by any clinical practice, fields that will be included in the SEDQ are questions targeting the patient's past experiences in order to gain a closer insight into specific factors that could have contributed to suspected dementia. Examples include "Where did the patient grow up? Where has the patient lived for the longest time as an adult?" "Has the patient had any past experiences with trauma or exposures?" and "Does the patient use tobacco in any form? If so, which form?"

As data accumulate in the data mart, it is expected that patterns of personal backgrounds, histories of substance usage, or toxic exposures will become apparent, and the Maier Institute Data Mart will be available for further investigation and analysis. Additionally, clinical questions that arise regarding patients of this demographic group whose information is not in the database (for a variety of reasons, for example, patients were evaluated prior to the institution of the SEDQ) can be evaluated and compared with the deidentified data of patients who are in this data mart.

The SEDQ itself will expedite the first office visit by collecting basic demographics and health history and evaluating the patient's current mental state, as would any good previsit data tool. The SEDQ will provide even more relevant patient functionality information that is often lacking in the care of patients with dementia and their caregivers. With the responses provided on the survey, the clinician will be able to gauge the patient's functional level through their ability to perform activities of daily living, including basic activities such as bathing and instrumental activities such as managing finances. This will improve the quality of patient care and allow the practitioner to formulate a treatment plan more efficiently. Ultimately, this process can also be examined by researchers to compare outcomes of persons evaluated at the Susan Edwards Drake Memory Clinic with the outcomes of persons evaluated and cared for elsewhere.

Conclusions

This paper serves as a leading example of the potential ways that informatics-based research powered by the Appalachian Informatics Platform can help enhance patient care for people with dementia in Appalachia. The platform has helped improve our understanding of certain problem areas within our elderly population. We hope that it will benefit care and treatment for future patients with dementia by way of improved understanding of dementia, enhancement of existing data using data collection instruments, risk identification, care gap analysis, and comparative analysis of treatment modalities.

Acknowledgments

This work was supported by the National Institutes of Health grants DK-67420, DK-108054, and P20GM121299-01A1; US Department of Veterans Affairs Merit Review grant BX003443-01 to U Sundaram and UL1TR00011719 to P Kern; and grants from the Claude Worthington Benedum Foundation and West Virginia Higher Education Policy Commission.

Conflicts of Interest

None declared.

Multimedia Appendix 1

List of diagnosis codes and drug names used in the study.
[DOC File , 30 KB - medinform_v8i8e18389_app1.doc]

References

1. Heron M. Deaths: Leading Causes for 2017. Natl Vital Stat Rep 2019 Jun;68(6):1-77 [FREE Full text] [Medline: 32501203]
2. Martin JA, Hamilton BE, Osterman MJK, Driscoll AK, Mathews TJ. Births: Final Data for 2015. Natl Vital Stat Rep 2017 Jan;66(1):1 [FREE Full text] [Medline: 28135188]
3. Smith A, Trevelyan E. The Older Population in Rural America: 2012-2016. American Community Survey Reports. Washington, DC: US Census Bureau; 2019 Sep 23. URL: https://www.census.gov/library/publications/2019/acs/acs-41.html [accessed 2020-07-06]
4. Alzheimer's Association. 2019 Alzheimer's disease facts and figures. Alzheimer's & Dementia 2019 Mar 01;15(3):321-387. [doi: 10.1016/j.jalz.2019.01.010]
5. de Toledo Ferraz Alves TC, Ferreira LK, Wajngarten M, Busatto GF. Cardiac disorders as risk factors for Alzheimer's disease. J Alzheimers Dis 2010;20(3):749-763. [doi: 10.3233/JAD-2010-091561] [Medline: 20413875]
6. West Virginia Behavioral Risk Factor Surveillance System Report, 2016. West Virginia Department of Health and Human Resources, Health Statistics Center. 2018. URL: http://www.wvdhhr.org/bph/hsc/pubs/brfss/2016/BRFSS2016.pdf [accessed 2020-07-06]
7. Rochon PA, Vozoris N, Gill SS. The harms of benzodiazepines for patients with dementia. CMAJ 2017 Apr 10;189(14):E517-E518 [FREE Full text] [doi: 10.1503/cmaj.170193] [Medline: 28396327]
8. Sharp ES, Gatz M. Relationship between education and dementia: an updated systematic review. Alzheimer Dis Assoc Disord 2011;25(4):289-304 [FREE Full text] [doi: 10.1097/WAD.0b013e318211c83c] [Medline: 21750453]
9. PDA Inc, The Cecil G. Sheps Center for Health Services Research, Appalachian Regional Commission. Health Disparities in Appalachia. Creating a Culture of Health in Appalachia: Disparities and Bright Spots. 2017 Aug. URL: https://www.arc.gov/assets/research_reports/Health_Disparities_in_Appalachia_August_2017.pdf [accessed 2020-07-06]
10. Boustani M, Schubert C, Sennour Y. The challenge of supporting care for dementia in primary care. Clin Interv Aging 2007;2(4):631-636 [FREE Full text] [doi: 10.2147/cia.s1802] [Medline: 18225464]
11. Center for Regional Economic Competitiveness, West Virginia University. Appalachia then and now: examining changes to the Appalachian region since 1965. Appalachian Regional Commission. 2015 Feb. URL: https://www.arc.gov/assets/research_reports/AppalachiaThenandNowExecutiveSummaryOmni-opt3.pdf [accessed 2020-07-06]
12. Dal Bello-Haas VPM, Cammer A, Morgan D, Stewart N, Kosteniuk J. Rural and remote dementia care challenges and needs: perspectives of formal and informal care providers residing in Saskatchewan, Canada. Rural Remote Health 2014;14(3):2747 [FREE Full text] [Medline: 25081857]
13. Peters SG, Buntrock JD. Big data and the electronic health record. J Ambul Care Manage 2014;37(3):206-210. [doi: 10.1097/JAC.0000000000000037] [Medline: 24887521]
14. Manyika J, Chui M, Brown B. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. 2011 May 01. URL: https://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation [accessed 2020-07-06]
15. Theoharidou M, Tsalis N, Gritzalis D. Smart home solutions: privacy issues. In: van Hoof J, Demiris G, Wouters EJM, editors. Handbook of Smart Homes, Health Care and Well-Being. Cham, Switzerland: Springer International Publishing; May 20, 2014:1-14.
16. Krysinska K, Sachdev PS, Breitner J, Kivipelto M, Kukull W, Brodaty H. Dementia registries around the globe and their applications: A systematic review. Alzheimers Dement 2017 Sep;13(9):1031-1047 [FREE Full text] [doi: 10.1016/j.jalz.2017.04.005] [Medline: 28576507]
17. Stein B, Morrison A. The enterprise data lake: Better integration and deeper analytics. PwC Technology Forecast: Rethinking Integration 2014;1:18 [FREE Full text]
18. Zhu F, Panwar B, Dodge HH, Li H, Hampstead BM, Albin RL, et al. COMPASS: A computational model to predict changes in MMSE scores 24-months after initial assessment of Alzheimer's disease. Sci Rep 2016 Oct 05;6:34567 [FREE Full text] [doi: 10.1038/srep34567] [Medline: 27703197]
19. Zhang H, Zhu F, Dodge HH, Higgins GA, Omenn GS, Guan Y, Alzheimer's Disease Neuroimaging Initiative. A similarity-based approach to leverage multi-cohort medical data on the diagnosis and prognosis of Alzheimer's disease. Gigascience 2018 Jul 01;7(7) [FREE Full text] [doi: 10.1093/gigascience/giy085] [Medline: 30010762]
20. Ronquillo JG, Baer MR, Lester WT. Sex-specific patterns and differences in dementia and Alzheimer's disease using informatics approaches. J Women Aging 2016;28(5):403-411 [FREE Full text] [doi: 10.1080/08952841.2015.1018038] [Medline: 27105335]
21. Nead KT, Gaskin G, Chester C, Swisher-McClure S, Dudley JT, Leeper NJ, et al. Influence of age on androgen deprivation therapy-associated Alzheimer's disease. Sci Rep 2016 Oct 18;6:35695 [FREE Full text] [doi: 10.1038/srep35695] [Medline: 27752112]
22. Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. JMIR Med Inform 2016 Nov 21;4(4):e38 [FREE Full text] [doi: 10.2196/medinform.5359] [Medline: 27872036]

23. Zhang R, Simon G, Yu F. Advancing Alzheimer's research: A review of big data promises. Int J Med Inform 2017 Oct;106:48-56 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.07.002] [Medline: 28870383]
24. Ienca M, Vayena E, Blasimme A. Big Data and Dementia: Charting the Route Ahead for Research, Ethics, and Policy. Front Med (Lausanne) 2018;5:13 [FREE Full text] [doi: 10.3389/fmed.2018.00013] [Medline: 29468161]
25. Albrecht JS, Hanna M, Kim D, Perfetto EM. Predicting Diagnosis of Alzheimer's Disease and Related Dementias Using Administrative Claims. J Manag Care Spec Pharm 2018 Nov;24(11):1138-1145 [FREE Full text] [doi: 10.18553/jmcp.2018.24.11.1138] [Medline: 30362918]
26. Vuong JT, Jacob SA, Alexander KM, Singh A, Liao R, Desai AS, et al. Mortality From Heart Failure and Dementia in the United States: CDC WONDER 1999-2016. J Card Fail 2019 Feb;25(2):125-129. [doi: 10.1016/j.cardfail.2018.11.012] [Medline: 30471348]
27. Defrancesco M, Marksteiner J, Fleischhacker WW, Blasko I. Use of Benzodiazepines in Alzheimer's Disease: A Systematic Review of Literature. Int J Neuropsychopharmacol 2015 May 19;18(10):pyv055 [FREE Full text] [doi: 10.1093/ijnp/pyv055] [Medline: 25991652]
28. Olfson M, King M, Schoenbaum M. Benzodiazepine use in the United States. JAMA Psychiatry 2015 Feb;72(2):136-142. [doi: 10.1001/jamapsychiatry.2014.1763] [Medline: 25517224]
29. Bachhuber MA, Hennessy S, Cunningham CO, Starrels JL. Increasing Benzodiazepine Prescriptions and Overdose Mortality in the United States, 1996-2013. Am J Public Health 2016 Apr;106(4):686-688. [doi: 10.2105/AJPH.2016.303061] [Medline: 26890165]
30. Sepassi A, Watanabe JH. Emergency Department Visits for Psychotropic-Related Adverse Drug Events in Older Adults With Alzheimer Disease, 2013-2014. Ann Pharmacother 2019 Dec;53(12):1173-1183. [doi: 10.1177/1060028019866927] [Medline: 31342766]

Abbreviations

ACTSI: Appalachian Clinical and Translational Science Institute
AD: Alzheimer’s disease
BZD: benzodiazepine
CDW: Clinical Data Warehouse
CHH: Cabell Huntington Hospital
ED: emergency department
EHR: electronic health record
MU/JCESOM: Marshall University Joan C Edwards School of Medicine
SEDQ: Susan Edwards Drake Memory Clinic Questionnaire

Edited by G Eysenbach; submitted 24.02.20; peer-reviewed by T Muto, A Aminbeidokhti; comments to author 20.04.20; revised version received 27.05.20; accepted 15.06.20; published 04.08.20.

Please cite as:
Bhardwaj N, Cecchetti AA, Murughiyan U, Neitch S
Analysis of Benzodiazepine Prescription Practices in Elderly Appalachians with Dementia via the Appalachian Informatics Platform: Longitudinal Study
JMIR Med Inform 2020;8(8):e18389
URL: https://medinform.jmir.org/2020/8/e18389
doi:10.2196/18389
PMID:32749226

Original Paper

Prediction of Cardiac Arrest in the Emergency Department Based on Machine Learning and Sequential Characteristics: Model Development and Retrospective Clinical Validation Study

Sungjun Hong[1], MS; Sungjoo Lee[1], BS; Jeonghoon Lee[1], MD; Won Chul Cha[1,2,3*], MD; Kyunga Kim[1,4*], PhD

[1]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea

[2]Department of Emergency Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[3]Health Information and Strategy Center, Samsung Medical Center, Seoul, Republic of Korea

[4]Statistics and Data Center, Research Institute for Future Medicine, Samsung Medical Center, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Won Chul Cha, MD
Department of Emergency Medicine
Samsung Medical Center
Sungkyunkwan University School of Medicine
81 Irwon-ro
Gangnam-gu
Seoul, 06351
Republic of Korea
Phone: 82 10 5386 6597
Fax: 82 2 2148 7899
Email: wc.cha@samsung.com

Abstract

Background: The development and application of clinical prediction models using machine learning in clinical decision support systems is attracting increasing attention.

Objective: The aims of this study were to develop a prediction model for cardiac arrest in the emergency department (ED) using machine learning and sequential characteristics and to validate its clinical usefulness.

Methods: This retrospective study was conducted with ED patients at a tertiary academic hospital who suffered cardiac arrest. To resolve the class imbalance problem, sampling was performed using propensity score matching. The data set was chronologically allocated to a development cohort (years 2013 to 2016) and a validation cohort (year 2017). We trained three machine learning algorithms with repeated 10-fold cross-validation.

Results: The main performance parameters were the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). The random forest algorithm (AUROC 0.97; AUPRC 0.86) outperformed the recurrent neural network (AUROC 0.95; AUPRC 0.82) and the logistic regression algorithm (AUROC 0.92; AUPRC=0.72). The performance of the model was maintained over time, with the AUROC remaining at least 80% across the monitored time points during the 24 hours before event occurrence.

Conclusions: We developed a prediction model of cardiac arrest in the ED using machine learning and sequential characteristics. The model was validated for clinical usefulness by chronological visualization focused on clinical usability.

(JMIR Med Inform 2020;8(8):e15932) doi:10.2196/15932

KEYWORDS

machine learning; cardiac arrest prediction; emergency department; sequential characteristics; clinical validity

Introduction

Clinical decision support systems (CDSSs) analyze data to assist health care providers in making decisions and improving service quality. Recently, artificial intelligence has been widely used in CDSSs, and its importance is increasing [1]. Previous studies have shown that CDSSs that use machine learning are actively applied worldwide and can be very helpful in clinical decision making. CDSSs enable clinicians to consider future possibilities and to develop and implement action plans for patient care. Recently, machine learning techniques have been widely used in various medical fields, including diagnosis or prognosis prediction, pattern recognition, and image classification [2,3].

It is difficult for emergency department (ED) staff to monitor all patients due to limited resources. Thus, precise triage systems that can identify high-risk patients are being considered. For this reason, information technology monitoring systems are important, and the application of machine learning techniques in such systems has been extensively studied [4,5]. These triage systems attempt to predict mortality or cardiac arrest based on patient characteristics. However, few studies of prediction modelling clearly reflect sequential characteristics due to the monitoring process. Moreover, the effectiveness of these systems and their applicability to real-world data have not been adequately investigated. For example, detailed analyses of data processing, imbalance adjustment, and dynamics of various factors are lacking. Accordingly, the clinical impact and usage of prediction models have not been sufficiently investigated. The aims of this study were to develop a prediction model of cardiac arrest in the ED using machine learning and sequential characteristics and to validate its clinical usefulness.

Methods

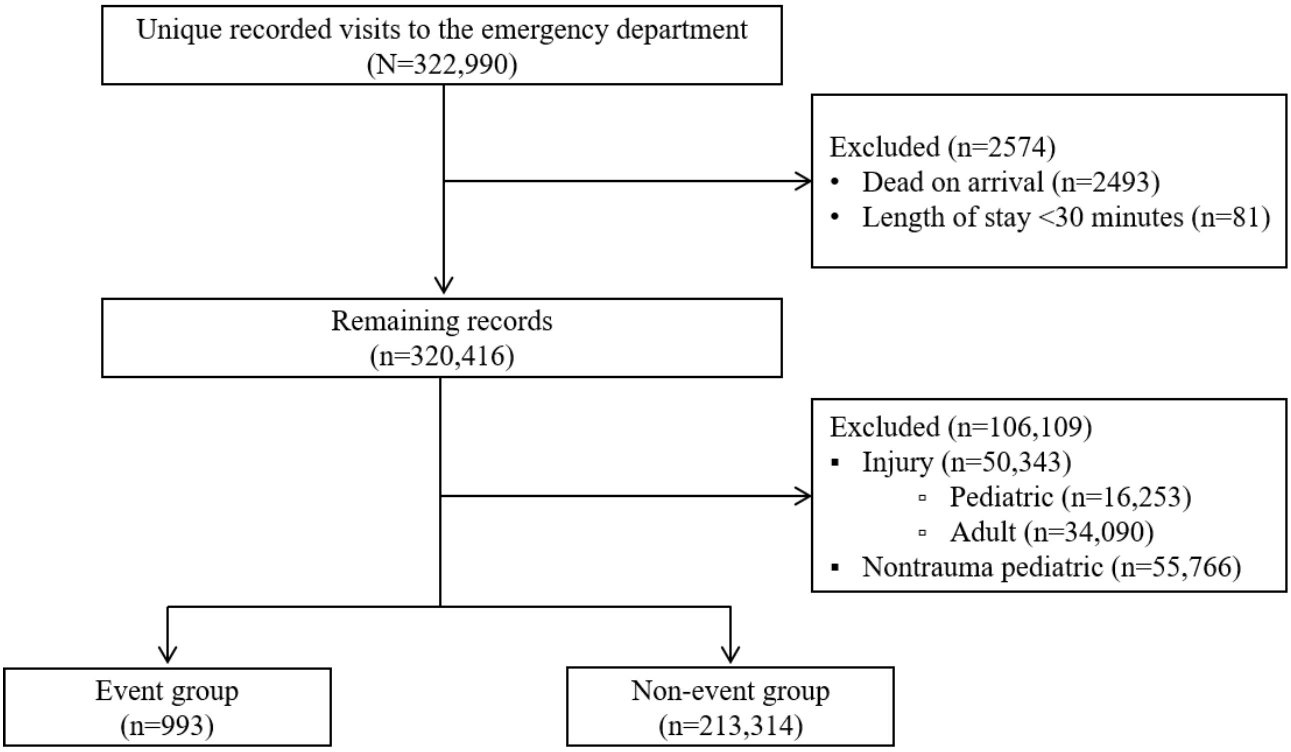
Study Setting

This retrospective study was conducted at Samsung Medical Center, a tertiary academic hospital in South Korea with approximately 2000 beds and an average of 200 ED patients per day. Data were obtained from the electronic medical record hospital database from January 1, 2013 to December 31, 2017. Moreover, data from the National Emergency Department Information System (NEDIS) were collected. NEDIS is a real-time management system for information on patients visiting emergency medical institutions. The NEDIS data contain patient demographics and clinical information, such as age, sex, and clinical outcomes. We followed the guidelines for transparent reporting of a multivariable prediction model for individual prognosis or diagnosis [6]. The study was approved by the institutional review board of Samsung Medical Center (IRB No. 2018-10-025).

Study Participants

The study population consisted of all ED patients in the study period. The following patients were excluded: those who were dead on arrival, pediatric patients aged <18 years, patients with injury, patients who suffered cardiac arrest or died within 30 min after visiting the ED, and patients who did not experience the outcome event within 30 days of admission. The remaining patients were chronologically divided into the model development cohort (years 2013 to 2016) and the model validation cohort (year 2017). The validation cohort was used to assess the model performance for temporal generalizability. Most patients only visited the ED once (147,303/208,415, 70.68%); fewer patients visited the ED multiple times, with an average of 3.24 visits per patient. Because emergency visits are mostly not scheduled and the reasons for the visits vary [7], each visit is often treated as an independent subject. Thus, we considered each visit as an independent study subject rather than each patient. Patient information was anonymized and deidentified. A flowchart of the study cohort is presented in Figure 1.

Figure 1. Flow diagram for the selection of the study cohort. Data were processed as unique records based on the date on which a patient visited the emergency department and may correspond to the same patient.



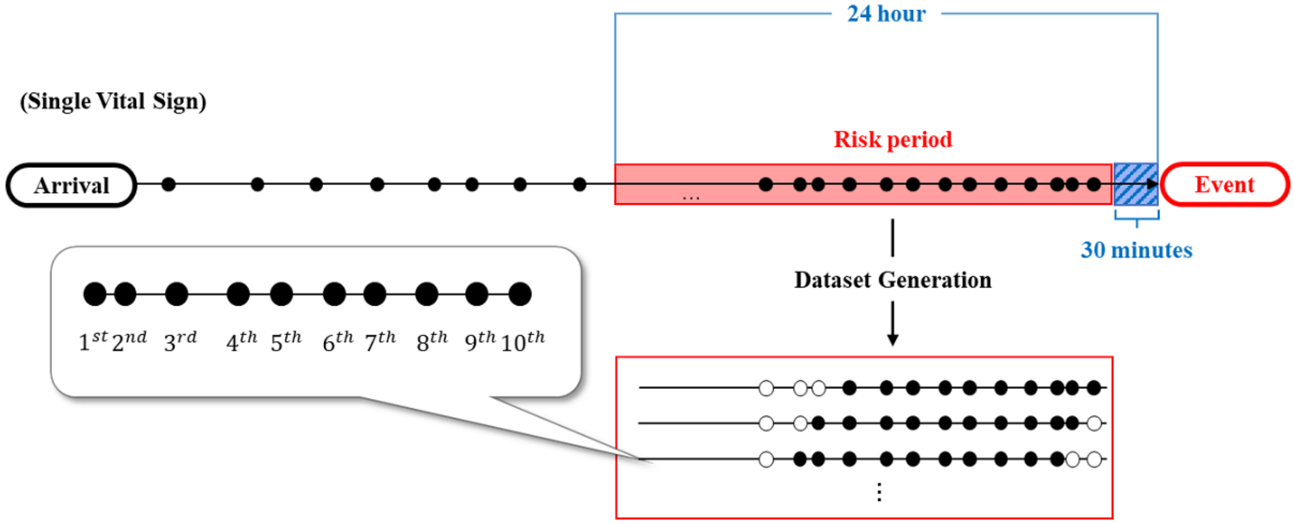
Study Outcome and Predictors

The primary outcome was cardiac arrest regardless of whether cardiopulmonary resuscitation was performed. We also included patients who suffered cardiac arrest after admission to the inpatient ward from the ED. If cardiac arrest occurred several times, we used the first cardiac arrest.

Two groups of predictors were used for the model: initially assessed predictors (sex, age, and chief concerns) and serially assessed predictors (systolic blood pressure, diastolic blood pressure, heart rate, body temperature, respiratory rate, and peripheral oxygen saturation [SpO_2]). The derived predictor for time (time interval) was the length of the interval (in minutes) between time points [8]. We set the value range for each vital sign as follows: 1 to 300 millimeters of mercury for systolic blood pressure and diastolic blood pressure, 1 to 200 beats per minute for heart rate, 30 to 44 degrees Celsius for body temperature, 1 to 60 breaths per minute for respiratory rate, and 1% to 100% for SpO_2. The chief concerns were extracted from the NEDIS data and were combined with the raw data. The main symptoms were classified into 39 codes as part of the initial nursing assessment.

The input vector was set to have 10 sequential measurement values for each time point. For example, if a patient's vital signs were measured 11 times, 11 sets were generated. If the length of the sequential measurements was less than 10, the insufficient data were treated as missing. The 1st and 10th sequence values represent the last and most recent observations, respectively, from the outcome occurrence. We defined the risk period as the interval from 0.5 to 24 hours before outcome occurrence [9]. If the 10th entry of each input vector belonged to the risk period, it was labelled as an event; otherwise, it was labeled as a non-event. These processes are shown in Figure 2.

Figure 2. Sequential dataset generation for a single vital sign of one patient. The risk period was defined as a 24-hour interval prior to the event. The 10 consecutive vital signs were grouped as a data set for prediction. Each point represents a single vital sign measurement. This process was applied to other vital signs in the same manner.



Data Preprocessing

Missing data in the sequential measurement values were imputed with the most recent value. If no previous value was available, zero was used [10]. The serially assessed predictors were standardized to have the same range or variability, and the initially assessed predictors were categorized. Our data are affected by the outcome class imbalance problem, which can reduce model performance. To address the imbalance problem, we used undersampling with propensity score matching. Because excessive adjustment may reduce representativeness, we considered various matching ratios, namely 1%, 5%, and 10%, between the event and the non-event groups [11] to determine a suitable ratio. Sex and age were used as matching variables in the propensity score matching based on the R package MatchIt (Multimedia Appendix 1). Data processing was performed using R version 3.4.3 (R Project). Then, statistical analysis was conducted using the Keras and scikit-learn libraries in Python version 3.6.6.

Analysis

Continuous data are expressed as mean values with the corresponding standard deviations. We performed *t* tests to determine the mean differences between groups. The standardized mean difference is a measure of the effect size for the comparison of two groups [12]. Categorical data were expressed as frequency and percentage. The chi-square test was performed to determine the relationships among categorical features. All tests were two-sided with a statistical significance level of *P*<.05.

To develop a cardiac arrest prediction model, we considered three popular machine learning algorithms, namely logistic regression (LR), random forest (RF), and a recurrent neural network (RNN) [3]. In LR, a ridge penalty was applied to increase the predictive performance and reduce the risk of overfitting [13]. In RF, an entropy criterion was used to measure the split quality [14]. An RNN is an artificial neural network with the advantage of processing sequential data; it is useful for time series analysis using a long short-term memory structure [15]. We used three-layer long short-term memory (the last layer with a sigmoid activation function), an Adam optimization algorithm, and a binary cross-entropy loss function. As a reference cardiac arrest prediction model, we employed the modified early warning score (MEWS) because it is a widely used monitoring tool in ED admission [16]. For optimization, all algorithms used the grid search method. Additionally, the RNN algorithm used the adaptive moment estimation, stochastic gradient descent, and root mean square propagation methods. The hyperparameters in each algorithm were tuned based on 10-fold cross-validation during the model development [17]. To avoid partition bias, the entire cross-validation process was repeated with 5 different partitions. Furthermore, a sensitivity analysis was conducted to assess the effects of the balancing ratio and the influence of the features on the variation of the results among models. More technical details of each algorithm are provided in Multimedia Appendix 2.

To assess the performance of the model, we used various measures, including the area under the receiver operating characteristic curve (AUROC) and the area under the precision recall curve (AUPRC). Also, we used the F1 score to assess class imbalance [18]. Balanced accuracy (BA) was used to determine the optimal cutoff values for the class prediction [19]. Moreover, we used the positive and negative likelihood ratios to assess the clinical usefulness of the prediction model as a diagnostic tool [20]. Calibration and decision curve analyses were conducted to assess the agreement between the observed and predicted values [21,22] and explore the practical threshold for clinical application [23], respectively.

Results

Patient Demographics

A total of 322,990 patients visited the ED during the study period. After the exclusion criteria were applied, the final number of patients was 214,307; among these, 993 (0.5%) had the primary outcome of cardiac arrest. We assigned 168,488/214,307 (78.6%) patients to the model development cohort and 45,819/214,307 (21.4%) patients to the model validation cohort. The patient demographics (divided into two groups for each cohort) are shown in Table 1. The number of female patients (114,280/214,307, 52.5%) was greater than the number of male patients. The mean age for the event group was 65.8 years (SD 15.3), whereas the mean age for the non-event group was 55.4 years (SD 17.8).

Table 1. Patient characteristics of the development and validation cohorts (N=214,307).

Characteristic	Development cohort			Validation cohort			SMD[a]
	Event (n=791)	Non-event (n=167,697)	*P* value	Event (n=202)	Non-event (n=45,617)	*P* value	
Demographic data							
Sex, n (%)			< 001			<.001	0.022
Male	472 (59.7)	78,631 (46.9)		133 (65.8)	22,591 (49.5)		
Female	319 (40.3)	89,066 (53.1)		69 (34.2)	23,026 (50.5)		
Age (years), mean (SD)	65.2 (15.6)	54.9 (17.8)	<.001	68.3 (13.6)	57.1 (17.6)	<.001	0.119
Vital signs, mean (SD)							
Blood pressure (millimeters of mercury)							
Systolic	112.6 (25.5)	120.7 (24.1)	<.001	112.9 (28.4)	121.4 (24.8)	<.001	0.033
Diastolic	65.0 (15.9)	72.7 (15.0)	<.001	64.3 (16.4)	72.7 (15.0)	<.001	0.002
Body temperature (degrees Celsius)	36.7 (2.4)	37.0 (1.7)	<.001	36.8 (2.1)	37.1 (2.1)	<.001	0.053
Heart rate (beats per minute)	99.9 (23.7)	88.8 (20.8)	<.001	99.0 (22.1)	88.3 (20.7)	<.001	0.033
Respiratory rate (breaths per minute)	21.2 (6.6)	19.8 (3.9)	<.001	20.6 (6.4)	19.1 (3.7)	<.001	0.174
SpO_2[b] (%)	94.9 (11.0)	90.2 (25.4)	<.001	95.2 (8.7)	96.8 (8.1)	<.001	0.331

[a] SMD (standardized mean difference) for comparison between the development and validation cohorts.

[b] SpO_2: peripheral oxygen saturation.

Figure 3 shows the average trends of the vital signs for the two groups. Compared to the non-event group, the heart and respiratory rates for the event group were higher on average, whereas the values of the other vital signs were lower. The starting points and the trends were clearly different, demonstrating that the two groups could be distinguished. The chief concern distributions of the groups were different, and dyspnea and abdominal pain were the most common chief concerns in the event and non-event groups, respectively. A comparison of the top 10 chief concerns for each group is shown in Table 2.

Figure 4 shows the frequency difference between the two groups over time and demonstrates that frequent measurements are performed for ED patients in serious condition. Figure 5 shows the model performance over time. It can be seen that the model performance was maintained, with the AUROC remaining at least 80% across the monitored time points during the 24 hours before event occurrence.

Figure 3. Trends in the event and non-event groups for the vital signs: A. systolic blood pressure; B. diastolic blood pressure; C. heart rate; D. respiratory rate; E. body temperature; F. peripheral oxygen saturation. The x-axis values are the 10 time points before event occurrence, and the y-axis values are the mean values of the vital signs. BT: body temperature; DBP: diastolic blood pressure; HR: heart rate; RR: respiratory rate; SBP: systolic blood pressure; SpO_2: peripheral oxygen saturation.

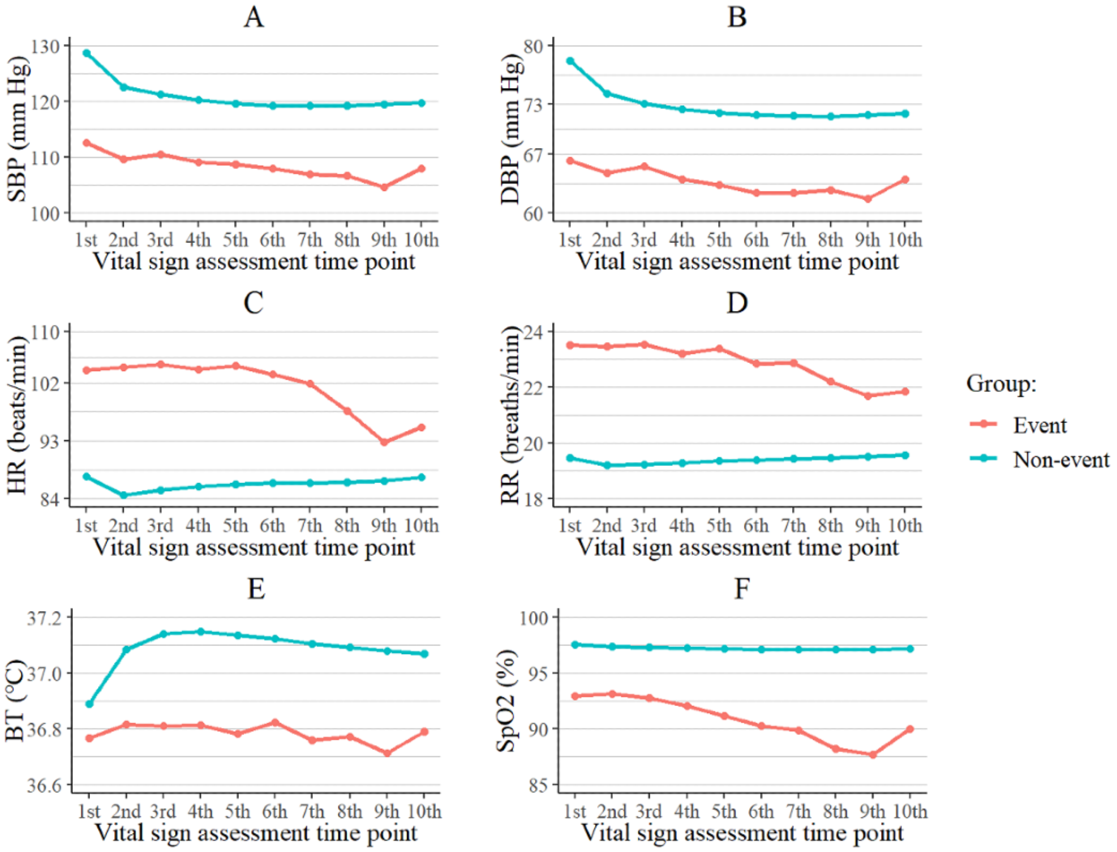


Table 2. Top 10 chief concerns in the event and non-event groups.

Rank	Event group (n=993)		Non-event group (n=213,314)	
	Chief concern	n (%)	Chief concern	n (%)
1	Dyspnea	350 (35.25)	Abdominal pain	32 996 (15.47)
2	Altered mentality	96 (9.67)	Fever	23 551 (11.04)
3	Fever	88 (8.86)	Dyspnea	16 887 (7.92)
4	Abdominal pain	64 (6.45)	Dizziness	13,718 (6.43)
5	Chest pain	60 (6.04)	Headache	9361 (4.39)
6	General weakness	23 (2.32)	Chest pain	6011 (2.82)
7	Dizziness	22 (2.22)	Skin rash, urticaria	5042 (2.36)
8	Chest discomfort	20 (2.01)	Altered mentality	3045 (1.43)
9	Hematemesis	14 (1.41)	Back pain	2937 (1.38)
10	Hemoptysis	13 (1.31)	Vomiting	2909 (1.36)

Figure 4. Average numbers of vital sign assessments at each prediction time point for the event and non-event groups. The lines and shaded 95% CIs show the trends for the vital assessments.

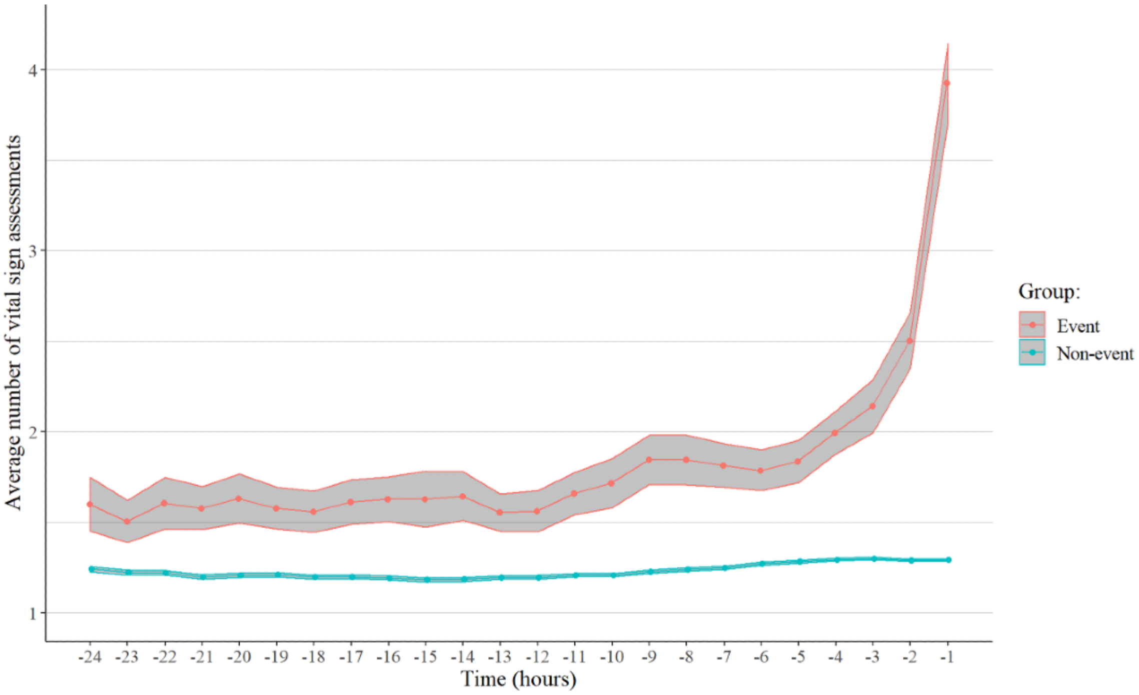
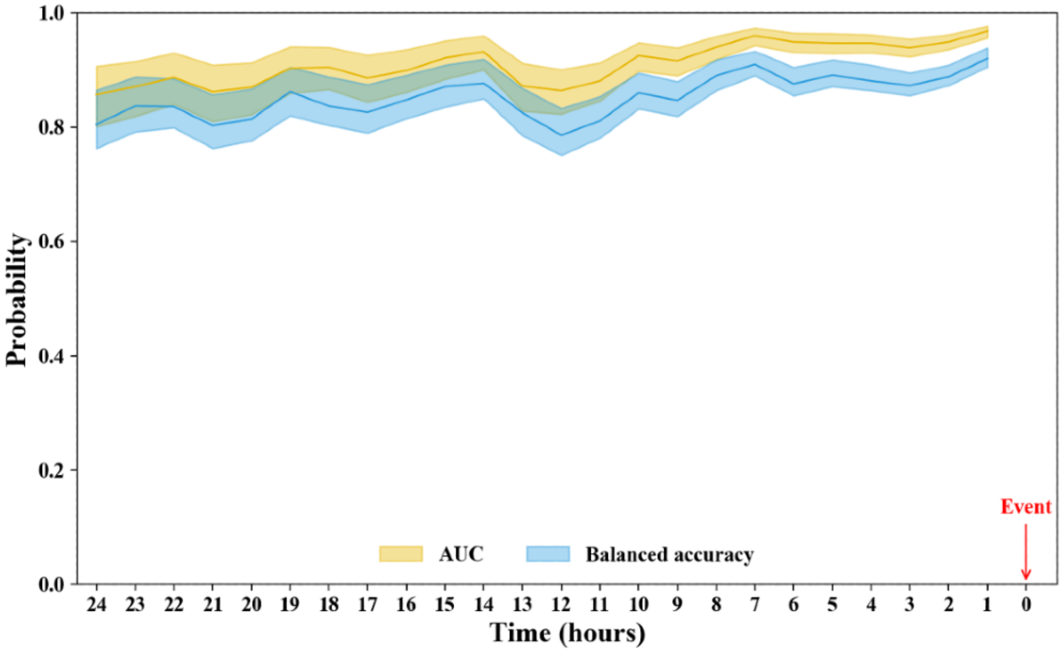


Figure 5. Time point performance in class prediction. The best model was selected based on Table 3, and the predictive performance was evaluated at each prediction time point from event occurrence. The lines and shaded 95% CIs show the trends for the predictive performance. AUC: area under the curve.



Model Performance

Table 3 and Multimedia Appendix 3 summarize the calibrations and overall prediction performance of each model when applying the different balancing ratios for imbalance adjustment, while Table 4 presents the class prediction performance. Model calibrations were described with the integrated calibration index (ICI) and calibration slope. Compared to the other models, the RF model had the smallest ICI in the validation cohort for each balancing ratio (eg, 0.04 for MEWS, 0.04 for LR, 0.02 for RNN, and 0.01 for RF with 10% balancing). The RF-based models showed better calibration in the validation cohort than in the development cohort across the various imbalance adjustments. All the other models showed poorer calibration in the validation cohort than in the development cohort; this suggests that overfitting occurred. As the class imbalance was adjusted with higher balancing ratios, overall improvement was observed for the calibration performance (see the bias-corrected curves in Figure A2 of Multimedia Appendix 3).

Table 3. Overall predictive performance for each machine learning algorithm with imbalance adjustment in the development and validation cohorts.

Matching ratio and model	Development cohort				Validation cohort			
	AUROC[a] (95% CI)	AUPRC[b] (95% CI)	ICI[c]	Calibration slope (95% CI)	AUROC (95% CI)	AUPRC (95% CI)	ICI	Calibration slope (95% CI)
0.5% (Real world)								
MEWS[d]	0.77 (0.77-0.77)	0.09 (0.08-0.09)	0.013	3.69 (3.64-3.74)	0.80 (0.80-0.81)	0.11 (0.10-0.12)	0.016	4.19 (4.09-4.29)
LR[e]	0.82 (0.82-0.83)	0.08 (0.08-0.09)	0.003	1.09 (1.08-1.10)	0.82 (0.81-0.83)	0.09 (0.09-0.10)	0.004	1.12 (1.09-1.15)
RNN[f]	0.96 (0.96-0.97)	0.47 (0.46-0.48)	0.002	1.13 (1.12-1.15)	0.91 (0.90-0.91)	0.17 (0.16-0.18)	0.006	0.70 (0.69-0.72)
RF[g]	1.00 (1.00-1.00)	1.00 (1.00-1.00)	0.007	6.71 (6.18-7.24)	0.94 (0.94-0.95)	0.37 (0.35-0.39)	0.003	1.09 (1.06-1.13)
1%								
MEWS	0.76 (0.76-0.77)	0.12 (0.12-0.12)	0.022	3.46 (3.41-3.51)	0.79 (0.79-0.80)	0.16 (0.15-0.17)	0.025	4.09 (3.97-4.20)
LR	0.88 (0.88-0.89)	0.26 (0.25-0.26)	0.008	1.07 (1.06-1.09)	0.88 (0.87-0.88)	0.28 (0.27-0.30)	0.007	1.09 (1.06-1.12)
RNN	0.96 (0.96-0.96)	0.52 (0.51-0.53)	0.003	1.03 (1.01-1.04)	0.91 (0.91-0.92)	0.33 (0.32-0.35)	0.010	0.79 (0.78-0.81)
RF	1.00 (1.00-1.00)	1.00 (1.00-1.00)	0.010	7.51 (6.86-8.15)	0.94 (0.93-0.94)	0.47 (0.45-0.49)	0.003	1.14 (1.11-1.18)
5%								
MEWS	0.73 (0.72-0.73)	0.25 (0.25-0.26)	0.052	3.08 (3.01-3.14)	0.77 (0.77-0.78)	0.35 (0.34-0.37)	0.066	4.22 (4.08-4.37)
LR	0.91 (0.90-0.91)	0.59 (0.58-0.60)	0.034	1.00 (0.99-1.02)	0.90 (0.89-0.90)	0.61 (0.60-0.63)	0.028	1.02 (0.99-1.04)
RNN	0.97 (0.97-0.97)	0.79 (0.79-0.80)	0.003	1.04 (1.02-1.06)	0.94 (0.93-0.94)	0.68 (0.66-0.69)	0.015	0.82 (0.80-0.84)
RF	1.00 (1.00-1.00)	1.00 (1.00-1.00)	0.025	9.88 (8.54-11.22)	0.96 (0.96-0.96)	0.78 (0.76-0.79)	0.012	1.21 (1.17-1.25)
10%								
MEWS	0.70 (0.70-0.71)	0.29 (0.29-0.30)	0.018	1.14 (1.11-1.17)	0.76 (0.75-0.77)	0.42 (0.41-0.44)	0.043	1.68 (1.62-1.75)
LR	0.93 (0.92-0.93)	0.71 (0.70-0.71)	0.043	1.00 (0.99-1.01)	0.92 (0.91-0.92)	0.72 (0.71-0.74)	0.039	0.98 (0.95-1.01)
RNN	0.98 (0.97-0.98)	0.87 (0.87-0.88)	0.002	1.02 (1.00-1.04)	0.95 (0.95-0.96)	0.82 (0.81-0.83)	0.015	0.81 (0.79-0.84)
RF	1.00 (1.00-1.00)	1.00 (1.00-1.00)	0.025	10.19 (8.73-11.65)	0.97 (0.97-0.97)	0.86 (0.84-0.87)	0.014	1.14 (1.09-1.18)

[a] AUROC: area under the receiver operating characteristic curve.

[b] AUPRC: area under the precision recall curve.

[c] ICI: integrated calibration index.

[d] MEWS: modified early warning score.

[e] LR: logistic regression.

[f] RNN: recurrent neural network.

[g] RF: random forest.

Table 4. Class prediction performance of each machine learning algorithm with imbalance adjustment in the validation cohort.

Matching ratio and model	BA[a] (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	F1 score	PLR[b] (95% CI)	NLR[c] (95% CI)
0.5% (Real world)						
MEWS[d]	0.75 (0.74-0.76)	0.72 (0.70-0.73)	0.78 (0.78-0.79)	0.096	3.31 (3.24-3.38)	0.36 (0.34-0.38)
LR[e]	0.76 (0.76-0.77)	0.75 (0.75-0.78)	0.76 (0.76-0.76)	0.093	3.21 (3.15-3.27)	0.31 (0.29-0.33)
RNN[f]	0.84 (0.84-0.85)	0.85 (0.84-0.86)	0.84 (0.83-0.84)	0.143	5.17 (5.09-5.26)	0.18 (0.17-0.19)
RF[g]	0.88 (0.88-0.89)	0.88 (0.87-0.89)	0.89 (0.88-0.89)	0.198	7.72 (7.61-7.85)	0.13 (0.12-0.14)
1%						
MEWS	0.74 (0.73-0.74)	0.72 (0.70-0.73)	0.76 (0.76-0.76)	0.148	2.97 (2.90-3.03)	0.37 (0.36-0.39)
LR	0.81 (0.80-0.81)	0.78 (0.76-0.79)	0.84 (0.84-0.84)	0.218	4.77 (4.67-4.88)	0.27 (0.25-0.28)
RNN	0.84 (0.83-0.85)	0.87 (0.85-0.88)	0.81 (0.81-0.82)	0.218	4.67 (4.59-4.76)	0.17 (0.15-0.18)
RF	0.88 (0.87-0.88)	0.90 (0.89-0.91)	0.86 (0.86-0.86)	0.278	6.49 (6.38-6.60)	0.12 (0.11-0.13)
5%						
MEWS	0.72 (0.71-0.73)	0.72 (0.70-0.73)	0.72 (0.72-0.73)	0.348	2.57 (2.50-2.63)	0.39 (0.37-0.41)
LR	0.85 (0.84-0.85)	0.83 (0.82-0.84)	0.87 (0.86-0.87)	0.555	6.15 (5.97-6.34)	0.20 (0.18-0.21)
RNN	0.87 (0.87-0.88)	0.89 (0.88-0.90)	0.85 (0.85-0.85)	0.562	5.96 (5.80-6.15)	0.12 (0.11-0.14)
RF	0.90 (0.90-0.91)	0.92 (0.91-0.93)	0.89 (0.88-0.89)	0.639	8.23 (7.97-8.49)	0.09 (0.08-0.10)
10%						
MEWS	0.70 (0.69-0.71)	0.72 (0.70-0.73)	0.69 (0.68-0.69)	0.419	2.29 (2.23-2.35)	0.41 (0.39-0.43)
LR	0.87 (0.86-0.87)	0.86 (0.85-0.87)	0.87 (0.87-0.88)	0.675	6.80 (6.54-7.07)	0.16 (0.15-0.17)
RNN	0.89 (0.89-0.90)	0.93 (0.92-0.94)	0.85 (0.85-0.86)	0.681	6.32 (6.11-6.54)	0.08 (0.07-0.09)
RF	0.92 (0.92-0.92)	0.94 (0.94-0.95)	0.90 (0.89-0.90)	0.756	9.31 (8.95-9.69)	0.06 (0.06-0.07)

[a] BA: balanced accuracy.

[b] PLR: positive likelihood ratio.

[c] NLR: negative likelihood ratio.

[d] MEWS: modified early warning score.

[e] LR: logistic regression.

[f] RNN: recurrent neural network.

[g] RF: random forest.

The RF models showed the best overall predictive performance in the validation cohort for each balancing ratio. For instance, the AUROC of RF was 0.97 and an AUPRC of 0.86, while RNN, LR, and MEWS had AUROCs of 0.95, 0.92, and 0.76 and AUPRCs of 0.82, 0.72, and 0.42, respectively, in the validation cohort with 10% balancing. The RF-based models outperformed the RNN- and LR-based models as well as MEWS in terms of all performance measures in class prediction (all *P*<.001). The RF-based models showed better overall prediction performance, and all performance measures for the class prediction improved as higher balancing ratios were applied for the class imbalance adjustment (eg, the AUPRC and F1 score improved from 0.37 to 0.86 and from 0.20 to 0.76, respectively).

After considering all the factors, the RF-based model with a 10% balancing ratio was selected as the best model. The best model had a sensitivity and specificity of 0.94 (95% CI 0.94-0.95) and 0.90 (95% CI 0.89-0.90), respectively. Moreover, the positive likelihood ratio value of 9.31 (95% CI 8.95-9.69) and the negative likelihood ratio value of 0.06 (95% CI 0.06-0.07) indicate that the model is clinically informative and very useful in practice. For the best model, the ICI and calibration slope were 0.01 and 1.14 (95% CI 1.09-1.18), respectively. Table 5 summarizes the importance of each predictor in the best model. Body temperature and SpO_2 were relatively important predictors.

Table 5. Predictor importance in the random forest model with 10% balancing.

Feature	Predictor importance, mean (SD)
Body temperature	0. 284 (0.014)
Peripheral oxygen saturation	0.232 (0.011)
Heart rate	0.127 (0.005)
Duration	0.096 (0.009)
Respiratory rate	0.084 (0.005)
Systolic blood pressure	0.081 (0.006)
Diastolic blood pressure	0.072 (0.005)
Chief concern	0.012 (N/A)[a]
Age	0.010 (N/A)
Sex	0.002 (N/A)

[a] N/A: not applicable.

Discussion

Principal Findings

Recent prediction model guidelines emphasize validation and clinical application [6,24]. Clinical usage of prediction models is important; therefore, these models should be clinically adaptable and persuasive. However, previous studies are lacking in these aspects. In the present study, we attempted to remedy this by verifying the suitability of the model using chronological visualization focused on clinical usability.

In this study, we developed the model using the method of generating sequential data vectors. The comprehensive model validation considered performance and various clinical relevance aspects. The clinical validity of the model was assessed through visualization of chronological characteristics.

Machine learning–based prediction models are often called “black boxes” because the algorithms provide answers without any “human” knowledge. When calculations and suggestions cannot be clinically explained, it is almost impossible to apply them in real-world settings. One reason for this is that it is not clear who or what is responsible for clinical decisions [25,26]. Another reason is that clinicians are not notified of the parameters on which they should focus; thus, applying machine learning–based prediction in a clinical setting may be confusing.

It can be practically important to suggest a single unified threshold for class prediction across all prediction time points. The best threshold chosen with the highest balanced accuracy at each prediction time point ranged from 0.30 to 0.40. Within this range, we considered several candidates for the unified threshold and investigated their performance in various aspects (Multimedia Appendix 4-6). A unified threshold of 0.35 was selected because of its stable performance and considerable net benefit. Clinicians can apply either a single unified threshold across all time points or the best threshold for each time point based on practicality and depending on their environment.

In practice, clinicians can apply the proposed prediction process as follows. When a new patient visits the ED, the initial assessment is conducted and the initially assessed predictors are recorded. Then, the patient’s vital signs are monitored and the sequential measurements are converted into a sequential record for serially assessed predictors. Then, the developed prediction model produces the predicted probability of cardiac arrest. When a vital sign is updated, the sequential record is promptly updated and used as a new input to update the predicted probability. Based on a prechosen threshold (eg, 0.35), the risk of the patient at the moment is classified as high if the prediction probability is greater than or equal to the threshold. This prediction process can be applied as a trigger alarm system, in which the high-risk prediction initiates more intensive care or closer monitoring. In this case, the missed rate and the false alarm rate are expected to be 6% and 10%, respectively. Therefore, the increase in the workload of the medical team is only 10%.

Significant efforts have been made to improve the explainability of prediction models so they can be applied in real-world settings [27-30]. Due to the nature of machine learning, it is difficult to explain individual decisions specifically; however, it is still possible to describe the overall decision process based on feature importance [31,32]. On average, body temperature and SpO_2 were important, especially in the 1st, 2nd, 3rd, and 10th measurements. The remaining features were relatively important only in the 10th measurement. Our findings demonstrate that there is a difference in the importance of these features over time. Thus, it is necessary to test the performance by narrowing the time interval.

Another factor that affects the clinical validity of prediction models is imbalance of the outcome parameters [33,34]. It is clinically valuable to know how performance changes in different settings, and this change was given little attention in previous studies. In this study, we used various structures and considered both model accuracy and realistic settings to demonstrate the statistical robustness of the model.

In the real world, serially assessed vital data often contain missing values for various reasons, and these data should be handled properly and efficiently. When choosing missing-handling methods, we focused on two factors: the nonrandomness of missing patterns in the ED data and the applicability to the risk prediction of a new patient in a real ED situation. In the process of sequential data set generation, nonrandom missing data naturally occurs due to the lack of information on vital signs at early prediction time points (ie, before assessing vital signs 10 times). We attempted to use this nonrandom missing pattern as additional information by zero imputation, which may reflect the low frequency of vital assessment to a certain degree. Moreover, at other prediction time points, missing values occur nonrandomly because vital assessments in the ED are ordered according to the patient's condition and are also monitored periodically. We suggested filling in the missing data with the most recent value because it is practically applicable to the prediction of risk for new patients based on our prediction procedure.

Limitations

Our study has several limitations. First, the use of machine learning algorithms was limited, and the study design (eg, risk period and number of sequential measurements) was set heuristically based on clinical experience in real clinical settings. However, a prediction model can be developed by applying the process in this study using other algorithms as well. Second, because this study was conducted in a single department of a single center, it is not representative. To use the model in other institutions, further external validation should be performed. Third, few features were used, and the results of other tests containing significant information (eg, laboratory tests) were not considered. Using this additional information may be advantageous, although the proposed model is already considerably accurate. Finally, the outcome was an ultimate result (ie, cardiac arrest) and did not include resuscitation efforts or prescription. In a clinical setting, resuscitation efforts should be considered. Therefore, it is necessary to extend the proposed method to include resuscitation and acute deterioration.

Conclusions

In this study, we developed a cardiac arrest prediction model in the ED using machine learning and sequential characteristics. The model was validated for clinical usefulness using chronological visualization focused on clinical usability.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI19C0275).

Authors' Contributions

SH collected and analyzed the data, interpreted the results, and drafted this manuscript. SL collected and analyzed the data. JL interpreted the results. WWC and KK conceived and designed the study and revised the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Patient characteristics after propensity score matching.
[PDF File (Adobe PDF File), 239 KB - medinform_v8i8e15932_app1.pdf]

Multimedia Appendix 2
Supplemental code for developing the models.
[PDF File (Adobe PDF File), 51 KB - medinform_v8i8e15932_app2.pdf]

Multimedia Appendix 3
ROC and calibration curves in the development and validation cohorts.
[PDF File (Adobe PDF File), 782 KB - medinform_v8i8e15932_app3.pdf]

Multimedia Appendix 4
Time-point performance in class prediction of the candidate threshold systems.
[PDF File (Adobe PDF File), 287 KB - medinform_v8i8e15932_app4.pdf]

Multimedia Appendix 5
Overall performance in class prediction of the candidate threshold systems.
[PDF File (Adobe PDF File), 135 KB - medinform_v8i8e15932_app5.pdf]

Multimedia Appendix 6
Decision curve for the best prediction model.

[PDF File (Adobe PDF File), 88 KB - medinform_v8i8e15932_app6.pdf]

References

1. Aljaaf A, Al-Jumeily D, Hussain A, Fergus P, Al-Jumaily M, Abdel-Aziz K. Toward an optimal use of artificial intelligence techniques within a clinical decision support system. 2015 Sep Presented at: Science and Information Conference (SAI); 2015; London, UK p. 548-554 URL: https://ieeexplore.ieee.org/document/7237196 [doi: 10.1109/sai.2015.7237196]
2. Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine Learning and Decision Support in Critical Care. Presented at: Proc IEEE Inst Electr Electron Eng. Feb;104(2). Medline; 2016 p. 444-466. [doi: 10.1109/JPROC.2015.2501978]
3. Safdar S, Zafar S, Zafar N, Khan NF. Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. Artif Intell Rev 2017 Mar 25;50(4):597-623. [doi: 10.1007/s10462-017-9552-8]
4. Hoot NR, Aronsky D. Systematic review of emergency department crowding: causes, effects, and solutions. Ann Emerg Med 2008 Aug;52(2):126-136 [FREE Full text] [doi: 10.1016/j.annemergmed.2008.03.014] [Medline: 18433933]
5. Kwon JM, Lee Y, Lee Y, Lee S, Park H, Park J. Validation of deep-learning-based triage and acuity score using a large national dataset. PLoS One 2018 Oct;13(10):e0205836 [FREE Full text] [doi: 10.1371/journal.pone.0205836] [Medline: 30321231]
6. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. Br J Surg 2015 Feb;102(3):148-158. [doi: 10.1002/bjs.9736] [Medline: 25627261]
7. Moore BJ, Stocks C, Owens PL. Trends in Emergency Department Visits, 2006-2014. Statistical Brief #227. 2017. URL: https://www.hcup-us.ahrq.gov/reports/statbriefs/sb227-Emergency-Department-Visit-Trends.jsp
8. Kyriacos U, Jelsma J, Jordan S. Monitoring vital signs using early warning scoring systems: a review of the literature. J Nurs Manag 2011 Apr;19(3):311-330. [doi: 10.1111/j.1365-2834.2011.01246.x] [Medline: 21507102]
9. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation 2013 Apr;84(4):465-470. [doi: 10.1016/j.resuscitation.2012.12.016] [Medline: 23295778]
10. Lipton ZC, Kale DC, Wetzel R. Modeling Missing Data in Clinical Time Series with RNNs. arXiv: 1606.04130.: arXiv; 2016 Jun. URL: https://arxiv.org/abs/1606.04130 [accessed 2016-11-11]
11. Poolsawad N, Kambhampati C, Cleland J. Balancing class for performance of classification with a clinical dataset. 2014 Presented at: Proceedings of the World Congress on Engineering; 2014; Longon, UK p. 1-6 URL: http://www.iaeng.org/publication/WCE2014/WCE2014_pp237-242.pdf
12. Faraone SV. Interpreting estimates of treatment effects: implications for managed care. Pharmacy and Therapeutics 2008 Dec;33(12):700-711 [FREE Full text] [Medline: 19750051]
13. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33(1):1-22 [FREE Full text] [Medline: 20808728]
14. Liaw A, Wiener M. Classification and regression by randomForest.: R news, 2.3; 2002. URL: https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf [accessed 2002-12-31]
15. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
16. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. QJM 2001 Oct;94(10):521-526. [doi: 10.1093/qjmed/94.10.521] [Medline: 11588210]
17. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. 1995 Presented at: Appears in the International Joint Conference on Articial Intelligence; 1995; Montreal, Quebec p. 1137-1145 URL: http://ai.stanford.edu/~ronnyk/accEst.pdf
18. Lever J, Krzywinski M, Altman N. Classification evaluation. Nat Methods 2016 Jul 28;13(8):603-604 [FREE Full text] [doi: 10.1038/nmeth.3945]
19. Akarachantachote N, Chadcham S, Saithanu K. CUTOFF THRESHOLD OF VARIABLE IMPORTANCE IN PROJECTION FOR VARIABLE SELECTION. Int J of Pure and Appl Math 2014;94(3):307-322 [FREE Full text] [doi: 10.12732/ijpam.v94i3.2]
20. Whiting P, Martin RM, Ben-Shlomo Y, Gunnell D, Sterne JAC. How to apply the results of a research paper on diagnosis to your patient. JRSM Short Rep 2013 Jan;4(1):7 [FREE Full text] [doi: 10.1258/shorts.2012.012089] [Medline: 23413409]
21. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016 Jun 22;353:i3140 [FREE Full text] [doi: 10.1136/bmj.i3140] [Medline: 27334381]
22. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med 2019 Sep 20;38(21):4051-4065 [FREE Full text] [doi: 10.1002/sim.8281] [Medline: 31270850]
23. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006 Sep;26(6):565-574 [FREE Full text] [doi: 10.1177/0272989X06295361] [Medline: 17099194]

24. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014 Aug 01;35(29):1925-1931 [FREE Full text] [doi: 10.1093/eurheartj/ehu207] [Medline: 24898551]
25. Deo RC. Machine Learning in Medicine. Circulation 2015 Nov 17;132(20):1920-1930 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.001593] [Medline: 26572668]
26. Samek W, Wiegand T, Muller KR. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. arXiv: 1708.08296.: arXiv; 2018 Aug. URL: https://arxiv.org/abs/1708.08296 [accessed 2017-08-28]
27. Shmueli G. To explain or to predict? Statistical science. . FREE Full text 2010;25(3):289-310. [doi: 10.1214/10-STS330]
28. Hofman JM, Sharma A, Watts DJ. Prediction and explanation in social systems. Science 2017 Feb 03;355(6324):486-488. [doi: 10.1126/science.aal3856] [Medline: 28154051]
29. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med 2016 Dec 08;375(23):2293-2297. [doi: 10.1056/NEJMsb1609216] [Medline: 27959688]
30. Corrigan-Curay J, Sacks L, Woodcock J. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. JAMA 2018 Sep 04;320(9):867-868. [doi: 10.1001/jama.2018.10136] [Medline: 30105359]
31. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics 2002 Oct;35(5-6):352-359. [doi: 10.1016/S1532-0464(03)00034-0]
32. Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Stat Comput 2016 Mar 23;27(3):659-678. [doi: 10.1007/s11222-016-9646-1]
33. Phoungphol P, Zhang Y, Zhao Y. Robust multiclass classification for learning from imbalanced biomedical data. Tinshhua Sci. Technol 2012 Dec;17(6):619-628. [doi: 10.1109/TST.2012.6374363]
34. Burez J. Van den Poel DJESwA. Handling class imbalance in customer churn prediction. Expert Systems with Applications 2009;36(3):4626-4636. [doi: 10.1016/j.eswa.2008.05.027]

Abbreviations

AUPRC: area under the precision recall curve
AUROC: area under the receiver operating characteristic curve
BA: balanced accuracy
CDSS: clinical decision support system
ED: emergency department
ICI: integrated calibration index
LR: logistic regression
MEWS: modified early warning score
NEDIS: National Emergency Department Information System
RF: random forest
RNN: recurrent neural network
SpO$_2$: peripheral oxygen saturation

Edited by G Eysenbach; submitted 20.08.19; peer-reviewed by JM Kwon, R Subhashini, K Cho, R Bajpai; comments to author 18.09.19; revised version received 30.12.19; accepted 14.07.20; published 04.08.20.

Please cite as:
Hong S, Lee S, Lee J, Cha WC, Kim K
Prediction of Cardiac Arrest in the Emergency Department Based on Machine Learning and Sequential Characteristics: Model Development and Retrospective Clinical Validation Study
JMIR Med Inform 2020;8(8):e15932
URL: http://medinform.jmir.org/2020/8/e15932/
doi:10.2196/15932
PMID:32749227

Original Paper

Assessment of the Robustness of Convolutional Neural Networks in Labeling Noise by Using Chest X-Ray Images From Multiple Centers

Ryoungwoo Jang[1], BA, MD, MSc; Namkug Kim[2], BA, MA, PhD; Miso Jang[1], BA, MA, MD; Kyung Hwa Lee[1], BA, MA, MD; Sang Min Lee[3], BA, MA, MD, PhD; Kyung Hee Lee[4], BA, MA, MD, PhD; Han Na Noh[5], BA, MA, MD, PhD; Joon Beom Seo[3], BA, MA, MD, PhD

[1]Department of Biomedical Engineering, College of Medicine, University of Ulsan, Seoul, Republic of Korea

[2]Department of Convergence Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[3]Department of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[4]Department of Radiology, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Seongnam, Republic of Korea

[5]Department of Health Screening and Promotion Center, Asan Medical Center, Seoul, Republic of Korea

Corresponding Author:
Namkug Kim, BA, MA, PhD
Department of Convergence Medicine
Asan Medical Center
University of Ulsan College of Medicine
88 Olympic-Ro 43-Gil, Songpa-Gu, Seoul, Korea
Seoul
Republic of Korea
Phone: 82 10 3017 4282
Email: namkugkim@gmail.com

Abstract

Background: Computer-aided diagnosis on chest x-ray images using deep learning is a widely studied modality in medicine. Many studies are based on public datasets, such as the National Institutes of Health (NIH) dataset and the Stanford CheXpert dataset. However, these datasets are preprocessed by classical natural language processing, which may cause a certain extent of label errors.

Objective: This study aimed to investigate the robustness of deep convolutional neural networks (CNNs) for binary classification of posteroanterior chest x-ray through random incorrect labeling.

Methods: We trained and validated the CNN architecture with different noise levels of labels in 3 datasets, namely, Asan Medical Center-Seoul National University Bundang Hospital (AMC-SNUBH), NIH, and CheXpert, and tested the models with each test set. Diseases of each chest x-ray in our dataset were confirmed by a thoracic radiologist using computed tomography (CT). Receiver operating characteristic (ROC) and area under the curve (AUC) were evaluated in each test. Randomly chosen chest x-rays of public datasets were evaluated by 3 physicians and 1 thoracic radiologist.

Results: In comparison with the public datasets of NIH and CheXpert, where AUCs did not significantly drop to 16%, the AUC of the AMC-SNUBH dataset significantly decreased from 2% label noise. Evaluation of the public datasets by 3 physicians and 1 thoracic radiologist showed an accuracy of 65%-80%.

Conclusions: The deep learning–based computer-aided diagnosis model is sensitive to label noise, and computer-aided diagnosis with inaccurate labels is not credible. Furthermore, open datasets such as NIH and CheXpert need to be distilled before being used for deep learning–based computer-aided diagnosis.

(JMIR Med Inform 2020;8(8):e18089) doi:10.2196/18089

KEYWORDS

deep learning; convolutional neural network; NIH dataset; CheXpert dataset; robustness

Introduction

Posteroanterior chest x-ray (CXR) is one of the most widely used methods to evaluate a subject's chest. CXR is low cost and easy to assess and acquire, and it provides a variety of information. Researchers developed computer-aided diagnosis (CAD) algorithms for CXRs because of the substantial presence of CXRs in large hospitals and medical centers [1]. At present, there are no widely used clinically meaningful CAD algorithms with classical image processing algorithms. However, the success of deep learning has led to the development of deep learning–based CXR CAD algorithms [2]. Among the various types of deep learning algorithms, the convolutional neural network (CNN) is the most widely used technique for CXR classification.

Before applying CNN to CAD development, we need to consider the robustness of CNN for inaccurate datasets. It is believed that CNN is robust to label noise [3]. Conversely, clean labels and accurate datasets are considered necessary conditions for CNN-based classification. However, the differences in complexity between datasets from Modified National Institute of Standards and Technology (MNIST) and CXRs were enormous. The MNIST images had a size of 28×28 pixels, whereas the image sizes in CXR datasets were mostly above 1024×1024 pixels. Therefore, relying on the robustness of deep learning alone for CXR datasets would be insufficient. Some [3] asserted that accuracy over 90% with 0% noisy labels is not very different from an approximate accuracy of 85% with 90% noisy labels. However, in medicine, an accurate diagnosis is essential for appropriate treatment, and even a 1% decrease in accuracy cannot be tolerated.

Since open CXR datasets from the National Institutes of Health (NIH) and Stanford CheXpert are preprocessed using natural language processing, they tend to contain [4] a certain extent of wrong and uncertain labels [5,6]. Several groups studied the effect of label noise in the CNN classification model. Rolnick et al [3] claimed that CNNs are robust to massive label noise. Beigman and Beigman [4], Guan et al [7], Lee et al [8], Choi et al [9], and Sukhbaatar and Fergus [10] attempted to develop models from noisy datasets directly. Others such as Brodley and Friedl [11] identified and reduced noisy data using majority voting before training. This research claims that they can make a model robust for up to 30% of label noise. This type of research is subject to the risk of classifying hard labels as noisy labels. To overcome this problem, some researchers attempted to combine noisy data with accurate datasets, as proposed by Zhu [12]. When the label noise was provided, Bootkrajang and Kabán [13] proposed a generic unbiased estimator for binary classification. Unlike electronic health records, images can be re-reported any time with domain experts' efforts. There are several studies that analyzed electronic health records using natural language processing techniques [14,15].

Many have attempted to classify CXR with deep learning techniques. Rajpurkar et al [5] proposed a CNN-based CXR classifier with an overall area under the curve (AUC) ranging between 0.8 and 0.93. Yao et al [16] used a similar method to classify multiclass CXR. Pesce et al [17] used over 430,000 CXRs and proposed an architecture with attention structure based on the evidence that deep learning is robust to label noise [3].

The questions raised were "Are noisy and wrong-labeled datasets credible?" and "Can we believe a CAD model that used these open datasets during training?" In this study, we contemplate the credibility of these datasets and the effect of label noise during training. The aim of this study is threefold: (1) to train computed tomography (CT)-confirmed CXR datasets from Asan Medical Center (AMC) and Seoul National University Bundang Hospital (SNUBH), which can be considered clean with an intentionally given label noise of 0%, 1%, 2%, 4%, 8%, 16%, and 32%; (2) to train NIH and CheXpert datasets, which are considered noisy with an intentionally given label noise of 0%, 1%, 2%, 4%, 8%, 16%, and 32%; and (3) to have the NIH and CheXpert datasets re-evaluated by 3 physicians and one radiologist.

Methods

Image Dataset

Our CXRs were collected from 2 hospitals, AMC and SNUBH in South Korea. Data from 2011 to 2016 were collected. Every CXR was confirmed with its nearest corresponding CT scan and was reevaluated by a chest radiologist with more than 20 years of experience. CXRs contained 5 clinically relevant disease categories, namely, nodule (ND), consolidation (CS), interstitial opacity (IO), pleural effusion (PLE), and pneumothorax (PT). These categories were classified into 2 classes, normal and abnormal. A detailed description of our dataset is provided in Multimedia Appendix 1.

Descriptions of the NIH and the CheXpert datasets can be found in Multimedia Appendices 2 and 3 [6,18]. To validate the NIH and CheXpert datasets, we randomly sampled the same number of normal and abnormal images from the NIH and CheXpert datasets as that from our dataset, that is, all 3 datasets were sampled to have 7103 no finding images and 8680 abnormal images. In the NIH dataset, images were classified into 15 categories including a "no finding" category. For the NIH dataset, we did not distinguish each disease category, but unified all the disease categories into 1 class, "abnormal". In the CheXpert dataset, images were classified into 14 categories including "no finding." In each image class, every image was subclassified as positive/uncertain/negative. We did not use positive/uncertain/negative because the uncertain class can be confusing and negative images were not clinically important. Instead, 14 positive-labeled disease categories were classified as "abnormal," and the "no finding" category was classified as "normal" in the CheXpert dataset. Because there were disease categories present in the CheXpert dataset, which were not in our dataset or the NIH dataset, we unified every disease class as "abnormal" and considered "no finding" as "normal." Furthermore, the "abnormal" class was randomly sampled to be the same number as our "abnormal" dataset without considering the number of each disease class. These "no finding" and "abnormal" dataset descriptions are presented in Table 1.

Table 1. Brief description of the datasets of Asan Medical Center and Seoul National University Bundang Hospital, National Institutes of Health, and CheXpert.

Distribution of images	AMC[a] and SNUBH[b] dataset	NIH[c] dataset	CheXpert dataset
Number of no-finding or normal images	7103	60,361	22,419
Number of abnormal images	8680	51,759	201,897
Number of total images	15,783	112,120	224,316

[a] AMC: Asan Medical Center.

[b] SNUBH: Seoul National University Bundang Hospital.

[c] NIH: National Institutes of Health.

After random shuffling, we analyzed the distribution of 3 randomly shuffled datasets. The distributions of these randomly shuffled datasets are shown in Multimedia Appendix 4.

The label quality of public data from open datasets was evaluated by 3 licensed nonradiologist physicians and 1 board-certified radiologist. For the 3 nonradiologists, in each of the CheXpert and the NIH dataset, we randomly sampled 100 images. In the NIH dataset, 25 images were “abnormal” and 75 images were “no finding.” In CheXpert, 85 images were “abnormal” and 15 images were “normal.” For the radiologist, we randomly selected 200 images from each public dataset. The board-certified radiologist evaluated each given dataset twice, and we recorded the concordance rate for the 2 evaluations. For each open dataset, these images were passed to 3 physicians and 1 radiologist, who reported whether each image belonged to the “no finding” or “abnormal” category.

Image Preprocessing

Every CXR image from the NIH and CheXpert datasets was stored in an 8-bit PNG format. To feed the images in the training model, we changed 3- or 4-channel PNG images to grayscale. The 12-bit DICOM (Digital Imaging and Communications in Medicine) files in our dataset were converted into 8-bit gray PNG format, for which we attempted to set a consistent training condition. In open datasets, sizes of images differed from image to image. To solve this problem, we unified the image size to be 1024×1024 pixels. Similarly, our DICOM images were resized from approximately 2000×2000 pixels to 1024×1024 pixels. Bilinear interpolation was used to resize images, and min-max scaling was applied to each image so that every pixel had a value in the range of 0-1. All the processing was performed using the opencv-python package by Olli-Pekka Heinisuo.

Training Details

Each dataset was classified into 3 groups: training, validation, and test sets. The detailed composition of our dataset including the training, validation, and test sets is presented in Multimedia Appendix 5. Among the various CNN models, CheXNet by Rajpurkar et al [5] was selected as the baseline model. CheXNet is a 121-layered Densenet [19] with 14 disease categories. We changed the last fully connected layer to 1 node to simplify the classification into normal and abnormal. We trained CheXNet from scratch without using the pretrained model. Labels of each training dataset were intentionally misrepresented with rates of 0%, 1%, 2%, 4%, 8%, 16%, and 32%. To generate a training set to have every label noise, we first randomly shuffled all the datasets and changed the label of images in the shuffled list in order from the front. The order was shuffled again to distribute the misrepresented label data evenly in the entire training set. We used Keras python package and Adam optimizer [20] with a learning rate of 0.0001. The loss was set to be binary cross-entropy, and we measured the accuracy with a threshold of 0.5. We trained 20 epochs for each label noise level and each dataset. The training was conducted with a NVIDIA GeForce RTX 2070 for approximately 3 days for each dataset. Moreover, we did not apply label noises for the validation and test sets.

Evaluation Metric and Statistics

For inference, we selected the model with the smallest validation loss in each dataset. In each test set of datasets, we evaluated receiver operating characteristics (ROC) and AUC. The inference results were compared using a semi-log plot. Subsequently, AUC of 0% was compared with each noise level, using standard error defined by Hanley and McNeil [21]. The SE is defined as follows:

where auc is AUC, n_a is the number of abnormal images, and n_n is the number of normal images, and

Results

Accuracies of Each Label Noise

After training 3 datasets with the CNN architecture, ROC curves were drawn as depicted in Figure 1.

Figure 2 illustrates a semilog plot of AUCs of ROC curves from our dataset, the NIH dataset, and the CheXpert dataset for every noise level. Each vertical line means standard error for given AUC.

In the NIH and the CheXpert datasets AUC was poorer than that in our dataset at 0% label noise. The AUC of our dataset was more sensitive to label noise than that of the NIH and the CheXpert datasets. F1 scores are plotted in Figure 3.

Figure 1. Receiver operating characteristic (ROC) curves for datasets of Asan Medical Center and Seoul National University Bundang Hospital, National Institutes of Health, and CheXpert (from left to right) with each label noise rate (0%, 1%, 2%, 4%, 8%, 16%, and 32%).

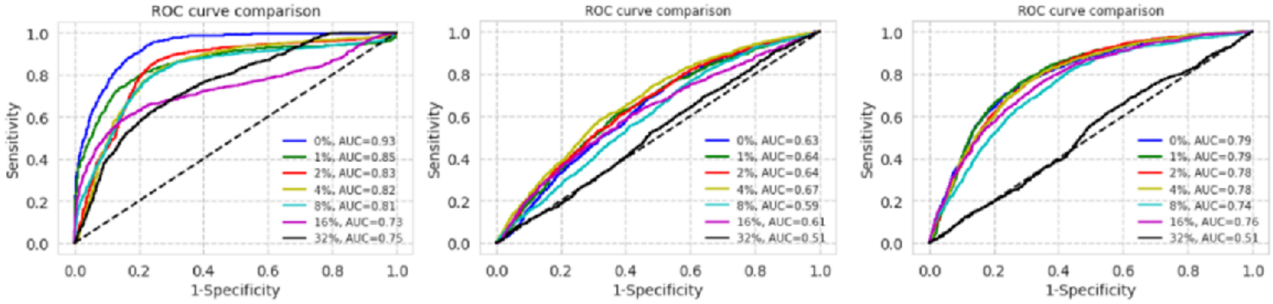


Figure 2. Semilog plot of area under the curves (AUC) of receiver operating characteristic (ROC) curves in the datasets of Asan Medical Center and Seoul National University Bundang Hospital, National Institutes of Health, and CheXpert (from left to right) with each label noise rate (0%, 1%, 2%, 4%, 8%, 16%, and 32%).

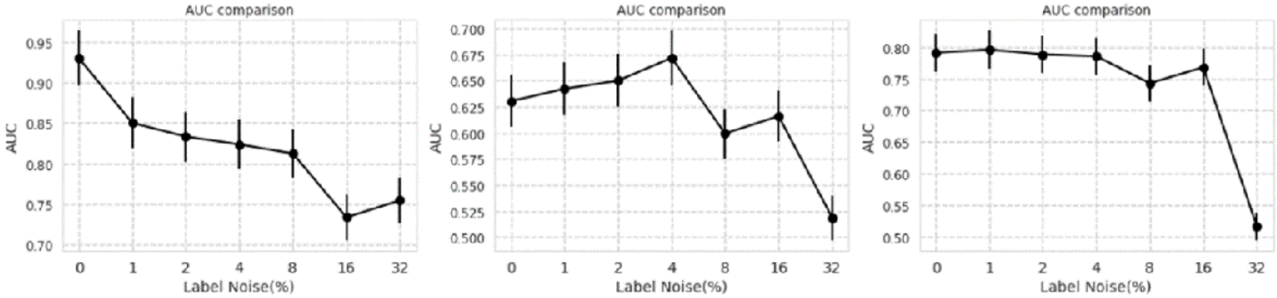
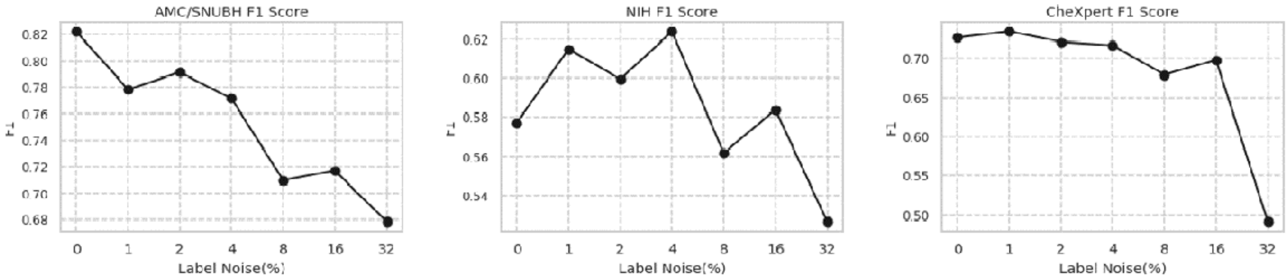


Figure 3. F1 scores of the datasets of Asan Medical Center and Seoul National University Bundang Hospital, National Institutes of Health, and CheXpert (from left to right).



The ROC comparisons for the 3 datasets are presented in Table 2. It became statistically significant when noise level became 2% in our dataset. However, in the NIH and CheXpert datasets, there was no statistical significance until 16% of noise was observable in the training set.

Table 2. Receiver operating characteristic (ROC) comparison for the datasets of Asan Medical Center and Seoul National University Bundang Hospital, National Institutes of Health, and CheXpert.

Dataset and label noise level (%)	Difference of AUC[a] with respect to 0%	*P* value
AMC[b] and SNUBH[c]		
1	0.08	.08
2	0.097	.04
4	0.107	.02
8	0.118	.007
16	0.197	<.001
32	0.176	<.001
NIH[d]		
1	–0.012	.74
2	–0.020	.58
4	–0.041	.24
8	0.031	.37
16	0.014	.68
32	0.111	<.001
CheXpert		
1	–0.005	.91
2	0.003	.99
4	0.005	.90
8	0.048	.86
16	0.022	.94
32	0.028	<.001

[a] AUC: area under the curve.

[b] AMC: Asan Medical Center.

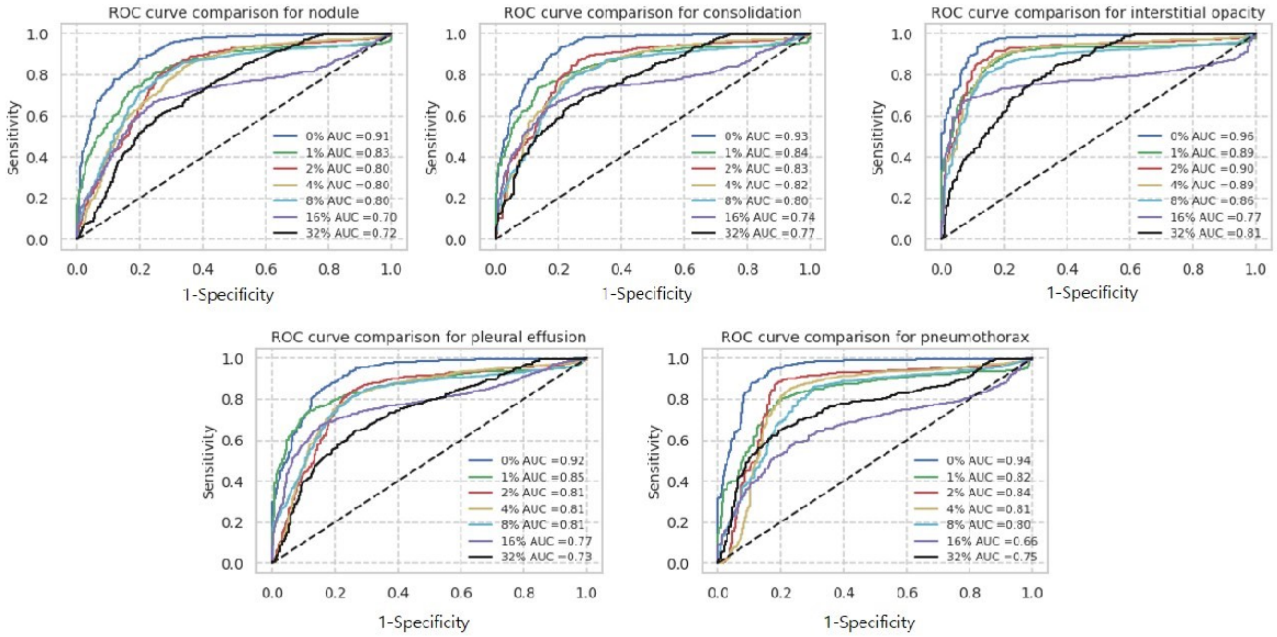
[c] SNUBH: Seoul National University Bundang Hospital.

[d] NIH: National Institutes of Health.

For our dataset, we analyzed subgroups of abnormal cases. It is shown in Figure 4.

There were 1413 normal CXRs, 449 ND CXRs, 322 CS CXRs, 261 IO CXRs, 548 PLE CXRs, 298 PT CXRs in our test set. We joined 1413 normal data with each disease subclass and performed ROC curve analysis. For overall subgroups including ND, CS, IO, PLE, PT, there was no distinguishing subgroup, which was much more sensitive to label noise. However, among these classes, IO was most robust to label noise, showing low decline of AUCs.

Figure 4. Subgroup analysis of abnormal cases in the dataset of Asan Medical Center and Seoul National University Bundang Hospital.

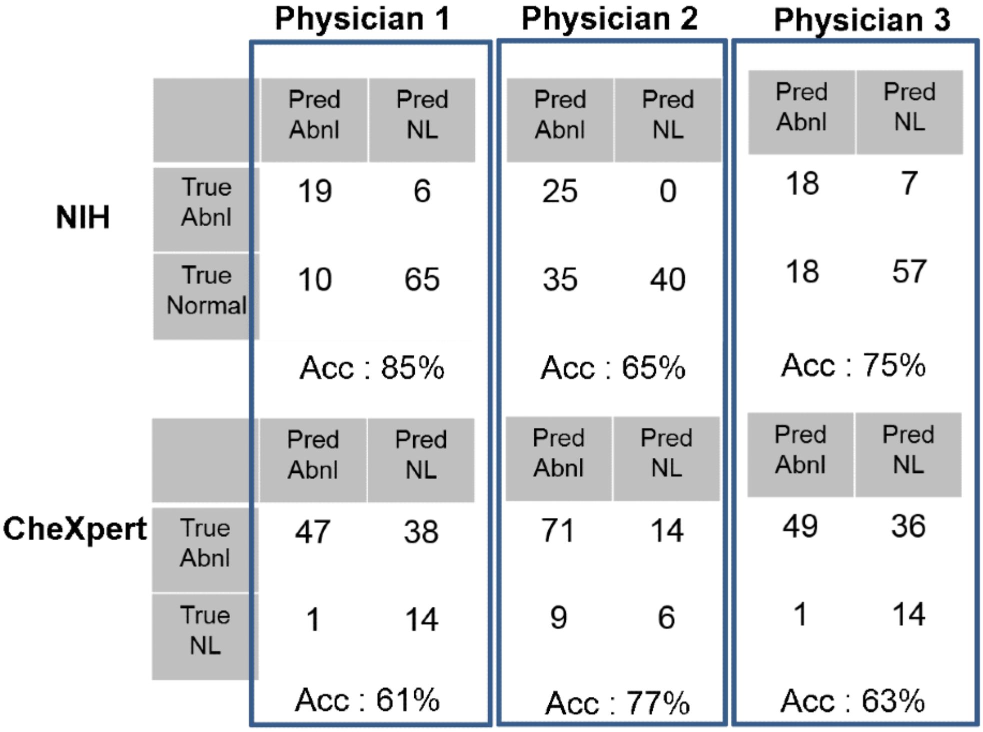


Visual Scoring of Open Dataset

The NIH and the CheXpert datasets were reevaluated by 3 nonradiologist licensed physicians and 1 radiologist. The physicians evaluated CXRs once for each doctor, and the radiologist evaluated CXRs twice. The 3 physicians rated the accuracy of the NIH dataset as 75% (75/100), 65% (65/100), and, 84% (84/100), and that of the CheXpert dataset as 65% (65/100), 77% (77/100) and 61% (61/100), respectively. The radiologist who evaluated CXRs twice rated the accuracy of NIH dataset as 67.5% (135/200) and 65 % (130/200) for each evaluation and rated the accuracy of CheXpert dataset as 81% (162/200) and 77% (154/200) for each evaluation. The concordance rates of 2 evaluations for 2 datasets were 92% (184/200) and 56% (112/200) for the NIH and CheXpert datasets, respectively. Figure 5 depicts the sensitivity and specificity of the report of the 3 physicians. First row is the result of visual scoring by 3 physicians for the NIH dataset, and the second row is the result of visual scoring by 3 physicians for the CheXpert (Stanford) dataset.

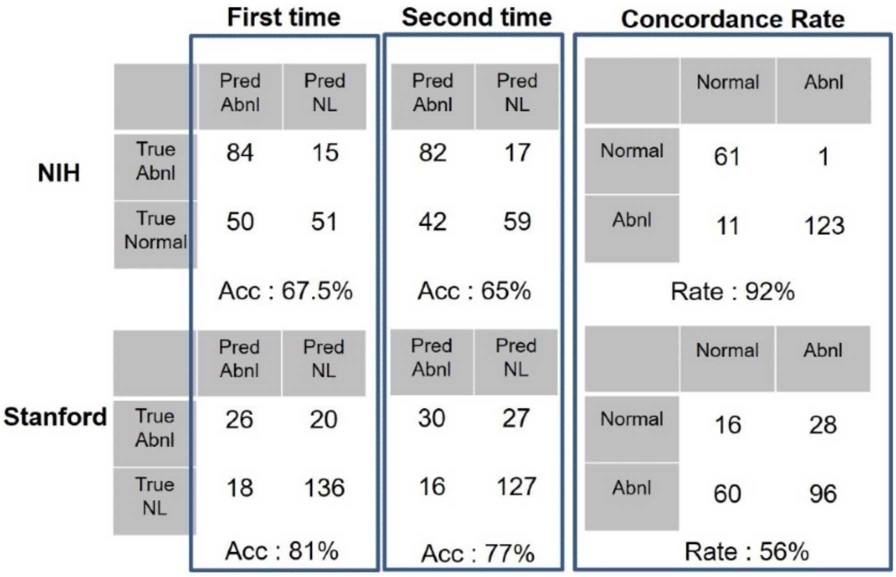
Figure 6 shows the accuracy, sensitivity, specificity of 2 evaluations of 1 radiologist with the concordance rate of 2 evaluations. One radiologist had visually scored 2 public datasets twice. First and second columns from the left show the result of visual scoring for the public datasets. The third column is about concordance rate for the 2 visual scorings for each dataset.

Figure 5. Visual scoring by 3 licensed physicians. Pred: predicted; Abnl: abnormal; NL: normal; NIH: National Institutes of Health; Acc: accuracy.



		Physician 1		Physician 2		Physician 3	
		Pred Abnl	Pred NL	Pred Abnl	Pred NL	Pred Abnl	Pred NL
NIH	True Abnl	19	6	25	0	18	7
	True Normal	10	65	35	40	18	57
		Acc : 85%		Acc : 65%		Acc : 75%	
		Pred Abnl	Pred NL	Pred Abnl	Pred NL	Pred Abnl	Pred NL
CheXpert	True Abnl	47	38	71	14	49	36
	True NL	1	14	9	6	1	14
		Acc : 61%		Acc : 77%		Acc : 63%	

Figure 6. Visual scoring of thoracic radiologist over a 20-year experience. Pred: predicted; Abnl: abnormal; NL: normal; NIH: National Institutes of Health; Acc: accuracy.



		First time		Second time		Concordance Rate		
		Pred Abnl	Pred NL	Pred Abnl	Pred NL		Normal	Abnl
NIH	True Abnl	84	15	82	17	Normal	61	1
	True Normal	50	51	42	59	Abnl	11	123
		Acc : 67.5%		Acc : 65%		Rate : 92%		
		Pred Abnl	Pred NL	Pred Abnl	Pred NL		Normal	Abnl
Stanford	True Abnl	26	20	30	27	Normal	16	28
	True NL	18	136	16	127	Abnl	60	96
		Acc : 81%		Acc : 77%		Rate : 56%		

Discussion

The results of our dataset reveal that the CNN architecture is extremely sensitive to label noise. However, the results of the NIH and CheXpert datasets demonstrate that open datasets are robust to label noise, suggesting that the NIH and CheXpert datasets essentially contain label noises. These datasets do not significantly change the label noise levels and yield robustness despite the label noise. Therefore, training open datasets with CNN architectures has several drawbacks. First, CheXNet cannot be trained in the NIH dataset, because of extensive noise level of NIH dataset. Since open datasets were processed with classical natural language processing, abnormal CXRs were reported to have "no interval change" can be categorized as "no findings." This can amplify label noise of open datasets.

Furthermore, the "no finding" category does not imply normal. There were 15 classes in NIH classified as "no finding," and 14 classes in CheXpert classified as "no finding," suggesting that other lesions may be categorized as "no finding." For example, cavity due to tuberculosis, reticular pattern due to diffuse interstitial lung diseases, hyperinflation due to chronic obstructive lung diseases could be classified as "no finding." Rajpurkar et al [5] reported the CheXNet performance to be similar to that of a radiologist in categorizing pneumonia, rather than a "no finding" category, possibly caused by label noises and/or due to the insufficient performance of CheXNet for differentiating “no finding” and “abnormal.” Therefore, labeling with natural language processing is not suitable for CXR CAD model development. Rating accuracies of our 3 physicians on “no finding” and “abnormal” was approximately 60%-80%, and the accuracy of confirmation by 1 radiologist for the NIH and CheXpert dataset was around 60% and 80%, respectively, which implies that these open datasets have a high occurrence of mislabeled data. The concordance rate of 1 radiologist was 92% (184/200) for NIH and 56% (112/200) for CheXpert. This low concordance rate for CheXpert may have originated from blurry texture of CheXpert images.

To analyze their performance, we experimented the ability of corrected test set of open datasets. First, after the radiologist’s 2-time confirmation, we tested corrected labels using weights of model that were trained with each label noise. The result is shown in Multimedia Appendix 6. Due to the massive label noise of NIH dataset, CheXNet does not work properly for each model of label noise. In CheXpert settings, situation is little bit better yet performance was poor as expected.

There could be an array of additional issues that affect the quality of the open datasets. The CheXpert and NIH datasets are 8-bit PNG image files. Therefore, information loss is unavoidable during conversion from 12-bit DICOM files to the PNG image format.

Robustness of the CheXNet model trained by the NIH and CheXpert datasets does not translate to the robustness of the CNN architecture. The results of our dataset show that CNN is not robust to the noise level. Rather, robustness of the models trained by open datasets can be considered a result of their original impurity. The open datasets are not well-preprocessed, leading them to contain label errors to a certain extent. A low level of label noise does not visibly affect the impurity, and accuracy seems to endure up to 16%.

Regardless of these drawbacks, CNN is considered the best tool for CAD development. Our study urges CAD developers to maximize their effort in accumulating extremely high-quality datasets.

Our study has several limitations. First, we considered only 1 network, CheXNet. Other networks such as ChoiceNet can be robust to label noise [9]. Second, a well-performing model that is robust to label noise is not indicative of its tolerability towards label noise in open datasets. Using open datasets commercially or for research must be seriously considered. Unlike MNIST, they have considerable impacts on the diagnosis of each patient.

Furthermore, it is interesting to speculate active learning with predicted images, which have low confidence levels. That is, predicted labels that have low confidence rate after final activation function, such as 0.4 to 0.6. We might consider them as mislabeled images. Therefore, using high-confidence images and their labels, we can re-label low confidence images assisted by radiologist if needed and train CNN again. This can be used as strategy for training the noisy dataset accurately. However, this strategy is beyond the scope of this study. In our future work, this kind of strategy will be used to train noisy dataset accurately.

As mentioned earlier, even a 1% decrease in accuracy can have an enormous effect on a large patient group. Additionally, categorizing data into “no finding” and “abnormal” may not be ideal as this could be a direct consequence of mislabels on “no finding.” There may be other disease patterns that were not labeled, resulting in an unfair comparison of the 3 datasets with the same criteria. Furthermore, there is a statistical limitation for this study. To compare CNN models exactly, we trained models with only 20 epochs for each label noise level. For some training steps, 20 epochs did not seem sufficient for accuracy saturation. However, we used the same network with the same hyperparameters for these comparisons. For further study, multiple and repetitive training needs to be performed.

In conclusion, the robustness of CAD to label noise with open datasets seems to be a result of their impurity caused by natural language processing. CNN is not robust to label noise in large-sized and complicated images. Therefore, it needs to be emphasized that clean labels and accurate datasets are a necessary condition for developing clinically relevant CAD in medicine.

Acknowledgments

This project was supported by grants of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (HI18C0022, HI18C2383).

Authors' Contributions

RJ conducted experiments, wrote the manuscript, and conducted visual scoring of public datasets (nonradiologist). MJ and Kyung Hwa Lee conducted visual scoring of public datasets (nonradiologist). HNN conducted visual scoring of public datasets (radiologist). SML and Kyung Hee Lee built chest x-ray datasets from Asan Medical Center and Seoul National University Bundang Hospital, respectively. JBS reified experiment instructions. As the project manager, NK contributed to manuscript editing and reified experiment instructions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Dataset description of the Asan Medical Center and Seoul National University Bundang Hospital dataset.

[DOCX File , 15 KB - medinform_v8i8e18089_app1.docx]

Multimedia Appendix 2
Dataset description of the National Institutes of Health (NIH) dataset.
[DOCX File , 15 KB - medinform_v8i8e18089_app2.docx]

Multimedia Appendix 3
Dataset description of CheXpert dataset.
[DOCX File , 15 KB - medinform_v8i8e18089_app3.docx]

Multimedia Appendix 4
Distribution of 3 randomly shuffled datasets.
[DOCX File , 17 KB - medinform_v8i8e18089_app4.docx]

Multimedia Appendix 5
Dataset description for training, validation, and test sets of the Asan Medical Center (AMC) and Seoul National University Bundang Hospital (SNUBH) dataset.
[DOCX File , 16 KB - medinform_v8i8e18089_app5.docx]

Multimedia Appendix 6
Receiver operating characteristic (ROC) curves of corrected test datasets. Left is for the NIH dataset, right is for the CheXpert dataset. One radiologist with over a 20-year experience confirmed 200 images from each dataset twice, and models that have been trained with each label noise were used to draw ROC curves.
[PNG File , 210 KB - medinform_v8i8e18089_app6.png]

References

1. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph 2007;31(4-5):198-211 [FREE Full text] [doi: 10.1016/j.compmedimag.2007.02.002] [Medline: 17349778]
2. Qin C, Yao D, Shi Y, Song Z. Computer-aided detection in chest radiography based on artificial intelligence: a survey. BioMed Eng OnLine 2018 Aug 22;17(1). [doi: 10.1186/s12938-018-0544-y]
3. Rolnick D, Veit A, Belongie S, Shavit N. Deep learning is robust to massive label noise. arXiv preprint arXiv 2017:170510694-170512017.
4. Beigman Klebanov B, Beigman E. From Annotator Agreement to Noise Models. Computational Linguistics 2009 Dec;35(4):495-503. [doi: 10.1162/coli.2009.35.4.35402]
5. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv 2017:171105225-171102017.
6. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers R. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017 Presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017; Honolulu p. 2097-2106. [doi: 10.1109/cvpr.2017.369]
7. Guan M, Gulshan V, Dai A, Hinton G. Who said what: Modeling individual labelers improves classification. 2018 Presented at: Thirty-Second AAAI Conference on Artificial Intelligence; 2018; New Orleans.
8. Lee KH, He X, Zhang L, Yang L. Cleannet: Transfer learning for scalable image classifier training with label noise. 2018 Presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018; Salt Lake City p. 5447-5456.
9. Choi S, Hong S, Lim S. ChoiceNet: robust learning by revealing output correlations. arXiv preprint arXiv 2018:180506431-180502018.
10. Sukhbaatar S, Fergus R. Learning from noisy labels with deep neural networks. arXiv preprint arXiv 2014;2(3):14062080-14062014.
11. Brodley CE, Friedl MA. Identifying Mislabeled Training Data. jair 1999 Aug 01;11:131-167. [doi: 10.1613/jair.606]
12. van Engelen J, Hoos H. A survey on semi-supervised learning. Mach Learn 2019 Nov 15;109(2):373-440. [doi: 10.1007/s10994-019-05855-6]
13. Bootkrajang J, Kabán A. Label-noise robust logistic regression and its applications. 2012 Presented at: Joint European conference on machine learning and knowledge discovery in databases: Springer; 2012; Bristol p. 143-158. [doi: 10.1007/978-3-642-33460-3_15]
14. Jin Y, Li F, Vimalananda VG, Yu H. Automatic Detection of Hypoglycemic Events From the Electronic Health Record Notes of Diabetes Patients: Empirical Study. JMIR Med Inform 2019 Nov 8;7(4):e14340. [doi: 10.2196/14340]

15. Li R, Hu B, Liu F, Liu W, Cunningham F, McManus DD, et al. Detection of Bleeding Events in Electronic Health Record Notes Using Convolutional Neural Network Models Enhanced With Recurrent Neural Network Autoencoders: Deep Learning Approach. JMIR Med Inform 2019 Feb 08;7(1):e10788. [doi: 10.2196/10788]
16. Yao L, Poblenz E, Dagunts D, Covington B, Bernard D, Lyman K. Learning to diagnose from scratch by exploiting dependencies among labels. In: arXiv preprint arXiv. 2018 Presented at: Sixth International Conference on Learning Representations; Mon Apr 30th through May 3rd, 2018; Vancouver p. 171010501-171012017.
17. Pesce E, Joseph Withey S, Ypsilantis P, Bakewell R, Goh V, Montana G. Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. Medical Image Analysis 2019 Apr;53:26-38. [doi: 10.1016/j.media.2018.12.007]
18. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. AAAI 2019 Jul 17;33:590-597. [doi: 10.1609/aaai.v33i01.3301590]
19. Huang G, Liu Z, Van DML, Weinberger K. Densely connected convolutional networks. 2017 Presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017; Hawai p. 4700-4708. [doi: 10.1109/CVPR.2017.243]
20. Kingma D, Ba J. Adam: A method for stochastic optimization. In: arXiv preprint arXiv. 2015 Presented at: 6th International Conference on Learning Representations; May 7 - 9, 2015; San Diego p. A.
21. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982 Apr;143(1):29-36. [doi: 10.1148/radiology.143.1.7063747]

Abbreviations

AMC: Asan Medical Center
AUC: area under the curve
CAD: computer-aided diagnosis
CNN: convolutional neural network
CS: consolidation
CT: computed tomography
CXR: chest x-ray
DICOM: Digital Imaging and Communications in Medicine
IO: interstitial opacity
MNIST: Modified National Institute of Standards and Technology
ND: nodule
NIH: National Institutes of Health
PLE: pleural effusion
PT: pneumothorax
ROC: receiver operating characteristic
SNUBH: Seoul National University Bundang Hospital

Edited by C Lovis; submitted 02.02.20; peer-reviewed by G Lim, C Lin; comments to author 06.05.20; revised version received 08.06.20; accepted 21.06.20; published 04.08.20.

Please cite as:
Jang R, Kim N, Jang M, Lee KH, Lee SM, Lee KH, Noh HN, Seo JB
Assessment of the Robustness of Convolutional Neural Networks in Labeling Noise by Using Chest X-Ray Images From Multiple Centers
JMIR Med Inform 2020;8(8):e18089
URL: https://medinform.jmir.org/2020/8/e18089
doi:10.2196/18089
PMID:32749222

Original Paper

Semantic Deep Learning: Prior Knowledge and a Type of Four-Term Embedding Analogy to Acquire Treatments for Well-Known Diseases

Mercedes Arguello Casteleiro[1*], DPhil; Julio Des Diz[2*], DPhil, MD; Nava Maroto[3*], DPhil; Maria Jesus Fernandez Prieto[4*], DPSI; Simon Peters[5*], DPhil; Chris Wroe[6*], BChir, MD; Carlos Sevillano Torrado[2*], MD; Diego Maseda Fernandez[7*], MD; Robert Stevens[1*], DPhil

[1]Department of Computer Science, University of Manchester, Manchester, United Kingdom

[2]Hospital do Salnés, Villagarcía de Arousa, Spain

[3]Departamento de Lingüística Aplicada a la Ciencia y a la Tecnología, Universidad Politécnica de Madrid, Madrid, Spain

[4]Salford Languages, University of Salford, Salford, United Kingdom

[5]School of Social Sciences, University of Manchester, Manchester, United Kingdom

[6]BMJ, London, United Kingdom

[7]Mid Cheshire Hospital Foundation Trust, NHS England, Crewe, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Robert Stevens, DPhil
Department of Computer Science
University of Manchester
Kilburn Building, Oxford Road, M13 9PL
Manchester,
United Kingdom
Phone: 44 161 275 6251
Email: robert.stevens@manchester.ac.uk

Abstract

Background: How to treat a disease remains to be the most common type of clinical question. Obtaining evidence-based answers from biomedical literature is difficult. Analogical reasoning with embeddings from deep learning (embedding analogies) may extract such biomedical facts, although the state-of-the-art focuses on pair-based proportional (pairwise) analogies such as man:woman::king:queen (“*queen = −man +king +woman*”).

Objective: This study aimed to systematically extract disease treatment statements with a Semantic Deep Learning (SemDeep) approach underpinned by prior knowledge and another type of 4-term analogy (other than pairwise).

Methods: As preliminaries, we investigated Continuous Bag-of-Words (CBOW) embedding analogies in a common-English corpus with five lines of text and observed a type of 4-term analogy (not pairwise) applying the 3CosAdd formula and relating the semantic fields *person* and *death*: “dagger = −Romeo +die +died” (search query: *−Romeo +die +died*). Our SemDeep approach worked with pre-existing items of knowledge (what is known) to make inferences sanctioned by a 4-term analogy (search query *−x +z1 +z2*) from CBOW and Skip-gram embeddings created with a PubMed systematic reviews subset (PMSB dataset). Stage1: Knowledge acquisition. Obtaining a set of terms, candidate y, from embeddings using vector arithmetic. Some n-gram pairs from the cosine and validated with evidence (prior knowledge) are the input for the 3cosAdd, seeking a type of 4-term analogy relating the semantic fields disease and treatment. Stage 2: Knowledge organization. Identification of candidates sanctioned by the analogy belonging to the semantic field treatment and mapping these candidates to unified medical language system Metathesaurus concepts with MetaMap. A concept pair is a brief disease treatment statement (biomedical fact). Stage 3: Knowledge validation. An evidence-based evaluation followed by human validation of biomedical facts potentially useful for clinicians.

Results: We obtained 5352 n-gram pairs from 446 search queries by applying the 3CosAdd. The microaveraging performance of MetaMap for candidate *y* belonging to the semantic field *treatment* was F-measure=80.00% (precision=77.00%, recall=83.25%). We developed an empirical heuristic with some predictive power for *clinical winners*, that is, search queries bringing candidate *y* with evidence of a therapeutic intent for target disease *x*. The search queries *-asthma +inhaled_corticosteroids*

+inhaled_corticosteroid and *-epilepsy +valproate +antiepileptic_drug* were *clinical winners*, finding eight evidence-based beneficial treatments.

Conclusions: Extracting treatments with therapeutic intent by analogical reasoning from embeddings (423K n-grams from the PMSB dataset) is an ambitious goal. Our SemDeep approach is knowledge-based, underpinned by embedding analogies that exploit prior knowledge. Biomedical facts from embedding analogies (4-term type, not pairwise) are potentially useful for clinicians. The heuristic offers a practical way to discover beneficial treatments for well-known diseases. Learning from deep learning models does not require a massive amount of data. Embedding analogies are not limited to pairwise analogies; hence, analogical reasoning with embeddings is underexploited.

(JMIR Med Inform 2020;8(8):e16948) doi:10.2196/16948

KEYWORDS

evidence-based practice; artificial intelligence; deep learning; semantic deep learning; analogical reasoning; embedding analogies; PubMed

Introduction

How to treat a disease or condition remains to be the most common type of clinical question [1]. It is difficult for clinicians to obtain comprehensive information on the clinical (and economic) worth of alternative drug choices for a given condition [2]. Evidence-based biomedical literature, although available in electronic form, primarily remains to be expert-to-expert communication—natural language statements intended for human consumption.

Analogical reasoning is basic relational reasoning without explicit representations of relations [3]. An acknowledged semantic property of embeddings (ie, vectors representing terms) from deep learning [4] is "*their ability to capture relational meanings*" [5], the so-called analogies [6]. Current efforts in analogical reasoning with embeddings focus on pair-based proportional analogies [5,7,8]. This is a type of "*the four-term analogy*" [6], also known as the cross-mapping analogy [6]. An example is *queen* = *−man +king +woman* [9], also represented as man:woman::king:queen [10], and read as "*man is to king as woman is to queen*" [11]. Examples for health care include the following:

- "*'acetaminophen' is as type of 'drug' as 'diabetes' is as type of 'disease'*" [12].
- "*(furosemide - kidney) + heart ~ fosinopril*" [13].

This study aimed to investigate embedding analogies (analogical reasoning with embeddings) [5] that are not pair-based proportional (pairwise for short) analogies. This study began by observing senior clinicians performing an analogical reasoning for sepsis (a major life-threatening condition) with embeddings and posing search queries such as *−sepsis +serum_albumin +fluid_therapy* to discover treatments with therapeutic intent. The clinical rationale behind this query is that "*current evidence suggests that resuscitation using albumin-containing solutions is safe*" [14], where *serum_albumin* is a shortened form of "*human serum albumin supplementation*" (extensively debated for sepsis [15]). We viewed this as another type of *the four-term analogy,* which is not pairwise.

This paper presents a semiautomatic approach to extract meaning (semantics) from the unstructured free text of biomedical literature (ie, PubMed systematic reviews [16]). The disease treatment statements systematically acquired from analogical reasoning are biomedical facts validated with evidence first and human audit afterward. The approach presented belongs to Semantic Deep Learning (SemDeep) [17], as we used embedding analogies (other than pairwise) and semantic knowledge representation paradigms [18] to provide meaning for the same.

Analogical Reasoning

Humans possess the ability to reason by analogy using abstract semantic relations such as synonyms or category membership [3]. For example, *common cold* and *influenza* are both types of illnesses with some common symptoms such as runny nose, sore throat, cough, and headache. As they share some key characteristics, we can possibly say they are near-synonyms, although they cannot be used interchangeably (as synonyms would) because of key medical differences. Our SemDeep approach acquires terms about treatments for a well-known disease using analogical reasoning that is underpinned by Aristotle's theory [19]:

- "*The strength of an analogy depends upon the number of similarities*" [19]. For example, "intravenous antibiotics" and "intravenous fluid resuscitation" are basic therapies that improve outcomes in patients with sepsis [14], that is, both are treatments with a therapeutic intent for sepsis. However, we cannot say that they are similar as "intravenous fluid resuscitation" is a procedure whereas an "intravenous antibiotic" is a substance, although both are "intravenous."
- "*Similarity reduces to identical properties and relations*" [19]. For example, "benzyl penicillin," "cefotaxime," or "amoxicillin/clavulanate" is similar as they belong to the same category, "antipseudomonal beta-lactam antibiotics" [14].
- "*Good analogies derive from underlying common causes or general laws*" [19]. This study investigated the systematic acquisition of treatments for a disease using the simple generic 3CosAdd formula [20,21].

The 3CosAdd Formula

Our work relied on vector semantics [5] and used the neural language models, Continuous Bag-of-Words (CBOW) and Skip-gram by Mikolov et al [20], from deep learning to create embeddings. Embeddings with vector semantics such as cosine or 3CosAdd can acquire a list of strings of characters (eg, n-grams), although they lack explicit semantic meaning. Until now, the 3CosAdd formula 1 [20,21] has been applied to analogies between 2 pairs of words a:a*::b:b* [7], where b* is the unknown (hidden) word.

We used the 3CosAdd formula as in Levy and Goldberg [21] with the rewording "*find the term y, which is similar to the term z1 and the term z2, while different from the term x*", where the target term *x* provides the semantic context and similar refers to the terms sharing "commonalities in structural features." In this study, a semantic field is a set of terms that "*belong together under the same conceptual heading*" [22] and is a form of knowledge representation that provides meaning to those terms. We rewrote the 3CosAdd formula as formula 2, with the search query *-x +z1 +z2*, and the type of 4-term analogy we sought had the following:

- The target term *x* belonging to the semantic field *disease* and representing a medical diagnosis mappable to a "type of" of systematized nomenclature of medicine - clinical terms (SNOMED CT) [23] concept called disorder.
- The 3 terms {*z1; z2; y*} belonging to the semantic field *treatment*(Tx for short), where Tx encompasses 3 textual definitions from Hart et al [24]. The candidate term *y* is the unknown.

The Research Questions

We adopted the view by Hill et al [25] by considering "relatedness" as "association" and synonymy as the strongest similarity. In this study, the association relationship of interest is "correlation," as defined in the semantic science integrated ontology (SIO) [26].

As preliminaries, we asked 2 research questions not specific to the health care domain:

- Q1: Can "good" embeddings be created with a small corpus?
- Q2: If the simple generic 3CosAdd formula [20,21] can capture a type of 4-term analogy as read in formula 2, can they be observed in embeddings created with a small corpus?

Our third research question (Q3) asked whether the 4-term type of analogy discovered in a small common-English corpus can also be discovered in a larger-scale biomedical corpus. To provide proof of such a generalization, we performed a real-world test with embeddings created with free text from PubMed systematic reviews [16]. We postulated that candidate inferences can be validated using evidence-based information resources. This study investigated the discovery of *clinical winners*, that is, search queries *-x +z1 +z2* bringing candidate treatments *y* with evidence of a therapeutic intent for target disease *x*; thus, enabling the most common type of clinical question, "how to treat a disease or condition" [1], to be answered.

Our final research question (Q4) asked for some predictive power over the *clinical winners* obtained (ie, an empirical heuristic) if our SemDeep approach worked, that is the type of analogy proposed finds disease treatment statements from PubMed systematic reviews (ie, a larger-scale biomedical corpus). This last question pursued a tacit preference and referred to the final characteristic of analogy: systematicity [6]. However, challenges have been acknowledged "for any vector space model that aims to make predictions about relational similarity" [27].

Between the semantic field *disease* and the semantic field *treatment*, "*few maximal structurally consistent interpretations (ie, mappings displaying one-to-one correspondences and parallel connectivity)*" [6] are to be expected. For example, aspirin treatment does not have a one-to-one correspondence with a disease as it can treat headache (common knowledge) and acute myocardial infarction [1]. In this study, "spontaneous unplanned inferences" [6] were also expected, and this propensity was captured with the notion of incremental mappings [6].

Methods

Overview

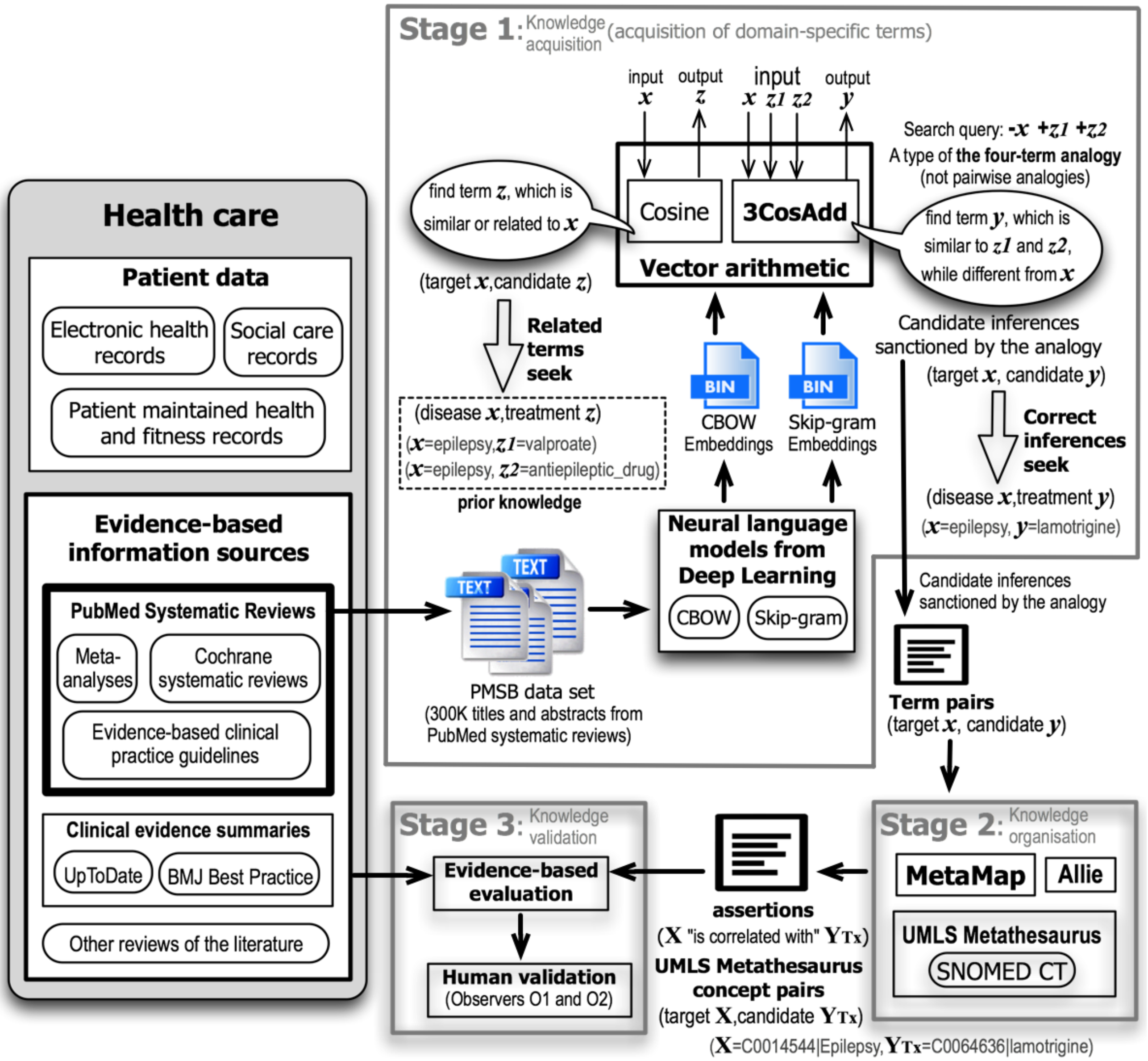
Our SemDeep approach answered Q3 and comprised the 3 stages depicted in Figure 1. The software package word2vec [28] implements the CBOW and Skip-gram algorithms along with the cosine and 3CosAdd formulas. The terms in this study are n-grams.

Stage 1 used prior knowledge (open-access reusable datasets [29]) consisting of n-gram pairs obtained by applying the cosine to embeddings, then mapped to the Unified Medical Language System (UMLS) Metathesaurus [30] concept pairs, and finally validated with evidence from biomedical literature using the British Medical Journal (BMJ) Best Practice [31] as the main information source.

BMJ Best Practice is separate from PubMed/MEDLINE [32] and is acknowledged for its editorial quality and evidence-based methodology [33]. In the United Kingdom, BMJ Best Practice is provided (free access) to all National Health Service (NHS) health care professionals in England, Scotland, and Wales [34]. BMJ Best Practice provides advice on symptom evaluation, tests to order, and treatment approach structured around the patient consultation.

We started by investigating embedding analogies in a small common-English corpus to answer Q1 and Q2.

Figure 1. Overview of our SemDeep approach.


Stage 1: Knowledge acquisition (acquisition of domain-specific terms)
input x
output z
input x z1 z2
output y
Search query: -x +z1 +z2
A type of the four-term analogy
(not pairwise analogies)
find term z, which is similar or related to x
Cosine
3CosAdd
Vector arithmetic
find term y, which is similar to z1 and z2, while different from x
(target x,candidate z)
Related terms seek
(disease x,treatment z)
(x=epilepsy,z1=valproate)
(x=epilepsy, z2=antiepileptic_drug)
prior knowledge
BIN
CBOW Embeddings
BIN
Skip-gram Embeddings
Candidate inferences sanctioned by the analogy
(target x, candidate y)
Correct inferences seek
(disease x,treatment y)
(x=epilepsy, y=lamotrigine)
Health care
Patient data
Electronic health records
Social care records
Patient maintained health and fitness records
Evidence-based information sources
PubMed Systematic Reviews
Meta-analyses
Cochrane systematic reviews
Evidence-based clinical practice guidelines
Clinical evidence summaries
UpToDate
BMJ Best Practice
Other reviews of the literature
TEXT
TEXT
PMSB data set
(300K titles and abstracts from PubMed systematic reviews)
Neural language models from Deep Learning
CBOW
Skip-gram
Candidate inferences sanctioned by the analogy
Term pairs
(target x, candidate y)
Stage 3: Knowledge validation
Evidence-based evaluation
Human validation
(Observers O1 and O2)
assertions
(X "is correlated with" YTx)
UMLS Metathesaurus concept pairs
(target X,candidate YTx)
(X=C0014544|Epilepsy, YTx=C0064636|lamotrigine)
Stage 2: Knowledge organisation
MetaMap
Allie
UMLS Metathesaurus
SNOMED CT

Preliminaries: Analogies for Shakespeare's Romeo in a Small Common-English Corpus

Topic models are related to semantic fields [5]. There are many small corpora and tutorials illustrating the inner workings of topic models, such as the spatially motivated Latent Semantic Analysis (LSA) method [35] and the probabilistic method latent Dirichlet allocation (LDA) [36]. We used a small common-English corpus appearing in an LSA tutorial [37]. Textbox 1 shows the corpus used to answer Q1 and Q2.

Textbox 1. A small common-English corpus consisting of 5 lines of text.

Romeo and Juliet

Juliet: Oh happy dagger!

Romeo died by dagger.

"Live free or die", that's the New-Hampshire's motto

Did you know New-Hampshire is in New-England?

In common English, punctuation marks can change the meaning of a sentence. For example, "prevail, not perish" versus "prevail not, perish." We did not transform routine letters into lowercase letters and did not remove punctuation marks, with the only exception of double quotations. Multimedia Appendix 1 contains the input text and the hyperparameter configuration for the CBOW model with word2vec [28].

Below, we summarize the answers to Q1 and Q2 (Multimedia Appendix 1):

- Answer to Q1: A “good” vector semantic model should find a candidate *y* that is “semantically similar” to the target *x* = *Romeo*. The candidate *y* with the highest cosine for the CBOW model is *you*: The terms *you* and *Romeo* are near-synonyms, that is “interchangeable in some contexts” [38]. Hence, the answer to Q1 is “yes.”
- Answer to Q2: We applied the 3CosAdd formula 2, where the target *x* = *Romeo* provides the semantic context. The terms *die* = *z1* and *died* = *z2* from the corpus are representative of inflectional morphology *infinitive:past*. The search query *–x +z1 +z2* is posed to the CBOW model, that is “find the term *y*, which is similar to *die* and *died*, while different from *Romeo*”. Candidate *y* with the highest 3CosAdd is “*dagger*.” The term *dagger* belongs to the semantic field *death* as “dagger is an instrument that causes death”; thus, the candidate inference is true. Hence, the answer to Q2 is also “yes.”

Stage 1: Knowledge Acquisition (Acquisition of Domain-Specific Terms)

The PubMed systematic reviews (in Figure 1) [16] is an evidence-based searching filter “AND (systematic [sb])”, intended for retrieving “best evidence” information sources from PubMed/MEDLINE [32] such as Cochrane systematic reviews [39]. Health care–related institutions such as the World Health Organization promote PubMed searches with this filter (examples in Prevention and Control of Noncommunicable Diseases: Guidelines for Primary Health Care in Low Resource Settings [40]).

This study used a subset of PubMed systematic reviews [16] of 301,201 PubMed/MEDLINE publications (titles and available abstracts), called the PubMed systematic reviews subset (PMSB dataset). The preprocessing of the input text for the PMSB dataset and the hyperparameter configuration for Skip-gram and CBOW are identical to those in our previous study [41] and detailed in the study by Arguello Casteleiro et al [42].

From the PMSB dataset, a total of 423K n-grams with a frequency count >5 have vector representations in both models, that is CBOW and Skip-gram. We considered “good” the Skip-gram and CBOW embeddings created in our previous study [41] as they both perform well (using conventional evaluation measure precision [43]) in semantic similarity and relatedness tasks with the cosine formula. The n-gram *z* reused in this study (ie, z1 and z2) is from our previous study [41].

Applying the 3CosAdd Formula to Acquire the Top 12 Ranked Term Pairs (x,y): A 4-Term Analogy

To address Q3 and apply the 3CosAdd formula 2, 2 n-gram pairs (disease *x*,treatment *z*) from our previous study (prior knowledge) [41] were needed. We kept only the 12 top-ranked candidate n-grams *y* for the 3CosAdd formula, that is, the 12-candidate *y* with CBOW and Skip-gram embeddings yielding the highest 3CosAdd values. We limited the list of candidates to 12, similar to Arguello Casteleiro et al [42], and following cognitive theories like Novak JD and Cañas AJ [44].

Stage 2: Knowledge Organization (Explicit Conceptualization of the Meaning of Terms)

This stage accomplishes a named entity recognition (NER) [45] task involving 3 domain experts (2 biomedical terminologists and 1 medical consultant who performs clinical coding). Every UMLS Metathesaurus concept has a concept unique identifier (CUI) and at least one UMLS Semantic Type (broad category) [30] assigned. The NER task consists of 3 sequential subtasks (Multimedia Appendix 1):

- First, disambiguation of n-grams *y* is difficult to interpret for being truncated strings of characters or containing short forms (eg, abbreviations or acronyms). String searches in the PMSB dataset and the web search the sense inventory, Allie [46], enabling disambiguation.
- Second, the manual binary classification of candidate n-gram *y* as to whether it belongs to the semantic field Tx (ie, y_{Tx}). Following Artstein R and Poesio M [47], we reported the interrater agreement with a Krippendorff alpha [48].
- Third, entity normalization (grounding) [49] with MetaMap [50], where 3 domain experts apply the NER guidelines for MetaMap's output [51] and together judge the automatic mapping of n-grams y_{Tx} to UMLS Metathesaurus concepts Y_{Tx}. MetaMap performance is calculated using precision, recall, and F-measure [43,52].

We took n1 as the number of different UMLS Metathesaurus concepts (represented as Z1 and Z2) mapped as *z1* and *z2* in the search query. Once the NER task was completed, we obtained the *NER winners*. An *NER winner* was a search query *-x +z1 +z2* with the maximum observed number for n2 or n3:

- n2 is the number of different 12 top-ranked candidate n-grams *y* belonging to Tx, that is, the number of y_{Tx}.
- n3 is the number of different UMLS Metathesaurus concepts Y_{Tx} excluding Z1 and Z2.

Stage 3: Knowledge Validation (Validating Statements)

We sought evidence for the Metathesaurus concept pairs (X,Y_{Tx}) acquired previously to determine the therapeutic intent of candidate Y_{Tx} for target disease *X*, where *X* was the UMLS Metathesaurus concept mapped to n-gram *x*.

The same 3 domain experts from Stage 2 triaged the results of manual literature searches considering the following:

1. The type of evidence-based information sources, seeking the “best evidence.” Evidence-based medicine [53] categorizes and ranks different types of clinical evidence [1]. For example, the Cochrane systematic reviews are at the forefront of “best evidence” [1], whereas studies of the physiological functions and clinicians’ observations are considered evidence of least value [1].
2. The publication date, seeking the “most recent papers published.”

The 3 domain experts introduced 6 evidence-based categories to further refine the correlations between the semantic field *disease* and *treatment* (Tx). Table 1 illustrates them with examples of evidence (quoted text) and references for the UMLS

Metathesaurus concepts Y_{Tx} related to the target concept disease X="C024302|Sepsis" with CUI=C024302. The rationale for the 7 evidence-based categories introduced is as follows:

- The name of 4 of the evidence-based categories (top rows in Table 1) resembles the categories "beneficial, likely beneficial, no known benefit, harmful" for health care interventions from the decommissioned BMJ Clinical Evidence (predecessor of BMJ Best Practice [31]).
- The evidence-based category "Tx ingredient" acknowledges that a complex treatment may have parts, that is "partitive relationships" [54].
- The evidence-based category "correlation" captures "spontaneous unplanned inferences" [6].
- The evidence-based category "general medical term" includes broad concepts of little value for clinicians that do not need further evidence (quotes and references).

This study distinguishes between *NER winners* (maximum observed number for n2 or n3 in Stage 2) and *clinical winners*. A *clinical winner* is a search query *-x +z1 +z2* (a type of 4-term analogy) for target disease *x* with a maximum observed value for n4, that is, the number of different concepts Y_{Tx} (excluding Z1 and Z2) assigned to the evidence-based category "Tx with therapeutic effect."

To audit the evidence-based categories assigned along with the evidence collected (quotes and references) for the concept pairs (X,Y_{Tx}), 2 more observers (O1 a medical consultant and O2 a BMJ health informatician who works with BMJ Best Practice content and has a junior doctor background) were asked to express agreement or disagreement with the evidence for the concept pairs (X,Y_{Tx}). Multimedia Appendix 1 has the evaluation guidelines given to the observers. Cohen kappa [55] was used to measure interobserver agreement.

Table 1. Evidence from the literature searches, that is quoted text and reference, for unified medical language system Metathesaurus concept pairs (X, Y_{Tx}) with X=C0243026|Sepsis.

Candidate concept Y_{Tx}; UMLS CUI|Concept name[a]	Evidence-based categories for concept Y_{Tx} correlated with concept X	Evidence (quoted text) [evidence source] [citation]
C0056562|crystalloid solutions	Tx with therapeutic effect	"Step-by-step treatment approach: ... Administer 30 mL/kg crystalloid for hypotension or lactate ≥4 mmol/L (≥36 mg/dL)" [BMJ BP topic: 245] [14]
C0001617|Adrenal Cortex Hormones	Tx with uncertain therapeutic effect	"Step-by-step treatment approach: Adjunctive therapies ... evidence for giving corticosteroids to patients with sepsis or septic shock is mixed." [BMJ BP topic: 245] [14]
C0020352|Hetastarch	Tx with unwanted or adverse effects (ie, nontherapeutic)	"Step-by-step treatment approach: Fluid resuscitation ... HES solutions for infusion have been significantly restricted across the European Union and are contraindicated in critically ill patients and those with sepsis or renal impairment." [BMJ BP topic: 245] [14]
C0677850|Adjuvant therapy	Potential Tx (under research and development)	"Adjuvant immune therapy to manipulate the hyper-inflammatory and/or immune-suppressive phase of sepsis is an attractive therapeutic option, which may improve outcome and ease the burden of antimicrobial resistance. However, before this can become a clinical reality, we must recognise that sepsis is a clinical syndrome, where significant heterogeneity exists." [PMID: 30515242] [56]
C3273371|CD4 Positive Memory T-Lymphocyte	Tx ingredient	"Administration of immune-modulatory therapy is a promising treatment approach for treating sepsis survivors. … these therapies can improve pathogen clearance, increase CD4 T cell responsiveness, and promote survival in sepsis." [PMID: 24791959] [57]
C0745442|Intravenous Catheters	Tx ingredient	"Recommendations: Monitoring ... Central venous catheters will be required to ensure reliable delivery of vasoactive medication." [BMJ BP topic: 245] [14]
C0812144|Medication administration: epidural	Correlation (epidural → potential sites of infection: epidural sites → sepsis: investigations)	"Investigations to identify causative organisms: ... If no localising signs are present, examination and culture of all potential sites of infection including wounds, catheters, prosthetic implants, epidural sites, and pleural or peritoneal fluid, as indicated by the clinical presentation and history, is required." [BMJ BP topic: 245] [14]
C0013227|Pharmaceutical Preparations	General medical term	__[b]

[a] The references shown are either the PubMed identifier (PMID) or the topic number in BMJ Best Practice ("BMJ BP topic" for short).

[b] The evidence-based category "general medical term" has no evidence (quoted text).

Results

We obtained 5352 n-gram pairs from 446 search queries by applying the 3CosAdd formula and taking the top 12 values. These are presented in Multimedia Appendix 2 (worksheet Stage 1). These n-gram pairs are enriched with domain knowledge meaning (Stage 2) and the biomedical evidence found from literature searches is ratified with an audit (Stage 3).

Stage 1: Knowledge Acquisition (Acquisition of Domain-Specific Terms)

To apply the 3CosAdd formula (and systematic creation of search queries), we reused 63 unique n-gram pairs (*x*,*z*) from our previous study [41] (open-access [29]). Every reused n-gram *z* was mapped to the UMLS concept *Z* with the UMLS Semantic Type "T061|Therapeutic or Preventive Procedure" or "T121|Pharmacologic Substance." Multimedia Appendix 1 has the UMLS CUI pairs (*X*,*Z*).

Applying the 3CosAdd Formula to Acquire the Top 12 Ranked Term Pairs (x,y): A 4-Term Analogy

With 63 n-gram pairs (*x*,*z*), we built 223 search queries *-x +z1 +z2* for the 3CosAdd formula. Multimedia Appendix 2 (worksheet Stage 1) contains the 223 search queries and the 5352 (*x*,*y*) n-gram pairs for 10 target diseases *x*, that is, the 12 top-ranked n-grams (highest 3CosAdd value) obtained per search query from the CBOW and Skip-gram embeddings. An n-gram pair with *y* as a non-ASCII character was discarded.

Stage 2: Knowledge Organization (Explicit Conceptualization of the Meaning of Terms)

Different search queries brought the same (target *x*,candidate *y*) n-gram pairs from applying the 3CosAdd formula. Multimedia Appendix 2 (worksheet Stage 2) has 1935 unique (*x*,*y*) n-gram pairs from the 5352 n-gram pairs. Among the 1935 unique (*x*,*y*) n-gram pairs, there were 954 n-gram pairs (x,y_{Tx}) with candidate *y* belonging to Tx. The Krippendorff alpha [48] was 0.86 for the 3 domain experts for the binary classification (Tx or non-Tx). Considering all candidates y_{Tx} mapped to Y_{Tx} for the 10 diseases (microaveraging) [43], MetaMap had an F-measure=80.00% with precision=77.00% and recall=83.25%. Multimedia Appendix 1 has the detailed results for NER subtasks, including an investigation of the UMLS semantic types for Y_{Tx}.

Table 2 contains the *NER winners*, that is, the search query *-x +z1 +z2* for the 3CosAdd formula per model and disease target *x* having the maximum observed values for n2 or n3.

- The maximum observed value for n2 was the highest possible value, that is, n2=12, for both CBOW and Skip-gram.
- The maximum observed value for n3 was for the search query, *−epilepsy +valproate +AED*. However, the number of different Y_{Tx} (excluding Z1 and Z2) differed, that is, n3=11 for Skip-gram and n3=10 for CBOW.

Stage 3: Knowledge Validation (Validating Statements)

Multimedia Appendix 2 (worksheet Stage 3) has the 569 unique UMLS Metathesaurus concept pairs (X,Y_{Tx}) mapped to the unique 954 n-gram pairs (x,y_{Tx}). Although the UMLS related concepts table (file=MRREL) [58] contains relationships asserted by source vocabularies between CUI pairs, only 68 of the 569 CUI pairs appeared within the MRREL table of 2019AA UMLS release.

Manual searches in the literature proved to be time-consuming and labor-intensive; thus, not all the concept pairs for the target disease anemia and hypertension had evidence. Hence, we limited the study to 408 UMLS CUI pairs (Multimedia Appendix 1), and only 59 of these were within the MRREL table (column J of Multimedia Appendix 2 worksheet Stage 3).

Table 2. NER winners per target disease x (search query -x +z1 +z2) for the 3CosAdd formula, that is, the highest value for n2 or n3 per model and per disease target x.

Disease target *x*	Model	NER max (n2)	NER max (n3)	Treatment *z1* search query	Treatment *z2* search query	n1	n2	n3
heart_failure	CBOW[a]	Yes	N/A[b]	angiotensin-converting_enzyme_(ACE)_inhibitors	aldosterone_antagonists	2	12	6
heart_failure	CBOW	N/A	Yes	cardiac_resynchronization_therapy_(CRT)	aldosterone_antagonists	2	10	9
heart_failure	Skip-gram	Yes	Yes	beta-blockers	aldosterone_antagonists	2	12	8
glaucoma	CBOW	Yes	Yes	trabeculectomy	cataract_surgery	2	5	5
glaucoma	Skip-gram	Yes	Yes	trabeculectomy	cataract_surgery	2	10	6
CKD[c]	CBOW	Yes	Yes	not_requiring_dialysis	dialysis	1	9	7
CKD	Skip-gram	Yes	Yes	not_requiring_dialysis	dialysis	1	8	5
diabetes	CBOW	Yes	Yes	glucose_variability	glucagon-like_peptide-1_receptor_agonists	2	10	6
diabetes	Skip-gram	Yes	Yes	glucose_variability	glucagon-like_peptide-1_receptor_agonists	2	10	5
asthma	CBOW	Yes	N/A	inhaled_corticosteroid	LABAs[d]	2	11	6
asthma	CBOW	N/A	Yes	inhaled_corticosteroids	inhaled_corticosteroid	1	10	8
asthma	Skip-gram	Yes	Yes	anti-LTs	LABAs	2	12	8
epilepsy	CBOW	Yes	Yes	valproate	AED[e]	2	12	10
epilepsy	Skip-gram	Yes	Yes	valproate	AED	2	12	11
arthritis	CBOW	Yes	Yes	plus_methotrexate	methotrexate	1	12	9
arthritis	Skip-gram	Yes	Yes	methotrexate	DMARDs[f]	2	11	6
osteoarthritis	CBOW	N/A	Yes	hyaluronic_acid	glucosamine	2	8	9
osteoarthritis	CBOW	Yes	N/A	knee_arthroplasty	hyaluronic_acid	2	9	7
osteoarthritis	CBOW	N/A	Yes	vs_acetaminophen	glucosamine	2	8	9
osteoarthritis	Skip-gram	Yes	Yes	vs_acetaminophen	hyaluronic_acid	2	11	8
anaemia	CBOW	Yes	Yes	iron	erythropoiesis-stimulating_agents	2	11	9
anaemia	Skip-gram	Yes	N/A	blood_transfusions	ESAs[g]	2	12	6
anaemia	Skip-gram	N/A	Yes	recombinant_human_erythropoietin	iron	2	11	8
hypertension	CBOW	Yes	N/A	antihypertensive_drugs	angiotensin_receptor_blockers	2	12	6
hypertension	CBOW	N/A	Yes	antihypertensive_therapy	antihypertensive	2	11	8
hypertension	Skip-gram	Yes	Yes	antihypertensive_drug_classes	antihypertensive	1	12	10

[a] CBOW: Continuous Bag-of-Words.

[b] N/A: not applicable.

[c] CKD: chronic kidney disease.

[d] LABA: long-acting beta2-agonist.

[e] AED: antiepileptic drug.

[f] DMARD: disease-modifying antirheumatic drug.

[g] ESA: erythropoiesis-stimulating agent.

Table 3 shows the 7 evidence-based categories assigned to the 408 UMLS CUI pairs investigated thoroughly. There are 19 concept pairs (X,Y_{Tx}) with more than 1 evidence-based category, such as the concept pair (X=C0014544|Epilepsy,Y_{Tx}=C0080356|Valproate). The evidence-based category “Tx with therapeutic effect” has the highest number of CUI pairs, with 190 pairs (X,Y_{Tx}), where 117 pairs have evidence (quotes) taken from BMJ Best Practice. The evidence-based category “correlation” has the highest number of evidence-based information sources with 108 uniform resource identifiers of the total 238. Multimedia Appendix 1 has further details.

Table 4 shows the *clinical winners*, that is, search query -x +z1 +z2 (a type of 4-term analogy) with the maximum observed number for n4 per target disease x. Table 4 reveals that an *NER winner* is not necessarily a *clinical winner*, that is, the maximum observed value for n4 does not always correspond to the maximum observed value for n3 or n2.

Table 3. The 408 unified medical language system concept unique identifier pairs investigated thoroughly and their evidence-based information sources per evidence-based category.

Evidence-based categories for concept Y_{Tx} correlated with concept X	Number of CUI[a] pairs	Number of evidence-based information sources (ie, URIs[b]) for CUI pairs	Number of CUI pairs with BMJ Best Practice as evidence source
Tx with therapeutic effect	190	73	117
Tx with uncertain therapeutic effect	38	22	11
Tx with unwanted or adverse effects (ie, nontherapeutic)	52	41	17
Potential Tx (under research and development)	5	5	0
Tx ingredient	22	21	6
General medical term	26	0	0
Correlation	94	108	19

[a] CUI: concept unique identifier.

[b] URI: Universal Resource Identifier.

In Table 4, there are two rows that are not *clinical winners* according to the observer O2. All rows except two are *clinical winners* according to the 3 domain experts and both observers.

Considering the 408 concept pairs (X,Y_{Tx}) with evidence, observer O1 disagrees with 25 of them, and observer O2 disagrees with 26 of them. The Cohen kappa of −0.023 is paradoxical [59-61], resolved in Multimedia Appendix 1 following Cicchetti DV and Feinstein AR [61].

Table 5 shows how the evidence-based category “Tx with therapeutic effect” assigned by an observer (when in disagreement) affects the *clinical winners* from Table 4. For observer O1, the only change was a decrease of n4 from 5 (Table 4) to 4 (Table 5) in the search query, *−anaemia +recombinant_human_erythropoietin +iron*, for Skip-gram. The observer O2 provided additional therapeutic evidence from BMJ Best Practice when in disagreement, typically increasing n4 or making “new” *clinical winners* (eg, search query, *−epilepsy +valproate +levetiracetam*).

Table 4. Clinical winners (highest value of n4) per model and disease target x considering the 3 domain experts.

Disease target *x*	Model	NER max (n2)	NER max (n3)	Treatment *z1* search query	Treatment *z2* search query	n1	n2	n3	n4
heart_failure	CBOW[a]	—	Yes	cardiac_resynchronization_therapy_(CRT)	aldosterone_antagonists	2	10	9	6
heart_failure	Skip-gram	Yes	Yes	beta-blockers	aldosterone_antagonists	2	12	8	5
glaucoma	CBOW	Yes	Yes	trabeculectomy	cataract_surgery	2	5	5	3
glaucoma	Skip-gram	Yes	Yes	trabeculectomy	cataract_surgery	2	10	6	3
CKD[b]	CBOW	Yes	Yes	not_requiring_dialysis	dialysis	1	9	7	5
CKD	Skip-gram	Yes	Yes	not_requiring_dialysis	dialysis	1	8	5	5
diabetes	CBOW	Yes	Yes	glucose_variability	glucagon-like_peptide-1_receptor_agonists	2	10	6	6
diabetes	Skip-gram	Yes	Yes	glucose_variability	glucagon-like_peptide-1_receptor_agonists	2	10	5	4
asthma	CBOW	—	Yes	inhaled_corticosteroids	inhaled_corticosteroid	1	10	8	8
asthma	Skip-gram	—	—	inhaled_corticosteroids	inhaled_corticosteroid	1	11	8	7
epilepsy	CBOW	—	—	valproate	antiepileptic_drug	2	11	10	8
epilepsy	CBOW	—	—	valproate	antiepileptic_drugs	2	11	10	8
epilepsy[c]	Skip-gram	Yes	Yes	valproate	AED[d]	2	12	11	7
arthritis	CBOW	Yes	Yes	plus_methotrexate	methotrexate	1	12	9	2
arthritis[c]	Skip-gram	—	—	plus_methotrexate	methotrexate	1	7	4	2
osteoarthritis	CBOW	Yes	—	knee_arthroplasty	hyaluronic_acid	2	9	7	5
osteoarthritis	Skip-gram	—	—	vs_acetaminophen	viscosupplementation	2	7	8	7
anaemia	CBOW	Yes	Yes	iron	erythropoiesis-stimulating_agents	2	11	9	4
anaemia	Skip-gram	—	Yes	recombinant_human_erythropoietin	iron	2	11	8	5
hypertension	CBOW	—	Yes	antihypertensive_therapy	antihypertensive	2	11	8	6
hypertension	Skip-gram	Yes	Yes	antihypertensive_drug_classes	antihypertensive	1	12	10	8

[a] CBOW: Continuous Bag-of-Words.

[b] CKD: chronic kidney disease.

[c] Not *clinical winners* according to O2.

[d] AED: antiepileptic drug.

Table 5. Changes in clinical winners (highest value of n4) per model and disease target x considering observer O1 and O2.

Disease target *x*	Model	Differences in *clinical winner* max (n4) according to observers	Treatment *z1* search query	Treatment *z2* search query	n1	n2	n3	n4
epilepsy	Skip-gram	Observer O2[a]: New	valproate	levetiracetam	2	10	9	7
arthritis	CBOW[b]	Observer O2: n4 different	plus_methotrexate	methotrexate	1	12	9	6
arthritis	CBOW	Observer O2: New	methotrexate	DMARDs[c]	2	11	9	6
arthritis	Skip-gram	Observer O2: New	IACI[d]	DMARDs	2	9	6	5
arthritis	Skip-gram	Observer O2: New	plus_methotrexate	DMARDs	2	10	6	5
anaemia	Skip-gram	Observer O1[e]: n4 different	recombinant_human_erythropoietin	iron	2	11	8	4

[a] O2: BMJ health informatician who works with BMJ Best Practice content and has a junior doctor background.

[b] CBOW: Continuous Bag-of-Words.

[c] DMARD: disease-modifying antirheumatic drug.

[d] IACI: intra-articular corticosteroid injection.

[e] O1: medical consultant.

Multimedia Appendix 1 has the best *clinical winner*, which is an *NER winner*. Table 6 shows the best *clinical winner* that is not an *NER winner*. Table 6 illustrates the enrichment of the candidate n-grams *y* with domain knowledge meaning (Stage 2 normalizes n-grams with UMLS CUIs) and biomedical evidence ratified with an audit (Stage 3). The evidence provided for the evidence-based categories (quotes with references from the biomedical literature) is presented in Multimedia Appendix 2 (worksheet Stage 3).

In conclusion, considering the *clinical winners* found (Table 4), the answer to Q3 is "yes," that is, the 4-term type of analogies discovered in a small common-English corpus can also be discovered in a large-scale biomedical corpus.

Table 6. Illustration of a Best clinical winner with max (n4)=8 for CBOW and disease target x = epilepsy, which is not an NER winner.

Rank[a]	Candidate *y*	3CosAdd	UMLS CUI for concept Y_{Tx} mapped to candidate y_{Tx}	Evidence-based categories for concept Y_{Tx} correlated with concept *X*
1	lamotrigine	0.385201	C0064636	Tx with therapeutic effect
2	carbamazepine	0.345227	C0006949	Tx with unwanted or adverse effects (ie, nontherapeutic)
3	low_propensity	0.324285	—[b]	—
4	clonazepam	0.310706	C0009011	Tx with uncertain therapeutic effect
5	topiramate	0.308402	C0076829	Tx with therapeutic effect
6	lithium_valproate	0.308223	C0023870|C0080356	Tx with therapeutic effect|Tx with unwanted or adverse effects (ie, nontherapeutic)
7	clobazam	0.306901	C0055891	Tx with therapeutic effect
8	sodium_valproate	0.300513	C0037567	Tx with therapeutic effect
9	lorazepam	0.29562	C0024002	Tx with therapeutic effect
10	lithium	0.294804	C0023870	Tx with therapeutic effect|Tx with unwanted or adverse effects (ie nontherapeutic)
11	gabapentin_pregabalin_topiramate	0.291698	C0657912|C0076829|C0060926	Tx with therapeutic effect
12	antiepileptic_drugs_other_than	0.290046	C0003299	Tx with therapeutic effect

[a] The search query −x +z1 +z2 is listed in Table 4, which is −epilepsy +valproate +antiepileptic_drug. The character "|" appears when there is more than 1 CUI or evidence-based category.

[b] The candidate y = "low_propensity" does not belong to the semantic field Tx, and so, it has no UMLS CUI assigned.

Answer Q4: An Empirical Heuristic with Some Predictive Power for Clinical Winners

Multimedia Appendix 2 (worksheet Q4) has the 304 search queries of the total of 446 (223 for CBOW and 223 for Skip-gram) queries, where all the candidates y_{Tx} mapped to concepts Y_{Tx} have at least one evidence-based category assigned. Textbox 2 summarizes the empirical heuristic developed by visual inspection, focusing on rows with the minimum (n4=0) and the maximum observed values of n4. The heuristic is programmatically implemented as a Boolean expression composed of 3 expressions with the Boolean AND.

Textbox 2. An empirical heuristic developed by visual inspection with some predictive power for the clinical winners.

1. Avoid n-grams *z1* and *z2* having short forms
2. Favor n-grams *z1* or *z2* (or both) not appearing among the 20 top-ranked candidates for target *x* with the highest value for cosine with Skip-gram embeddings
3. Favor n-gram *z2* with frequency counts in the corpus >100

The heuristic selects 93 of the 304 search queries, which brings 126 of the 190 UMLS Metathesaurus concepts Y_{Tx} with the evidence-based category “Tx with therapeutic effect,” that is, Y_{Tx} with therapeutic intent.

Table 7 (source data in Multimedia Appendix 1) shows the performance of the heuristic considering (1) the values of n4 (the last 3 yellow columns in Multimedia Appendix 2 worksheet Q4), (2) the different thresholds for n4, and (3) precision and recall as metric.

Considering the precision and recall values for the empirical heuristic (Table 7), the answer to Q4 is also “yes,” that is, some predictive power over the *clinical winners* obtained is possible.

Table 7. Precision and recall for the empirical heuristic developed using Multimedia Appendix 2 (worksheet Q4).

Threshold	True positive (TP)	False positive (FP)	False negative (FN)	Precision[a] %	Recall[b] %
n4>0	91	2	189	97.85	32.5
n4>1	84	9	151	90.32	35.74
n4>2	73	20	111	78.49	39.67
n4>3	48	45	76	51.61	38.71
n4>4	28	65	52	30.11	35
n4>5	14	79	20	15.05	41.18
n4>6	9	84	5	9.68	64.29
n4>7	4	89	0	4.3	100

[a] Precision: calculated as TP/(TP+FP).

[b] Recall: calculated as TP/(TP+FN).

Discussion

Principal Findings

Humans can agree that the semantic field *person* {you; Romeo} is related to the semantic field *death* {die; died; dagger} in the context of Shakespeare’s Romeo. Hence, we answer Q1 and Q2 with a “yes”; therefore, analogical reasoning with CBOW embeddings seems feasible with a small common-English corpus. This challenges the current assumption that “learning in current deep learning models relies on massive data” [3].

We answered Q3 by demonstrating that there is proof of the generalization; thus, the 3CosAdd formula can discover another type of 4-term analogy that is not a pair-based proportional analogy. Furthermore, we have proven that the analogical inferences sanctioned by the 3CosAdd formula with embeddings could extract treatments with therapeutic intent from free text. Indeed, there were strong examples of analogical reasoning with abstract semantic relations between *z1* and *z2* among *clinical winners* (Table 4):

- *Antonym*. The search query, −*CKD* +*not_requiring_dialysis* +*dialysis*, with n4=5 for CBOW and Skip-gram.
- *Synonym*. The search query, −*asthma* +*inhaled_corticosteroids* +*inhaled_corticosteroid*, with n4=8 for CBOW (the best *clinical winner*) and n4=7 for Skip-gram, where the relation between *z2* and *z1* was inflectional morphology *singular:plural*. This query resembled the search query, −*Romeo* +*die* +*died*.
- *Category membership*. The search query, −*epilepsy* +*valproate* +*antiepileptic_drug*, with n4=8 for CBOW. The search query, −*hypertension* +*antihypertensive_drug_classes* +*antihypertensive*, with n4=8 for Skip-gram. Both search queries were the best *clinical winners* (maximum observed value for n4).
- *Commonalities in structural features*. All search queries focused on the therapeutic intent of *z1* and *z2* for target disease *x*. However, some queries did not have the above abstract semantic relationships between *z1* and *z2*. For example, the search queries −*osteoarthritis* +*knee_arthroplasty* +*hyaluronic_acid* with n4=5 for

CBOW and −*heart_failure* +*beta-blockers* +*aldosterone_antagonists* with n4=5 for Skip-gram.

We answered Q4 by demonstrating that it is feasible to gain some predictive power for the *clinical winners*; therefore, a tactic preference was latent promising systematicity [6]. Textbox 3 highlights the precision and recall values for 3 n4 thresholds of the overall performance of the empirical heuristic developed by visual inspection.

Textbox 3. Empirical heuristic performance for 304 search queries with all candidate concepts YTx with evidence.

- With a threshold n4 >7, the recall is 100%. All search queries with n4 >7 (the best *clinical winners*) are selected by the heuristic. The precision was 4.30% (the lowest value).
- With a threshold n4 >0 (at least one Y_{Tx} with therapeutic intent), the precision was 97.85% (the highest value) and the recall was 32.50%.
- With a threshold n4 >2, where 3 was the lowest value among the *clinical winners* (Tables 3 and 4), the precision was 78.49% and the recall was 39.67%.

Limitations

Our work relies on semantic fields and has 2 main limitations [62]: (1) there are overlaps of meaning and (2) there are gaps in meaning. This has 2 clear implications for the lists of concepts Y_{Tx} per disease x:

- The lists may not comprise mutually exclusive concepts in meaning. For example, “C0060657|formoterol” and “C1276807|Budesonide/formoterol” are both treatments with evidence of therapeutic intent for asthma [63].
- The lists were incomplete. For example, “C0772501|Levalbuterol” and “C0907850|ciclesonide” are both treatments with evidence of therapeutic intent for asthma [63] and not among the Y_{Tx} for asthma.

We did not use Skip-gram with negative sampling (also known as SGNS); therefore, it can be argued that we did not use the best configuration of a word2vec model [21]. The effect of hyperparameter configurations appeared in studies by Levy et al and Chiu et al [64,65], and Allen and Hospedales [66] reviewed mathematical proofs and equations with an emphasis on SGNS for pair-based proportional analogies.

For Stage 1, the 3CosAdd formula needed at least two n-gram pairs (disease *x*, treatment *z*) [29]. Only one search query could be made for the target disease, chronic kidney disease and diabetes, and none for obesity. Other studies that replicated the application to the 3CosAdd formula for target disease *x* could suffer the same limitation. For example, in Appendix B in the study by Pakhomov et al [67], among the 100 top-ranked candidate terms (highest cosine value) “semantically similar or related” to target disease “heart failure”, there were no treatments (ie, Tx encompassing 3 textual definitions from Hart et al [24]).

For Stage 2, the MetaMap version was 2016v2 (with a 2016 UMLS release), and few n-grams were considered as clear terminological gaps. The n-gram “anti-VEGF_agents” was manually mapped to CUI=C4727875, which exists in the 2019AA UMLS release. Five n-grams were mapped to very broad CUIs as they had the character “*” in Multimedia Appendix 2 (worksheet Stage 3).

The NER task (Stage 2) and the searchers in the literature seeking evidence for concept pairs (Stage 3) were time-consuming and required highly trained domain experts. The appraisal of the literature was not performed by a review team as proficient as the ones conducting Cochrane systematic reviews.

The heuristic developed by visual inspection lacked finesse, and its improvement calls for further investigation.

Comparison With Prior Work

The UMLS CUIs were mapped to SNOMED CT identifiers [30]. From a “digital health care” perspective [68], the UK NHS is moving toward the adoption of SNOMED CT as the only terminology for all care settings [69]. A subset of SNOMED CT concepts under worldwide adoption is the CORE Problem List Subset of SNOMED CT [70], and the UK NHS has developed 2 human-readable SNOMED CT subsets [71]: UK Clinical Extension and UK Drug Extension. However, SNOMED CT lacks statements representing the treatments that can be considered for a disease (eg, inhaled corticosteroid treats asthma) and, to the best of the authors’ knowledge, there are no SNOMED CT subsets for well-known diseases.

There are reusable datasets for evaluating relatedness made of UMLS CUI pairs:

- Medical coders set [72]: 101 CUI pairs mapped to terms, typically multiple words. Only 29 pairs have a high interrater agreement.
- Medical Residents Relatedness Set [73]: 588 CUI pairs mapped to terms, typically single words. Using single words is a severe limitation as “most medical terms consist of more than one word” [67].
- UMLS MRREL table [58]: It has relationships asserted by source vocabularies between CUI pairs. Among the relationship attributes appear the following: “may_prevent”, “may_treat”, and “has_contraindicated_drug”.

All reusable datasets mentioned above lack evidence (quotes with references) from the biomedical literature. Multimedia Appendix 1 cross-compares these reusable datasets and the 408 UMLS CUI pairs investigated thoroughly in this study.

Conclusions

Extracting clinically useful information automatically from free text in PubMed/MEDLINE may require a natural language understanding of statements containing relevant relations for health care. Hence, extracting treatments with therapeutic intent by analogical reasoning from embeddings (423K n-grams from the PMSB dataset) is an ambitious goal. Our SemDeep approach is knowledge-based, underpinned by embedding analogies that exploit prior knowledge. Biomedical facts from embedding analogies (a 4-term type, not pairwise) are potentially useful for clinicians. The heuristic offers a practical way to discover beneficial treatments for well-known diseases. Learning from deep learning models does not require a massive amount of data. Embedding analogies are not limited to pairwise analogies; hence, analogical reasoning with embeddings is underexploited.

Acknowledgments

The authors thank Tim Furmston (IT Research, University of Manchester, United Kingdom) for help with the software and electronic infrastructure.

Conflicts of Interest

CW works for BMJ which produces the clinical decision support tool BMJ Best Practice. All other authors have no conflict of interest to declare.

Multimedia Appendix 1
Additional material.
[PDF File (Adobe PDF File), 378 KB - medinform_v8i8e16948_app1.pdf]

Multimedia Appendix 2
Data to replicate the results reported.
[XLS File (Microsoft Excel File), 708 KB - medinform_v8i8e16948_app2.xls]

References

1. Masic I, Miokovic M, Muhamedagic B. Evidence based medicine - new approaches and challenges. Acta Inform Med 2008;16(4):219-225 [FREE Full text] [doi: 10.5455/aim.2008.16.219-225] [Medline: 24109156]
2. Avorn J. The psychology of clinical decision making - implications for medication use. N Engl J Med 2018 Mar 22;378(8):689-691. [doi: 10.1056/NEJMp1714987] [Medline: 29466158]
3. Lu H, Wu YN, Holyoak KJ. Emergence of analogy from relation learning. Proc Natl Acad Sci U S A 2019 Mar 5;116(10):4176-4181 [FREE Full text] [doi: 10.1073/pnas.1814779116] [Medline: 30770443]
4. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
5. Jurafsky D, Martin J. Stanford University. 2019. Speech and Language Processing URL: https://web.stanford.edu/~jurafsky/slp3/ [accessed 2020-06-01]
6. Gentner D, Markman AB. Structure mapping in analogy and similarity. Am Psychol 1997;52(1):45-56. [doi: 10.1037//0003-066x.52.1.45]
7. ACL Anthology. Analogy (State of the Art) URL: https://aclweb.org/aclwiki/Analogy_(State_of_the_art) [accessed 2020-06-01]
8. Schnabel T, Labutov I, Mimno D, Joachims T. Evaluation Methods for Unsupervised Word Embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 Presented at: EMNLP'15; September 17-21, 2015; Lisbon, Portugal p. 298-307. [doi: 10.18653/v1/d15-1036]
9. Mikolov T, Yih W, Zweig G. Linguistic Regularities in Continuous Space Word Representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013 Presented at: NAACL'13; June 9-14, 2013; Atlanta, Georgia, US p. 746-751 URL: https://www.aclweb.org/anthology/N13-1090/
10. Drozd A, Gladkova A, Matsuoka S. Word Embeddings, Analogies, and Machine Learning: Beyond King - Man + Woman = Queen. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016 Presented at: COLING'16; December 11-16, 2016; Osaka, Japan p. 3519-3530 URL: https://www.aclweb.org/anthology/C16-1332/
11. Liu H, Wu Y, Yang Y. Analogical inference for multi-relational embeddings. Arxiv 2017:- epub ahead of print (1705.02426) [FREE Full text]
12. Rather NN, Patel CO, Khan SA. Using deep learning towards biomedical knowledge discovery. Int J Math Sci Comput 2017 Apr 8;3(2):1-10. [doi: 10.5815/ijmsc.2017.02.01]
13. Dynomant E, Lelong R, Dahamna B, Massonnaud C, Kerdelhué G, Grosjean J, et al. Word embedding for the French natural language in health care: comparative study. JMIR Med Inform 2019 Jul 29;7(3):e12310 [FREE Full text] [doi: 10.2196/12310] [Medline: 31359873]
14. BMJ Best Practice. 2018. Sepsis in Adults URL: http://bestpractice.bmj.com/topics/en-gb/245 [accessed 2020-06-01]

15. Gatta A, Verardo A, Bolognesi M. Hypoalbuminemia. Intern Emerg Med 2012 Oct;7(Suppl 3):S193-S199. [doi: 10.1007/s11739-012-0802-0] [Medline: 23073857]
16. National Library of Medicine - National Institutes of Health - NIH. Search Strategy Used to Create the PubMed Systematic Reviews Filter URL: https://www.nlm.nih.gov/bsd/pubmed_subsets/sysreviews_strategy.html [accessed 2020-06-01]
17. Special Issue on Semantic Deep Learning. URL: http://www.semantic-web-journal.net/content/special-issue-semantic-deep-learning [accessed 2020-06-01]
18. Stevens R, Goble CA, Bechhofer S. Ontology-based knowledge representation for bioinformatics. Brief Bioinform 2000 Dec;1(4):398-414. [doi: 10.1093/bib/1.4.398] [Medline: 11465057]
19. Bartha P. Stanford Encyclopedia of Philosophy. 2019. Analogy and Analogical Reasoning URL: https://plato.stanford.edu/archives/spr2019/entries/reasoning-analogy/ [accessed 2020-06-01]
20. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. In: Proceedings of 1st International Conference on Learning Representations. 2013 Presented at: ICLR'13; May 2-4, 2013; Scottsdale, Arizona, US URL: https://arxiv.org/abs/1301.3781
21. Levy O, Goldberg Y. Linguistic Regularities in Sparse and Explicit Word Representations. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning. 2014 Presented at: ACL'14; June 26-27, 2014; Baltimore, Maryland, USA p. 171-180. [doi: 10.3115/v1/w14-1618]
22. Nerlich B, Clarke DD. Semantic fields and frames: historical explorations of the interface between language, action, and cognition. J Pragmat 2000 Jan;32(2):125-150. [doi: 10.1016/s0378-2166(99)00042-9]
23. SNOMED Confluence. Technical Implementation Guide URL: http://snomed.org/tig [accessed 2020-06-01]
24. Hart T, Tsaousides T, Zanca JM, Whyte J, Packel A, Ferraro M, et al. Toward a theory-driven classification of rehabilitation treatments. Arch Phys Med Rehabil 2014 Jan;95(1 Suppl):S33-44.e2. [doi: 10.1016/j.apmr.2013.05.032] [Medline: 24370323]
25. Hill F, Reichart R, Korhonen A. SimLex-999: evaluating semantic models with (genuine) similarity estimation. Comput Linguist 2015 Dec;41(4):665-695. [doi: 10.1162/coli_a_00237]
26. NCBO BioPortal. 2020. Semanticscience Integrated Ontology URL: https://bioportal.bioontology.org/ontologies/SIO [accessed 2020-06-01]
27. Chen D, Peterson J, Griffiths T. Evaluating Vector-Space Models of Analogy. In: Proceedings of the 39th Annual Meeting of the Cognitive Science Society. 2017 Presented at: CogSci'17; July 26-29, 2017; London, UK.
28. Google Code. Word2vec - Default URL: https://code.google.com/archive/p/word2vec/source/default/source [accessed 2020-06-01]
29. Arguello-Casteleiro M, Stevens R, Des-Diz J, Wroe C, Fernandez-Prieto MJ, Maroto N, et al. Exploring semantic deep learning for building reliable and reusable one health knowledge from PubMed systematic reviews and veterinary clinical notes. J Biomed Semantics 2019 Nov 12;10(Suppl 1):22 [FREE Full text] [doi: 10.1186/s13326-019-0212-6] [Medline: 31711540]
30. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
31. BMJ Best Practice. URL: https://bestpractice.bmj.com/ [accessed 2020-06-01]
32. PubMed/MEDLINE. URL: https://www.ncbi.nlm.nih.gov/pubmed/ [accessed 2020-06-10]
33. Kwag KH, González-Lorenzo M, Banzi R, Bonovas S, Moja L. Providing doctors with high-quality information: an updated evaluation of web-based point-of-care information summaries. J Med Internet Res 2016 Jan 19;18(1):e15 [FREE Full text] [doi: 10.2196/jmir.5234] [Medline: 26786976]
34. BMJ Best Practice. Do You Work for the NHS in England, Scotland or Wales? URL: https://bestpractice.bmj.com/info/bmagp/ [accessed 2020-06-01]
35. Landauer TK, Dumais ST. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev 1997;104(2):211-240. [doi: 10.1037/0033-295X.104.2.211]
36. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3(4-5):993-1002. [doi: 10.1162/jmlr.2003.3.4-5.993]
37. Thomo A. Engineering and Computer Science - University of Victoria. Latent Semantic Analysis URL: https://www.engr.uvic.ca/~seng474/svd.pdf [accessed 2020-06-01]
38. Ardila J. Meaning in language. An introduction to semantics and pragmatics. J Pragma 2011 Aug;43(10):2670-2672. [doi: 10.1016/j.pragma.2011.03.014]
39. Cochrane Library: Cochrane Reviews. Cochrane Database of Systematic Reviews URL: https://www.cochranelibrary.com/cdsr/about-cdsr [accessed 2020-06-01]
40. World Health Organization. Prevention and Control of Noncommunicable Diseases: Guidelines for Primary Health Care in Low Resource Settings. Geneva, Switzerland: World Health Organization; 2012.
41. Arguello-Casteleiro M, Stevens R, des-Diz J, Wroe C, Fernandez-Prieto MJ, Maroto N, et al. Exploring semantic deep learning for building reliable and reusable one health knowledge from PubMed systematic reviews and veterinary clinical notes. J Biomed Semantics 2019 Nov 12;10(Suppl 1):22 [FREE Full text] [doi: 10.1186/s13326-019-0212-6] [Medline: 31711540]

42. Casteleiro MA, Demetriou G, Read W, Prieto MJ, Maroto N, Fernandez DM, et al. Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature. J Biomed Semantics 2018 Apr 12;9(1):13 [FREE Full text] [doi: 10.1186/s13326-018-0181-1] [Medline: 29650041]
43. Manning C, Schütze H. Foundations of Statistical Natural Language Processing. New York, USA: MIT Press; 1999.
44. Novak J, Cañas A. CMap. 2008. The Theory Underlying Concept Maps and How to Construct and Use Them URL: http://cmap.ihmc.us/Publications/ResearchPapers/TheoryUnderlyingConceptMaps.pdf [accessed 2020-06-01]
45. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011;18(5):544-551 [FREE Full text] [doi: 10.1136/amiajnl-2011-000464] [Medline: 21846786]
46. Yamamoto Y, Yamaguchi A, Bono H, Takagi T. Allie: a database and a search service of abbreviations and long forms. Database (Oxford) 2011;2011:bar013 [FREE Full text] [doi: 10.1093/database/bar013] [Medline: 21498548]
47. Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Comput Linguist 2008 Dec;34(4):555-596. [doi: 10.1162/coli.07-034-r2]
48. Shelley M, Krippendorff K. Content analysis: an introduction to its methodology. J Am Stat Assoc 1984 Mar;79(385):240. [doi: 10.2307/2288384]
49. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet 2012 Dec;13(12):829-839. [doi: 10.1038/nrg3337] [Medline: 23150036]
50. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
51. Guidelines Developed for Step 4: Named Entity Recognition Task. URL: https://static-content.springer.com/esm/art%3A10.1186%2Fs13326-019-0212-6/MediaObjects/13326_2019_212_MOESM2_ESM.pdf [accessed 2020-06-01]
52. Pratt W, Yetisgen-Yildiz M. A study of biomedical concept identification: MetaMap vs people. AMIA Annu Symp Proc 2003:529-533 [FREE Full text] [Medline: 14728229]
53. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. Br Med J 1996 Jan 13;312(7023):71-72 [FREE Full text] [doi: 10.1136/bmj.312.7023.71] [Medline: 8555924]
54. Sager J. A Practical Course in Terminology Processing. Amsterdam, The Netherlands: John Benjamins Publishing; 1990.
55. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Jul 2;20(1):37-46. [doi: 10.1177/001316446002000104]
56. Davies R, O'Dea K, Gordon A. Immune therapy in sepsis: are we ready to try again? J Intensive Care Soc 2018 Nov;19(4):326-344 [FREE Full text] [doi: 10.1177/1751143718765407] [Medline: 30515242]
57. Cabrera-Perez J, Condotta SA, Badovinac VP, Griffith TS. Impact of sepsis on CD4 T cell immunity. J Leukoc Biol 2014 Nov;96(5):767-777 [FREE Full text] [doi: 10.1189/jlb.5MR0114-067R] [Medline: 24791959]
58. NCBI Bookshelf. 2019. UMLS Reference Manual URL: https://www.ncbi.nlm.nih.gov/books/NBK9685/table/ch03.T.related_concepts_file_mrrel_rrf/ [accessed 2020-06-10]
59. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005 May;37(5):360-363 [FREE Full text] [Medline: 15883903]
60. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43(6):543-549. [doi: 10.1016/0895-4356(90)90158-l] [Medline: 2348207]
61. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43(6):551-558. [doi: 10.1016/0895-4356(90)90159-m] [Medline: 2189948]
62. Lehrer A. The Influence of Semantic Fields on Semantic Change. Berlin, Germany: De Gruyter Mouton; 1985.
63. BMJ Best Practice. Asthma in Adults URL: http://bestpractice.bmj.com/topics/en-gb/44 [accessed 2020-06-01]
64. Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings. Trans Assoc Comput Linguist 2015 Dec;3:211-225. [doi: 10.1162/tacl_a_00134]
65. Chiu B, Crichton G, Korhonen A, Pyysalo S. How to Train good Word Embeddings for Biomedical NLP. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. 2016 Presented at: ACL'16; August 12, 2016; Berlin, Germany. [doi: 10.18653/v1/w16-2922]
66. Allen C, Hospedales T. Analogies Explained: Towards Understanding Word Embeddings. In: Proceedings of the 36th International Conference on Machine Learning. 2019 Presented at: ICML'19; June 9-15, 2019; Long Beach, California, USA p. 223-231. [doi: 10.1201/9780429469275-13]
67. Pakhomov SV, Finley G, McEwan R, Wang Y, Melton GB. Corpus domain effects on distributional semantic modeling of medical terms. Bioinformatics 2016 Dec 1;32(23):3635-3644 [FREE Full text] [doi: 10.1093/bioinformatics/btw529] [Medline: 27531100]
68. Tresp V, Overhage JM, Bundschus M, Rabizadeh S, Fasching PA, Yu S. Going digital: a survey on digitalization and large-scale data analytics in healthcare. Proc IEEE 2016 Nov;104(11):2180-2206. [doi: 10.1109/JPROC.2016.2615052]
69. Government of UK. Personalised Health and Care 2020 URL: https://www.gov.uk/government/publications/personalised-health-and-care-2020 [accessed 2020-06-01]
70. National Library of Medicine - National Institutes of Health - NIH. The CORE Problem List Subset of SNOMED CT URL: https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html [accessed 2020-06-01]

71. NHS Digital. SNOMED CT Subset Members in Spreadsheet View URL: https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/40 [accessed 2020-06-01]
72. Pedersen T, Pakhomov SV, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007 Jul;40(3):288-299 [FREE Full text] [doi: 10.1016/j.jbi.2006.06.004] [Medline: 16875881]
73. Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic similarity and relatedness between clinical terms: an experimental study. AMIA Annu Symp Proc 2010 Dec 13;2010:572-576 [FREE Full text] [Medline: 21347043]

Abbreviations

BMJ: British Medical Journal
CBOW: Continuous Bag-of-Words
CUI: concept unique identifier
MRREL: the UMLS related concepts table (file=MRREL)
NER: named entity recognition
NHS: National Health Service
PMSB: PubMed systematic reviews subset
SemDeep: Semantic Deep Learning
SGNS: skip-gram with negative sampling
SNOMED CT: systematized nomenclature of medicine - clinical terms
Tx: treatment
UMLS: unified medical language system
URIs: Universal Resource Identifiers

Edited by G Eysenbach; submitted 06.11.19; peer-reviewed by Z He; comments to author 28.11.19; revised version received 27.02.20; accepted 27.02.20; published 06.08.20.

Please cite as:
Arguello Casteleiro M, Des Diz J, Maroto N, Fernandez Prieto MJ, Peters S, Wroe C, Sevillano Torrado C, Maseda Fernandez D, Stevens R
Semantic Deep Learning: Prior Knowledge and a Type of Four-Term Embedding Analogy to Acquire Treatments for Well-Known Diseases
JMIR Med Inform 2020;8(8):e16948
URL: https://medinform.jmir.org/2020/8/e16948
doi:10.2196/16948
PMID:32759099

Original Paper

Decompensation in Critical Care: Early Prediction of Acute Heart Failure Onset

Patrick Essay[1], MS; Baran Balkan[1], BS; Vignesh Subbian[2], PhD

[1]College of Engineering, The University of Arizona, Tucson, AZ, United States

[2]Department of Systems and Industrial Engineering, Department of Biomedical Engineering, The University of Arizona, Tucson, AZ, United States

Corresponding Author:
Patrick Essay, MS
College of Engineering
The University of Arizona
1127 E James E Rogers Way
Tucson, AZ, 85721-0020
United States
Phone: 1 4024305524
Email: p.essay@icloud.com

Abstract

Background: Heart failure is a leading cause of mortality and morbidity worldwide. Acute heart failure, broadly defined as rapid onset of new or worsening signs and symptoms of heart failure, often requires hospitalization and admission to the intensive care unit (ICU). This acute condition is highly heterogeneous and less well-understood as compared to chronic heart failure. The ICU, through detailed and continuously monitored patient data, provides an opportunity to retrospectively analyze decompensation and heart failure to evaluate physiological states and patient outcomes.

Objective: The goal of this study is to examine the prevalence of cardiovascular risk factors among those admitted to ICUs and to evaluate combinations of clinical features that are predictive of decompensation events, such as the onset of acute heart failure, using machine learning techniques. To accomplish this objective, we leveraged tele-ICU data from over 200 hospitals across the United States.

Methods: We evaluated the feasibility of predicting decompensation soon after ICU admission for 26,534 patients admitted without a history of heart failure with specific heart failure risk factors (ie, coronary artery disease, hypertension, and myocardial infarction) and 96,350 patients admitted without risk factors using remotely monitored laboratory, vital signs, and discrete physiological measurements. Multivariate logistic regression and random forest models were applied to predict decompensation and highlight important features from combinations of model inputs from dissimilar data.

Results: The most prevalent risk factor in our data set was hypertension, although most patients diagnosed with heart failure were admitted to the ICU without a risk factor. The highest heart failure prediction accuracy was 0.951, and the highest area under the receiver operating characteristic curve was 0.9503 with random forest and combined vital signs, laboratory values, and discrete physiological measurements. Random forest feature importance also highlighted combinations of several discrete physiological features and laboratory measures as most indicative of decompensation. Timeline analysis of aggregate vital signs revealed a point of diminishing returns where additional vital signs data did not continue to improve results.

Conclusions: Heart failure risk factors are common in tele-ICU data, although most patients that are diagnosed with heart failure later in an ICU stay presented without risk factors making a prediction of decompensation critical. Decompensation was predicted with reasonable accuracy using tele-ICU data, and optimal data extraction for time series vital signs data was identified near a 200-minute window size. Overall, results suggest combinations of laboratory measurements and vital signs are viable for early and continuous prediction of patient decompensation.

(JMIR Med Inform 2020;8(8):e19892) doi:10.2196/19892

KEYWORDS

critical care; heart failure; intensive care units; machine learning; time series; heart; cardiology; prediction; chronic disease; ICU; intensive care unit

Introduction

Background

Intensive care units (ICUs) are data-rich clinical environments involving complex decision-making for patients who are critically ill making them a major area of health care innovation [1]. The ability to continuously monitor patients in the ICU provides unique opportunities for analytics such as estimation of physiological states and prediction of decompensation (ie, clinical deterioration) or patient outcomes [2]. There has been substantial progress in terms of predicting longer-term outcomes such as mortality and readmission rates in patients with heart failure, but there is limited work around predicting shorter-term clinical events in the ICU, such as acute heart failure onset [3-5]. Predicting such decompensation events allows for prevention and mitigation steps while patients are in the ICU and promotes a proactive decision-making process for clinicians, potentially resulting in timely interventions and improved patient outcomes.

In this work, we present the application of machine learning techniques for predicting decompensation in critical care settings using acute heart failure onset as the prediction outcome [6]. The objectives of this study are to examine the prevalence of three heart failure risk factors (ie, coronary artery disease, hypertension, or myocardial infarction); to apply and evaluate machine learning techniques to predict heart failure onset in patients with and without one of the three known risk factors; and to evaluate features of interest including aggregate time series vital signs data, laboratory values, and other physiological inputs used in traditional clinical scoring systems.

Heart failure is a major cause of mortality and morbidity worldwide, and a major public health concern. It is a complex clinical syndrome where cardiac dysfunction impairs the ability of the ventricle to fill and eject blood, leading to a wide range of signs and symptoms and unspecific diagnosis [7-9]. Although there have been advances in therapies, further understanding of prognosis and management of acute heart failure is needed [10]. This is particularly true in critical care where heart failure may be of secondary concern to clinicians relative to primary ICU diagnosis.

There has been interest in shifting prognostication of decompensation events such as onset of heart failure to a remote monitoring team (tele-ICU) [11]. Although such telemedicine-based efforts have become increasingly common in cardiovascular ICUs, risk of acute heart failure onset has not been extensively investigated through a machine learning and tele-ICU lens [12]. Additionally, there are several known risk factors of heart failure, including hypertension, coronary artery disease, myocardial infarction, obesity, diabetes, and other lifestyle factors such as alcohol intake, smoking, and leisure activity [13]. Of these, hypertension, coronary artery disease, and myocardial infarction are identifiable key risk factors of acute heart failure and relevant to remote ICU monitoring.

Significance

Multiple prior studies related to heart failure in different settings (eg, inpatient vs outpatient) using dissimilar data sources (eg, home-based monitoring data vs in-hospital clinical data) have been conducted [14,15]. These studies used features such as change in body weight, heart rate, and blood pressure under the hypothesis that hemodynamic changes in patients can be characterized in continuous physiological data collected by the patient at home. In critical care settings, many of the variables used by the bedside clinical team are readily available to the remote tele-ICU team as well for deeper analytics.

Previous studies have modeled risk of hospitalization, long-term survival rates, and mode of death prediction as a result of heart failure [16-18]. Models used features related to clinical status, therapy, and laboratory parameters including home-based physiological telemonitoring [19]. Generally, these studies use temporal data to make longer-term (ie, months to years) predictions [20].

These and other studies illustrate potential and previous accomplishments in heart failure prediction, but to our knowledge, models have not been developed in the context of critical care and the fast-paced ICU environment or used the expansive capabilities of tele-ICU data. These previous studies do, however, suggest that trends in patient physiology and hemodynamics may be leveraged for early heart failure prediction.

Our study attempts to predict onset of acute heart failure by examining readily available physiological discrete and time series data on a truncated scale near the time of ICU admission. We applied data extraction methods similar to approaches used in longer-term prediction models and comparable physiological measurements, in addition to potentially more extensive and reliable tele-ICU data as compared to home-based measurements.

Methods

Data Source and Preprocessing

In this study, we used the eICU Collaborative Research Database [21], which contains remotely monitored critical care data from adult patients admitted to over 200 hospitals in the United States from 2014-2015 [22]. The database includes basic patient characteristics as well as medications, laboratory values, vital signs, and other discrete physiological variables measured at the bedside ICU and interfaced with the tele-ICU. We selected both multivariate logistic regression and decision tree models for predicting acute heart failure, given their interpretable nature.

Patient ICU stays were extracted based on primary admission diagnosis and subsequent diagnostic codes during the same unit stay. Inclusion criteria were such that each ICU stay must not have a primary admission diagnosis of heart failure (ie, the patient was admitted to the ICU for a reason other than heart failure). Readmissions were included unless the subsequent stays were primarily due to heart failure.

Patient stays were segregated based on three heart failure risk factors: coronary artery disease, hypertension, and myocardial infarction. In each risk factor group, patients were categorized by heart failure onset after primary admission diagnosis. A fourth group of *nonrisk factor patients* was extracted including all patients admitted for reasons other than heart failure and did not have record of one of the three risk factors. The International Classification of Diseases version 9 (ICD-9) codes were used to determine heart failure and risk factors (Table 1).

Table 1. Heart failure ICD-9 codes for cohort discovery.

ICD-9[a] code	Description
Heart failure	
398.91	Rheumatic heart failure (congestive)
428.0	Congestive heart failure, unspecified
428.1	Left heart failure
428.20	Systolic heart failure, unspecified
428.21	Acute systolic heart failure
428.22	Chronic systolic heart failure
428.23	Acute on chronic systolic heart failure
428.30	Diastolic heart failure, unspecified
428.31	Acute diastolic heart failure
428.32	Chronic diastolic heart failure
428.33	Acute on chronic diastolic heart failure
428.40	Combined systolic and diastolic heart failure, unspecified
428.41	Acute combined systolic and diastolic heart failure
428.42	Chronic combined systolic and diastolic heart failure
428.43	Acute on chronic combined systolic and diastolic heart failure
428.9	Heart failure, unspecified
Coronary Artery Disease	
414.0	Coronary atherosclerosis
Hypertension[b]	
401	Essential hypertension
402.00	Malignant hypertensive heart disease without heart failure
402.10	Benign hypertensive heart disease without heart failure
402.90	Unspecified hypertensive heart disease without heart failure
Myocardial Infarction	
410	Acute myocardial infarction
412	Old myocardial infarction

[a] ICD-9: International Classification of Diseases version 9.

[b] ICD-9 codes for hypertensive conditions *with* heart failure were not included because heart failure onset later in the intensive care unit stay is used as the prediction outcome.

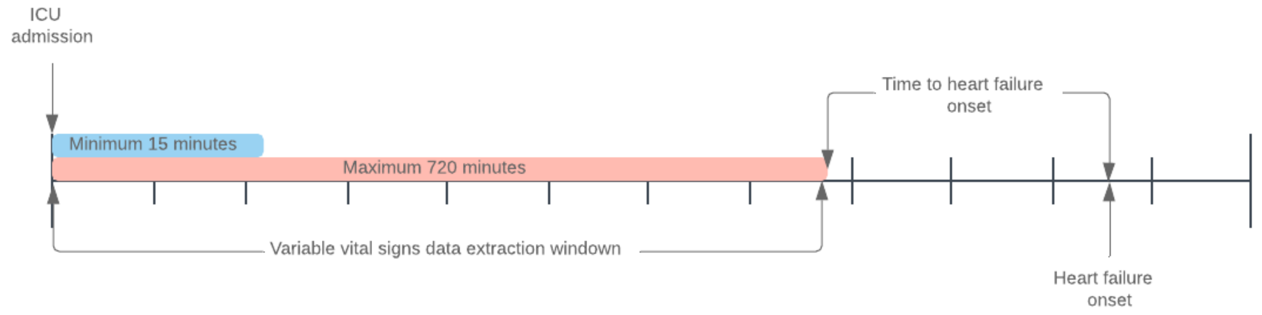
Vital signs, laboratory values, and Acute Physiology and Chronic Health Evaluation (APACHE) IVa variables were extracted for all four patient groups (three risk factor groups and the *nonrisk factor* patients). APACHE variables included features such as age and gender, admission diagnoses, and worst physiological values in the first 24 hours of ICU admission (eg, white blood count, temperature, respiratory rate) [23]. In total, 35 APACHE variables were extracted for each patient stay. Discrete APACHE variables such as *admission diagnosis* and *admission source* that do not reflect an ordinal or hierarchical relationship were encoded using the one-hot vector method.

Laboratory variables were selected based on those measurements that are routinely performed under normal ICU operations. We found overlap with our extracted lab values and those used in previous studies to predict heart failure [24]. In total, we used seven lab measurements: bedside glucose, potassium, sodium, glucose, hemoglobin, creatinine, and blood urea nitrogen. All of which were within the ten most frequently performed laboratory measurements in our data set. To predict decompensation as early in the ICU as possible, only the first measurement for each of the selected lab values was retained for model input.

Vital signs included data collected at both regular and irregular intervals. For example, temperature, heart rate, and respiratory rate tend to be regularly recorded in clinical practice and subsequently archived to the database, while cardiac output and noninvasive blood pressure may be recorded at irregular time intervals. When available at the bedside, vital signs data are collected from bedside monitoring devices at a frequency of 1-minute averages and archived as 5-minute median values. A total of 23 physiological vital signs features were extracted and are listed in Multimedia Appendix 1.

To predict heart failure onset as early as possible, vital signs were extracted at variable time windows based on number of minutes from ICU admission (Figure 1). For example, a time window of 180 minutes results in vital signs extraction from the time of ICU admission to 180 minutes after admission. The extraction window was varied from 15 minutes to 720 minutes (12 hours) from the time of admission. All available vital signs data were aggregated to mean, median, minimum, maximum, and standard deviation for each feature. This eliminated variations in the time series length between unit stays caused by irregular data sampling and missing data within each series.

Figure 1. Timeline illustrating vital signs data extraction window from the time of ICU admission. ICU: intensive care unit.


ICU admission
Minimum 15 minutes
Maximum 720 minutes
Time to heart failure onset
Variable vital signs data extraction windown
Heart failure onset

Multivariate Logistic Regression

We applied multivariate logistic regression using a binary L2 penalized minimization cost function where the target class prediction ($\hat{y}$) is a linear combination of the input features with a coefficient vector $w = (w_1, ..., w_p)$ and intercept w_0 (1), where input vectors $x = (x_1, ..., x_p)$ consist of discrete physiological variables and aggregate vital signs measurements.

$$\hat{y}(w,x) = w_0 + w_1x_1 + ... + w_px_p \mathbf{(1)}$$

Model input features minimize the cost variable (*c*) and coefficients (*w*) in the minimization cost function (2).

Combinations of input variables were tested for each *risk factor* and *nonrisk factor* cohort.

Random Forest

The random forest model was applied with the Gini impurity measure for each cohort and compared to logistic regression performance. Random forest is an ensemble method that uses a collection of tree-structured classifiers to calculate the average prediction over all individual decision tree classifiers. Inputs to each tree consist of randomly split combinations of input feature vectors $x_p \in R^n$, $i = 1, \ldots, l$ and target labels (heart failure or not heart failure) $y \in R^l$. The data (*Q*) at each node (*m*) was used to calculate Gini impurity by multiplying node importance by $H(X_m)$ through (3), where $\theta = (j, t_m)$ for each data split consisting of a feature *j* and threshold t_m. Node importance was denoted as $n_{left\ or\ right}$, and the equation is recursed for each node subset until the maximum depth is reached (ie, $N_m<min_{samples}$ or N_m=1).

A minimum split requirement of two samples was used with no maximum depth parameter, meaning all tree nodes were expanded until leaves contained less than two samples. The maximum number of estimators (number of trees in the forest) was chosen empirically during testing and held constant at 150 estimators for all input combinations.

Test and Evaluation

All model input variables were standardized centering the data around zero by subtracting the mean of each feature and dividing by the standard deviation. Model inputs consisted of lab values, APACHE variables, or aggregate vital signs as individual sets of inputs or as combinations of input features (ie, labs and vitals, labs and APACHE, vitals and APACHE, all three input data types). Each logistic regression and random forest model was tested with each data type and combination of inputs.

More extensive testing was performed using vital signs only as the data extraction window was varied to determine the impact of aggregating longer time series. Vital signs inputs were tested from the minimum to maximum data extraction window (15-720 minutes from ICU admission).

We then used the random forest model to identify the most important input features for predicting heart failure. The ensemble tree structure of random forest is easily interpretable and allows for the calculation of the relative importance of each feature.

Model performance was evaluated across all four patient cohorts. In addition, we combined coronary artery disease, hypertension, and patients with myocardial infarction into a single *risk factor* cohort for side-by-side comparison with the *nonrisk factor* patients. Results are included for individual patient groups and the combined *risk factor patients*.

Training and testing were performed with 67% train and 33% test split allowing for a sufficient number of patients to return statistically meaningful results and a test group which was representative of each cohort as a whole. Model performance was evaluated by accuracy and area under the receiver operating characteristic curve (AUC). Precision (true positives divided by the sum of true positives and false positives) and recall (true positives divided by the sum of true positives and false negatives) are also calculated along with precision-recall (P-R) curves to describe how good the models are at predicting heart failure correctly as opposed to correctly predicting patients with nonheart failure. Data preprocessing and prediction modeling was performed in Python (v.2.7.14; Python Software Foundation) using the Pandas (v.0.23.4) [25], Seaborn (v.0.9.0) [26], and sci-kit learn package (v.0.19) [27] libraries.

Results

Our study sample consisted of 145,913 adult ICU stays from 122,884 unique patients with a slightly higher number of male than female patients covering a wide range of diagnoses. Additional patient characteristics within each risk factor cohort and *nonrisk factor patients* are shown in Table 2.

Table 2. Heart failure and nonheart failure patient characteristics.

Risk factor cohort	Coronary artery disease	Hypertension	Myocardial infarction	Nonrisk patients
Patients, n	2885	17,376	6273	96,350
ICU[a] stays, n	3161	19,424	6689	116,639
Readmissions, n (%)	276 (8.73)	2048 (10.54)	416 (6.22)	20,289 (17.39)
Heart failure rate, n (%)	715 (22.62)	3058 (15.74)	799 (11.95)	7571 (6.49)
Age (years), median (IQR)	71 (16)	67 (21)	66 (20)	64 (24)
Gender (male), n (%)	2154 (68.14)	10,304 (53.04)	4255 (63.61)	62,387 (53.49)
Ethnicity, n (%)				
Caucasian	2605 (82.41)	13,161 (67.76)	5366 (80.22)	91,176 (78.17)
African American	263 (8.32)	3333 (17.16)	533 (7.97)	12,461 (10.68)
Hispanic	137 (4.33)	1549 (7.97)	196 (2.93)	3817 (3.27)
Asian	21 (0.66)	333 (1.71)	91 (1.36)	1628 (1.40)
Native American	11 (0.35)	69 (0.36)	21 (0.31)	926 (0.79)
Other/unknown	124 (3.93)	979 (5.04)	482 (7.20)	1426 (5.68)
APACHE[b] score, median (IQR)	54 (29)	50 (28)	46 (30)	51 (32)
ICU LOS[c] (days), median (IQR)	1.99 (2.69)	1.86 (2.51)	1.69 (2.06)	1.80 (2.29)
ICU mortality, n (%)	146 (4.62)	737 (3.79)	432 (6.46)	7127 (6.11)
Hospital LOS (days), median (IQR)	6.32 (7.39)	5.43 (6.99)	3.86 (5.86)	5.61 (7.06)
Hospital mortality, n (%)	245 (7.75)	1319 (6.79)	632 (9.45)	11,255 (9.65)

[a] ICU: intensive care unit.

[b] APACHE: Acute Physiology and Chronic Health Evaluation.

[c] LOS: length of stay.

Patients with hypertension were much more prevalent than patients with myocardial infarction or coronary artery disease, as might be expected. Coronary artery disease, hypertension, and myocardial infarction account for a total of 4572 (37.65%) of 12,143 total heart failure unit stays, suggesting that most patients present to the ICU without diagnosis of one of these three risk factors. It is important to note, however, that we are examining remote monitoring critical care data only. Risk factors may be captured in hospital bedside records prior to ICU admission. Readmissions to the ICU for illnesses other than heart failure account for 2740 of 29,274 (9.36%) ICU stays in the three risk factor cohorts and 20,289 of 116,639 (17.39%) stays of *nonrisk factor patients*.

The AUC and P-R curves for the *risk factor* and *nonrisk factor patients* for both logistic regression and random forest are shown in Figures 2 and 3. Additional AUC and P-R curves for each risk factor group individually are included in Multimedia Appendix 2. For all AUC and P-R curves, the vital signs data extraction window was held constant at 360 minutes from ICU admission. Clearly, discrete APACHE variables outperform lab values and vital signs individually; however, combining inputs with APACHE variables improves results. Additionally, it appears lab values had a greater impact on performance than vital signs alone as seen by the "APACHE + labs" curves relative to other combinations of input variables.

Figure 2. Nonrisk factor patients (patients presenting to the intensive care unit without risk factor of heart failure) area under receiver operating characteristic curve and precision-recall curve for both multivariate logistic regression and random forest models. Each curve represents a different model input combination. Vital signs data extraction window was held constant at 360 minutes for all inputs. APACHE: Acute Physiology and Chronic Health Evaluation.

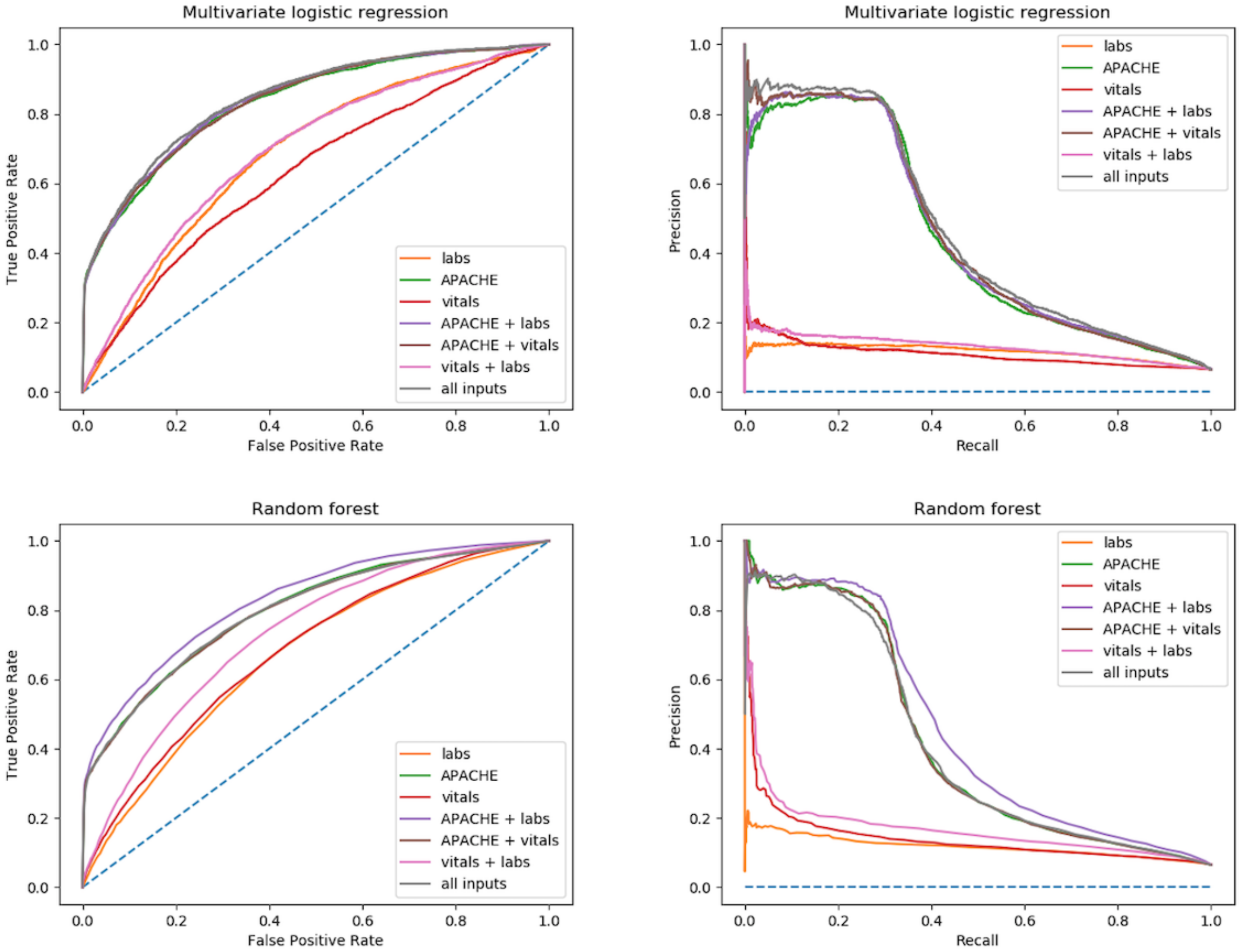


Figure 3. Risk factor patients (patients presenting to the intensive care unit with coronary artery disease, hypertension, or myocardial infarction) area under receiver operating characteristic curve and precision-recall curve for both multivariate logistic regression and random forest models. Each curve represents a different model input combination. The vital signs data extraction window was held constant at 360 minutes for all inputs. APACHE: Acute Physiology and Chronic Health Evaluation.

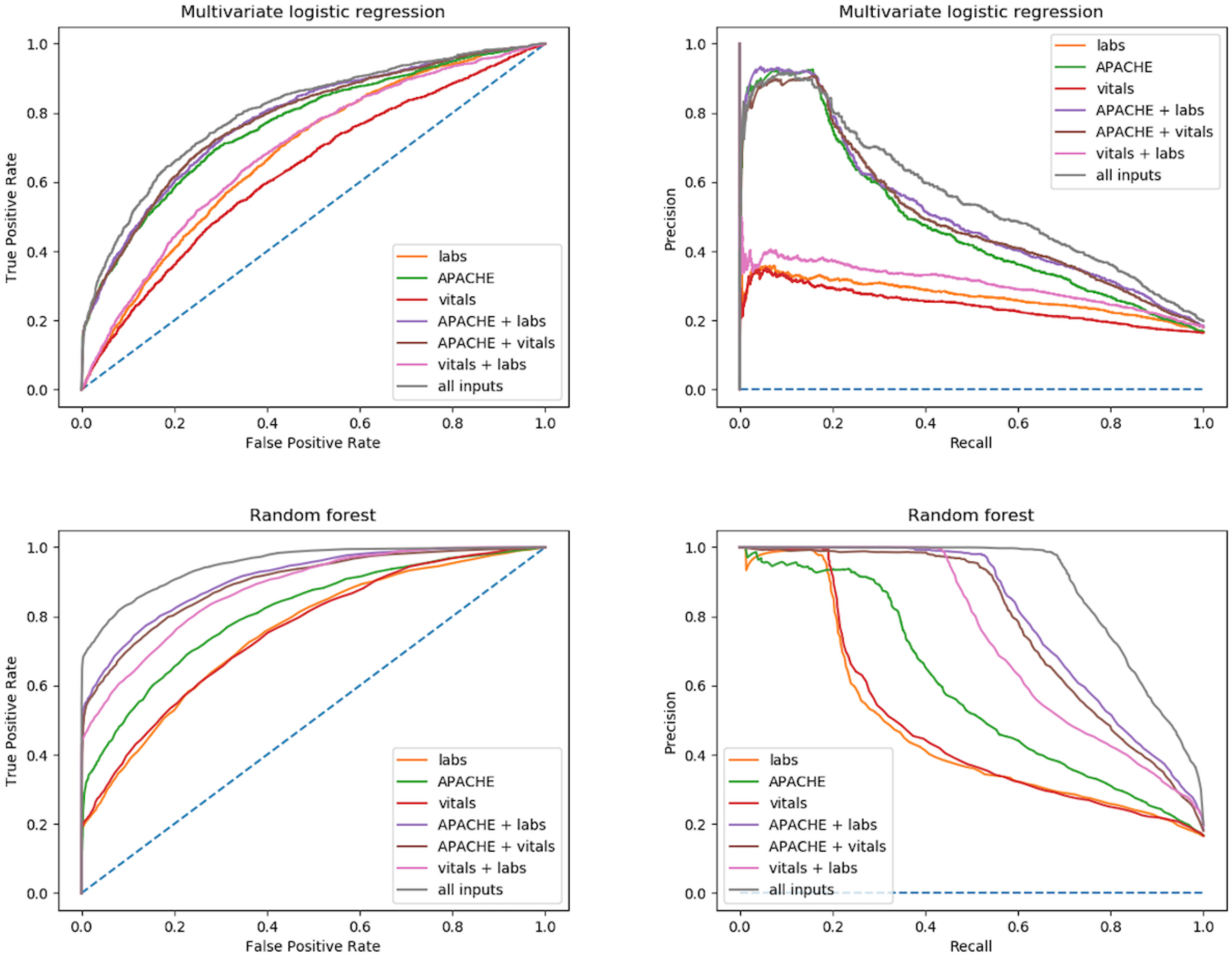


Table 3. Logistic regression and random forest F1 scores across model input combinations. Vital signs data extraction window held constant at 360 minutes for all trials.

Patients	Logistic Regression	Random Forest
Risk factor patients		
APACHE[a]	0.82	0.85
Labs	0.76	0.82
Vitals	0.76	0.83
APACHE + labs	0.81	0.90
APACHE +vitals[b]	0.81	0.90
Labs + vitals	0.75	0.88
APACHE + labs + vitals	0.81	0.93
Nonrisk factor patients		
APACHE	0.94	0.94
Labs	0.90	0.90
Vitals	0.90	0.90
APACHE + labs	0.94	0.94
APACHE +vitals	0.94	0.94
Labs + vitals	0.90	0.90
APACHE + labs + vitals	0.94	0.94

[a] APACHE: Acute Physiology and Chronic Health Evaluation.

[b] Vital signs extraction window of 360 minutes from intensive care unit admission.

Both models were compared across input combinations for *risk factor* and *nonrisk factor patients* using the F1 score (Table 3). Interestingly, logistic regression with APACHE and labs inputs had the highest F1 score, while, in general, random forest has higher AUC, accuracy, and weighted average precision and recall (Tables 4 and 5). In this application, precision shows what proportion of heart failure identifications were actually heart failure, and recall is the proportion of heart failure stays that were correctly identified [28]. Random forest with APACHE, laboratory measurements, and vital signs combined model inputs had the highest performance metrics at an AUC of 0.9503, accuracy of 93.15%, and micro- and macroweighted average precision and recall of 0.93 and 0.93, respectively. It is important to note that, although the weighted average precision and recall are fairly high, the P-R curves exhibit a steep drop in precision as recall increases.

Table 4. Heart failure prediction accuracy and AUC.

Models	Risk factor patients		Nonrisk factor patients	
	AUC[a]	Accuracy	AUC	Accuracy
Logistic regression				
APACHE[b] + labs	0.7790	0.8417	0.8396	0.9501
APACHE + vitals[c]	0.7775	0.8456	0.8374	0.9512
Labs + vitals[c]	0.6859	0.8125	0.6947	0.9333
APACHE + labs + vitals[c]	0.8005	0.8357	0.8458	0.9502
Random forest				
APACHE + labs	0.9081	0.9112	0.8285	0.9499
APACHE +vitals[c]	0.8956	0.9080	0.7967	0.9488
Labs + vitals[c]	0.8794	0.8965	0.7318	0.9343
APACHE + labs + vitals[c]	0.9503	0.9315	0.7999	0.9471

[a] AUC: area under the receiver operating characteristic curve.

[b] APACHE: Acute Physiology and Chronic Health Evaluation.

[c] Vital signs extraction window of 360 minutes from intensive care unit admission.

Table 5. Logistic regression and random forest precision and recall.

Models	Risk factor patients		Nonrisk factor patients	
	Precision[a]	Recall[b]	Precision[a]	Recall[b]
Logistic regression				
APACHE[c] + labs	0.82	0.84	0.94	0.95
APACHE +vitals[d]	0.83	0.85	0.95	0.95
Labs + vitals[d]	0.74	0.81	0.89	0.93
APACHE + labs + vitals[d]	0.82	0.84	0.95	0.95
Random forest				
APACHE + labs	0.92	0.91	0.95	0.95
APACHE +vitals[d]	0.91	0.91	0.94	0.95
Labs + vitals[d]	0.91	0.90	0.92	0.93
APACHE + labs + vitals[d]	0.93	0.93	0.94	0.95

[a] Weighted average microprecision and macroprecision.

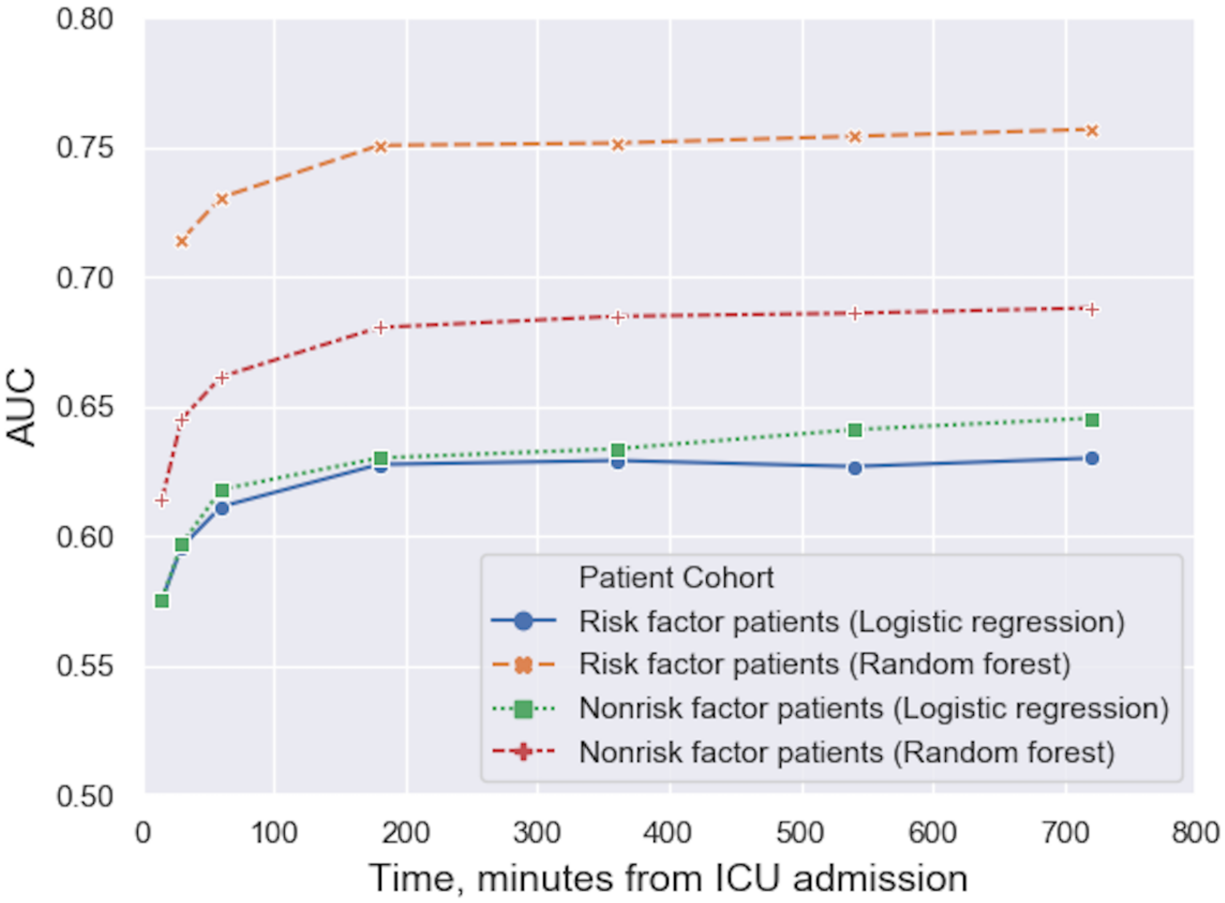
[b] Weighted average microrecall and macrorecall.

[c] APACHE: Acute Physiology and Chronic Health Evaluation.

[d] Vital signs model inputs at 360 minutes from intensive care unit admission.

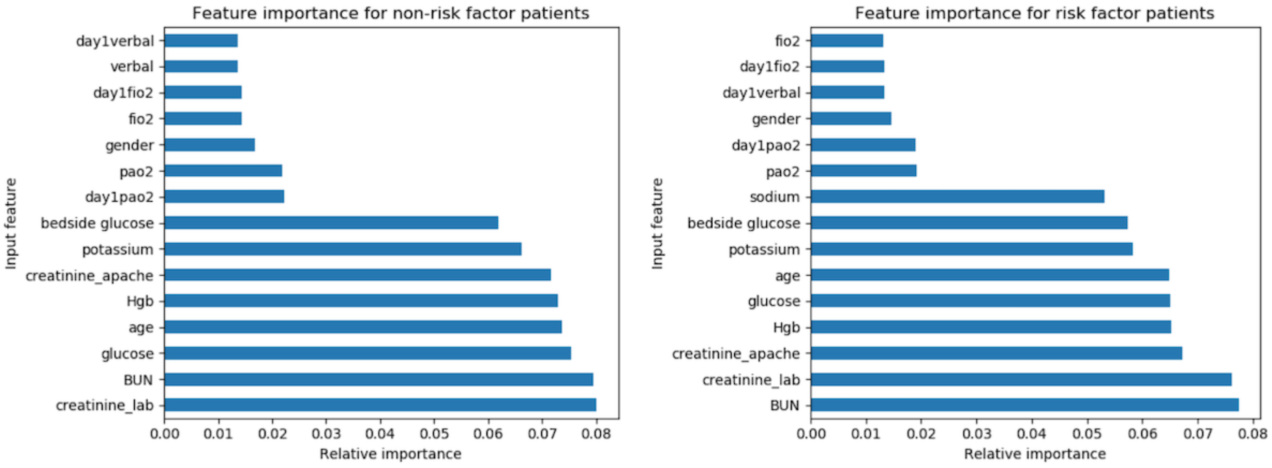
Using only aggregate vital signs as data inputs we evaluated model performance across variable vitals data extraction windows. Figure 4 illustrates AUC values (y-axis) of each model at different extraction window sizes (x-axis). In both models, there appears a point of diminishing returns around 200 minutes where additional vital signs data do not continue to improve results. This behavior is seen in both prediction models across all patient cohorts.

Figure 4. Predication AUC for risk factor and nonrisk factor patients with variable vital signs extraction time windows from 15 minutes to 720 minutes using only vital signs as model inputs. The x-axis represents the total number of minutes from ICU admission that vital signs were extracted from the database, meaning at higher time values more data was extracted. AUC: area under receiver operating characteristic curve; ICU: intensive care unit.



We then used the random forest model to identify which discrete features were most influential in predicting heart failure by plotting the relative feature importance. We applied the same number of estimators (n_estimators=150) and calculated feature importance for all lab values and APACHE variables (Figure 5). The selected top features were similar between *risk factor* and *nonrisk factor patients*. In addition, many of the top 10 features are laboratory values, even though, when used as individual inputs, APACHE variables outperformed laboratory measurements.

Figure 5. Random forest feature importance with 150 estimators for nonrisk factor and risk factor patients. BUN: blood urea nitrogen.



Discussion

Performance and Clinical Relevance

In this study, we evaluated two interpretable prediction models for decompensation in critical care using heart failure onset as a target outcome. Both logistic regression and random forest were evaluated as close to the time of ICU admission as possible using multiple types of input features.

We found that results across all four cohorts showed reasonable prediction accuracy. Generally, random forest outperformed multivariate logistic regression. On an individual basis, APACHE variables predicted heart failure onset better than laboratory measurements or vital signs; however, the best performance was achieved when model inputs were combined. Trials consisting of APACHE and laboratory measurements or all three data inputs (APACHE, labs, and vitals) had the highest performance metrics compared to any individual trial. This was corroborated by random forest feature selection highlighting several laboratory measurements as important to heart failure prediction relative to other input features.

Although vital signs near the time of ICU admission improve heart failure predictions when combined with other inputs, overall, vital signs results individually were not strong. Methodologically, vital signs and laboratory measurements, however, are promising for future prediction models. Traditional severity scoring models, such as APACHE, use data from only the first 24 hours of an ICU stay. Laboratory measurements and vital signs, however, are typically monitored on a continuous or semicontinuous basis throughout the length of an ICU stay. This would allow for future iterations of our prediction models to make predictions closer to the time of heart failure rather than being limited to ICU admission time. The continuous monitoring of vital signs and temporal value of laboratory measurements could also allow predictions to be made prospectively on a semicontinuous basis (eg, prediction output every 3 hours).

In addition, vital signs AUC values in Figure 4 suggest that there is an optimal threshold in the size of data extraction window for both predictive performance and computational load, and could inform future prediction models. If not enough data are extracted, results are diminished. Similarly, a data extraction time window too large increases computational load and does not necessarily improve performance.

Prediction window variation has been applied over longer time periods and multiple hospital visits for heart failure detection. We applied a similar methodology over a much shorter time frame more appropriate for ICU visits. Earlier predictions allow clinicians to determine patient prognosis and begin appropriate intervention. Clinicians may also revisit disease state predictions throughout a patient stay based on treatments or emergence of comorbidities.

Higher frequency continuous vital signs data in conjunction with laboratory measurements are a feasible option for predicting heart failure or other patient decompensation events in critical care through tele-ICU data early in an ICU stay. Vital signs tend to be available upon admission and continue through the majority of a patient ICU stay allowing for semicontinuous predictions. Real-time predictions throughout a patient stay are particularly useful for illnesses such as heart failure where poor outcomes can range from chronic to acute onset. In addition, heart failure mode of death assessments illustrate high variability as well and require predictions that facilitate timely interventions specific to the associated risks [17].

Results were similar between *risk factor* and *nonrisk factor patients* meaning accurate heart failure prediction will likely be made for patients not presenting with an indication of apparent risk of heart failure. This is supported by the similar AUC, precision, recall, and F1 scores across both models for *nonrisk factor patients* and could be used to inform ICU clinicians of impending failure for patients not initially deemed at risk.

Challenges and Limitations

The prediction models in this study demonstrate the viability of machine learning applications leveraging remote monitoring data to further alleviate the challenges imposed by complex and data-intensive critical care environments, and contribute to the prognostication of cardiovascular diseases in the ICU. Our prediction models, however, may be partially influenced by and do not compensate for potential bias due to ICD-9 coding practices. Heart failure is not an explicitly defined event but rather a patient state in which the heart is struggling to function properly and as such is difficult to diagnose.

Moreover, vital signs data were collected using bedside monitoring systems as 1-minute averages and archived into the database as 5-minute median values. This decreased granularity over varying time windows of vital signs data extraction. Data may miss critical, subclinical cardiovascular events. Additional information loss occurs by reducing vital signs from time series data to discrete aggregate values. Data collection frequencies, however, are generally dependent upon what measurements are being taken from each patient at the bedside and at what times during their ICU stay. This can also cause high variability in time intervals between data points for each patient unit stay and total length of each time series.

Lastly, our approach does not account for the temporal relationship between vital signs data extraction or laboratory measurements and the prediction event. In an attempt to predict patient decompensation soon after ICU admission our variable data window begins at time of admission regardless of when heart failure onset may have occurred. Similarly, laboratory measurements are taken throughout a patient ICU stay, yet we retained only the first measurement in the interest of early decompensation prediction. An alternative approach to data aggregation is time series analysis of continuous, more granular, and physiologic data. This is corroborated by a recent study that showed the importance of temporal relations in recurrent neural network model inputs and is a possible future avenue for this work [29].

Future Work

Logistic regression and random forest methods were selected based on interpretability and previous critical care applications using similar data inputs [30]. Model inputs, however, were limited to discrete variables. Alternatively, handling vital signs data as time series model inputs without overaggregating may yield improved results. A sliding window approach with real time series data and more powerful machine learning methods would allow for subsequent predictions to be made well after admission and throughout a patient stay [31]. This alternative approach would address the temporal relationship between the decompensation event (heat failure onset) and the input data used to make the prediction.

Ongoing and future studies also include analysis and machine learning application to specific events, which contribute to risk of heart failure onset (eg, myocardial infarction and pulmonary embolism). The ability to predict and potentially prevent these distinct events may subsequently avoid patient decompensation rather than predicting heart failure itself. In conjunction with feature selection, events or physiologic features most relevant to heart failure onset in critical care could be refined, thus, improving results. Model inputs could also be altered such that the heart failure risk factors are used as additional inputs rather than using risk factors for cohort segregation.

There are many different ICU types including cardiac ICUs. Heart failure may be managed differently in different critical care settings. Further research in this area could give insight to heart failure management variation. Our modeling approach may alleviate variations across ICUs by acting as a support system for clinicians focused on diagnoses other than heart failure.

Conclusions

Remotely monitored critical care data offers opportunity for machine learning applications and deeper analysis than what may be possible at the bedside. Handling of disparate clinical data sources, data cleaning, preprocessing, and leveraging machine learning techniques may take place remotely so as to not disrupt existing ICU workflow and to provide complex clinical decision support. Risk factors for patient decompensation, or clinical deterioration, are prevalent in tele-ICU data as are clinical features sufficient for clinically relevant patient decompensation predictions with interpretable machine learning methods. Both logistic regression and random forest models were able to identify appropriate input features and narrowed data extraction time windows and thresholds for computational limitations at roughly 200 minutes after ICU admission. Our approach validates the feasibility of identifying decompensation events and patient risk factors, and making predictions using dissimilar data from variable timelines. More powerful machine learning approaches beyond regression and ensemble methods with alteration of our data extraction time window approach to avoid data aggregation could yield improved results in predicting heart failure onset or other patient decompensation events in critical care, albeit at the expense of interpretability.

Acknowledgments

This work was supported in part by the National Science Foundation under grant #1838745 and the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number 5T32HL007955-19. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, National Science Foundation, or other organizations.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Vital signs measurements used in all trials where model input variables included vitals measurements.

[PPTX File , 39 KB - medinform_v8i8e19892_app1.pptx]

Multimedia Appendix 2

Supplementary multivariate logistic regression and random forest area under receiver operating characteristic curve and precision-recall curve for three individual risk factor groups: coronary artery disease, hypertension, and myocardial infarction.

[PPTX File , 1834 KB - medinform_v8i8e19892_app2.pptx]

References

1. Celi LA, Csete M, Stone D. Optimal data systems. Curr Opinion Crit Care 2014;20(5):573-580. [doi: 10.1097/mcc.0000000000000137]
2. Ghassemi M, Celi L, Stone DJ. State of the art review: the data revolution in critical care. Crit Care 2015 Mar 16;19:118 [FREE Full text] [doi: 10.1186/s13054-015-0801-4] [Medline: 25886756]
3. Smith D. Predicting poor outcomes in heart failure. Permanente J 2011 Sep 1;15(4). [doi: 10.7812/tpp/11-100]
4. Sahle BW, Owen AJ, Chin KL, Reid CM. Risk prediction models for incident heart failure: a systematic review of methodology and model performance. J Card Fail 2017 Sep;23(9):680-687. [doi: 10.1016/j.cardfail.2017.03.005] [Medline: 28336380]

5. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, et al. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. JACC Heart Fail 2020 Jan;8(1):12-21 [FREE Full text] [doi: 10.1016/j.jchf.2019.06.013] [Medline: 31606361]
6. Hollenberg SM, Warner Stevenson L, Ahmad T, Amin VJ, Bozkurt B, Butler J, et al. 2019 ACC expert consensus decision pathway on risk assessment, management, and clinical trajectory of patients hospitalized with heart failure: a report of the American College of Cardiology Solution Set Oversight Committee. J Am Coll Cardiol 2019 Oct 15;74(15):1966-2011. [doi: 10.1016/j.jacc.2019.08.001] [Medline: 31526538]
7. Rathi S, Deedwania PC. The epidemiology and pathophysiology of heart failure. Med Clin North Am 2012 Sep;96(5):881-890. [doi: 10.1016/j.mcna.2012.07.011] [Medline: 22980052]
8. Nieminen MS, Harjola V. Definition and epidemiology of acute heart failure syndromes. Am J Cardiol 2005 Sep 19;96(6A):5G-10G. [doi: 10.1016/j.amjcard.2005.07.015] [Medline: 16181818]
9. Tanai E, Frantz S. Pathophysiology of heart failure. Compr Physiol 2015 Dec 15;6(1):187-214. [doi: 10.1002/cphy.c140055] [Medline: 26756631]
10. Roger VL, Weston SA, Redfield MM, Hellermann-Homan JP, Killian J, Yawn BP, et al. Trends in heart failure incidence and survival in a community-based population. JAMA 2004 Jul 21;292(3):344-350. [doi: 10.1001/jama.292.3.344] [Medline: 15265849]
11. Anker SD, Koehler F, Abraham WT. Telemedicine and remote management of patients with heart failure. Lancet 2011 Aug;378(9792):731-739. [doi: 10.1016/s0140-6736(11)61229-4]
12. Bashi N, Karunanithi M, Fatehi F, Ding H, Walters D. Remote monitoring of patients with heart failure: an overview of systematic reviews. J Med Internet Res 2017 Jan 20;19(1):e18 [FREE Full text] [doi: 10.2196/jmir.6571] [Medline: 28108430]
13. Del Gobbo LC, Kalantarian S, Imamura F, Lemaitre R, Siscovick DS, Psaty BM, et al. Contribution of major lifestyle risk factors for incident heart failure in older adults: the cardiovascular health study. JACC Heart Fail 2015 Jul;3(7):520-528 [FREE Full text] [doi: 10.1016/j.jchf.2015.02.009] [Medline: 26160366]
14. Henriques J, Carvalho P, Rocha T, Paredes S, Morais J. Multi-parametric trends analysis and events prediction in the context of a cardiac rehabilitation system. 2015 Presented at: 6th European Conference of the International Federation for Medical and Biological Engineering; 2015; Switzerland p. 678-681. [doi: 10.1007/978-3-319-11128-5_169]
15. Henriques J, Carvalho P, Paredes S, Rocha T, Habetha J, Antunes M, et al. Prediction of heart failure decompensation events by trend analysis of telemonitoring data. IEEE J Biomed Health Inform 2015 Sep;19(5):1757-1769. [doi: 10.1109/jbhi.2014.2358715]
16. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, et al. The Seattle Heart Failure Model. Circulation 2006 Mar 21;113(11):1424-1433. [doi: 10.1161/circulationaha.105.584102]
17. Mozaffarian D, Anker SD, Anand I, Linker DT, Sullivan MD, Cleland JG, et al. Prediction of mode of death in heart failure. Circulation 2007 Jul 24;116(4):392-398. [doi: 10.1161/circulationaha.106.687103]
18. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Dwivedi G. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. ESC Heart Fail 2019 Apr;6(2):428-435. [doi: 10.1002/ehf2.12419] [Medline: 30810291]
19. Koulaouzidis G, Iakovidis D, Clark A. Telemonitoring predicts in advance heart failure admissions. Int J Cardiol 2016 Aug 01;216:78-84. [doi: 10.1016/j.ijcard.2016.04.149] [Medline: 27140340]
20. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017;12(4):e0174944 [FREE Full text] [doi: 10.1371/journal.pone.0174944] [Medline: 28376093]
21. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018 Sep 11;5:180178 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
22. Essay P, Shahin TB, Balkan B, Mosier J, Subbian V. The connected intensive care unit patient: exploratory analyses and cohort discovery from a critical care telemedicine database. JMIR Med Inform 2019 Jan 24;7(1):e13006 [FREE Full text] [doi: 10.2196/13006] [Medline: 30679148]
23. Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 1981 Aug;9(8):591-597. [doi: 10.1097/00003246-198108000-00008] [Medline: 7261642]
24. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA Netw Open 2020 Jan 03;3(1):e1918962 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.18962] [Medline: 31922560]
25. van der Walt S, Millman J. Python in science. 2010 Presented at: 9th Python in Science Conference; 2010; Austin, Texas URL: http://conference.scipy.org/proceedings/scipy2010/pdfs/proceedings.pdf
26. Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Gemperline D, et al. mwaskom/seaborn: v0.8.1 (September 2017). Zenodo 2017 Jul 16. [doi: 10.5281/zenodo.883859]

27. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B. Scikit-learn: machine Learning in Python. J Machine Learning Res 2011;12:2825.
28. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. 2006 Presented at: 23rd International Conference on Machine learning; 2006; Pittsburgh, PA p. 233-240. [doi: 10.1145/1143844.1143874]
29. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc 2017 Mar 01;24(2):361-370 [FREE Full text] [doi: 10.1093/jamia/ocw112] [Medline: 27521897]
30. Balkan B, Essay P, Subbian V. Evaluating ICU clinical severity scoring systems and machine learning applications: APACHE IV/IVa case study. Conf Proc IEEE Eng Med Biol Soc 2018 Jul;2018:4073-4076. [doi: 10.1109/EMBC.2018.8513324] [Medline: 30441251]
31. Maragatham G, Devi S. LSTM model for prediction of heart failure in big data. J Med Syst 2019 Mar 19;43(5):111. [doi: 10.1007/s10916-019-1243-3] [Medline: 30888519]

Abbreviations

APACHE: Acute Physiology and Chronic Health Evaluation
AUC: area under the receiver operating characteristic curve
ICD-9: International Classification of Diseases version 9
ICU: intensive care unit
P-R: precision-recall

Edited by G Eysenbach; submitted 05.05.20; peer-reviewed by E van der Velde, H Ross, I Kedan, NM Trofenciuc; comments to author 27.05.20; revised version received 12.06.20; accepted 07.07.20; published 07.08.20.

Please cite as:
Essay P, Balkan B, Subbian V
Decompensation in Critical Care: Early Prediction of Acute Heart Failure Onset
JMIR Med Inform 2020;8(8):e19892
URL: http://medinform.jmir.org/2020/8/e19892/
doi:10.2196/19892
PMID:32663162

Original Paper

Machine Learning Model Based on Transthoracic Bioimpedance and Heart Rate Variability for Lung Fluid Accumulation Detection: Prospective Clinical Study

Natasa Reljin[1], PhD; Hugo F Posada-Quintero[1], PhD; Caitlin Eaton-Robb[1], MSc; Sophia Binici[2], BSc; Emily Ensom[3], BSc; Eric Ding[3], BSc; Anna Hayes[3], MD; Jarno Riistama[4], PhD; Chad Darling[5], MD; David McManus[3], MD; Ki H Chon[1], PhD

[1]Department of Biomedical Engineering, University of Connecticut, Mansfield, CT, United States

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

[3]University of Massachusetts Memorial Hospital Care, Worcester, MA, United States

[4]Philips Research, Eindhoven, The Netherlands

[5]Department of Emergency Medicine, University of Massachusetts Medical School, Worcester, MA, United States

Corresponding Author:
Hugo F Posada-Quintero, PhD
Department of Biomedical Engineering, University of Connecticut
67 N Eagleville Rd, Storrs
Mansfield, CT, 06269
United States
Phone: 1 5088739247
Email: h.posada@uconn.edu

Abstract

Background: Accumulation of excess body fluid and autonomic dysregulation are clinically important characteristics of acute decompensated heart failure. We hypothesized that transthoracic bioimpedance, a noninvasive, simple method for measuring fluid retention in lungs, and heart rate variability, an assessment of autonomic function, can be used for detection of fluid accumulation in patients with acute decompensated heart failure.

Objective: We aimed to evaluate the performance of transthoracic bioimpedance and heart rate variability parameters obtained using a fluid accumulation vest with carbon black–polydimethylsiloxane dry electrodes in a prospective clinical study (System for Heart Failure Identification Using an External Lung Fluid Device; SHIELD).

Methods: We computed 15 parameters: 8 were calculated from the model to fit Cole-Cole plots from transthoracic bioimpedance measurements (extracellular, intracellular, intracellular-extracellular difference, and intracellular-extracellular parallel circuit resistances as well as fitting error, resonance frequency, tissue heterogeneity, and cellular membrane capacitance), and 7 were based on linear (mean heart rate, low-frequency components of heart rate variability, high-frequency components of heart rate variability, normalized low-frequency components of heart rate variability, normalized high-frequency components of heart rate variability) and nonlinear (principal dynamic mode index of sympathetic function, and principal dynamic mode index of parasympathetic function) analysis of heart rate variability. We compared the values of these parameters between 3 participant data sets: control (n=32, patients who did not have heart failure), baseline (n=23, patients with acute decompensated heart failure taken at the time of admittance to the hospital), and discharge (n=17, patients with acute decompensated heart failure taken at the time of discharge from hospital). We used several machine learning approaches to classify participants with fluid accumulation (baseline) and without fluid accumulation (control and discharge), termed *with fluid and without fluid* groups, respectively.

Results: Among the 15 parameters, 3 transthoracic bioimpedance (extracellular resistance, R_0; difference in extracellular-intracellular resistance, $R_0 – R_\infty$, and tissue heterogeneity, α) and 3 heart rate variability (high-frequency, normalized low-frequency, and normalized high-frequency components) parameters were found to be the most discriminatory between groups (patients with and patients without heart failure). R_0 and $R_0 – R_\infty$ had significantly lower values for patients with heart failure than for those without heart failure (R_0: *P*=.006; $R_0 – R_\infty$: *P*=.001), indicating that a higher volume of fluids accumulated in the lungs of patients with heart failure. A cubic support vector machine model using the 5 parameters achieved an accuracy of 92%

for with fluid and without fluid group classification. The transthoracic bioimpedance parameters were related to intra- and extracellular fluid, whereas the heart rate variability parameters were mostly related to sympathetic activation.

Conclusions: This is useful, for instance, for an in-home diagnostic wearable to detect fluid accumulation. Results suggest that fluid accumulation, and subsequently acute decompensated heart failure detection, could be performed using transthoracic bioimpedance and heart rate variability measurements acquired with a wearable vest.

(JMIR Med Inform 2020;8(8):e18715) doi:10.2196/18715

KEYWORDS

heart failure; transthoracic bioimpedance; heart rate variability; fluid accumulation; autonomic nervous system; machine learning; cardiology

Introduction

Heart failure is estimated to affect more than 25 million people worldwide and over 6 million people in the United States [1-4]. Acute decompensated heart failure frequently results in hospitalization and can also increase risk for arrhythmia, stroke, and death [5,6]. The most clinically apparent features associated with acute decompensated heart failure include pulmonary or peripheral edema [5,7,8]. Several validated biomarkers for acute decompensated heart failure detection exist, including body weight, B-type natriuretic protein, invasive pulmonary pressure measurement, and intrathoracic bioimpedance from cardiac implantable devices [9]. The simplest, least costly, and most widely used measure for ambulatory patients with chronic heart failure is body weight; however, body weight monitoring is not an ideal approach, since weight change correlates poorly with acute heart failure worsening, thus limiting the impact of existing home-based heart failure management programs [10].

Transthoracic bioimpedance can measure intrathoracic volume, a surrogate biomarker of pulmonary edema [11-13]. For years, it has been applied for lung fluid abnormality detection and fluid management after heart failure [14,15]. Transthoracic bioimpedance injects a small alternating current into the tissue via electrodes and measures the voltage response. By doing so, and by using Ohm’s law, the electrical resistance of the thorax can be calculated. Higher values of resistance suggest lower volumes of fluid accumulated in the lungs, and vice versa (for a detailed technical explanation of transthoracic bioimpedance, please see the Methods section). Electrocardiographic (ECG) signals are used to compute parameters of heart rate variability [16], which has been shown to be dysregulated in patients with heart failure and provides information about the autonomic nervous system [16-18].

Traditionally, various types of electrodes have been used for transthoracic bioimpedance and ECG measurements using fluid accumulation vests: adhesive Ag-AgCl electrodes, which often result in skin irritation and are often misaligned when positioned; textile electrodes, which need to be wetted prior to every use; and recently proposed reusable carbon black–polydimethylsiloxane (PDMS) dry electrodes [19,20]. In our previous work [19,20], we showed that carbon black–PDMS electrodes could be a suitable alternative to textile electrodes for measuring transthoracic bioimpedance and ECG signals using customized fluid accumulation vests. Since these electrodes are biocompatible, do not cause skin irritations, do not need to be wetted prior to use, and show comparable results to those of textile and adhesive electrodes, we decided to use carbon black–PDMS dry electrodes.

There are several studies [12,21] that have explored bioimpedance to detect acute decompensated heart failure. Our group has shown that transthoracic bioimpedance can be measured daily with fluid accumulation vests using conventional electrodes, and a predictive algorithm analyzing daily bioimpedance parameters showed reasonable performance in predicting recurrent heart failure events, including hospitalization, diuretic uptitration, and worsening heart failure symptoms [12]. Lindholm et al [22] determined that leg bioimpedance was inversely correlated with heart failure incidence, and by combining leg bioimpedance with demographic information, they obtained accurate heart failure predictions. Sato et al [23] evaluated parameters from bioelectrical impedance analysis in participants with congenital heart disease and determined that the edema index obtained from bioelectrical impedance analysis could also be a marker for heart failure severity.

In this prospective clinical study (System for Heart Failure Identification using an External Lung Fluid Device; SHIELD) to examine the performance of transthoracic bioimpedance and heart rate variability measured using carbon black–PDMS electrodes embedded in fluid accumulation vests for detection of acute decompensated heart failure, we hypothesized that (1) participants without acute decompensated heart failure should have resistance measurements that are higher than those of participants with acute decompensated heart failure at the time of admittance to the hospital; (2) participants with acute decompensated heart failure at the time of discharge from hospital should have smaller amount of accumulated lung fluid and therefore higher resistance measurements than those of participants with acute decompensated heart failure at the time of admission; and (3) autonomic function assessed by heart rate variability would provide additional information about the dysregulation of heart failure patients, hence, it would detect acute decompensated heart failure.

Methods

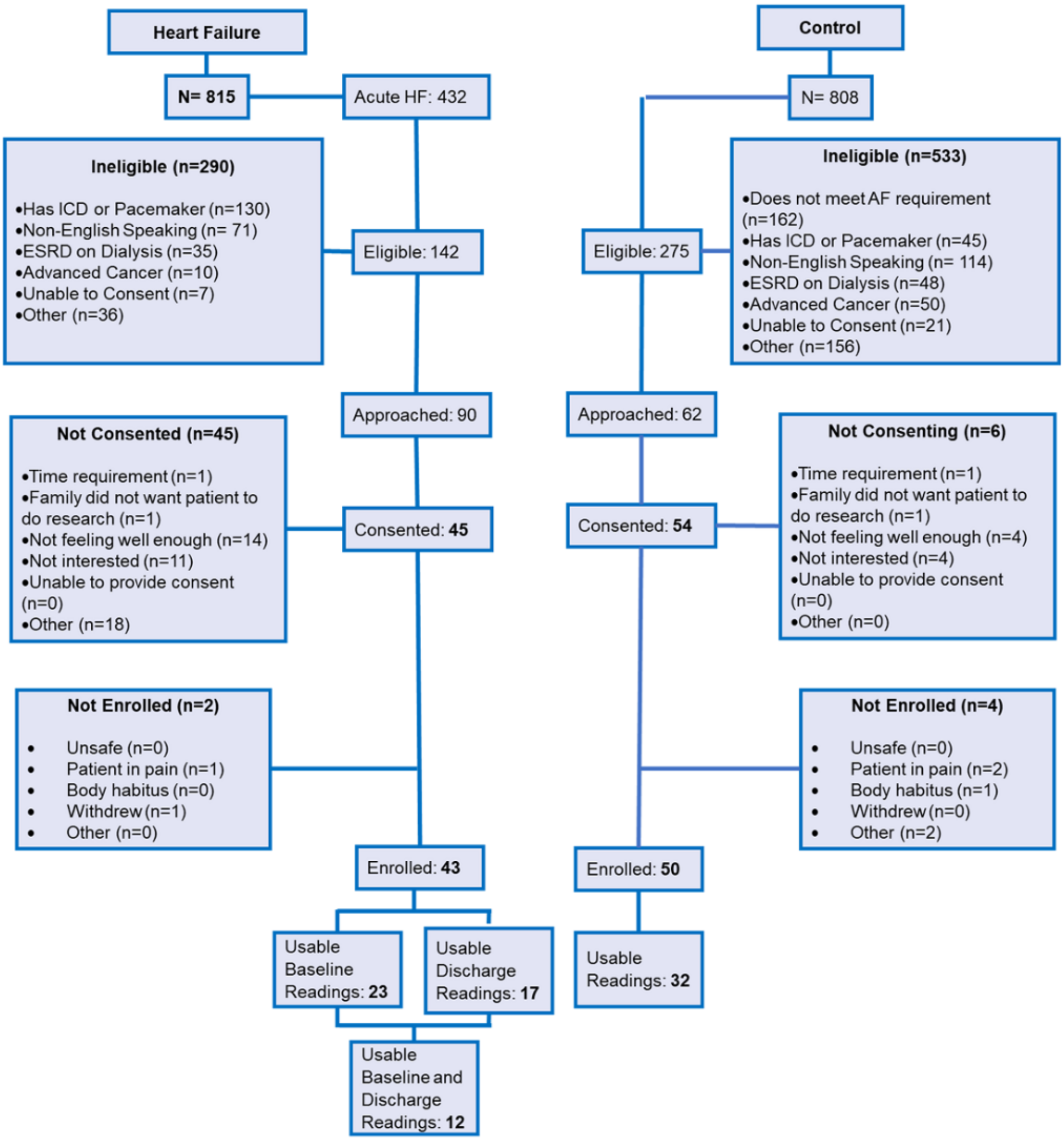
Experimental Setup

A total of 93 hospitalized individuals were prospectively enrolled in our observational study at the University of Massachusetts Medical Center. We acquired recordings from participants with acute decompensated heart failure taken within the first few hours of hospital arrival (baseline) and taken prior to discharge from hospital (discharge). We also acquired recordings from a group of patients without acute decompensated heart failure (control). All participants gave written informed consent before participating in the study, in accordance with the Declaration of Helsinki. The protocol was approved by the institutional review board of the University of Massachusetts Memorial Hospital (docket number H00014714).

The CONSORT diagram in Figure 1 depicts the screening and enrollment process for this study. We screened over 800 people for the heart failure group alone, which resulted in 432 people identified with acute heart failure. Of these 432 people, only 142 were eligible. We had strict eligibility criteria for this study. Exclusion criteria were patients with an implantable cardioverter defibrillator or pacemaker, who were non-English speaking, who were on dialysis, who had advanced cancer requiring chemotherapy, or who did not have the ability to consent. Most people were excluded from the study due to the presence of an implantable cardioverter-defibrillator or pacemaker (130/290, 44.8%). Our inclusion criteria consisted of patients who were aged over 40 years (50 years if enrolled before June 28, 2018); who were on hospital-based telemetry; who had New York Heart Association functional class II, III, or IV heart failure; and whose skin was intact.

For this study, we used Philips prototype fluid accumulation vests [12], which provide transthoracic bioimpedance measurements at 16 frequencies in the range from 10 kHz to 999 kHz and ECG recordings at 256 Hz. Participants wore the vest without clothing, so that its 4 electrodes were affixed to their left and right abdomen. Copper mesh carbon black–PDMS electrodes were used [24]. These electrodes have been proven to provide consistent transthoracic bioimpedance and ECG measurements when used with this vest [19]. For each recording, participants were asked to sit still for 10 minutes while seated on a chair with their legs resting on the floor. Once the recording was completed, a device attached to the vest wirelessly transmitted the data via a secure Bluetooth connection to a mobile phone (Samsung Galaxy Gio GT-S5660). The data were saved on an extractable secure digital memory card on the mobile phone and subsequently transferred to a PC for processing and analysis. Patients needed to be able to remain seated for at least 15 minutes to participate in the study.

Figure 1. CONSORT diagram. AF: atrial fibrillation; ESRD: end-stage renal disease; HF: heart failure; ICD: implantable cardioverter-defibrillator.



Transthoracic Bioimpedance Measurements

Transthoracic bioimpedance is a noninvasive method that measures the impedance of the tissue at a series of frequencies. A small alternating current, typically ranging from 100 μA to 10 mA, is injected into the tissue via electrodes, while the voltage drop is measured as the output. By applying Ohm's law, the resistance of the body tissue can be calculated. Biological tissue is typically modeled with a resistance R_0 to represent the extracellular fluid, in parallel with a resistance R_I to represent intracellular fluid, and a capacitance C_m to represent cell membranes [17]. Electrical current with a frequency f=0 Hz will pass around all cells, and the total resistance is equal to the resistance from the extracellular fluid only, R_0. At the other extreme, when the frequency is infinite, f=∞, the current will pass through the cells, and the total resistance can be calculated as the parallel circuit of R_0 and R_I,

where R_I can be represented as

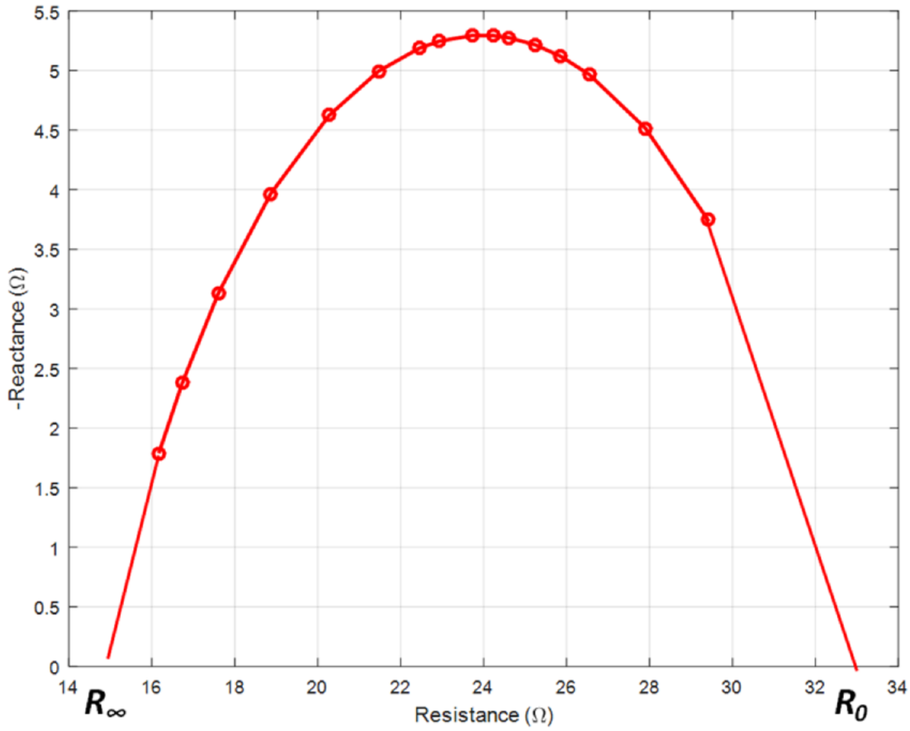
If we measure impedance for frequencies between these two extreme cases, we obtain an arc-like Cole-Cole plot in the impedance plane [25-27]. The equation for the model of the Cole-Cole plot [28,29] is

The parameters of the model can be extrapolated from a set of measurements made at a predefined set of frequencies. The exponent α represents the heterogeneity of the tissue in the model. For each frequency, the real (resistance) and imaginary (reactance) part of the electrical impedance is estimated. The Taubin algorithm [30] is used to fit a circle onto the measured impedance data. From the data computed using the Taubin algorithm, parameters of the Cole model are estimated using a heuristic search method, the Nelder-Mead algorithm [31].

Figure 2 shows an illustrative example of a Cole-Cole plot for one of the participants. The value of R_0 is obtained as the x-axis intercept at the far right side of the Cole-Cole plot, while the value of R_∞ is the x-axis intercept at the far left side of the same plot. The frequency that corresponds to the upper point of the circle is called the resonance frequency, f_c,

The sum of the square error is minimized in the fitting process. The *fitting error* was calculated as the sum of the square error at the optimal parameters. We calculated 8 transthoracic bioimpedance measurements in this study: R_0, R_I, R_∞, $R_0 - R_\infty$, f_c, C_m, α, and fitting error.

Figure 2. Illustrative example of the Cole-Cole plot of one patient.



Heart Rate Variability Measurements

To compute heart rate variability parameters, 4 minutes of clean ECG data were extracted from each 5-minute recording of ECG acquired simultaneously with transthoracic bioimpedance measurements. Noise and motion artifacts were removed from the ECG signals using a bandpass filter (0.05 Hz-40 Hz). The R peaks were detected using a validated algorithm [32,33]. Segments were visually inspected to ensure correct heartbeat detection. The R-R intervals were computed, and the time series was transformed to an evenly sampled signal (sampling frequency: 4 Hz) using cubic-spline interpolation. Mean heart rate was computed as a parameter. A 256-point Blackman window was applied to each segment, and the fast Fourier transform was calculated for each windowed segment. Finally, the power spectra of the segments were averaged.

We computed the indices of low frequencies of heart rate variability (low-frequency components of heart rate variability: 0.045 Hz to 0.15 Hz), high frequencies of heart rate variability (high-frequency components of heart rate variability: 0.15 Hz to 0.4 Hz), and the indices normalized to the total power of heart rate variability (normalized low-frequency components of heart rate variability, normalized high-frequency components of heart rate variability) [16]. Indices obtained from the low-frequency components of heart rate variability represent sympathetic control, and indices from the high-frequency components of heart rate variability power represent parasympathetic control. Furthermore, we derived 2 more parameters of heart rate variability based on principal dynamic modes, a nonlinear method designed to extract only the principal dynamic components of the signal via eigendecomposition [18]. The principal dynamic mode technique is able to separate sympathetic (principal dynamic mode index of sympathetic function) and parasympathetic (principal dynamic mode index of parasympathetic function) dynamics from heart rate variability [17,18]. Table 1 includes the parameters computed in this study.

Table 1. Transthoracic bioimpedance and heart rate variability parameters computed in this study.

Parameter	Description
Transthoracic bioimpedance	
R_0	Model resistance of biological tissue—extracellular fluid or resistance when *f*=0
R_I	Model resistance of biological tissue—intracellular fluid
R_∞	Resistance of biological tissue when *f*=∞
$R_0 - R_\infty$	Range of *x* values in Cole-Cole plot
f_c	Characteristic frequency, ie, frequency corresponding to the upper point of Cole-Cole plot circle
C_m	Cell membrane capacitance
α	Exponent of the model representing tissue heterogeneity
Fitting error	Sum of squared error of the optimal Cole-Cole plot model
Heart rate variability	
LF[a] HRV[b]	Low-frequency components of heart rate variability power
Normalized LF HRV	Normalized low-frequency components of heart rate variability power
HF[c] HRV	High-frequency components of heart rate variability power
Normalized HF HRV	Normalized high-frequency components of heart rate variability power
PDMI sympathetic[d]	Sympathetic function heart rate variability dynamics
PDMI parasympathetic[e]	Parasympathetic function heart rate variability dynamics

[a] LF: low-frequency.

[b] HRV: heart rate variability.

[c] HF: high-frequency.

[d] PDMI sympathetic: principal dynamic mode index of sympathetic function.

[e] PDMI parasympathetic: principal dynamic mode index of parasympathetic function.

Statistical Analysis and Machine Learning Classification

The normality of each parameter was tested using the Kolmogorov-Smirnov test [34-36]. We tested the differences in the parameters of transthoracic bioimpedance and heart rate variability between control, baseline, and discharge groups, using one-way ANOVA with Tukey posthoc for normally distributed data and the Dunn test for nonnormally distributed data (MATLAB, version 9.6; The Mathworks). The Dunn test is a nonparametric analog to multiple pairwise *t* tests following rejection of an ANOVA null hypothesis [37]. A *P* value<.05 was considered statistically significant for ANOVA and Dunn tests.

Statistical analysis of the differences between groups provides insight into the suitability of the measures of transthoracic bioimpedance and heart rate variability to detect fluid accumulation, which is used as an indication of heart failure exacerbation. However, measurement results have nonlinear characteristics and cannot be completely described with linear statistical methods. Hence, we used nonlinear methods such as machine learning to examine 15 features derived from transthoracic bioimpedance and heart rate variability for classification between groups (control, baseline, and discharge). Furthermore, participants in the discharge group were partially recovered, so they could be considered similar to the participants in control group. We tested the feasibility of classifying participants without fluid accumulation in the lung, termed *patients without fluid* (control and discharge groups) and participants with increased fluid in the lungs, termed *patients with fluid* (baseline group)

For these classification analyses, 3 algorithms were used: support vector machines [38], a *k*-nearest neighbor classifier (*k*=1) [39], and decision trees [40]. Cubic, quadratic, and Gaussian (C=1, γ=2.6) kernels were used for the support vector machine algorithm. All combinations of the 15 parameters were tested with the abovementioned classifiers to discriminate control/baseline/discharge groups, and patients with fluid/patients without fluid conditions. To compensate for the imbalance of the classes, the prior probabilities of the classes were set to be uniform in the training process. Leave-one-subject-out cross-validation was used to evaluate the performance of the machine learning models to prevent overfitting. Accuracy was computed as the number of correct classifications, divided by the total number of classifications performed, which corresponds to the number of participants in this case (N=60). Furthermore, the confusion matrices of the best models were obtained for a more detailed analysis.

Results

We approached 90 patients with heart failure who were eligible, and 43 were enrolled in this study. Out of the 43 enrolled participants, we were able to collect data from 28 participants with heart failure; 23 were included in the baseline group (mean 72, SD 10.7 years), and 17 were included in the discharge group (mean 72.4, SD 9 years). Only 12 participants were included in both baseline and discharge groups. There were several reasons for the lower number of participants in the discharge group: (1) in some cases, the recordings were of poor quality (n=14); (2) some participants (n=5) were lost to follow-up (ie, owing to a late night or weekend discharge); (3) some participants (n=7) could not provide the second recording owing to illness or refusal.

We enrolled 50 participants without acute decompensated heart failure (mean 71.5, SD 8.5 years) in the control group. Of the recordings taken on the 50 enrolled participants 32 of them were usable. It should be noted that participants from both groups were well matched with respect to age.

The demographic and medical characteristics of study participants are shown in Table 2. There were no significant differences in the demographic characteristics of the control group compared with those of participants with heart failure (age: *P*=.70; sex: *P*=.70; race: *P*=.52). Participants with acute decompensated heart failure were more likely to have a history of cardiovascular disease risk factors (coronary artery disease: *P*=.04; myocardial infarction: *P*=.03), prior heart failure (*P*<.001), and atrial fibrillation (*P*<.001). All transthoracic bioimpedance and heart rate variability parameters were found to be normally distributed, except for low-frequency components of heart rate variability and high-frequency components of heart rate variability.

Table 2. Demographic and clinical characteristics.

Characteristic	Control (n=32)	Acute decompensated heart failure (n=28)	*P* value
Age, mean (SD)	71.5 (8.5)	72.4 (10.3)	.70
Sex, n (%)			
Male	19 (59)	18 (64)	.70
Female	13 (41)	10 (36)	
Race, n (%)			.52
White	29 (91)	26 (93)	
Black	1 (3)	2 (7)	
Other[a]	2 (6)	0 (0)	
Chest circumference (cm), mean (SD)	105.4 (14.1)	107.8 (13.1)	.57
BMI (kg/m^2), mean (SD)	27.7 (5.1)	29.3 (6.6)	.28
Medical history, n (%)			
Myocardial infarction	3 (9)	9 (32)	.03
Coronary artery disease	7 (22)	13 (46)	.04
Hypertension	20 (63)	23 (82)	.09
Stroke/transient ischemic attack	2 (6)	3 (11)	.50
Previous diagnosis of heart failure	1 (3)	17 (61)	<.001
Diabetes	6 (19)	7 (25)	.56
Dyslipidemia	23 (72)	20 (71)	.97
Chronic lung disease	4 (13)	9 (32)	.06
Renal failure	2 (6)	3 (11)	.53
Atrial fibrillation	0 (0)	13 (46)	<.001
Vital signs and serum laboratories, mean (SD)			
Heart rate (beats/min)	75.4 (13.2)	84.4 (25.1)	.09
Systolic blood pressure	141.1 (28.6)	146.1 (28.7)	.51
Diastolic blood pressure	79.9 (13.1)	81.3 (17.3)	.72
Respiratory rate (breaths/min)	18.3 (2.2)	20.7 (2.8)	<.001
Sodium (mg/dL)	138.8 (2.4)	138.9 (2.8)	.97
Potassium (mg/dL)	4.1 (0.4)	4.1 (0.8)	.79
Glucose (mg/dL)	121.6 (45.4)	143.5 (80.9)	.20
Blood urea nitrogen (mg/dL)	19.2 (6.7)	26.3 (18.9)	.06
Creatinine (mg/dL)	1.1 (0.4)	1.3 (0.6)	.15
B-type natriuretic peptide[b]	112.0 (76.2)	1013.9 (1004.5)	.14
Troponin[b]	0.2 (1.0)	0.2 (0.9)	.96
INR	1.3 (0.7)	1.4 (0.5)	.95
Medication use, n (%)			
Beta blocker	2 (6)	2 (7)	.89
Angiotensin converting enzyme inhibitor	5 (16)	1 (4)	.12
Diuretic	2 (6)	3 (11)	.53
Statin	6 (19)	3 (11)	.38
Oral anticoagulant	2 (6)	0 (0)	.18

[a] Asian; American Indian, or Alaska Native; Native Hawaiian or other Pacific Islander.

[b] Data for the control group is for 6 patients only.

We compared values of 15 parameters from transthoracic bioimpedance and heart rate variability measurements between participants in control, baseline, and discharge groups (Table 3). As can be noted, values of 2 parameters, R_0 and $R_0 – R_\infty$, for the baseline group had statistically significantly lower values than those for the control group, with P=.006 and P=.001, respectively. Even though values of these 2 parameters for the discharge group were higher than those for the baseline group, there were no statistically significant differences (R_0: P=.99; $R_0 – R_\infty$: P=.57). Possible reasons could be the lower number of participants in the discharge group (discharge: n=17; baseline: n=23), and one possibility is that, at the time of discharge, some of the participants still had excess fluid in their lungs. The parameter α for the baseline group had statistically significantly higher values than those of the control group (P=.003),

Table 3. Values of transthoracic bioimpedance and heart rate variability parameters.

Parameters	Control (n=32), mean (SD)	Baseline (n=23), mean (SD)	*P* value	Discharge (n=17), mean (SD)	*P* value
Transthoracic bioimpedance					
R_0 (Ω)	38.1 (10.8)	26.5 (12.8)[a]	.006	34.2 (17.4)	.99
R_I (Ω)	52.0 (17.0)	52.0 (24.7)	>.999	54.3 (23.3)	>.999
C_m (F)	$4.08\cdot10^{-8}$ ($2.96\cdot10^{-8}$)	$4.60\cdot10^{-8}$ ($1.71\cdot10^{-8}$)	>.999	$4.42\cdot10^{-8}$ ($1.85\cdot10^{-8}$)	>.999
α	0.609 (0.0881)	0.716 (0.121)[a]	.003	0.646 (0.144)	.87
f_c (Hz)	$6.11\cdot10^{-4}$ ($3.45\cdot10^{-4}$)	5.34·10–4 ($1.51\cdot10^{-4}$)	.83	$5.07\cdot10^{-4}$ ($1.72\cdot10^{-4}$)	.56
Fitting error (Hz)	334 (669)	232 (389)	.51	347 (374)	.35
R_∞ (Ω)	21.5 (6.0)	17.0 (7.5)	.08	20.3 (9.1)	>.999
$R_0– R_\infty$ (Ω)	16.6 (6.1)	9.54 (6.0)[a]	.001	13.9 (8.8)	.57
Heart rate variability					
LF[b] HRV[c]	3.5 (4.2)	19.3 (43.4)	.06	19.2 (51.3)	.09
Normalized LF HRV	7.4 (14.4)	32.9 (55.7)[a]	.02	34.6 (57.0)[a]	.01
HF[d] HRV	0.225 (0.134)	0.178 (0.092)	.38	0.127 (0.085)[a]	.01
Normalized HF HRV	0.255 (0.154)	0.391 (0.134)[a]	.003	0.371 (0.129)[a]	.02
PDMI sympathetic[e]	11.8 (5.52)	17.2 (12.4)	.06	15.3 (5.98)	.52
PDMI parasympathetic[f]	13.2 (5.47)	17.1 (10.4)	.20	17.9 (7.56)	.14
Mean heart rate	72.3 (11.9)	74.1 (18.0)	>.999	74.7 (15.9)	>.999

[a] Denotes a statistically significant difference with respect to control group.

[b] LF: low-frequency.

[c] HRV: heart rate variability.

[d] HF: high-frequency.

[e] PDMI sympathetic: principal dynamic mode index of sympathetic function.

[f] PDMI parasympathetic: principal dynamic mode index of parasympathetic function.

As for the heart rate variability parameters, for the baseline and discharge groups, high-frequency components of heart rate variability (baseline: P=.02; discharge: P=.13) and normalized high-frequency components of heart rate variability (baseline: P=.003, discharge: P=.02) had significantly higher values than those for the control group. Normalized low-frequency components of heart rate variability exhibited a significantly lower value in the discharge group, when compared to those in the control group (P=.01). None of the other parameters of heart rate variability exhibited significant differences between groups.

Tables 4 and 5 include the results for the machine learning classification analysis. First, only transthoracic bioimpedance parameters were used for control/baseline/discharge classification and with fluid/without fluid classification. The most accurate model for transthoracic bioimpedance parameters only for classification of control/baseline/discharge was the Gaussian support vector machine, which reached an overall accuracy of 68% using R_0, R_I, and α. For patients without fluid/patients with fluid classification using only transthoracic bioimpedance parameters, cubic support vector machine and gaussian support vector machine models achieved 82% accuracy, although the cubic support vector machine required less parameters (R_0, α, fitting error, R_∞). Incorporating the heart rate variability parameters improved the accuracy of most models. The quadratic support vector machine model achieved 75% accuracy using 8 parameters (C_m, f_c, fitting error, R_∞, $R_0 – R_\infty$, normalized low-frequency components of heart rate variability, normalized high-frequency components of heart rate variability, mean heart rate). As for patients without fluid/patients with fluid classification, the overall best model was the cubic support vector machine, which achieved an accuracy of 92% using 6 parameters (R_0, R_I, C_m, low-frequency components of heart rate variability, principal dynamic mode index of parasympathetic function, mean heart rate).

Table 6 shows the confusion matrix for the most accurate model for control/baseline/discharge classification (quadratic support vector machine), and Table 7 shows the confusion matrix for the most accurate model for patients without fluid/patients with fluid classification (cubic support vector machine). In control/baseline/discharge classification, the control and baseline groups were correctly classified 78% and 83%, respectively. However, the discharge group was accurately classified only in 59% of the cases. It is worth highlighting that this group was misclassified 29% of the time as the control group. In the patients without fluid/patients with fluid classification, the patients without fluid condition (control and discharge groups) were classified correctly 96% of the time, and patients with fluid (baseline group) condition was correctly classified in 82% of the time.

Table 4. Highest accuracy and parameters included for control/baseline/discharge classification in each machine learning algorithm.

Type	Cubic SVM[a]	Quadratic SVM	Gaussian SVM	*k*-Nearest neighbor	Decision tree
Transthoracic bioimpedance					
Accuracy, %	63	61	68	67	72
Parameters					
R_0	x	x	x	x	x
R_I	x		x		
C_m		x		x	
α			x	x	x
f_c				x	x
Fitting error	x	x			
R_∞		x		x	
$R_0 - R_\infty$					x
Heart rate variability					
Accuracy, %	58	63	56	57	53
Parameters					
LF[b] HRV[c]	x	x			
Normalized LF HRV	x	x	x	x	x
HF[d] HRV	x	x		x	x
Normalized HF HRV		x			x
PDMI sympathetic[e]		x		x	
PDMI parasympathetic[f]	x	x	x	x	
Mean heart rate		x		x	x
Transthoracic bioimpedance and heart rate variability					
Accuracy, %	74	75	68	74	72
Parameters					
R_0	x	x	x		x
R_I	x	x	x		
C_m	x			x	
α	x		x	x	x
f_c		x			x
Fitting error	x	x		x	
R_∞		x			
$R_0 - R_\infty$		x			x
LF HRV	x				
Normalized LF HRV					
HF HRV	x	x		x	
Normalized HF HRV		x			
PDMI sympathetic				x	
PDMI parasympathetic					
Mean heart rate	x	x		x	

[a] SVM: support vector machine.

[b] LF: low-frequency.

[c] HRV: heart rate variability.

[d] HF: high-frequency.

[e] PDMI sympathetic: principal dynamic mode index of sympathetic function.

[f] PDMI parasympathetic: principal dynamic mode index of parasympathetic function.

Table 5. Highest accuracy and parameters included for patients without fluid/patients with fluid classification on each machine learning algorithm

Type	Cubic SVM	Quadratic SVM	Gaussian SVM	*k*-Nearest neighbor	Decision tree
Transthoracic bioimpedance					
Accuracy, %	82	75	82	78	79
Parameters					
R_0	x	x	x	x	
R_I					
C_m			x	x	
α	x				
f_c			x		x
Fitting error	x		x		
R_∞	x	x	x	x	x
$R_0 - R_\infty$					x
Heart rate variability					
Accuracy, %	75	76	75	71	72
Parameters					
LF[b] HRV[c]		x			
Normalized LF HRV	x		x		x
HF[d] HRV		x	x	x	
Normalized HF HRV	x				x
PDMI sympathetic[e]	x		x	x	
PDMI parasympathetic[f]		x			
Mean heart rate		x	x	x	
Transthoracic bioimpedance and heart rate variability					
Accuracy, %	92	88	83	85	81
Parameters					
R_0	x		x	x	
R_I	x				x
C_m	x	x		x	
α					
f_c		x			
Fitting error			x	x	
R_∞		x	x		x
$R_0 - R_\infty$		x			
LF HRV	x			x	
Normalized LF HRV				x	x
HF HRV		x	x		
Normalized HF HRV		x		x	
PDMI sympathetic					
PDMI parasympathetic	x			x	
Mean heart rate	x	x		x	

[a]SVM: support vector machine.

[b]LF: low-frequency.

[c]HRV: heart rate variability.

[d]HF: high-frequency.

[e]PDMI sympathetic: principal dynamic mode index of sympathetic function.

[f]PDMI parasympathetic: principal dynamic mode index of parasympathetic function.

Table 6. Confusion matrix for quadratic support vector machine—the most accurate model for control/baseline/discharge classification.

Actual	Predicted, %		
	Control	Baseline	Discharge
Control	78.1	6.3	15.6
Baseline	13.0	82.6	4.3
Discharge	29.4	11.8	58.8

Table 7. Confusion matrix for cubic support vector machine—the most accurate model for patients without fluid/patients with fluid classification.

Actual	Predicted, %	
	Patients with fluid	Patients without fluid
Patients with fluid	82.6	17.4
Patients without fluid	4.1	95.9

Discussion

Principal Findings

In this prospective observational study, we successfully trained machine learning models to classify participants with and without fluid accumulation using parameters obtained with a fluid accumulation vest, specifically transthoracic bioimpedance and heart rate variability parameters. We achieved a cross-validation accuracy of 92% using a cubic support vector machine model. The transthoracic bioimpedance parameters that contributed to this accuracy were related to intra- and extracellular fluid, whereas the heart rate variability parameters were mostly related to sympathetic activation. Our results suggest that the transthoracic bioimpedance and heart rate variability signals acquired with a wearable vest with carbon black–PDMS dry electrodes are suitable for detecting fluid accumulation and can potentially help with prediction and management of clinical worsening in heart failure patients.

In the past, transthoracic bioimpedance has been used for lung fluid abnormality detection [14,15]. In this study, we aimed to test the feasibility of a more accurate detection method for fluid accumulation by combining transthoracic bioimpedance and heart rate variability, given the autonomic dysregulation observed in heart failure patients. We used fluid accumulation vests to capture transthoracic bioimpedance and heart rate variability simultaneously. The accuracy of lung fluid abnormality detection using both transthoracic bioimpedance and heart rate variability was 92%, which is considerably higher than the maximum accuracy achieved using either only transthoracic bioimpedance (82%) or only heart rate variability (76%). Although the maximum accuracy of transthoracic bioimpedance was higher than that of heart rate variability, both contributed to the even higher accuracy of the model that combined them. We hypothesized that acute decompensated heart failure participants at the time of admission (baseline group) would have significantly lower resistances than participants in the control and acute decompensated heart failure discharge groups. Our results showed statistically significantly lower R_0 and $R_0 - R_\infty$ resistances in the baseline group (mean 27 Ω, SD 13 Ω; mean 10 Ω, SD 6 Ω, respectively) than those in the control group (mean 38, SD 11 Ω; mean 17, SD 6.1 Ω, respectively), with *P* values of .006 and .001, respectively. This suggests that participants in the baseline group had higher fluid volumes retained in the lungs than participants in the control group did. Moreover, the same parameters R_0 and $R_0 - R_\infty$ for discharge participants (mean 34, SD 17 Ω; mean 14, SD 9 Ω, respectively) were higher than those for the baseline participants. However, this difference did not reach statistical significance (*P*=.99; *P*=.57, respectively). Since predischarge assessments could not be performed in all participants, our findings may be attributable to a relatively small sample size. Alternatively, significant variability in the amount of intrathoracic fluid remaining before discharge may also explain our findings.

Bioimpedance is a proven biomarker of acute decompensated heart failure. Our group previously performed a clinical study of 106 hospitalized patients discharged after an admission for acute decompensated heart failure. Participants were sent home with a fluid accumulation vests and we determined that it was feasible to measure transthoracic bioimpedance on a daily basis [12]. We also demonstrated that a predictive algorithm analyzing daily bioimpedance measures achieved good performance for predicting recurrent acute decompensated heart failure [12]. Lindholm et al [22] also performed a longitudinal study including over 500,000 participants and determined that leg bioimpedance was inversely correlated with new-onset heart failure and that by combining the leg bioimpedance with clinical parameters such as age, sex, and history of myocardial infarction, accurate prediction of heart failure could be achieved.

In another study [23] on participants with congenital heart disease, bioelectrical impedance correlated with heart failure severity. In contrast to these prior studies, we sought to evaluate the performance of intrathoracic bioimpedance measured using a novel dry electrode for detecting acute decompensated heart failure. We observed that participants hospitalized with acute decompensated heart failure had lower values of intrathoracic resistance due to higher intrathoracic fluid volume.

As for the heart rate variability, high-frequency components of heart rate variability (at admission: *P*=.02; at discharge: *P*=.13) and normalized high-frequency components of heart rate variability parameters (at admission: *P*=.003, at discharge: *P*=.02) were significantly higher in acute decompensated heart failure participants when compared to control participants without acute decompensated heart failure. This is possibly the result of more labored breathing exhibited by the participants with acute decompensated heart failure [41]. Although not statistically significant, we observed overall higher sympathetic activation in the acute decompensated heart failure participants, as evidenced by higher low-frequency components of heart rate variability (control: mean 3.5, SD 4.2; at admission: mean 19.3, SD 43.4, *P*=.06; at discharge: mean 19.2, SD 51.3 *P*=.09). The activation of the sympathetic nervous system is a known countermeasure of the body aiming to restore cardiac output in the case of heart failure [42]. Conversely, acute decompensated heart failure participants exhibited a significantly lower normalized low-frequency components of heart rate variability but only in the discharge group. This was produced by the highly elevated parasympathetic tone (high-frequency components of heart rate variability), which affected the computation of the normalized indices (normalized low-frequency components of heart rate variability and normalized high-frequency components of heart rate variability). These results corroborate the alteration of the autonomic nervous functions produced by acute decompensated heart failure and explain why the parameters of the autonomic function are valuable for detecting acute decompensated heart failure and its subsequent consequences.

In the machine learning classifications, R_0 was consistently chosen in most of the optimal models and was present in both the most accurate models for both classifications tested in this study (control/baseline/discharge and patients without fluid/patients with fluid classification). This is in agreement with the between-group statistical differences, in which this parameter was found to be the most sensitive to heart failure. Using the set of transthoracic bioimpedance parameters only, machine learning models were able to provide moderate classification accuracy for both types of classification: an accuracy of 68% was found for 3-class classification (control/baseline/discharge classification) model, and an accuracy of 82% was found for 2-class models (patients without fluid/patients with fluid classification), which are acceptable performances, considering that the bottom line accuracy for 3- and 2-class models are 33% and 50%, respectively. However, adding heart rate variability parameters (the model was trained with transthoracic bioimpedance and heart rate variability parameters together) further increased the accuracy of the models. The control/baseline/discharge classification, with 75% accuracy, was acceptable. Furthermore, 92% accuracy for classifying of patients without fluid and patients with fluid suggested the feasibility of such an algorithm to potentially detect the healthy condition (control group) or recovering (at least partially) of a patient from excess fluid accumulation. This model used parameters from transthoracic bioimpedance (extracellular resistance, intracellular resistance, cell membrane capacitance), as well as parameters from heart rate variability (low-frequency components, principal dynamic mode index of parasympathetic function, mean heart rate). The transthoracic bioimpedance parameters that were included are related to intra- and extracellular fluid, whereas the heart rate variability parameters are mostly related to the sympathetic activation. This finding is useful in developing in-home diagnostic tools that can detect or predict fluid accumulation in heart failure participants.

Statistical analysis and machine learning analysis showed similar results for a reduced set of features. For instance, extracellular resistance and low-frequency components of heart rate variability exhibited significant differences between non–heart failure (control) and heart failure groups (baseline and discharge), and these features were present in the most accurate model for fluid accumulation detection. However, other features including intracellular resistance, cell membrane capacitance, principal dynamic mode index of parasympathetic function, and mean heart rate did not exhibit significant differences between groups but were relevant for improving accuracy of the machine learning algorithms.

Limitations

As for the limitations of the study, many recordings were not usable, mostly in the acute decompensated heart failure group. This is related to technical issues with the fluid accumulation vests, which can be partially attributed to the carbon black–PDMS electrodes. From the 28 participants with acute decompensated heart failure, we obtained reliable measures from only 23 participants at baseline and from 17 participants at discharge. We obtained data from both baseline and discharge for only 12 participants. Even in the control group, we collected usable data from only 32 out of the 50 participants. In some instances, applying a layer of hydrating lotion helped with data collection. This limitation could potentially diminish the clinical use of the device and must be addressed in the near future. A more robust hardware design, tailored to match the impedance of the carbon black–PDMS electrodes, is a potential improvement. Configurations that enable collection of transthoracic bioimpedance data from several locations on the thorax could help the quality and usability of the data, as accumulation of fluid does not occur always in the same location. Furthermore, given the limited data set, we have reported leave-one-subject-out cross-validation accuracy, and the results cannot be interpreted as conclusive concerning the efficacy of the transthoracic bioimpedance device and features derived from it. Instead, the results can be interpreted as promising, based on the validation of the transthoracic bioimpedance and its associated features and machine learning. A larger testing data set is required for further evaluation of transthoracic bioimpedance to allow for more definite conclusions about its efficacy.

There are several potential clinical applications of transthoracic bioimpedance measurements in patients with heart failure. Wearable technologies such as fluid accumulation vests could allow for rapid point-of-care diagnostics that could be used in the emergency setting to help identify heart failure decompensation. In addition, fluid accumulation vest measurements in different clinical states such as decompensated heart failure, predischarge, and in outpatient setting, could be used to establish a profile for a given patient that could improve diagnostic certainty and guide treatment. Moreover, triaging medical severity is a necessary and time-consuming step of the patient care process, but this is often difficult due to limitations in both the number of available medical personnel and individual provider time.

The device and algorithm in this study can be used in a longitudinal study with patients with heart failure, extending monitoring into the home. The system could be used to monitor a patient's fluid accumulation daily and generate early warnings of heart failure decompensation, provide guidance on therapeutic changes to improve quality of life, and reduce heart failure readmissions. Alternatively, the system can be used to monitor either the discharge readiness of a patient from the hospital or the home treatment regime effectiveness on the patient. Wearable sensors such as the fluid accumulation vest can potentially provide an ideal avenue for patient monitoring over time, allowing for rapid action in response to acute decompensation. Garments integrating vital sign sensors have been utilized in acute medical settings to monitor patients with high medical risk profiles [43]. In addition, wearable sensor-based systems for vital sign monitoring are well-received by both patients and nursing staff with regards to usability, further highlighting their potential role in clinical implementation [44].

Conclusions

The main goal of this study was to evaluate the performance of biologically relevant parameters measured by a fluid accumulation vests with carbon black–PDMS dry electrodes. In our clinical study (SHIELD), transthoracic bioimpedance and heart rate variability parameters were considered for statistical analysis and discrimination between patients with nonacute decompensated heart failure and acute decompensated heart failure. As expected, our results show that among the 15 parameters, 2 (extracellular resistance and intracellular-extracellular difference in resistance) showed statistically significantly lower values (*P*=.006; *P*=.001, respectively), and 3 (tissue heterogeneity exponent, high-frequency components of heart rate variability, and normalized high-frequency components of heart rate variability) showed statistically significantly higher values (*P*=.01, *P*=.02, *P*=.003, respectively) for participants with acute decompensated heart failure at hospital admission than those for participants in the control group. A significant difference in the sympathetic control (assessed with the normalized low-frequency components, *P*=.01) was observed between acute decompensated heart failure participants at the time of discharge and the control participants. Transthoracic bioimpedance and heart rate variability exhibited promising results for classifying participants with excess intrathoracic fluid versus those with normal intrathoracic fluid. Further clinical studies will be undertaken to refine our approach and determine the optimal application of this monitoring technology in acute medical settings.

Acknowledgments

This work was supported by the Smart and Connect Health of the National Science Foundation grants: 1522087, 1522084, and 1522052.

Conflicts of Interest

DDM has received honorary, speaking/consulting fee or grants from Flexcon, Rose Consulting, Bristol-Myers Squibb, Pfizer, Boston Biomedical Associates, Samsung, Phillips, Mobile Sense, Care Evolution, Flexcon Boehringer Ingelheim, Biotronik, Otsuka Pharmaceuticals, and Sanofi.

References

1. Ambrosy AP, Gheorghiade M, Chioncel O, Mentz RJ, Butler J. Global Perspectives in Hospitalized Heart Failure: Regional and Ethnic Variation in Patient Characteristics, Management, and Outcomes. Curr Heart Fail Rep 2014 Sep 25;11(4):416-427. [doi: 10.1007/s11897-014-0221-9]
2. Anter E, Jessup M, Callans DJ. Atrial fibrillation and heart failure: treatment considerations for a dual epidemic. Circulation 2009 May 12;119(18):2516-2525. [doi: 10.1161/CIRCULATIONAHA.108.821306] [Medline: 19433768]
3. Fang J, Mensah GA, Croft JB, Keenan NL. Heart failure-related hospitalization in the U.S., 1979 to 2004. J Am Coll Cardiol 2008 Aug 05;52(6):428-434 [FREE Full text] [doi: 10.1016/j.jacc.2008.03.061] [Medline: 18672162]
4. McMurray J, Petrie M, Murdoch D, Davie A. Clinical epidemiology of heart failure: public and private health burden. European heart journal 1998;19:9-16.
5. Lloyd-Jones DM, Wang TJ, Leip EP, Larson MG, Levy D, Vasan RS, et al. Lifetime risk for development of atrial fibrillation: the Framingham Heart Study. Circulation 2004 Aug 31;110(9):1042-1046. [doi: 10.1161/01.CIR.0000140263.20897.42] [Medline: 15313941]

6. Stevenson WG, Stevenson LW, Middlekauff HR, Fonarow GC, Hamilton MA, Woo MA, et al. Improving survival for patients with atrial fibrillation and advanced heart failure. J Am Coll Cardiol 1996 Nov 15;28(6):1458-1463 [FREE Full text] [doi: 10.1016/s0735-1097(96)00358-0] [Medline: 8917258]
7. Lloyd-Jones D, Adams RJ, Brown TM, Carnethon M, Dai S, De SG, et al. Executive summary: heart disease and stroke statistics--2010 update: a report from the American Heart Association. Circulation 2010 Feb 23;121(7):948-954 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.109.192666] [Medline: 20177011]
8. Vasan RS, Levy D. Defining diastolic heart failure: a call for standardized diagnostic criteria. Circulation 2000 May 02;101(17):2118-2121. [doi: 10.1161/01.cir.101.17.2118] [Medline: 10790356]
9. Boehmer J, Hariharan R, Devecchi F, Smith A, Molon G, Capucci A, et al. A Multisensor Algorithm Predicts Heart Failure Events in Patients With Implanted Devices: Results From the MultiSENSE Study. In: JACC Heart Fail. JACC: Heart Failure JACC: Heart Failure; Mar 2017:216-225.
10. Reiter H, Muehlsteff J, Sipilä A. Medical application and clinical validation for reliable and trustworthy physiological monitoring using functional textiles: experience from the HeartCycle and MyHeart project. Conf Proc IEEE Eng Med Biol Soc 2011;2011:3270-3273. [doi: 10.1109/IEMBS.2011.6090888] [Medline: 22255037]
11. Baumgartner RN, Chumlea WC, Roche AF. Estimation of body composition from bioelectric impedance of body segments. Am J Clin Nutr 1989 Aug;50(2):221-226. [doi: 10.1093/ajcn/50.2.221] [Medline: 2756908]
12. Darling CE, Dovancescu S, Saczynski JS, Riistama J, Sert Kuniyoshi F, Rock J, et al. Bioimpedance-Based Heart Failure Deterioration Prediction Using a Prototype Fluid Accumulation Vest-Mobile Phone Dyad: An Observational Study. JMIR Cardio 2017 Mar 13;1(1):e1. [doi: 10.2196/cardio.6057]
13. Seppä V, Viik J, Hyttinen J. Assessment of pulmonary flow using impedance pneumography. IEEE Trans Biomed Eng 2010 Sep;57(9):2277-2285. [doi: 10.1109/TBME.2010.2051668] [Medline: 20542759]
14. Albert NM. Bioimpedance to prevent heart failure hospitalization. Curr Heart Fail Rep 2006 Sep;3(3):136-142. [doi: 10.1007/s11897-006-0013-y] [Medline: 16914106]
15. Martindale JL, Wakai A, Collins SP, Levy PD, Diercks D, Hiestand BC, et al. Diagnosing Acute Heart Failure in the Emergency Department: A Systematic Review and Meta-analysis. Acad Emerg Med 2016 Feb 13;23(3):223-242. [doi: 10.1111/acem.12878]
16. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Eur Heart J 1996 Mar;17(3):354-381. [Medline: 8737210]
17. Zhong Y, Jan K, Ju KH, Chon KH. Quantifying cardiac sympathetic and parasympathetic nervous activities using principal dynamic modes analysis of heart rate variability. American Journal of Physiology-Heart and Circulatory Physiology 2006 Sep;291(3):H1475-H1483. [doi: 10.1152/ajpheart.00005.2006]
18. Zhong Y, Wang H, Ju KH, Jan K, Chon KH. Nonlinear analysis of the separate contributions of autonomic nervous systems to heart rate variability using principal dynamic modes. IEEE Trans Biomed Eng 2004 Feb;51(2):255-262. [doi: 10.1109/TBME.2003.820401] [Medline: 14765698]
19. Posada-Quintero H, Reljin N, Eaton-Robb C, Noh Y, Riistama J, Chon K. Analysis of Consistency of Transthoracic Bioimpedance Measurements Acquired with Dry Carbon Black PDMS Electrodes, Adhesive Electrodes, and Wet Textile Electrodes. Sensors 2018 May 26;18(6):1719. [doi: 10.3390/s18061719]
20. Reljin N, Posada-Quintero H, Noh Y, Robb C, Dimitrov T, Murphy L, et al. Preliminary results on transthoracic bioimpedance measurements with a variety of electrode materials. 2018 Presented at: IEEE EMBS International Conference on Biomedical Health Informatics (BHI); 2018; Las Vegas, Nevada, USA p. 62-65. [doi: 10.1109/bhi.2018.8333370]
21. Dovancescu S, Saczynski JS, Darling CE, Riistama J, Sert Kuniyoshi F, Meyer T, et al. Detecting Heart Failure Decompensation by Measuring Transthoracic Bioimpedance in the Outpatient Setting: Rationale and Design of the SENTINEL-HF Study. JMIR Res Protoc 2015 Oct 09;4(4):e121 [FREE Full text] [doi: 10.2196/resprot.4899] [Medline: 26453479]
22. Lindholm D, Fukaya E, Leeper NJ, Ingelsson E. Bioimpedance and New - Onset Heart Failure: A Longitudinal Study of >500 000 Individuals From the General Population. J Am Heart Assoc 2018 Jul 03;7(13). [doi: 10.1161/jaha.118.008970]
23. Sato M, Inai K, Shimizu M, Sugiyama H, Nakanishi T. Bioelectrical impedance analysis in the management of heart failure in adult patients with congenital heart disease. Congenital Heart Disease 2018 Oct 23;14(2):167-175. [doi: 10.1111/chd.12683]
24. Noh Y, Bales JR, Reyes BA, Molignano J, Clement AL, Pins GD, et al. Novel Conductive Carbon Black and Polydimethlysiloxane ECG Electrode: A Comparison with Commercial Electrodes in Fresh, Chlorinated, and Salt Water. Ann Biomed Eng 2016 Aug;44(8):2464-2479. [doi: 10.1007/s10439-015-1528-8] [Medline: 26769718]
25. Ayllón D, Gil-Pita R, Seoane F. Detection and Classification of Measurement Errors in Bioimpedance Spectroscopy. PLoS ONE 2016 Jun 30;11(6):e0156522. [doi: 10.1371/journal.pone.0156522]
26. Cole KS. PERMEABILITY AND IMPERMEABILITY OF CELL MEMBRANES FOR IONS. Cold Spring Harbor Symposia on Quantitative Biology 1940 Jan 01;8:110-122. [doi: 10.1101/sqb.1940.008.01.013]
27. Cornish BH, Thomas BJ, Ward LC. Improved prediction of extracellular and total body water using impedance loci generated by multiple frequency bioelectrical impedance analysis. Phys Med Biol 1993 Mar;38(3):337-346. [doi: 10.1088/0031-9155/38/3/001] [Medline: 8451277]

28. Buendia R, Gil-Pita R, Seoane F. Cole parameter estimation from the modulus of the electrical bioimpeadance for assessment of body composition. A full spectroscopy approach. Journal of Electrical Bioimpedance 2019;2(1):72-78. [doi: 10.5617/jeb.197]
29. Cole KS, Cole RH. Dispersion and Absorption in Dielectrics I. Alternating Current Characteristics. The Journal of Chemical Physics 1941 Apr;9(4):341-351. [doi: 10.1063/1.1750906]
30. Taubin G. Estimation of planar curves, surfaces, and nonplanar space curves defined by implicit equations with applications to edge and range image segmentation. IEEE Trans. Pattern Anal. Machine Intell 1991;13(11):1115-1138. [doi: 10.1109/34.103273]
31. Nelder JA, Mead R. A Simplex Method for Function Minimization. The Computer Journal 1965 Jan 01;7(4):308-313. [doi: 10.1093/comjnl/7.4.308]
32. Nygårds ME, Sörnmo L. Delineation of the QRS complex using the envelope of the e.c.g. Med Biol Eng Comput 1983 Sep;21(5):538-547. [doi: 10.1007/bf02442378] [Medline: 6633003]
33. Vidaurre C, Sander TH, Schlögl A. BioSig: the free and open source software library for biomedical signal processing. Comput Intell Neurosci 2011;2011:935364 [FREE Full text] [doi: 10.1155/2011/935364] [Medline: 21437227]
34. Massey FJ. The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association 1951 Mar;46(253):68-78. [doi: 10.1080/01621459.1951.10500769]
35. Miller LH. Table of Percentage Points of Kolmogorov Statistics. Journal of the American Statistical Association 1956 Mar;51(273):111-121. [doi: 10.1080/01621459.1956.10501314]
36. Marsaglia G, Tsang WW, Wang J. Evaluating Kolmogorov's Distribution. J. Stat. Soft 2003;8(18). [doi: 10.18637/jss.v008.i18]
37. Dinno A. Nonparametric Pairwise Multiple Comparisons in Independent Groups using Dunn's Test. The Stata Journal 2018 Nov 19;15(1):292-300 [FREE Full text] [doi: 10.1177/1536867X1501500117]
38. Shawe-Taylor J, Cristianini N. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press Cambridge 2000. [doi: 10.1017/cbo9780511801389]
39. Friedman JH, Bentley JL, Finkel RA. An Algorithm for Finding Best Matches in Logarithmic Expected Time. ACM Trans. Math. Softw 1977 Sep;3(3):209-226. [doi: 10.1145/355744.355745]
40. Breiman L. Classification and regression trees. Routledge 2017. [doi: 10.1201/9781315139470-8]
41. Sroufe LA. EFFECTS OF DEPTH AND RATE OF BREATHING ON HEART RATE AND HEART RATE VARIABILITY. Psychophysiology 1971 Sep;8(5):648-655. [doi: 10.1111/j.1469-8986.1971.tb00500.x]
42. Triposkiadis F, Karayannis G, Giamouzis G, Skoularigis J, Louridas G, Butler J. The Sympathetic Nervous System in Heart Failure. Journal of the American College of Cardiology 2009 Nov;54(19):1747-1762. [doi: 10.1016/j.jacc.2009.05.015]
43. Wearable Technology Applications in Healthcare: A Literature Review Internet. Wu M, PhD, Luo J, Contributors POJ of NI. 2019. URL: https://www.himss.org/resources/wearable-technology-applications-healthcare-literature-review [accessed 2019-05-15]
44. Claudio D, Velázquez MA, Bravo-Llerena W, Okudan GE, Freivalds A. Perceived Usefulness and Ease of Use of Wearable Sensor-Based Systems in Emergency Departments. IIE Transactions on Occupational Ergonomics and Human Factors 2015 Sep 11;3(3-4):177-187. [doi: 10.1080/21577323.2015.1040559]

Abbreviations

AF: atrial fibrillation
ECG: electrocardiography
ESRD: end-stage renal disease
HRV: heart rate variability
ICD: implantable cardioverter-defibrillator
PDMS: polydimethylsiloxane

Edited by G Eysenbach; submitted 13.03.20; peer-reviewed by E van der Velde, P Athilingam, Y Cai; comments to author 29.06.20; revised version received 10.07.20; accepted 23.07.20; published 27.08.20.

Please cite as:
Reljin N, Posada-Quintero HF, Eaton-Robb C, Binici S, Ensom E, Ding E, Hayes A, Riistama J, Darling C, McManus D, Chon KH
Machine Learning Model Based on Transthoracic Bioimpedance and Heart Rate Variability for Lung Fluid Accumulation Detection: Prospective Clinical Study
JMIR Med Inform 2020;8(8):e18715
URL: http://medinform.jmir.org/2020/8/e18715/
doi:10.2196/18715
PMID:32852277

Original Paper

Using Dual Neural Network Architecture to Detect the Risk of Dementia With Community Health Data: Algorithm Development and Validation Study

Xiao Shen[1], PhD; Guanjin Wang[2], PhD; Rick Yiu-Cho Kwan[3], PhD; Kup-Sze Choi[1], PhD

[1]Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

[2]Murdoch University, Western Australia, Australia

[3]School of Nursing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Corresponding Author:
Kup-Sze Choi, PhD
Centre for Smart Health
School of Nursing
The Hong Kong Polytechnic University
Hung Hom
Kowloon
Hong Kong
Phone: 852 3400 3214
Email: hskschoi@polyu.edu.hk

Abstract

Background: Recent studies have revealed lifestyle behavioral risk factors that can be modified to reduce the risk of dementia. As modification of lifestyle takes time, early identification of people with high dementia risk is important for timely intervention and support. As cognitive impairment is a diagnostic criterion of dementia, cognitive assessment tools are used in primary care to screen for clinically unevaluated cases. Among them, Mini-Mental State Examination (MMSE) is a very common instrument. However, MMSE is a questionnaire that is administered when symptoms of memory decline have occurred. Early administration at the asymptomatic stage and repeated measurements would lead to a practice effect that degrades the effectiveness of MMSE when it is used at later stages.

Objective: The aim of this study was to exploit machine learning techniques to assist health care professionals in detecting high-risk individuals by predicting the results of MMSE using elderly health data collected from community-based primary care services.

Methods: A health data set of 2299 samples was adopted in the study. The input data were divided into two groups of different characteristics (ie, client profile data and health assessment data). The predictive output was the result of two-class classification of the normal and high-risk cases that were defined based on MMSE. A dual neural network (DNN) model was proposed to obtain the latent representations of the two groups of input data separately, which were then concatenated for the two-class classification. Mean and *k*-nearest neighbor were used separately to tackle missing data, whereas a cost-sensitive learning (CSL) algorithm was proposed to deal with class imbalance. The performance of the DNN was evaluated by comparing it with that of conventional machine learning methods.

Results: A total of 16 predictive models were built using the elderly health data set. Among them, the proposed DNN with CSL outperformed in the detection of high-risk cases. The area under the receiver operating characteristic curve, average precision, sensitivity, and specificity reached 0.84, 0.88, 0.73, and 0.80, respectively.

Conclusions: The proposed method has the potential to serve as a tool to screen for elderly people with cognitive impairment and predict high-risk cases of dementia at the asymptomatic stage, providing health care professionals with early signals that can prompt suggestions for a follow-up or a detailed diagnosis.

(JMIR Med Inform 2020;8(8):e19870) doi:10.2196/19870

KEYWORDS

cognitive screening; dementia risk; dual neural network; predictive models; primary care

Introduction

Background

Dementia is a collective term referring to a group of diseases that cause a decline in cognitive function owing to brain cell damage. The symptoms include degradation in memory, communication, or reasoning ability, which can seriously interfere with activities of daily living [1]. Dementia is aging related. The incidence doubles with every increase in age of 5.9 years [2], and the number of people living with dementia worldwide is estimated to increase almost three times from 47 million in 2015 to 135 million in 2050 [3]. Thus, dementia is not only an overwhelming issue among elderly people and their families, but also an unprecedented burden on the health social care system and the society at large [4].

It has been reported that 35% of dementia cases are attributable to modifiable risk factors, such as hypertension, obesity, depression, and smoking [5], which concern physical, cognitive, and social inactivity and can be countered through lifestyle interventions [6]. As it takes time to modify lifestyle, early detection of people with high risk of dementia is important to enable timely diagnosis and intervention, which may halt or delay the development of dementia [7-9]. However, underdiagnosis of dementia at the early stage is common since the symptoms are subtle and the progression of cognitive impairment is insidious and cannot be easily observed by the person, family members, or even health care professionals [10,11].

Apart from cognitive symptoms, dementia risk is also associated with many noncognitive conditions (eg, cardiovascular conditions, nutrition, mobility, and depression) [12], which are routinely and vastly obtained from primary care settings, such as elderly community centers. While these routinely collected data provide good potential for the risk prediction of dementia, there is a lack of formulae in the literature to estimate the risk of dementia by using these data.

With the advance of artificial intelligence, machine learning offers a promising approach for the intelligent detection of dementia risk, particularly when the causal connections with risk factors remain unclear. A “school of methods” is to apply machine learning techniques to the data collected from population or community-based settings [13], such as the results of neuropsychology tests or physical examinations, to screen for people with high risk of dementia. While statistical analysis methods like logistic regression and Cox proportional hazard regression are commonly used for analyzing community-acquired elderly health data [14-16], various machine learning techniques have been employed. Among the techniques, supervised machine learning methods represent a majority [17-19], and they include naïve Bayes, decision tree (DT), random forest (RF), artificial neural network, and support vector machine (SVM), whereas their unsupervised counterparts have also been exploited for dementia risk prediction [20]. Nevertheless, missing data is a common problem with data collected from population or community-based settings. Data may be lost owing to noncompliance with appointment schedules, unwillingness to respond to specific questions, or inadvertence of interviewers. Discarding records with missing data and imputation with population means are conventionally used methods to deal with missing data [17,20,21]. Another issue of data analysis is class imbalance, where samples of the target (ie, high-risk cases) and nontarget (ie, normal cases) are disproportionate. When learning from imbalanced data, supervised machine learning algorithms are usually overwhelmed by majority class examples [22]. Other than simply reducing the size of sample-abundant data sets, oversampling methods, such as the synthetic minority oversampling technique, can be used to balance the data sets [23]. In addition, cost-sensitive learning (CSL) is an effective method to handle class imbalance, which is employed in machine learning algorithms to set the cost ratio according to prior class distributions [22,24-27].

In primary care, Mini-Mental State Examination (MMSE) [28] is a commonly used tool for screening cognitive impairment, which is a strong diagnostic criterion of dementia. However, MMSE is a questionnaire that is administered when symptoms of memory decline have occurred, and early administration at the asymptomatic stage or repeated measurements would lead to a practice effect [29] (the questions could be remembered), degrading the effectiveness of MMSE when used at later stages.

Objective

The aim of this study was to develop an alternative machine learning approach based on MMSE that can be used for screening cognitive impairment and the early detection of people with high dementia risk at the asymptomatic stage. The data adopted were collected through the delivery of elderly care services in the community, which included a wide range of health assessments. A dual neural network (DNN) model was proposed to learn latent representations by utilizing the health profiles of elderly clients and the results of health assessment questionnaires as two types of input features. The predictive output of the model was the result of two-class classification of normal versus high-risk cases, which were defined based on MMSE. Furthermore, the mean and *k*-nearest neighbor (KNN) imputation methods were used to deal with missing data, whereas CSL was used to deal with class imbalance. The performance of the DNN model was evaluated experimentally and compared with that of conventional machine learning algorithms. It was hypothesized that with CSL, the proposed DNN would outperform the algorithms under comparison.

The major contributions of the study are as follows: (1) the community-based health data that were collected for 10 years during elderly care services could provide useful information to meet the increasing emphasis on primary care development (the data set can be shared by request from a qualified researcher; the request should be directed to the corresponding author); (2) the study explored the use of the data set for predicting the risk of dementia, which is a new approach to the best of our knowledge; (3) as the data set has two different characteristics, innovative use of the contemporary DNN architecture was proposed as a new informatics method to fit the specific application scenario; (4) KNN and CSL were incorporated to solve the problems of missing data and class imbalance; and (5) extensive experiments were conducted for comparisons with classical algorithms that are commonly used in health care research to demonstrate outperformance and provide evidence to support the feasibility for dementia risk prediction.

Methods

Community Health Data

The data set used was obtained through mobile health care services offered in collaboration with elderly care centers run by local nongovernmental organizations. The health care services were provided for community-dwelling elderly people living in various districts of Hong Kong for free during the period from 2008 to 2018. The services included a wide range of elderly-specific health assessments, where follow-up appointments, workshops, and programs were arranged to promote health care and self-management. The data set included demographic information of elderly clients (eg, gender, age, marital status, type of residency, relationship with roommates, and social participation), bio-measurements (eg, body temperature, pulse rate, oxygen saturation, blood pressure, and waist-hip ratio), and medical history (eg, records of health problems or past diseases), as well as comprehensive information collected using a battery of health assessment questionnaires (ie, MMSE [28], brief pain inventory [BPI] [30], elderly mobility scale [EMS] [31], geriatric depression scale [GDS] [32], mini-nutrition assessment [MNA] [33], constipation questionnaire [CQ] [34], and a questionnaire based on the Roper-Logan-Tierney model of nursing [RLT] [35]), the records of gross oral hygiene and visual acuity assessments, and a survey of the favorite activities of the elderly clients. The health assessment questionnaires will be discussed further in the next section. The elderly health care services were provided by registered nurses and advanced practice nurses or student nurses under supervision, who were also responsible for recording the data while conducting health assessments or administering the questionnaires.

Health Assessments

The data set adopted contained the results of 10 health assessments, which are described below.

Mini-Mental State Examination

MMSE is a quick and reliable assessment of cognitive impairment in older adults. The use of MMSE as part of the process for diagnosing dementia is supported by a Cochrane review of 24,310 citations [36]. MMSE consists of six sets of questions focusing on the cognitive aspects of mental function. For example, elderly clients were asked to give the date of the day, perform arithmetic operations, and perform hand drawing. The assessment can be completed within 10 minutes. The maximum score is 30. A score between 24 and 30 indicates normal cognition, whereas a score below 24 suggests various degrees of impairment, with a lower score indicating greater impairment. In this study, two-class classification was adopted (ie, normal [score ≥24] and high risk [score <24]).

Brief Pain Inventory

The BPI is a questionnaire used to assess the severity of pain and its influence on elderly people [30]. The short-form BPI was administered, and it has nine items concerning the location and degree of pain in the last 24 hours, treatments being applied, and their influences on functioning, such as walking ability, mood, and sleep.

Elderly Mobility Scale

The EMS is a seven-item assessment tool used to evaluate the mobility of elderly people through functional tests (eg, transition between sitting and lying, gait, timed walk, and functional reach) [31]. The maximum score is 20. A score of 14 or above indicates normal mobility and independent living; a score between 10 and 13 indicates a borderline case; and a score below 10 indicates the necessity of assistance to perform activities of daily living.

Geriatric Depression Scale

The GDS is a measure of depression in older adults [32]. The short-form GDS was administered in the clinic. It contains 15 yes or no questions, each carrying one point, on the feelings about and attitudes toward various aspects of life. The maximum score is 15. A score greater than five indicates depression.

Mini-Nutrition Assessment

The MNA is a tool used to assess the nutritional status of older people [33]. It is administered in two steps. The short form of MNA (MNA-SF), which has six items with a maximum score of 14, is first used to detect signs of decline in ingestion. The questions concern appetite loss, weight loss, and psychological stress in the last 3 months; mobility; and BMI. A score of 11 or below indicates possible malnutrition, and follow-up with the full MNA is required in the second step. The full MNA consists of 12 items with a maximum score of 16, and it involves further details such as independent living, medication, ulcers, diet, feeding modes, and mid-arm and calf circumferences. The maximum total score of the MNA is 30, with a score below 17 indicating malnourishment.

Constipation Questionnaire

The CQ is used to assess the severity of functional constipation [34]. The questionnaire administered contains six items with questions concerning frequency of evacuation, difficulty to evacuate, incomplete evacuation, stool and abdominal symptoms, and medication.

RLT-Based Questionnaire

Based on the RLT [35], a questionnaire with 36 items was designed to assess the independence of older adults in 12 categories of activities of daily living, including maintaining a safe environment, communication, breathing, eating and drinking, elimination, washing and cleaning, controlling body temperature, mobilization, working and playing, sleeping, expressing sexuality, and dying. The results of the questionnaire can be used to determine the interventions required to enable elderly people to remain independent in activities of daily living.

Gross Oral Hygiene Assessment

The assessment tool consists of 20 items concerning various oral hygiene conditions of elderly clients, including teeth cleansing, tooth decay, tooth mobility, denture use, denture care, missing or remaining teeth, calculus, gum bleeding, and oral candidiasis, with which the corresponding tooth locations and symptoms are recoded.

Visual Acuity Assessment

Visual acuity of elderly clients was measured at the mobile clinic. The data collected included the distance at which measurement was performed, the visual aid used, and the results of measurements using the Snellen chart and the chart of the logarithm of the minimum angle of resolution (LogMAR chart).

Survey of Favorite Activities

The survey involves binary yes or no questions, each recording a favorite activity of the elderly clients. The questions cover a wide range of over 40 activities (eg, playing chess, watching television, listening to radio, fishing, hiking, calligraphy, dancing, and shopping).

Data Set

The data set contained the records of 2299 elderly clients, with one record per client. Each record had a total of 567 features that were the inputs of the models. The features originated from demographic data, bio-measurements, and medical history, as well as the data collected from the various health assessment questionnaires described above, except MMSE. The scores of MMSE were utilized to generate the output labels of the models. If the score of an elderly client was lower than 24, the corresponding sample was labeled as a "high-risk case;" otherwise, the sample was labeled as a "normal case."

As shown in Table 1, among the 567 features, complete values were only available from 96 features for all 2299 records. In addition, 49 features had a data missing rate of no more than 10% (ie, the values for these 49 features were missing in less than 10% of the records). The data missing rate of 140 features was over 60%. Besides, the data set was imbalanced, with 1872 normal cases versus 427 high-risk cases.

Table 1. Statistics of the features with missing data.

Percentage of missing data	Number of features
0%	96
1%-9%	49
10%-19%	22
20%-29%	6
30%-39%	97
40%-49%	5
50%-59%	152
60%-69%	140

KNN Imputation

To address the missing data problem, mean and KNN imputations were used in the study. For mean imputation, the missing values of a client record were filled by the average values of other records with observable feature values. For the KNN imputation method, the missing values of each client record were filled based on the observable values of its KNN. The idea is to assign a higher degree of importance to neighbors that are more similar to the target client record when filling the missing values. With regard to Figure 1, let be the set of features with complete values for all records, denoted as *complete features*, and be the set of features with missing values, denoted as *incomplete features*, where n_c and n_s represent the number of complete and incomplete features, respectively. In our data set, n_c was 96 and n_s was 471. Specifically, c_t represents a complete feature where all the client records in the data set have an entry value for the feature *t*. In contrast, s_b indicates an incomplete feature where at least one of the client records in the data set does not provide an entry value for the feature *b*. Furthermore, s_{bi} represents the entry value of the feature *b* in client record *i*, where s_{bi} is null if the value of feature *b* in client record *i* is missing. Let $D \in R^{m \times m}$ be a distance matrix that measures the distance between each pair of client records based on all complete features, where *m* is the number of client records in the data set, and D_{ij} represents the distance between client records *i* and *j*. In this work, we employed Euclidean distance as the distance metric; however, other distance metrics (eg, City Block Distance, Cosine similarity, L1 distance, L2 distance, and Manhattan distance [37,38]) can also be used.

The algorithm of the KNN imputation method is shown in Figure 2. First, the distance between each pair of client records was measured based on all 96 complete features. Thereafter, the missing values of the incomplete features in a client record were filled with the weighted average of the observable feature values of the *k* nearest records to that client record. After imputation, all the features were treated as "complete" and then utilized as input features of the proposed DNN model for dementia prediction.

Figure 1. Organization of features into the complete feature set C (left) and the incomplete feature set S (right). The checkmark symbol indicates that a value for a feature is present in a client record, whereas the cross symbol indicates a value is missing (empty).

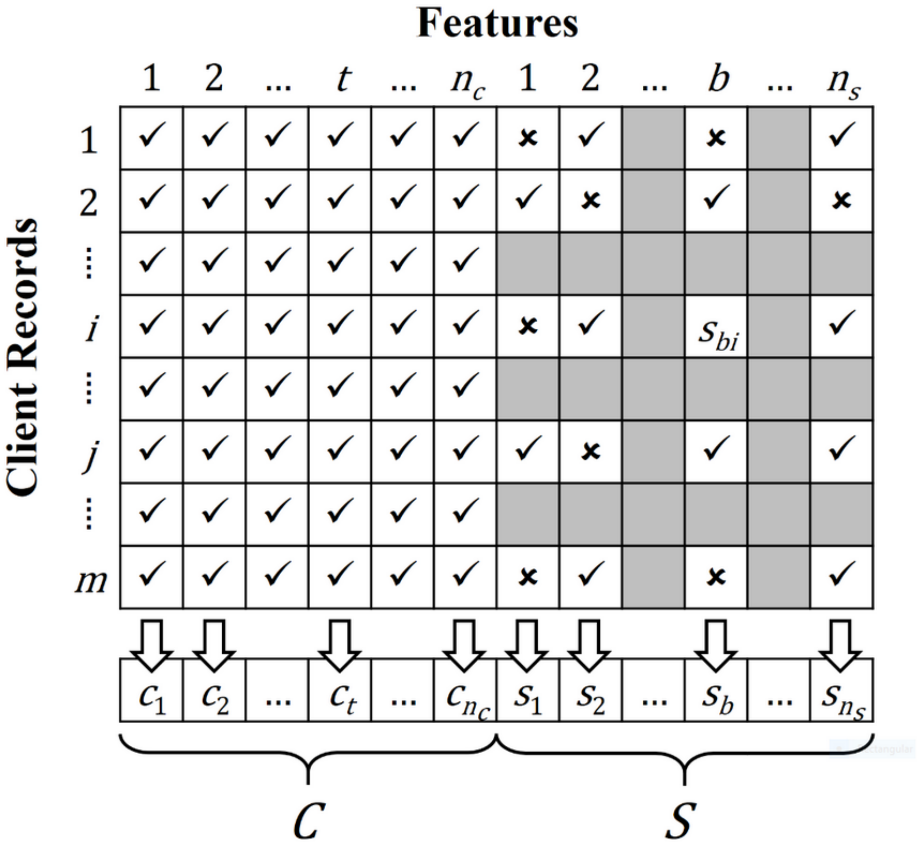
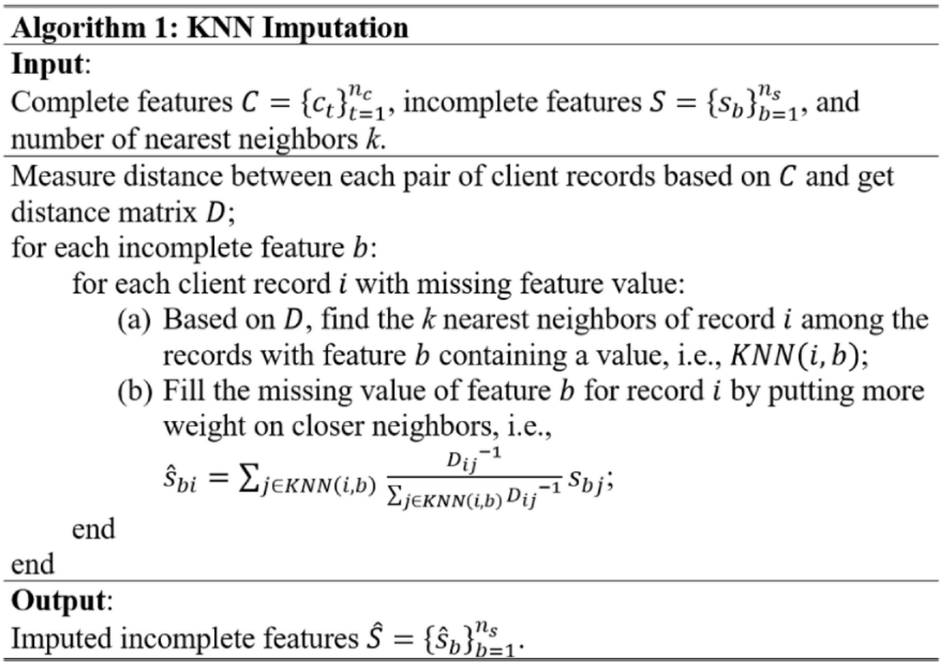

Features
1 2 ... t ... n_c 1 2 ... b ... n_s
Client Records
1
2
i
j
m
s_{bi}
c_1 c_2 ... c_t ... c_{n_c} s_1 s_2 ... s_b ... s_{n_s}
C
S

Figure 2. Algorithm of k-nearest neighbor imputation. KNN: k-nearest neighbor.



```
Algorithm 1: KNN Imputation
Input:
Complete features $C = \{c_t\}_{t=1}^{n_c}$, incomplete features $S = \{s_b\}_{b=1}^{n_s}$, and
number of nearest neighbors $k$.
Measure distance between each pair of client records based on $C$ and get
distance matrix $D$;
for each incomplete feature $b$:
    for each client record $i$ with missing feature value:
        (a) Based on $D$, find the $k$ nearest neighbors of record $i$ among the
            records with feature $b$ containing a value, i.e., $KNN(i,b)$;
        (b) Fill the missing value of feature $b$ for record $i$ by putting more
            weight on closer neighbors, i.e.,
            $\hat{s}_{bi} = \sum_{j\in KNN(i,b)} \frac{D_{ij}^{-1}}{\sum_{j\in KNN(i,b)} D_{ij}^{-1}} s_{bj}$;
    end
end
Output:
Imputed incomplete features $\hat{S} = \{\hat{s}_b\}_{b=1}^{n_s}$.
```

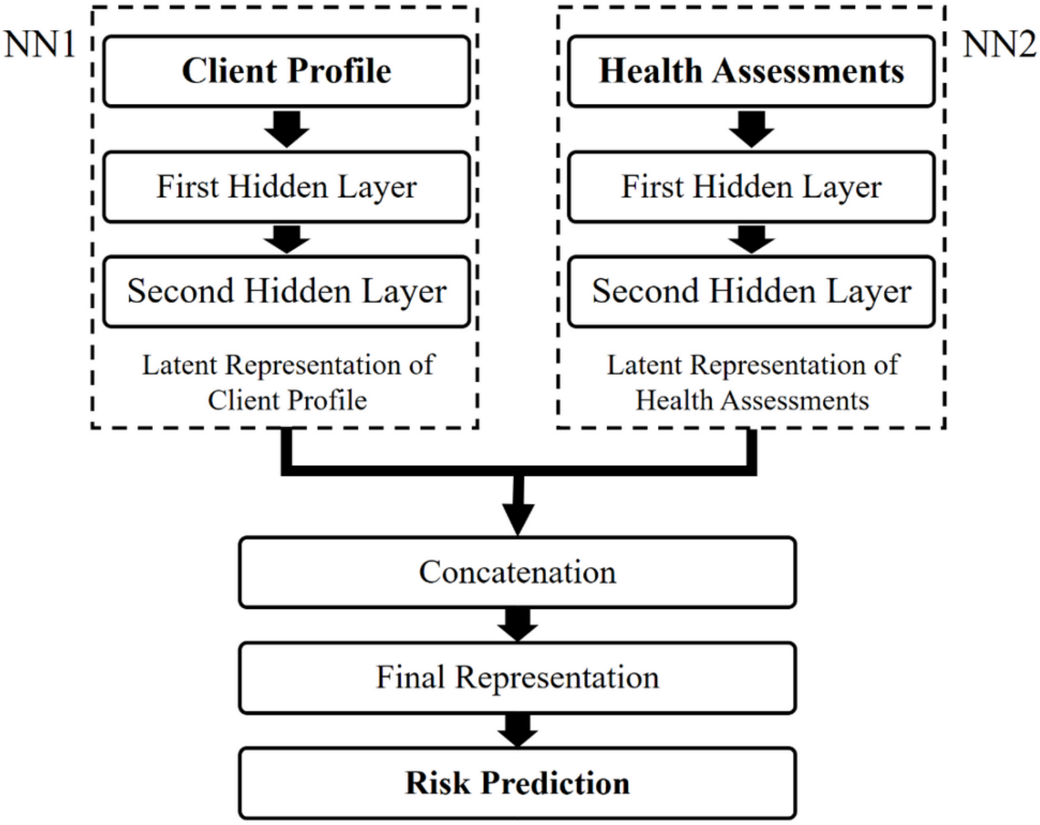
DNN Architecture

In the proposed DNN model, the input features were categorized into two types as follows: client profile and health assessment. The former included demographic information, medical history, and bio-measurements of the elderly clients. The latter included the information collected from nine health assessment questionnaires (ie, BPI, EMS, GDS, MNA, CQ, RLT, gross oral hygiene assessment, visual acuity assessment, and survey of favorite activities).

Recently, DNN architecture has been proposed and utilized in state-of-the-art feature representation learning models to learn latent representations based on two types of input features

[39-41]. The two types of latent representations are then integrated to learn the final representation for the classification tasks. The approach has demonstrated promising performance in feature representation learning and the ability to capture different kinds of information relevant to the classification task when the two types of input features convey different information and have varied data distributions. Motivated by this approach, we proposed a DNN architecture for screening people with high dementia risk. It learned the latent representations based on the two types of input features concerned in this study. Figure 3 shows the main architecture of the proposed model. With reference to a previous report [40], we employed two neural networks, namely, neural network 1 (NN1) and neural network 2 (NN2), each with two hidden layers, to learn the latent representations for each client from the client profile and health assessments, respectively. The representations were referred to as *latent profile representation* and *latent health assessment representation*. The two latent representations were learned with the two distinct types of features fed as inputs to NN1 and NN2, which were then concatenated to yield the final representation for predicting the dementia risk.

Figure 3. The architecture of the dual neural network. NN1: neural network 1; NN2: neural network 2.



Let $p_i \in R^{1\times np}$ be a vector representing the profile information associated with client *i*, where n_p is the number of features in the profile. Let $q_i \in R^{1\times nq}$ be a vector representing the health assessment information associated with client *i*, where n_q is the number of features in the assessment questionnaires. Additionally, $n=n_p+n_q$ is the total number of input features. In our data set, n_p was 132, n_q was 435, and *n* was 567.

In NN1, with the client profile information as the input, the latent profile representation was learned layer by layer as follows:

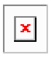
where $ReLU(\cdot)$is the rectified linear unit activation function characterized by $ReLU(x)=\max(0,x)$, p_i is the input profile feature associated with client *i*, $h_i^{p(1)} \in R^{1\times d1}$ and $h_i^{p(2)} \in R^{1\times d2}$ represent the latent profile representation of client *i*, learned by the first and second hidden layers of NN1, respectively, and d_1 and d_2 are the dimensionalities of the first and second hidden layers of NN1, respectively. Additionally, $W^{p(1)} \in R^{np\times d1}$ and $b^{p(1)} \in R^{1\times d1}$ are the trainable weight and bias parameters associated with the first hidden layer of NN1. Moreover, $W^{p(2)} \in R^{d1\times d2}$ and $b^{p(2)} \in R^{1\times d2}$ are the trainable parameters associated with the second hidden layer of NN1.

Similarly, in NN2, with the information from the health assessment as the input, the latent health assessment representation was learned layer by layer as follows:

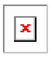
where q_i is the feature of health assessment of client *i* and $h_i^{q(1)} \in R^{1\times d1}$ and $h_i^{q(2)} \in R^{1\times d2}$ are the latent health assessment representations of client *i* learned by the first and second hidden layers of NN2. Additionally, $W^{q(1)} \in R^{nq\times d1}$, $b^{q(1)} \in R^{1\times d1}$,

$W^{q(2)} \in R^{d1\times d2}$, and $b^{q(2)} \in R^{1\times d2}$ are the trainable parameters of NN2. In the proposed DNN model, the hidden dimensionalities for NN1 and NN2 were set to be the same.

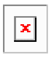
Thereafter, the deepest latent profile representation learned by NN1 (ie, $h_i^{p(2)}$) and the deepest latent health assessment representation learned by NN2 (ie, $h_i^{q(2)}$) were concatenated to give the final representation as follows:



where $h_i \in R^{1\times 2d2}$ is the final representation of client i. The final representation is then fed into the classification layer to predict whether an elderly client is high risk or normal as follows:

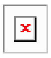


where $\hat{y}_i$ denotes the predicted probability that client i is at high risk. W^y and b^y are the trainable parameters associated with the dementia classification. Given the ground truth labels of the client records that are used as training samples, the supervised classification loss L is defined as follows:



where m_r is the number of training samples. The ground truth label is y_i=1 if the training sample corresponding to client record i is a high-risk case and y_i=0 if it is a normal case.

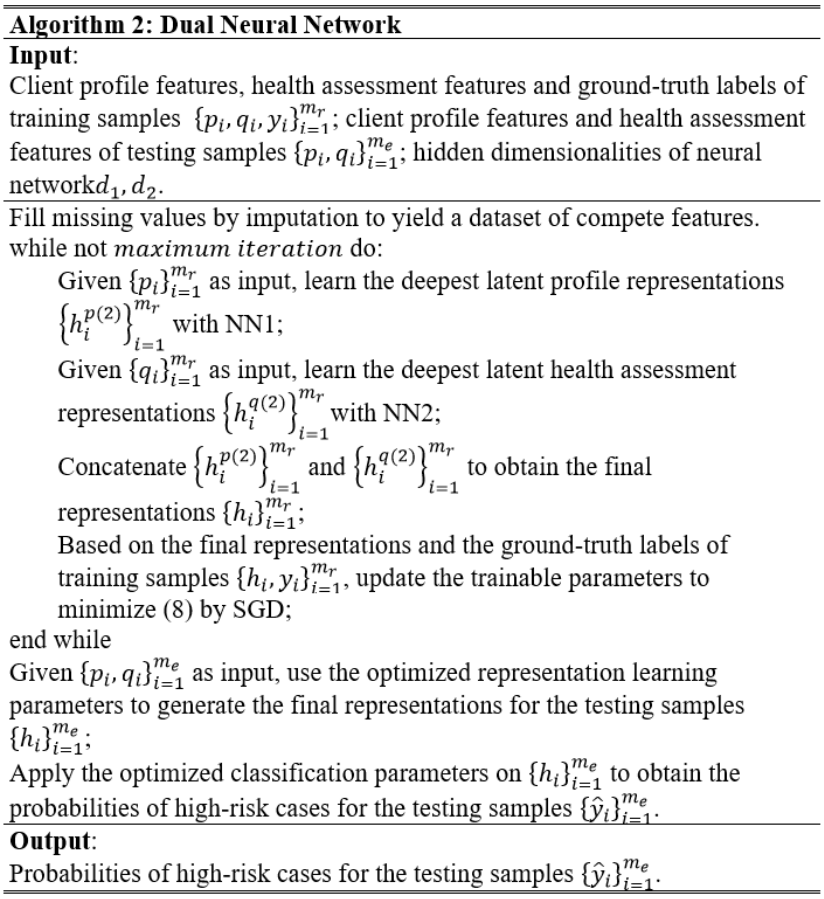
As the data set adopted in the study was imbalanced, with 1872 normal cases and only 427 high-risk cases, the classifiers in supervised machine learning could be biased toward the majority class samples (ie, normal cases). As a screening tool that is used to identify possible cases of high dementia risk, it is important to accurately detect the minority class (high-risk cases). To make the proposed DNN model focus more on high-risk cases, a CSL method was employed by introducing the cost ratio w into the classification loss in equation 7 as follows:



where $w_i=m_r^n/m_r^d$ if y_i=1 (ie, high-risk case) and w_i=1 if y_i=0 (ie, normal case). Additionally, m_r^n and m_r^d are the numbers of normal cases and high-risk cases in the training samples, respectively.

The proposed DNN model was trained following the algorithm shown in Figure 4. First, the missing values in the data set were filled by imputation. For KNN imputation, algorithm 1 was used. Thereafter, NN1 and NN2 were used to learn the latent profile representation and latent health assessment representation, respectively, which were concatenated to yield the final representation for classification. The trainable parameters of NN1 and NN2 that minimize the cost-sensitive classification loss in equation 8 were identified using the stochastic gradient descent (SGD) algorithm [42]. After the model converged, the optimized parameters were employed to generate the final representations and predict the probabilities of high-risk cases for the testing samples.

Figure 4. Algorithm of the dual neural network. NN1: neural network 1; NN2: neural network 2; SGD: stochastic gradient descent.



```
Algorithm 2: Dual Neural Network
Input:
Client profile features, health assessment features and ground-truth labels of training samples $\{p_i, q_i, y_i\}_{i=1}^{m_r}$; client profile features and health assessment features of testing samples $\{p_i, q_i\}_{i=1}^{m_e}$; hidden dimensionalities of neural network$d_1, d_2$.
Fill missing values by imputation to yield a dataset of compete features.
while not $maximum\ iteration$ do:
    Given $\{p_i\}_{i=1}^{m_r}$ as input, learn the deepest latent profile representations $\left\{h_i^{p(2)}\right\}_{i=1}^{m_r}$ with NN1;
    Given $\{q_i\}_{i=1}^{m_r}$ as input, learn the deepest latent health assessment representations $\left\{h_i^{q(2)}\right\}_{i=1}^{m_r}$ with NN2;
    Concatenate $\left\{h_i^{p(2)}\right\}_{i=1}^{m_r}$ and $\left\{h_i^{q(2)}\right\}_{i=1}^{m_r}$ to obtain the final representations $\{h_i\}_{i=1}^{m_r}$;
    Based on the final representations and the ground-truth labels of training samples $\{h_i, y_i\}_{i=1}^{m_r}$, update the trainable parameters to minimize (8) by SGD;
end while
Given $\{p_i, q_i\}_{i=1}^{m_e}$ as input, use the optimized representation learning parameters to generate the final representations for the testing samples $\{h_i\}_{i=1}^{m_e}$;
Apply the optimized classification parameters on $\{h_i\}_{i=1}^{m_e}$ to obtain the probabilities of high-risk cases for the testing samples $\{\hat{y}_i\}_{i=1}^{m_e}$.
Output:
Probabilities of high-risk cases for the testing samples $\{\hat{y}_i\}_{i=1}^{m_e}$.
```

Experiments and Settings

The performance of the proposed DNN model was evaluated by making comparisons with five kinds of conventional algorithms (ie, logistic regression [LR], DT, RF, SVM, and single neural network [SNN]). For SVM, three kernel functions were used (ie, linear, polynomial, and radial basis functions, denoted as SVM (linear), SVM (poly), and SVM (RBF), respectively. The SNN, employing all features in one shot as the input, was used to evaluate the effect of introducing an additional neural network in the proposed DNN on classification performance. Moreover, the effect of using CSL to tackle class imbalance was studied by applying it to the algorithms. The corresponding algorithms were denoted as LR+CSL, DT+CSL, RF+CSL, SVM (linear)+CSL, SVM (poly)+CSL, SVM (RBF)+CSL, SNN+CSL, and DNN+CSL. In summary, there were 16 algorithms overall under testing.

In the experiments, mean and KNN imputations were utilized to fill the missing data before model learning. The number of neighbors was set as k=5 for the KNN imputation. The LR, DT, RF, and SVM algorithms were implemented using the Scikit-Learn toolkit [43], where default settings were adopted for LR, DT, and the three versions of SVM models with different kernel functions. In RF, the number of trees was set as 100 and the maximum depth of the trees was set as 3. For the DNN, the hidden dimensionalities for both NN1 and NN2 were set with the typical values of d_1=128 and d_2=32. Note that in the DNN, we concatenated the latent representations of NN1 and NN2 as the final representations. To make the SNN and DNN have the same final dimensionality, we set the hidden dimensionalities of SNN as twice of NN1 and NN2 (ie, d_1=256 and d_2=64). All the neural network models were trained by the SGD with a momentum rate of 0.9 following common practice [40]. While normalization to the range of 0 to 1 was initially applied to preprocess the input features, it turned out that the performance degraded instead. Hence, preprocessing methods were not applied in the experiments.

The algorithms under comparison were evaluated with 10-fold cross-validation. The client records were randomly split into 10 folds of equal size. For each of the 10 runs, nine folds of records were employed as training samples and the remaining one fold of records was utilized as testing samples to evaluate prediction performance.

Four performance metrics were adopted, including area under the receiver operating characteristic curve (AUC) [44], average precision (AP) [45], sensitivity, and specificity. For imbalanced data sets, using classification accuracy as an evaluation metric would produce misleading results [46]. Here, AUC was used instead as it is insensitive to class imbalance. The metric AP summarized the precision-recall curve by weighting the precision achieved at each threshold with the increase in recall at the previous threshold. Sensitivity is the recall of high-risk cases (ie, the proportion of "high-risk" testing samples accurately identified). Specificity is the recall of normal cases (ie, the proportion of "normal" testing samples accurately identified).

It was hypothesized that the performance of DNN+CSL would be better than that of the algorithms under comparison, which was tested by running pairwise one-sided *t* tests between DNN+CSL and each algorithm separately in terms of the four metrics. Furthermore, experiments were conducted to investigate variation in the performance of the DNN in terms of AUC and AP with the number of neighbors when KNN imputation was used and with the dimensionalities d_1 and d_2 of the hidden layers in NN1 and NN2.

In addition, the effect of adding fully connected layers (FCLs) between the concatenated representation and the final prediction results was investigated. The experiment was conducted by adding one and two FCLs separately to the proposed DNN+CSL approach and evaluating the performance in terms of the four metrics.

Results

Classification Performance

The results of the experiments conducted to evaluate the performance of the algorithms under comparison are shown in Tables 2 and 3, where the mean and SD of the four metrics over 10 runs are provided. In addition, the performance of the proposed DNN+CSL model was compared with that of the other algorithms using a pair-wise *t* test, and the corresponding *P* values are shown in the tables.

As shown in Table 2, when mean imputation was applied, for the metrics AUC and AP, RF+CSL, RF, DNN, and DNN+CSL were the top performing algorithms. For sensitivity, DNN+CSL was among the top three algorithms, with SVM (poly)+CSL and SVM (RBF)+CSL being the first and second algorithms, respectively, and RF exhibited the worst sensitivity (0.01). For specificity, RF, SVM (RBF), and DNN were the top three algorithms. The specificity of DNN+CSL reached 0.80.

Similar results were obtained for KNN imputation. As shown in Table 3, DNN+CSL, DNN, and RF were the top performing algorithms in terms of AUC and AP. DNN+CSL ranked third in sensitivity after SVM (RBF)+CSL and SVM (poly)+CSL. The sensitivity of RF was the worst (0.03). The specificities of RF, SVM (RBF), and DNN were the best and that of DNN+CSL was 0.79.

The results also indicated that the performance of the algorithms evaluated by using mean imputation to tackle missing data was similar to that using KNN imputation. It can also be seen that when CSL was applied to tackle class imbalance, the sensitivity of the algorithms increased and specificity decreased.

Table 2. Performance of algorithms with missing data handled by mean imputation.

Algorithm	Mean imputation							
	AUC[a], mean (SD)	*P* value	AP[b], mean (SD)	*P* value	Sensitivity, mean (SD)	*P* value	Specificity, mean (SD)	*P* value
LR[c]	0.82 (0.04)	.02	0.87 (0.03)	.047	0.50 (0.10)	<.001	0.91 (0.03)	>.99
LR+CSL[d]	0.82 (0.04)	.02	0.87 (0.03)	.03	0.67 (0.07)	.002	0.82 (0.02)	.98
DT[e]	0.65 (0.05)	<.001	0.76 (0.03)	<.001	0.43 (0.09)	<.001	0.87 (0.03)	>.99
DT+CSL	0.64 (0.02)	<.001	0.75 (0.03)	<.001	0.41 (0.05)	<.001	0.87 (0.02)	>.99
RF[f]	0.84 (0.05)	.52	0.89 (0.03)	.90	0.01 (0.01)	<.001	1.00 (0.00)	>.99
RF+CSL	0.84 (0.05)	.67	0.89 (0.03)	.93	0.64 (0.09)	.001	0.84 (0.03)	>.99
SVM[g] (RBF[h])	0.78 (0.06)	<.001	0.85 (0.03)	<.001	0.12 (0.04)	<.001	0.99 (0.01)	>.99
SVM (RBF)+CSL	0.81 (0.05)	<.001	0.86 (0.03)	<.001	0.76 (0.08)	.98	0.73 (0.03)	<.001
SVM (poly[i])	0.74 (0.06)	<.001	0.83 (0.03)	<.001	0.50 (0.07)	<.001	0.84 (0.03)	>.99
SVM (poly)+CSL	0.81 (0.05)	<.001	0.87 (0.03)	<.001	0.77 (0.08)	.99	0.73 (0.03)	<.001
SVM (linear)	0.79 (0.04)	<.001	0.85 (0.03)	<.001	0.48 (0.07)	<.001	0.89 (0.02)	>.99
SVM (linear)+CSL	0.80 (0.04)	.004	0.86 (0.03)	.005	0.65 (0.07)	<.001	0.81 (0.02)	.94
SNN[j]	0.81 (0.05)	<.001	0.87 (0.03)	.003	0.32 (0.09)	<.001	0.95 (0.01)	>.99
SNN+CSL	0.81 (0.05)	<.001	0.87 (0.03)	<.001	0.65 (0.11)	.002	0.83 (0.02)	>.99
DNN[k]	0.83 (0.05)	.045	0.88 (0.03)	.13	0.33 (0.09)	<.001	0.96 (0.02)	>.99
DNN+CSL	0.84 (0.04)	N/A[l]	0.88 (0.03)	N/A	0.73 (0.09)	N/A	0.80 (0.03)	N/A

[a] AUC: area under the receiver operating characteristic curve.

[b] AP: average precision.

[c] LR: logistic regression.

[d] CSL: cost-sensitive learning.

[e] DT: decision tree.

[f] RF: random forest.

[g] SVM: support vector machine.

[h] RBF: radial basis function kernel.

[i] poly: polynomial kernel.

[j] SNN: single neural network.

[k] DNN: dual neural network.

[l] N/A: not applicable.

Table 3. Performance of algorithms with missing data handled by k-nearest neighbor imputation.

Algorithm	KNN[a] imputation							
	AUC[b], mean (SD)	*P* value	AP[c], mean (SD)	*P* value	Sensitivity, mean (SD)	*P* value	Specificity, mean (SD)	*P* value
LR[d]	0.81 (0.04)	.02	0.87 (0.03)	.10	0.46 (0.10)	<.001	0.91 (0.02)	>.99
LR+CSL[e]	0.81 (0.04)	.005	0.87 (0.02)	.03	0.65 (0.09)	.001	0.81 (0.03)	.90
DT[f]	0.67 (0.05)	<.001	0.77 (0.04)	<.001	0.48 (0.09)	<.001	0.86 (0.03)	>.99
DT+CSL	0.66 (0.07)	<.001	0.76 (0.05)	<.001	0.45 (0.13)	<.001	0.86 (0.02)	>.99
RF[g]	0.81 (0.04)	.004	0.87 (0.03)	.13	0.03 (0.04)	<.001	1.00 (0.00)	>.99
RF+CSL	0.81 (0.04)	.004	0.87 (0.03)	.02	0.68 (0.08)	.04	0.78 (0.02)	.10
SVM[h] (RBF[i])	0.77 (0.06)	<.001	0.84 (0.03)	<.001	0.08 (0.04)	<.001	0.99 (0.01)	>.99
SVM (RBF)+CSL	0.80 (0.05)	<.001	0.86 (0.03)	<.001	0.75 (0.09)	.90	0.73 (0.02)	<.001
SVM (poly[j])	0.75 (0.04)	<.001	0.83 (0.03)	<.001	0.50 (0.10)	<.001	0.86 (0.02)	>.99
SVM (poly)+CSL	0.81 (0.05)	<.001	0.86 (0.03)	<.001	0.74 (0.09)	.83	0.73 (0.02)	<.001
SVM (linear)	0.80 (0.04)	.001	0.86 (0.03)	.005	0.48 (0.11)	<.001	0.89 (0.02)	>.99
SVM (linear)+CSL	0.77 (0.04)	<.001	0.85 (0.02)	<.001	0.58 (0.11)	<.001	0.80 (0.02)	.75
SNN[k]	0.81 (0.06)	<.001	0.87 (0.04)	.006	0.33 (0.10)	<.001	0.95 (0.01)	>.99
SNN+CSL	0.80 (0.06)	<.001	0.86 (0.03)	<.001	0.65 (0.11)	.01	0.81 (0.03)	.90
DNN[l]	0.83 (0.05)	.04	0.88 (0.03)	.08	0.35 (0.09)	<.001	0.96 (0.01)	>.99
DNN+CSL	0.84 (0.04)	N/A[m]	0.88 (0.03)	N/A	0.72 (0.10)	N/A	0.79 (0.04)	N/A

[a] KNN: *k*-nearest neighbor.

[b] AUC: area under the receiver operating characteristic curve.

[c] AP: average precision.

[d] LR: logistic regression.

[e] CSL: cost-sensitive learning.

[f] DT: decision tree.

[g] RF: random forest.

[h] SVM: support vector machine.

[i] RBF: radial basis function kernel.

[j] poly: polynomial kernel.

[k] SNN: single neural network.

[l] DNN: dual neural network.

[m] N/A: not applicable.

Optimal Parameter Setting for the DNN

The effects of the parameters k, d_1, and d_2 on the performance of the proposed DNN in terms of AUC and AP are shown in Figure 5, Figure 6, and Figure 7 respectively. It can be seen from Figure 5 that when KNN imputation was used, both AUC and AP increased with k for $k<5$. When k was further increased, AUC exhibited a decreasing trend, whereas AP remained at about the same level. This suggests that it is appropriate to set the number of neighbors as $k=5$ for KNN imputation. For the number of dimensions d_1 of the first hidden layer of NN1 and NN2, as shown in Figure 6, a relatively large value (ie, 128 or 256) would yield a higher AUC and AP. In contrast, Figure 7 shows that setting the number of dimensions d_2 of the second hidden layer of NN1 and NN2 to a relatively small value (ie, 64 or 32) would achieve a higher AUC, while AP was insensitive to d_2.

Figure 5. Variation in the area under the receiver operating characteristic curve (AUC) and average precision (AP) with the number of neighbors k in k-nearest neighbor.

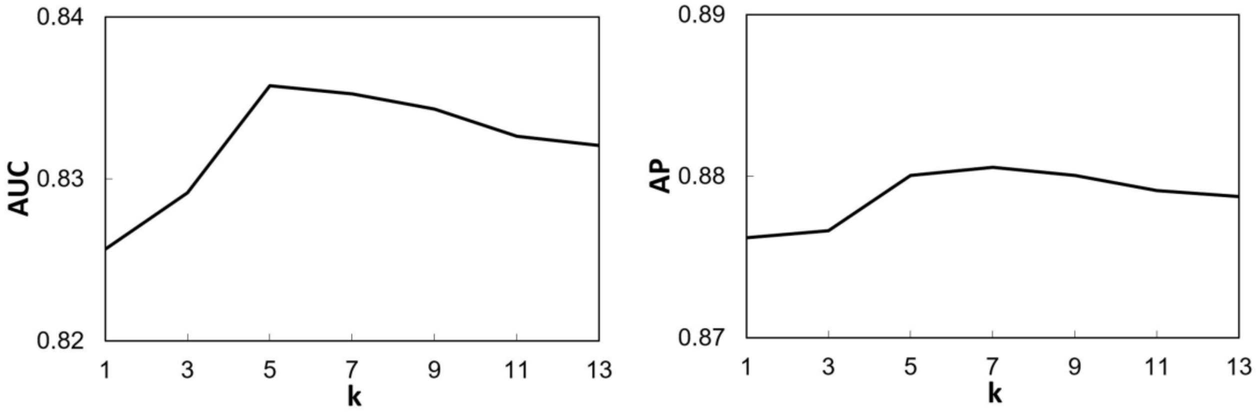


Figure 6. Variation in the area under the receiver operating characteristic curve (AUC) and average precision (AP) with the dimensionality d1 of the first hidden layer in neural network 1 and neural network 2.

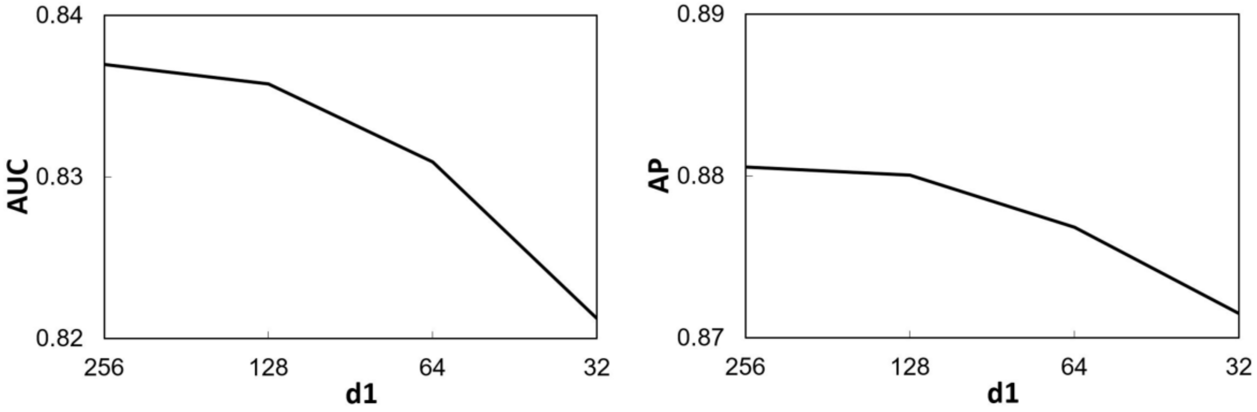
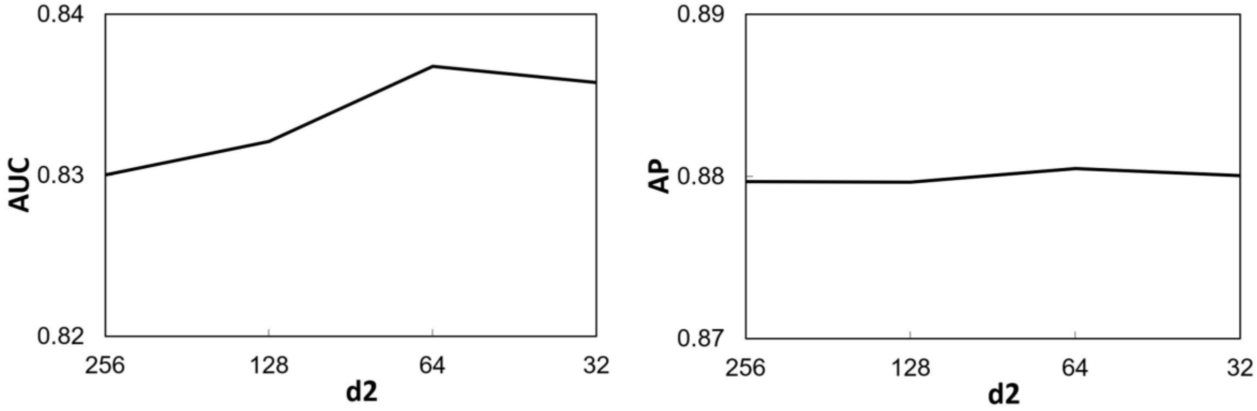


Figure 7. Variation in the area under the receiver operating characteristic curve (AUC) and average precision (AP) with the dimensionality d2 of the second hidden layer in neural network 1 and neural network 2.



Effect of FCLs

The effect of adding FCLs to the proposed DNN+CSL model is shown in Table 4. For both mean and KNN imputations, it was found that adding one FCL lowered the AUC and specificity as compared with the finding without an FCL, whereas adding two FCLs lowered the AUC and specificity while increasing sensitivity.

Table 4. Effect of fully connected layers on the proposed dual neural network plus cost-sensitive learning model.

Imputation and algorithm	AUC[a], mean (SD)	AP[b], mean (SD)	Sensitivity, mean (SD)	Specificity, mean (SD)
Mean				
DNN[c]+CSL[d]	0.84 (0.04)	0.88 (0.03)	0.73 (0.09)	0.80 (0.03)
DNN+CSL with one FCL[e]	0.83 (0.04)	0.88 (0.03)	0.73 (0.09)	0.79 (0.07)
DNN+CSL with two FCLs	0.83 (0.05)	0.88 (0.03)	0.77 (0.11)	0.75 (0.04)
KNN[f]				
DNN+CSL	0.84 (0.04)	0.88 (0.03)	0.72 (0.10)	0.79 (0.04)
DNN+CSL with one FCL	0.83 (0.04)	0.88 (0.03)	0.71 (0.10)	0.77 (0.09)
DNN+CSL with two FCLs	0.82 (0.05)	0.87 (0.03)	0.77 (0.12)	0.74 (0.03)

[a]AUC: area under the receiver operating characteristic curve.

[b]AP: average precision.

[c]DNN: dual neural network.

[d]CSL: cost-sensitive learning.

[e]FCL: fully connected layer.

[f]KNN: *k*-nearest neighbor.

Discussion

Principal Findings

Among the 16 algorithms under testing, DNN+CSL outperformed and consistently ranked among the top three algorithms in terms of AUC, AP, and sensitivity for both mean and KNN imputations. In the case of KNN imputation, DNN+CSL indeed showed the best AUC (mean 0.84, SD 0.04) and AP (mean 0.88, SD 0.03), and ranked third in sensitivity (mean 0.72, SD 0.10). The mean specificity of DNN+CSL reached 0.79 (SD 0.10). Although RF was competitive and ranked among the top three algorithms in terms of AUC, AP, and specificity, the sensitivity was almost zero.

The results suggest that the proposed approach of deep learning with DNNs is promising for screening cognitive impairment in elderly people and thus high-risk cases of dementia. This is attributed to the ability of the DNN to learn hierarchical latent representations from two types of data with different characteristics. The DNN approach is able to capture complex nonlinear relationships between input features and the output.

For both mean and KNN imputations, the performance of using two neural networks in the proposed DNN was much better than that using a SNN in terms of AUC, AP, and sensitivity. While the same features were adopted in both the DNN and SNN, the main difference was that for the DNN, the features were divided into two groups and fed into the two separate neural networks NN1 and NN2. The inputs for NN1 were features concerning the client profile, whereas the inputs for NN2 were features pertaining to the health assessment questionnaires. In the data set adopted, the client profile features were more complete than the health assessment features, that is, more than 72% of the client profile features were complete, while all the features from the health assessment questionnaires contained missing values, with the missing rate ranging from 4.9% to as much as 69.6%. This shows that the elderly clients in general had high acceptance toward the collection of demographic data, information about their medical history, and bio-measurement data, thereby resulting in a low data missing rate for client profile features. On the other hand, the high data missing rate for health assessment features is consistent with the general situation in primary care. According to the frontline health care staff of the clinic, data were missed because clients were absent from scheduled appointments, unable to recall specific events that happened in the past, or declined to respond to questions that they felt uncomfortable to answer or considered sensitive. Given the different characteristics of the two kinds of features, it was beneficial to employ two separate neural networks with different trainable weights to learn the corresponding latent representations.

Furthermore, since all the features were used indiscriminately in the SNN as the input, the characteristics of the two types of features could be interfered or diffused. More importantly, it was likely that the health assessment features, whose quality was affected by missing data, could contaminate the client profile features that were more complete and of better quality. This could be a reason for the suboptimal performance of the SNN as compared with the proposed DNN.

In the data set adopted, the ratio of high-risk cases to normal cases was 1 to 4.4. If the issue of class imbalance was ignored, the classification result would have been biased toward the majority class (ie, normal cases in this study). As a screening tool, high sensitivity is desirable as it is important to identify possible true positives (high-risk cases) and generate early signals, suggesting the potential need for a follow-up. CSL was thus proposed to remedy class imbalance. The effectiveness was evident from the result that the sensitivity of most algorithms improved. For example, when mean imputation was applied, sensitivity increased by 118% for the DNN, 537% for SVM (RBF), and over 70 times for RF, whose sensitivity was almost zero (from 0.01 to 0.64). For missing data imputed using KNN, sensitivity increased by 109% for the DNN, 818% for SVM (RBF), and over 18 times for RF (from 0.03 to 0.68). The increase in sensitivity was achieved at the expense of specificity, with a moderate decrease of less than 26% for data imputed with both imputation methods. Nevertheless, the specificities of the algorithms were still above 0.73 when CSL was applied.

Limitations

The study presents a machine learning method to screen for elderly people with cognitive impairment and identify high-risk cases of dementia simply by two-class classification. The method can be extended to four-class classification, that is, normal, mild, moderate, and severe, according to MMSE score ranges of 24-30, 19-23, 10-18, and 0-9, respectively. However, the problem of class imbalance would become more relevant. A balanced number of samples for the four classes would be required to construct a fair classification model to avoid predilection for the majority class.

In the proposed DNN architecture, KNN-based imputation was incorporated to tackle missing data, where the nearest neighbors were simply calculated by treating all features with the same weight. Future work will be conducted to design a scheme to assign different weights to different features during KNN imputation.

The elderly health data used in the study were collected from a specific setting of primary care services. Some of the data may not be available from elderly care centers in general, which precludes the use of the proposed DNN-based screening tool. Nevertheless, the client profile data involved and the health assessments adopted were indeed relatively conventional and could be readily integrated into existing health care services in order to make use of the proposed screening tool. On the other hand, future work will be conducted to evaluate and rank the importance of input features, so that less critical features can be dropped to reduce the variety of health data required while still maintaining classification performance.

Conclusions

This study proposed a DNN approach to screen for elderly people with high risk of dementia using data collected from health care services provided in primary care. Imputation techniques were used to deal with missing data, whereas CSL was adopted to tackle class imbalance. The proposed approach overall outperformed conventional machine learning techniques. It has the potential to serve as an assistive tool for health care professionals to identify people with high risk of dementia at the asymptomatic stage, thereby generating early signals to prompt suggestions for follow-up or the need for a detailed diagnosis of dementia.

Acknowledgments

The work was supported in part by the Innovation and Technology Fund of the Hong Kong Special Administrative Region under grant MRP/015/18.

Conflicts of Interest

None declared.

References

1. ICD-11 for Mortality and Morbidity Statistics. World Health Organization. URL: https://icd.who.int/browse11/l-m/en [accessed 2020-04-10]
2. Dementia: A Public Health Priority. Geneva, Switzerland: World Health Organization; 2012.
3. Policy Brief for Heads of Government: The Global Impact of Dementia 2013-2050. Alzheimer's Disease International. URL: https://www.alz.co.uk/research/GlobalImpactDementia2013.pdf [accessed 2020-04-10]
4. Nichols E, Szoeke CE, Vollset SE, Abbasi N, Abd-Allah F, Abdela J, et al. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet Neurology 2019 Jan;18(1):88-106. [doi: 10.1016/S1474-4422(18)30403-4]
5. Livingston G, Sommerlad A, Orgeta V, Costafreda SG, Huntley J, Ames D, et al. Dementia prevention, intervention, and care. The Lancet 2017 Dec;390(10113):2673-2734. [doi: 10.1016/S0140-6736(17)31363-6]
6. Kivipelto M, Mangialasche F, Ngandu T. Lifestyle interventions to prevent cognitive impairment, dementia and Alzheimer disease. Nat Rev Neurol 2018 Nov;14(11):653-666. [doi: 10.1038/s41582-018-0070-3] [Medline: 30291317]
7. Dubois B, Padovani A, Scheltens P, Rossi A, Dell'Agnello G. Timely Diagnosis for Alzheimer's Disease: A Literature Review on Benefits and Challenges. J Alzheimers Dis 2016;49(3):617-631 [FREE Full text] [doi: 10.3233/JAD-150692] [Medline: 26484931]
8. De Lepeleire J, Wind A, Iliffe S, Moniz-Cook E, Wilcock J, Gonzalez VM, Interdem Group. The primary care diagnosis of dementia in Europe: an analysis using multidisciplinary, multinational expert groups. Aging Ment Health 2008 Sep;12(5):568-576. [doi: 10.1080/13607860802343043] [Medline: 18855172]
9. Robinson L, Tang E, Taylor J. Dementia: timely diagnosis and early intervention. BMJ 2015 Jun 16;350:h3029 [FREE Full text] [doi: 10.1136/bmj.h3029] [Medline: 26079686]
10. Amjad H, Roth DL, Sheehan OC, Lyketsos CG, Wolff JL, Samus QM. Underdiagnosis of Dementia: an Observational Study of Patterns in Diagnosis and Awareness in US Older Adults. J Gen Intern Med 2018 Jul;33(7):1131-1138 [FREE Full text] [doi: 10.1007/s11606-018-4377-y] [Medline: 29508259]

11. Connolly A, Gaehl E, Martin H, Morris J, Purandare N. Underdiagnosis of dementia in primary care: variations in the observed prevalence and comparisons to the expected prevalence. Aging Ment Health 2011 Nov;15(8):978-984. [doi: 10.1080/13607863.2011.596805] [Medline: 21777080]
12. Patterson C, Feighther J, Garcia A, MacKnight C. General risk factors for dementia: a systematic evidence review. Alzheimers Dement 2007 Oct;3(4):341-347. [doi: 10.1016/j.jalz.2007.07.001] [Medline: 19595956]
13. Agarwal V, Zhang L, Zhu J, Fang S, Cheng T, Hong C, et al. Impact of Predicting Health Care Utilization Via Web Search Behavior: A Data-Driven Analysis. J Med Internet Res 2016 Sep 21;18(9):e251 [FREE Full text] [doi: 10.2196/jmir.6240] [Medline: 27655225]
14. Tang EY, Harrison SL, Errington L, Gordon MF, Visser PJ, Novak G, et al. Current Developments in Dementia Risk Prediction Modelling: An Updated Systematic Review. PLoS One 2015;10(9):e0136181 [FREE Full text] [doi: 10.1371/journal.pone.0136181] [Medline: 26334524]
15. Nori VS, Hane CA, Martin DC, Kravetz AD, Sanghavi DM. Identifying incident dementia by applying machine learning to a very large administrative claims dataset. PLoS One 2019;14(7):e0203246 [FREE Full text] [doi: 10.1371/journal.pone.0203246] [Medline: 31276468]
16. Barnes D, Covinsky K, Whitmer R, Kuller L, Lopez O, Yaffe K. Predicting risk of dementia in older adults: The late-life dementia risk index. Neurology 2009 Jul 21;73(3):173-179 [FREE Full text] [doi: 10.1212/WNL.0b013e3181a81636] [Medline: 19439724]
17. Williams J, Weakley A, Cook D, Schmitter-Edgecombe M. Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. 2013 Presented at: The 27th Conference on Artificial Intelligence; July 14-18, 2013; Bellevue, Washington, USA.
18. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonça A. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res Notes 2011 Aug 17;4:299 [FREE Full text] [doi: 10.1186/1756-0500-4-299] [Medline: 21849043]
19. So A, Hooshyar D, Park K, Lim H. Early Diagnosis of Dementia from Clinical Data by Machine Learning Techniques. Applied Sciences 2017 Jun 23;7(7):651. [doi: 10.3390/app7070651]
20. Cleret de Langavant L, Bayen E, Yaffe K. Unsupervised Machine Learning to Identify High Likelihood of Dementia in Population-Based Surveys: Development and Validation Study. J Med Internet Res 2018 Jul 09;20(7):e10493 [FREE Full text] [doi: 10.2196/10493] [Medline: 29986849]
21. Pekkala T, Hall A, Lötjönen J, Mattila J, Soininen H, Ngandu T, et al. Development of a Late-Life Dementia Prediction Index with Supervised Machine Learning in the Population-Based CAIDE Study. J Alzheimers Dis 2017;55(3):1055-1067 [FREE Full text] [doi: 10.3233/JAD-160560] [Medline: 27802228]
22. Thai-Nghe N, Gantner Z, Schmidt-Thieme L. Cost-sensitive learning methods for imbalanced data. 2010 Presented at: International Joint Conference on Neural Networks; July 18-20, 2010; Barcelona, Spain.
23. Pereira T, Lemos L, Cardoso S, Silva D, Rodrigues A, Santana I, et al. Predicting progression of mild cognitive impairment to dementia using neuropsychological data: a supervised learning approach using time windows. BMC Med Inform Decis Mak 2017 Jul 19;17(1):110 [FREE Full text] [doi: 10.1186/s12911-017-0497-2] [Medline: 28724366]
24. Raskutti B, Kowalczyk A. Extreme re-balancing for SVMs. SIGKDD Explor. Newsl 2004 Jun;6(1):60-69. [doi: 10.1145/1007730.1007739]
25. Zhang Y, Wang D. A Cost-Sensitive Ensemble Method for Class-Imbalanced Datasets. Abstract and Applied Analysis 2013;2013:1-6. [doi: 10.1155/2013/196256]
26. Shen X, Chung F. Deep Network Embedding for Graph Representation Learning in Signed Networks. IEEE Trans. Cybern 2020 Apr;50(4):1556-1568. [doi: 10.1109/tcyb.2018.2871503]
27. Margineantu D. When does imbalanced data require more than cost-sensitive learning. 2000 Presented at: The 17th Conference on Artificial Intelligence; July 30-August 3, 2000; Texas, Austin, USA.
28. Folstein M, Folstein S, McHugh P. “Mini-mental state”. Journal of Psychiatric Research 1975 Nov;12(3):189-198. [doi: 10.1016/0022-3956(75)90026-6]
29. Galasko D, Abramson I, Corey-Bloom J, Thal LJ. Repeated exposure to the Mini-Mental State Examination and the Information-Memory-Concentration Test results in a practice effect in Alzheimer's disease. Neurology 1993 Aug;43(8):1559-1563. [doi: 10.1212/wnl.43.8.1559] [Medline: 8351011]
30. Cleeland CS, Ryan KM. Pain assessment: global use of the Brief Pain Inventory. Ann Acad Med Singapore 1994 Mar;23(2):129-138. [Medline: 8080219]
31. Smith R. Validation and Reliability of the Elderly Mobility Scale. Physiotherapy 1994 Nov;80(11):744-747. [doi: 10.1016/s0031-9406(10)60612-8]
32. Yesavage JA, Sheikh JI. 9/Geriatric Depression Scale (GDS). Clinical Gerontologist 2008 Oct 25;5(1-2):165-173. [doi: 10.1300/J018v05n01_09]
33. Guigoz Y, Vellas B, Garry P. Mini nutritional assessment: a practical assessment tool for grading the nutritional state of elderly patients. In: Vellas BJ, Albarede L, Garry PJ, editors. Facts and Research in Gerontology. Paris: Serdi; 1994:15-60.

34. Chan AO, Lam KF, Hui WM, Hu WH, Li J, Lai KC, et al. Validated questionnaire on diagnosis and symptom severity for functional constipation in the Chinese population. Aliment Pharmacol Ther 2005 Sep 01;22(5):483-488 [FREE Full text] [doi: 10.1111/j.1365-2036.2005.02621.x] [Medline: 16128687]
35. Roper N, Logan W, Tierney A. The Roper-Logan-Tierney model of nursing: based on activities of living. Edinburgh, Scotland: Churchill Livingstone; 2000.
36. Creavin ST, Wisniewski S, Noel-Storr AH, Trevelyan CM, Hampton T, Rayment D, et al. Mini-Mental State Examination (MMSE) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations. Cochrane Database Syst Rev 2016 Jan 13(1):CD011145. [doi: 10.1002/14651858.CD011145.pub2] [Medline: 26760674]
37. Weinberger KQ, Blitzer J, Saul LK. Distance metric learning for large margin nearest neighbor classification. 2005 Presented at: The 18th Annual Conference on Neural Information Processing Systems; December 5-8, 2005; Vancouver, British Columbia, Canada.
38. Felsenstein J. An alternating least squares approach to inferring phylogenies from pairwise distances. Syst Biol 1997 Mar;46(1):101-111. [doi: 10.1093/sysbio/46.1.101] [Medline: 11975348]
39. Liang J, Jacobs P, Sun J, Parthasarathy S. Semi-supervised embedding in attributed networks with outliers. 2018 Presented at: The SIAM International Conference on Data Mining; May 3-5, 2018; San Diego, California p. 3-5. [doi: 10.1137/1.9781611975321.18]
40. Shen X, Dai Q, Chung F, Lu W, Choi K. Adversarial Deep Network Embedding for Cross-Network Node Classification. 2020 Apr 03 Presented at: The 34th Conference on Artificial Intelligence; February 7-12, 2020; New York, USA p. 2991-2999. [doi: 10.1609/aaai.v34i03.5692]
41. Zhuang C, Ma Q. Dual graph convolutional networks for graph-based semi-supervised classification. 2018 Presented at: The 2018 World Wide Web Conference; April 23-27, 2018; Lyon, France p. 499-508. [doi: 10.1145/3178876.3186116]
42. Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. 2004 Presented at: The 21st International Conference on Machine learning; July 4-8, 2004; Banff, Alberta, Canada. [doi: 10.1145/1015330.1015332]
43. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011 Oct;12(85):2825-2830.
44. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters 2006 Jun;27(8):861-874. [doi: 10.1016/j.patrec.2005.10.010]
45. Manning C, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press; 2008.
46. Jeni L, Cohn J, De LT. Facing imbalanced data - recommendations for the use of performance metrics. 2013 Presented at: The 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction; September 2-5, 2013; Geneva, Switzerland p. 245-251. [doi: 10.1109/acii.2013.47]

Abbreviations

AP: average precision
AUC: area under the receiver operating characteristic curve
BPI: brief pain inventory
CQ: constipation questionnaire
CSL: cost-sensitive learning
DNN: dual neural network
DT: decision tree
EMS: elderly mobility scale
FCL: fully connected layer
GDS: geriatric depression scale
KNN: k-nearest neighbors
LR: logistic regression
MMSE: Mini-Mental State Examination
MNA: mini-nutrition assessment
NN1: neural network 1
NN2: neural network 2
Poly: polynomial kernel
RBF: radial basis function kernel
RF: random forest
RLT: Roper-Logan-Tierney model of nursing
SGD: stochastic gradient descent
SNN: single neural network
SVM: support vector machine

Edited by G Eysenbach; submitted 06.05.20; peer-reviewed by G Long, L Zhang; comments to author 01.06.20; revised version received 10.06.20; accepted 26.07.20; published 31.08.20.

Please cite as:
Shen X, Wang G, Kwan RYC, Choi KS
Using Dual Neural Network Architecture to Detect the Risk of Dementia With Community Health Data: Algorithm Development and Validation Study
JMIR Med Inform 2020;8(8):e19870
URL: https://medinform.jmir.org/2020/8/e19870
doi:10.2196/19870
PMID:32865498

Corrigenda and Addenda

Correction: Undergraduate Medical Students' Search for Health Information Online: Explanatory Cross-Sectional Study

Teresa Loda[1], MSc; Rebecca Erschens[1], PhD; Florian Junne[1], MD; Andreas Stengel[1,2,3], MD; Stephan Zipfel[1,4], MD; Anne Herrmann-Werner[1], MD

[1]Medical Department VI/Psychosomatic Medicine and Psychotherapy, University Hospital Tuebingen, Tuebingen, Germany
[2]Charité Center for Internal Medicine and Dermatology, Department for Psychosomatic Medicine, Charité-Universitätsmedizin Berlin, Berlin, Germany
[3]Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany
[4]Deanery of Students' Affairs, Faculty of Medicine, University Hospital Tuebingen, Tuebingen, Germany

Corresponding Author:
Rebecca Erschens, PhD
Medical Department VI/Psychosomatic Medicine and Psychotherapy
University Hospital Tuebingen
Osianderstr 5
Tuebingen, 72076
Germany
Phone: 49 07071 ext 2986719
Email: rebecca.erschens@med.uni-tuebingen.de

Related Article:

Correction of: https://medinform.jmir.org/2020/3/e16279

(JMIR Med Inform 2020;8(8):e23253) doi:10.2196/23253

In "Undergraduate Medical Students' Search for Health Information Online: Explanatory Cross-Sectional Study" (JMIR Med Inform 2020;8(3):e16279) the authors noted errors in the presentation of the *P* values in the text of the Results section and in Table 1 of the published manuscript.

For the effected text in the Results section, under the "Sample" subheading, the following sentence was revised from:

> *There were 50 students randomly assigned to Google, 46 to Medisuch, and 44 to the free choice group (χ^2_{278}=280.0 ,P P=).*

To:

> *There were 50 students randomly assigned to Google, 46 to Medisuch, and 44 to the free choice group (χ^2_{278}=280.0,P=.46).*

And this sentence was revised from:

> *There were no significant differences between the groups with regards to age ($F_{2,135}$=5.04,PP=), gender (χ^2_4=4.5,PP=), and previous formal medical or information technology (IT) training (χ^2_2=1.5,PP=).*

To:

> *There were no significant differences between the groups with regards to age ($F_{2,135}$=5.04,P=.008), gender (χ^2_4=4.5,P=.34), and previous formal medical or information technology (IT) training (χ^2_2=1.5,P=.23).*

Under the "Information-Seeking Behavior" subheading, the following sentence was revised from:

> *However, students of the free choice group (mean 0.88, SD 0.79) reported significantly fewer pages as recommendable to patients than the other two groups ($F_{2,133}$=5.04,PP=; M_{Google} 1.55, SD 0.91; $M_{Medisuch}$ 1.52, SD 1.53).*

To:

> *However, students of the free choice group (mean 0.88, SD 0.79) reported significantly fewer pages as recommendable to patients than the other two groups ($F_{2,133}$=5.04,P=.008; M_{Google} 1.55, SD 0.91; $M_{Medisuch}$ 1.52, SD 1.53).*

This sentence was revised from:

> *Students in the free choice group opened significantly fewer recommendable pages ($F_{2,133}$=5.04,PP=).*

To:

> *Students in the free choice group opened significantly fewer recommendable pages ($F_{2,133}$=5.04,P=.008).*

And this sentence was revised from:

> *There was a highly significant difference between groups in whether or not the students entered specific*

> *medical terminology in the search engine (χ^2_4=16.6,PP=).*

To:

> *There was a highly significant difference between groups in whether or not the students entered specific medical terminology in the search engine (χ^2_4=16.6,P=.005).*

Under the "Quality of Webpages" subheading, the following sentence was revised from:

> *There were significantly high Pearson correlations between the number of webpages and the number of reliable webpages for all three groups (Google: r=.895; free group: r=.912; Medisuch: r=.860; allP P<).*

To:

> *There were significantly high Pearson correlations between the number of webpages and the number of reliable webpages for all three groups (Google: r=.895; free group: r=.912; Medisuch: r=.860; all P<.001).*

This sentence was revised from:

> *There were no significant differences in the frequencies of trustworthy webpages found among the three groups with χ^2_{14}=16.45,PP=.*

To:

> *There were no significant differences in the frequencies of trustworthy webpages found among the three groups with χ^2_{14}=16.45,P=.29.*

And this sentence was revised from:

> *With regard to the quotient of reliable webpages and all webpages found by students, again, no significant difference was shown ($F_{2,121}$=1.68,PP=) between the groups.*

To:

> *With regard to the quotient of reliable webpages and all webpages found by students, again, no significant difference was shown ($F_{2,121}$=1.68,P=.19) between the groups.*

Additionally, for Table 1, the P values in the far right column, under the heading "Chi-square (*df*)" have also been revised.

The "Histamine testing (wrong)" row was revised from:

> *1.03 (2),PP=*

To:

> *1.03 (2),P=.60*

The "Assessment of diaminoxydase (wrong)" row was revised from:

> *3.55 (2),PP=*

To:

> *3.55 (2),P=.17*

The "Test of urine and feces (wrong)" row was from:

> *0.84 (2), PP=*

To:

> *0.84 (2),P=.66*

The "Nutrition diary (correct)" row was revised from:

> *0.02 (2), PP=*

To:

> *0.02 (2),P=.99*

The "Elimination diet (correct)" row was revised from:

> *7.87 (2), PP=*

To:

> *7.87 (2),P=.02*

The "Provocation test (correct)" row was revised from:

> *0.06 (2), PP=*

To:

> *0.06 (2),P=.97*

The correction will appear in the online version of the paper on the JMIR Publications website on August 11, 2020, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 05.08.20; this is a non–peer-reviewed article;accepted 05.08.20; published 11.08.20.

Please cite as:
Loda T, Erschens R, Junne F, Stengel A, Zipfel S, Herrmann-Werner A
Correction: Undergraduate Medical Students' Search for Health Information Online: Explanatory Cross-Sectional Study
JMIR Med Inform 2020;8(8):e23253
URL: http://medinform.jmir.org/2020/8/e23253/
doi:10.2196/23253
PMID:32780713

Review

Artificial Intelligence for Caregivers of Persons With Alzheimer's Disease and Related Dementias: Systematic Literature Review

Bo Xie[1,2], BSc, MSc, PhD; Cui Tao[3], PhD; Juan Li[4], PhD; Robin C Hilsabeck[5], PhD; Alyssa Aguirre[5], LCSW

[1]School of Nursing, The University of Texas at Austin, Austin, TX, United States

[2]School of Information, The University of Texas at Austin, Austin, TX, United States

[3]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States

[4]Department of Computer Science, North Dakota State University, Fargo, ND, United States

[5]Department of Neurology, Dell Medical School, The University of Texas at Austin, Austin, TX, United States

Corresponding Author:
Bo Xie, BSc, MSc, PhD
School of Nursing
The University of Texas at Austin
1710 Red River
Austin, TX, 78712
United States
Phone: 1 512 232 5788
Email: boxie@utexas.edu

Abstract

Background: Artificial intelligence (AI) has great potential for improving the care of persons with Alzheimer's disease and related dementias (ADRD) and the quality of life of their family caregivers. To date, however, systematic review of the literature on the impact of AI on ADRD management has been lacking.

Objective: This paper aims to (1) identify and examine literature on AI that provides information to facilitate ADRD management by caregivers of individuals diagnosed with ADRD and (2) identify gaps in the literature that suggest future directions for research.

Methods: Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for conducting systematic literature reviews, during August and September 2019, we performed 3 rounds of selection. First, we searched predetermined keywords in PubMed, Cumulative Index to Nursing and Allied Health Literature Plus with Full Text, PsycINFO, IEEE Xplore Digital Library, and the ACM Digital Library. This step generated 113 nonduplicate results. Next, we screened the titles and abstracts of the 113 papers according to inclusion and exclusion criteria, after which 52 papers were excluded and 61 remained. Finally, we screened the full text of the remaining papers to ensure that they met the inclusion or exclusion criteria; 31 papers were excluded, leaving a final sample of 30 papers for analysis.

Results: Of the 30 papers, 20 reported studies that focused on using AI to assist in activities of daily living. A limited number of specific daily activities were targeted. The studies' aims suggested three major purposes: (1) to test the feasibility, usability, or perceptions of prototype AI technology; (2) to generate preliminary data on the technology's performance (primarily accuracy in detecting target events, such as falls); and (3) to understand user needs and preferences for the design and functionality of to-be-developed technology. The majority of the studies were qualitative, with interviews, focus groups, and observation being their most common methods. Cross-sectional surveys were also common, but with small convenience samples. Sample sizes ranged from 6 to 106, with the vast majority on the low end. The majority of the studies were descriptive, exploratory, and lacking theoretical guidance. Many studies reported positive outcomes in favor of their AI technology's feasibility and satisfaction; some studies reported mixed results on these measures. Performance of the technology varied widely across tasks.

Conclusions: These findings call for more systematic designs and evaluations of the feasibility and efficacy of AI-based interventions for caregivers of people with ADRD. These gaps in the research would be best addressed through interdisciplinary collaboration, incorporating complementary expertise from the health sciences and computer science/engineering–related fields.

(JMIR Med Inform 2020;8(8):e18189) doi:10.2196/18189

KEYWORDS

Alzheimer disease; dementia; caregiving; technology; artificial intelligence

Introduction

Alzheimer’s disease and related dementias (ADRD) have become a major public health concern in the United States. An estimated 5.6 million Americans aged 65 and older (10% of the US population) were living with ADRD in 2019, and this number is expected to grow dramatically as the population continues to age. By 2025, the number of Americans aged 65 or older with ADRD is expected to reach 7.1 million, nearly a 27% increase from 2019, and by 2050, this population is projected to be 13.8 million, with the highest growth among those in ADRD’s advanced stage [1].

Persons with ADRD require progressively extensive assistance in their daily lives, the majority of which is provided by family members, friends, and other unpaid caregivers [1]. It is estimated that in 2018, American caregivers of persons with ADRD provided 18.5 billion hours of informal unpaid assistance, valued at $233.9 billion [1]. Family caregivers (hereafter “caregivers”) of persons with ADRD are expected to make important care decisions for their family members with ADRD on a daily basis. However, these caregivers report being unprepared for their roles and responsibilities, uninformed about care options, and unsupported by professionals in their decision making [2-5]. Caregiving for persons with ADRD is stressful [6-10], and it can severely affect the caregiver’s own health and well-being [7]. There is an urgent need to better prepare caregivers to manage their daily lives and those of their family members with ADRD, yet there are critical knowledge gaps regarding the types and amounts of information that caregivers may want to have in order to better manage ADRD. To provide patient-centered care for people with ADRD and enhance caregivers’ quality of life, we must address those gaps.

Artificial intelligence (AI) is showing great promise in areas of health care—in precision treatments, patient education, virtual assistance, and cost reduction [11]. Some attempts have been made to apply AI for persons with ADRD and their caregivers in order to improve patients’ daily functioning, quality of life, and well-being, as well as reduce caregiver burden (eg, social robots to facilitate social interaction and engagement, assistive robots to facilitate daily activities such as handwashing, tea making, or dressing) [12-16]. To date, however, there has been little systematic review to identify research on AI for ADRD management by caregivers and gaps that remain in our understanding of AI for ADRD management. We have conducted this systematic review to identify and examine literature on AI that provides information to facilitate ADRD management by caregivers of individuals diagnosed with ADRD and to identify gaps in the literature that suggest future directions for research.

Methods

Overview

Following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for conducting systematic literature reviews and following procedures used in previous systematic literature reviews [17-19], we performed 3 rounds of search in selected databases. Because this review focuses on AI and ADRD management, we searched databases commonly used for research not only in the health sciences but also in computer science and engineering: PubMed, Cumulative Index to Nursing and Allied Health Literature (CINAHL) Plus with Full Text, PsycINFO, IEEE Xplore Digital Library, and ACM Digital Library. First, we searched titles and abstracts using keywords. Next, we screened the titles and abstracts using inclusion and exclusion criteria. Finally, we screened the papers’ full texts to ensure that they met the inclusion or exclusion criteria.

Round 1: Keyword Search

On August 23, 2019, we searched titles and abstracts in PubMed using the following 3 sets of keywords: (“dementia” OR “Alzheimer”) AND (“caregiver*” OR “proxy” OR “proxies” OR “surrogate*”) AND (“artificial intelligence” OR “intelligent”). These sets of keywords were inclusive but in line with our study’s aims. For the same reason, we did not use built-in limiters in PubMed. This yielded 16 papers. Next, we performed the same search of medical subject heading (MeSH) terms in PubMed, excluding “proxies” and “intelligent,” which are not MeSH terms: (“dementia” OR “Alzheimer”) AND (“caregiver*” OR “proxy” OR “surrogate*”) AND (“artificial intelligence”). This yielded 10 papers. Of the 26 papers from these two searches, 1 was a duplicate, yielding a combined total of 25 nonduplicate papers (4 were reviews; the other 21 reported original data). In addition, we searched both CINAHL Plus with Full Text and PsycINFO, using the same 3 sets of keywords that we used for titles and abstracts in PubMed. CINAHL yielded 10 papers, including 7 duplicates. PsycINFO yielded 10 papers, including 6 duplicates. Excluding duplicates, 7 papers remained, for a total of 32 papers across the 3 health sciences databases.

Again, on September 9, 2019, using the same sets of keywords, a search of all metadata (titles, abstracts, and indexing terms) for all available years in the IEEE Xplore Digital Library yielded 47 papers. We also searched the ACM Digital Library (ACM Full-Text Collection) for abstracts or titles that matched any of the following words or phrases: “Alzheimer’s,” “dementia,” “caregiver,” “proxy,” “proxies,” “surrogate,” “artificial intelligence,” “intelligent.” These results were sorted by relevance, and the first 200 records were manually inspected; this generated 36 papers. No duplicates were found between the ACM and IEEE databases. However, when merged with the first 32 papers from PubMed, CINAHL, and PsycINFO, 2 duplicates were found, yielding a total of 113 nonduplicate papers.

Round 2: Screening of Titles and Abstracts

Next, 3 of the authors (BX, CT, JL) each screened approximately one-third of the titles and abstracts of the 113 papers. The results were cross-examined by the other 2 authors to ensure accuracy and consistency. Differences were resolved through several rounds of discussion. This round of screening was based on the rationale that the focus of our systematic literature review was AI tools that could provide information and service to facilitate ADRD management by caregivers of persons diagnosed with ADRD. Other topics were outside of the scope of our review. Specifically, we removed any paper that met at least one of the following exclusion criteria: (1) primary focus on using artificial intelligence to automatically collect information from users (eg, via sensors), not to provide information to users (n=32); (2) paper did not report empirical data from human participants (eg, literature review, book review, column/commentary, system architecture; n=9); (3) primary focus on screening, identification, or diagnosis of dementia or detecting or modeling anxiety or burnout in caregivers instead of providing services to persons already diagnosed with dementia or their caregivers (n=5); (4) study participants were paid or volunteer caregivers and did not include any family caregivers (n=3); (5) full text not in English (n=3).

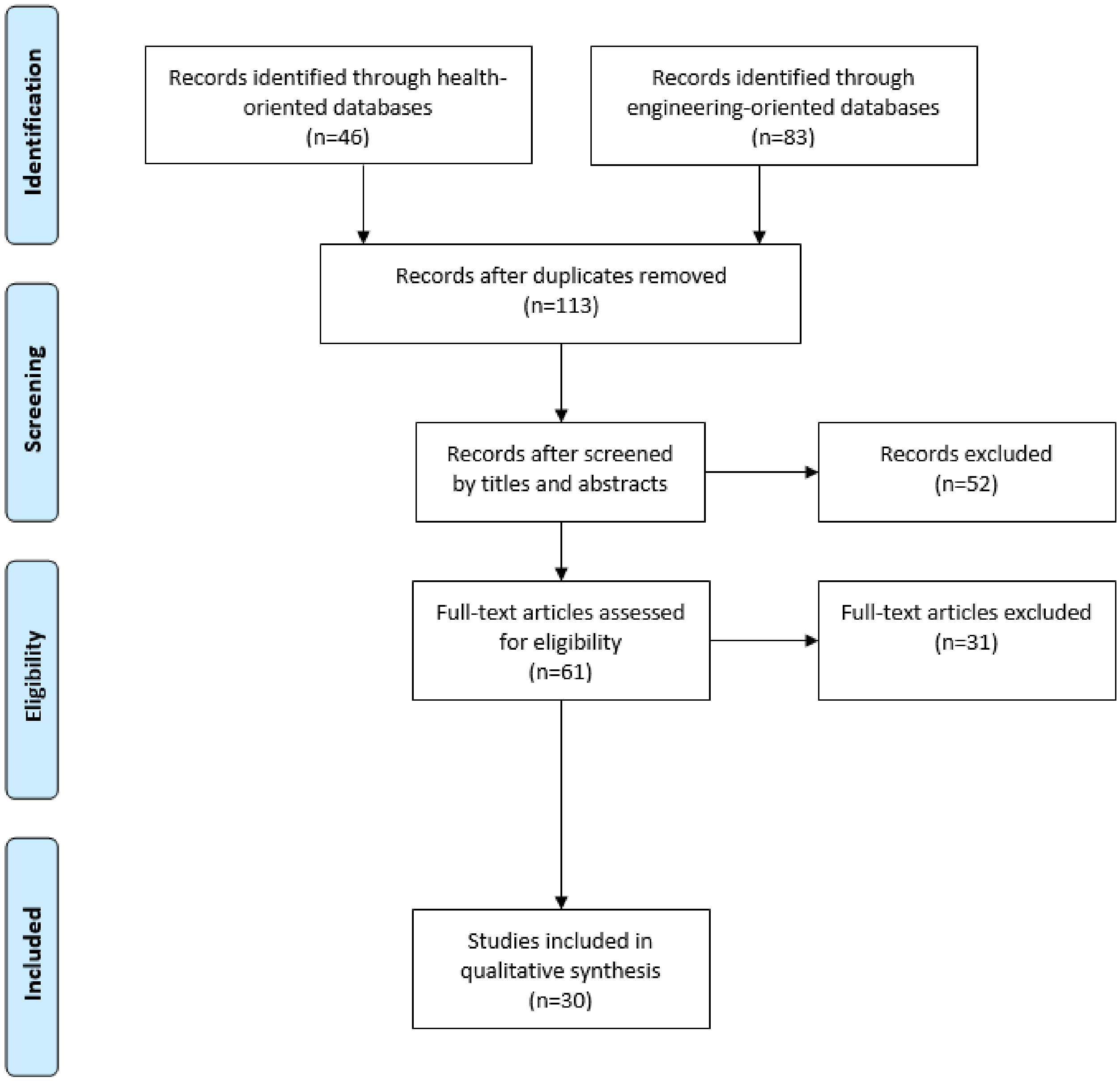
This round of screening resulted in the removal of 52 papers, with 61 papers remaining.

Round 3: Screening of Full Text

In the next round of screening, we eliminated 31 more papers because they met at least one of the aforementioned exclusion criteria: (1) did not report empirical data from human participants (n=18); (2) study participants did not include family caregivers (n=4); (3) primary focus on using artificial intelligence to automatically collect information from users (eg, via sensors), not to provide information to users (n=4); (4) technology under investigation was not artificial intelligence (eg, videogames; n=3); (5) primary focus on screening, identification, or diagnosis of dementia or detecting or modeling anxiety or burnout in caregivers (n=1); (6) report of essentially the same content as in another paper (n=1).

A total of 30 papers remained in the final sample [20-49]. The selection process is summarized in Figure 1 according to the PRISMA guidelines [50].

Figure 1. Search and screening process.


Identification
Screening
Eligibility
Included
Records identified through health-oriented databases (n=46)
Records identified through engineering-oriented databases (n=83)
Records after duplicates removed (n=113)
Records after screened by titles and abstracts
Records excluded (n=52)
Full-text articles assessed for eligibility (n=61)
Full-text articles excluded (n=31)
Studies included in qualitative synthesis (n=30)

Round 4: Coding of Full Text

The 30 papers in our final sample were coded using a framework consistent with our prior work [17-19], summarizing key information from each paper. The coding included each study's publication year, study aim, research method, participant characteristics, sample size, country/area where data collection took place, dosage of AI technology (ie, amount and frequency of time exposed to the AI technology), outcome measures, and key findings. The results of the coding are presented in Multimedia Appendix 1. In addition, we assessed levels of evidence reported in the 30 papers [51].

Results

Our initial searches yielded 113 papers. Through multiple rounds of screening, we removed 83 of them to arrive at our final sample. The reasons for excluding these 83 papers are summarized in Table 1.

Table 1. Summary of the reasons for excluded papers.

Reason for exclusion	n[a]
Primary focus was using AI[b] to automatically collect information from users (eg, sensors), not to provide information to users	36
Did not report empirical data from human participants (eg, literature review, book review, column/commentary, system architecture)	27
Study participants did not include any family caregivers	7
Primary focus was on screening/identification/diagnosis of dementia or detecting/modeling anxiety or burnout in caregivers (instead of providing services to persons already diagnosed with dementia or their caregivers)	6
The technology under investigation was not AI	3
Full text not in English	3
Reporting essentially the same content as another paper (that was already included in the final sample)	1
Total	83

[a] Number of excluded papers.

[b] AI: artificial intelligence.

Key characteristics of the final 30 papers are summarized in Multimedia Appendix 1. The papers were published from 2001 to 2019, averaging 2 per year. The number of publications was consistently low (1 per year) until rising in 2008. The year 2018 had the most papers (4), suggesting an increasing interest in our topic.

AI technologies included in the 30 papers varied. We categorized these technologies according to their intended use. As Table 2 shows, the majority (20/30, 67%) focused on using AI to assist in activities of daily living. A limited number of specific daily activities were targeted in these studies, particularly handwashing, tea making, and dressing.

Table 2. Summary of artificial intelligence technology's intended use.

AI[a] technology use	n[b]
Assist in activities of daily living (eg, assistive robots to aid handwashing, tea making, or dressing)	20
Facilitate social interaction (eg, social robots)	2
Provide cognitive stimulation (eg, computerized activities to stimulate cognition)	2
Ensure safe home environments (eg, smart homes)	2
Educate (eg, through a teleconferencing program or virtual reality platform)	2
Assist in reminiscence therapy	2
Total, N	30

[a] AI: artificial intelligence.

[b] Number of papers.

Aims of the 30 studies fell into one of three major categories: (1) to test the feasibility, usability, or perceptions of a prototype AI technology; (2) to generate preliminary data on the technology's performance (primarily accuracy in detecting target events, such as falls); and (3) to understand user needs and preferences for the design and functionality of to-be-developed technologies.

The majority of these studies used qualitative research methods, with interviews, focus groups, and observation being the 3 most common methods. Cross-sectional surveys were also common, but with small convenience samples. The majority of the studies were descriptive, exploratory, and lacking in theoretical guidance; they were not intended to test theory-informed hypotheses.

The sample sizes of all 30 studies were small, ranging from 6 to 106, with the vast majority on the low end. A total of 7 studies reported data from healthy volunteers or health care professionals but did not include actual patients or caregivers as research participants. We included these studies in our analysis because the technologies under consideration were intended for use by patients and caregivers. One study did not report any participant characteristics.

Nearly half of the studies were conducted in Canada (14/30, 47%), 8 of the 30 (27%) in Europe, and 4 of the 30 (13%) in the United States. Israel, Japan, Mexico, and Taiwan each had one study (1/30, 3%). At least 7 of the 30 studies (23%) were conducted in a research lab; 6 others did not report the setting for data collection. The remaining studies took place in a facility (eg, senior living facility, hospital) or private home.

We also analyzed the AI technology's dosage (ie, the amount of time and frequency that users were exposed to the AI technology in each study). A total of 9 of the 30 studies used interviews or surveys, so the dosage criterion was not applicable to them. Among the 21 studies that involved exposure to AI technology, 5 did not report dosage. The dosages reported in the remaining 16 studies varied widely in terms of both total time and frequency of exposure, ranging from as much as 24/7 access for 4 to 6 weeks or 2 hours per week over 12 weeks to as little as 15 to 20 minutes in a single session.

Outcome measures varied widely as well. Overall, they included both objective and subjective measures. Outcome measures included (1) feasibility, satisfaction, and stress, which were subjective measures; (2) performance, such as the accuracy of AI technology in completing its intended task, measured objectively; (3) usability (self-reported ease of use and perceptions of usefulness); (4) usage patterns (eg, which AI features were used, frequency/duration of usage), also measured objectively; and (5) user needs and requirements for the technology, another set of subjective measures.

Many studies reported positive outcomes in favor of the AI technology being studied (or to be developed) in terms of the technology's feasibility (with acceptability used as the most common measure of feasibility) and satisfaction (positive perceptions of the technology). One of those studies reported a high dropout rate (65%) [22], making it difficult to interpret the study's reported positive outcomes. One study reported preliminary evidence supporting limited efficacy of a social robot in reducing patients' stress [29]. Some studies reported mixed results for feasibility and satisfaction, with some participants reporting that they liked the AI technology but others reporting that they did not [23,26,47,49]. Notably, in 2 separate studies, caregivers reported more positive attitudes than did patients toward the use of AI technology in home care [23,49].

Performance of the technology, measured primarily by accuracy in detecting target events, varied widely across different tasks, ranging from as low as 23% in detecting incorrect dressing events [28] to as high as 98% in detection of falls [46]. In assisting with daily activities, assistive AI devices helped reduce patients' dependence on caregivers [42,43]. Usage patterns also varied widely, ranging from continuous active use to inactive use. Mixed results were reported for the AI technology's features, with some features easier to use and more popular than others [33].

A range of user needs was identified, including needs for assistance in home care, getting information (about time, schedule, care options, etc), and communication and social interactions. There is a great need for AI technology to provide tailored assistance to meet these user needs [37]. However, several factors make it challenging to design tailored technology. These include variation in patients' needs and abilities from day to day and even during the day [39], patients' varying and evolving identities and preferences for a technology's styles and features [40], users' diverse technology literacy levels [41], and challenges associated with ethical issues [48], particularly conflicting needs between caregivers and patients [36] and privacy concerns in assisting in private tasks [31,32].

Regardless of the findings, the levels of evidence [51] of all studies in our final sample were low due to their small convenience samples and exploratory research methods.

Discussion

AI has great potential for improving the care for persons with ADRD and the quality of life of family caregivers. To date, however, there has been little effort to systematically review literature on AI for ADRD management by caregivers and to determine what still needs to be done to understand the impact of AI on ADRD management. In this study, we have addressed those gaps. We have identified work on AI that provides information to facilitate ADRD management by family caregivers of patients diagnosed with ADRD, and we have identified gaps in existing work, which suggest future directions for research. The majority of the AI studies included in our final sample (20/30, 67%) focused on using AI to assist in activities of daily living. A limited number of specific daily activities were targeted. The aims of the 30 studies suggested three major purposes: (1) to test the feasibility, usability, or perceptions of a prototype AI technology; (2) to generate preliminary data on the technology's performance (primarily accuracy in detecting target events, such as falls); and (3) to understand user needs and preferences for the design and functionality of to-be-developed technologies. The majority of these studies used qualitative research methods, with interviews, focus groups, and observation being the 3 most common methods. Cross-sectional surveys were also common, but with small convenience samples. The sample sizes of the 30 studies were small, ranging from 6 to 106, with the vast majority on the low end. The majority of the studies were descriptive, exploratory, and lacking in theoretical guidance. Many studies reported positive outcomes in favor of AI technology's feasibility and user satisfaction; some reported mixed results for these measures. Performance of technology varied widely across different tasks.

Our findings illustrate important characteristics of research to date on the use of AI that provides information to aid ADRD management by family caregivers. First, only a few studies (N=30) have focused on this topic. Given the topic's interdisciplinary nature, we intentionally searched databases commonly used in the health sciences and in computer science/engineering. We found only 2 duplicates between these 2 sets of databases, with more than two-thirds of the studies in the computer science/engineering databases (32 from the health sciences databases, 83 from the computer science/engineering databases). On the topic of AI in ADRD management by caregivers, there was little overlap between the health sciences and computer sciences/engineering databases, suggesting that the latter databases currently contain the majority of existing research. To review developments on this topic, one must examine both sets of databases. Future systematic literature reviews should also track potential changes in the ratio of work found between these sets of databases as an indicator of the maturity of the technology and its applications in health care. It is likely that, as time goes by, when AI technology and its applications in health care are more mature, the research found in the health sciences databases will increase, while that in computer science/engineering databases may decrease (in absolute number or relative ratio).

We also found that a large number of studies (n=36) had the primary focus of using AI to automatically collect information from users (eg, via sensors), which would be used by health care professionals to make care decisions. We did not include those studies in our final sample because our review was meant as a basis for the development of AI-based interventions to provide information to family caregivers (our interdisciplinary team is currently working on such an intervention). However, acknowledging that collecting user information is necessary for providing tailored information, we did include studies that both collected information from and provided information to caregivers. It was beyond the scope of the present review to include studies that focused only on collecting information. Researchers interested in obtaining a full list of those studies may contact the first author for that list.

We also found a large number of papers (n=27) that did not report empirical data from human participants. Some were common types of papers reporting nonempirical data (eg, literature reviews, book reviews, and columns/commentaries), which one would typically expect from searches of health sciences databases (as in prior reviews [17-19]). Characteristically for the present systematic literature review, however, we also found a number of nonempirical studies reporting technical specifics or system architecture for designing AI systems. This is typical of technology development–related work commonly reported in computer science/engineering databases but uncommon in health sciences databases. Further, of the studies that did report empirical data, the majority were descriptive, exploratory, with small convenience samples, and lacking theoretical guidance. Such studies have their own merit and are appropriate for the current stage of research. However, they also show that research on AI for ADRD management is still in the stage of technological development and far from ripe for clinical evaluation. It is premature at this point to systematically examine the efficacy of AI interventions for patients and caregivers.

Consistent with the early stage of research in this area, the aims of the 30 studies in our sample focused on testing the feasibility, usability, or perceptions of prototype AI technologies; generating preliminary data on the technology's performance; and understanding user needs and preferences for the design and functionality of to-be-developed or to-be-revised technologies. Key study findings showed mixed results. Some studies reported promising signs for the acceptability and feasibility of AI tools, but others found challenges that must be addressed before large-scale rollout of AI tools for ADRD management. Notably, the studies in our sample frequently did not report key pieces of information necessary for extraction in health science–oriented systematic reviews, including research participants' demographics, research settings, or even locations where data collection took place. Many of the studies may have been conducted by researchers with training in non–health science fields, such as engineering and computer science, in which reporting norms differ from those commonly used in the health sciences. As a result, systematic review methods and quality criteria commonly used in the health sciences, such as levels of evidence [51], are not easily applicable to current research on AI for ADRD management. This presents an opportunity for interdisciplinary collaboration between researchers in the health sciences and in computer science/engineering–related fields (as is the case for our interdisciplinary team, with expertise in nursing, medicine, and social work on the one hand and in computer science and informatics on the other).

Our systematic review has limitations. We selected only papers with full text in English, so we might have missed cutting-edge studies in other languages. The selection of our initial search terms was also not exhaustive; AI is a broad concept that includes technologies that may be labeled under different terms but are nonetheless still AI based. By using only "artificial intelligence" or "intelligent" as our AI-related search terms, we might have missed technologies that did not use these terms but did use AI (eg, expert systems, decision aids). However, a merit of our approach is that it allowed us to focus on publications self-labeled by their authors as AI-related work. By using "artificial intelligence" and "intelligent" as our AI-related search terms, we were able to focus on studies defined by their authors as reporting AI-related technology and thus to identify researchers who self-identify as AI researchers. Overall, our review has identified work on AI that provides information to facilitate ADRD management by caregivers, as well as gaps in the literature that require future research. These findings call for more systematic designs and evaluations of the feasibility and efficacy of AI-based interventions for caregivers. Such tasks will be best addressed through interdisciplinary collaboration incorporating complementary expertise from the health sciences and computer science/engineering–related fields.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of the 30 studies in the final sample.
[DOCX File , 27 KB - medinform_v8i8e18189_app1.docx]

References

1. Alzheimer's Association. 2019 Alzheimer's disease facts and figures. Alzheimer's & Dementia 2019 Mar 01;15(3):321-387 [FREE Full text] [doi: 10.1016/j.jalz.2019.01.010]
2. Collopy B. The moral underpinning of the proxy-provider relationship: issues of trust and distrust. J Law Med Ethics 1999;27(1):37-45 [FREE Full text] [doi: 10.1111/j.1748-720x.1999.tb01434.x] [Medline: 11657141]
3. Ditto P, Danks J, Smucker W, Bookwala J, Coppola KM, Dresser R, et al. Advance directives as acts of communication: a randomized controlled trial. Arch Intern Med 2001 Mar 12;161(3):421-430. [doi: 10.1001/archinte.161.3.421] [Medline: 11176768]
4. Gessert C, Forbes S, Bern-Klug M. Planning end-of-life care for patients with dementia: roles of families and health professionals. Omega (Westport) 2000;42(4):273-291. [doi: 10.2190/2mt2-5gyu-gxvv-95ne] [Medline: 12569923]
5. Swigart V, Lidz C, Butterworth V, Arnold R. Letting go: family willingness to forgo life support. Heart Lung 1996;25(6):483-494. [doi: 10.1016/s0147-9563(96)80051-3] [Medline: 8950128]
6. Hanson L, Carey T, Caprio A, Lee TJ, Ersek M, Garrett J, et al. Improving decision-making for feeding options in advanced dementia: a randomized, controlled trial. J Am Geriatr Soc 2011 Nov;59(11):2009-2016 [FREE Full text] [doi: 10.1111/j.1532-5415.2011.03629.x] [Medline: 22091750]
7. Stirling C, Leggett S, Lloyd B, Scott J, Blizzard L, Quinn S, et al. Decision aids for respite service choices by carers of people with dementia: development and pilot RCT. BMC Med Inform Decis Mak 2012 Mar 19;12:21 [FREE Full text] [doi: 10.1186/1472-6947-12-21] [Medline: 22429384]
8. Bonner G, Wang E, Wilkie D, Ferrans C, Dancy B, Watkins Y. Advance care treatment plan (ACT-Plan) for African American family caregivers: a pilot study. Dementia (London) 2014 Jan;13(1):79-95 [FREE Full text] [doi: 10.1177/1471301212449408] [Medline: 24381040]
9. Einterz S, Gilliam R, Lin F, McBride J, Hanson L. Development and testing of a decision aid on goals of care for advanced dementia. J Am Med Dir Assoc 2014 Apr;15(4):251-255 [FREE Full text] [doi: 10.1016/j.jamda.2013.11.020] [Medline: 24508326]
10. Jox R, Denke E, Hamann J, Mendel R, Förstl H, Borasio G. Surrogate decision making for patients with end-stage dementia. Int J Geriatr Psychiatry 2012 Oct;27(10):1045-1052 [FREE Full text] [doi: 10.1002/gps.2820] [Medline: 22139621]
11. Matheny M, Whicher D, Thadaney Israni S. Artificial Intelligence in Health Care: A Report From the National Academy of Medicine. JAMA 2019 Dec 17;323(6):509-510. [doi: 10.1001/jama.2019.21579] [Medline: 31845963]
12. Maresova P, Tomsone S, Lameski P, Madureira J, Mendes A, Zdravevski E, et al. Technological Solutions for Older People with Alzheimer's Disease: Review. Curr Alzheimer Res 2018;15(10):975-983 [FREE Full text] [doi: 10.2174/1567205015666180427124547] [Medline: 29701154]
13. Moyle W, Arnautovska U, Ownsworth T, Jones C. Potential of telepresence robots to enhance social connectedness in older adults with dementia: an integrative review of feasibility. Int Psychogeriatr 2017 Dec;29(12):1951-1964. [doi: 10.1017/S1041610217001776] [Medline: 28879828]
14. Gagnon-Roy M, Bourget A, Stocco S, Courchesne A, Kuhne N, Provencher V. Assistive Technology Addressing Safety Issues in Dementia: A Scoping Review. Am J Occup Ther 2017;71(5):7105190020p1-7105190020p10. [doi: 10.5014/ajot.2017.025817] [Medline: 28809655]
15. Chang S, Sung H. The effectiveness of seal-like robot therapy on mood and social interactions of older adults: a systematic review protocol. JBI Database of Systematic Reviews and Implementation Reports 2013;11(10):68-75 [FREE Full text] [doi: 10.11124/jbisrir-2013-914]
16. Bharucha A, Anand V, Forlizzi J, Dew MA, Reynolds CF, Stevens S, et al. Intelligent assistive technology applications to dementia care: current capabilities, limitations, and future challenges. Am J Geriatr Psychiatry 2009 Mar;17(2):88-104 [FREE Full text] [doi: 10.1097/JGP.0b013e318187dde5] [Medline: 18849532]
17. Xie B, Berkley A, Kwak J, Fleischmann K, Champion J, Koltai K. End-of-life decision making by family caregivers of persons with advanced dementia: A literature review of decision aids. SAGE Open Med 2018;6:2050312118777517 [FREE Full text] [doi: 10.1177/2050312118777517] [Medline: 29844911]
18. Watkins I, Xie B. eHealth literacy interventions for older adults: a systematic review of the literature. J Med Internet Res 2014 Nov 10;16(11):e225 [FREE Full text] [doi: 10.2196/jmir.3318] [Medline: 25386719]
19. Xie B, Huang M, Watkins I. Technology and retirement life: a systematic review of the literature on older adults and social media. In: Wang M, editor. The Oxford Handbook of Retirement. Oxford, UK: Oxford University Press; 2012:493-509.
20. Abdollahi H, Mollahosseini A, Lane J, Mahoor M. A pilot study on using an intelligent life-like robot as a companion for elderly individuals with dementia and depression. 2017 Presented at: 2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids); Nov 15-17, 2017; Birmingham, UK URL: https://doi.org/10.1109/HUMANOIDS.2017.8246925 [doi: 10.1109/humanoids.2017.8246925]

21. Amiribesheli M, Bouchachia A. Smart homes design for people with dementia. 2015 Presented at: International Conference on Intelligent Environments; July 15-17, 2015; Prague, Czech Republic URL: https://doi.org/10.1109/IE.2015.33 [doi: 10.1109/ie.2015.33]
22. Apostolidis I, Karakostas A, Dimitriou T, Tsiatsos T, Tsolaki M. Advanced bio-feedback and collaborative techniques to support caregivers of Alzheimer patients. 2014 Presented at: 2014 International Conference on Intelligent Networking and Collaborative Systems; Sep 10-12, 2014; Salerno, Italy URL: https://doi.org/10.1109/INCoS.2014.88 [doi: 10.1109/incos.2014.88]
23. Begum M, Wang R, Huq R, Mihalidis A. Performance of daily activities by older adults with dementia: The role of an assistive robot. 2013 Presented at: 2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR); June 24-26, 2013; Seattle, WA URL: https://doi.org/10.1109/ICORR.2013.6650405 [doi: 10.1109/icorr.2013.6650405]
24. Begum M, Huq R, Wang R, Mihailidis A. Collaboration of an assistive robot and older adults with dementia. Gerontechnology 2015;13(4):405-419 [FREE Full text] [doi: 10.4017/gt.2015.13.4.005.00]
25. Berenbaum R, Lange Y, Abramowitz L. Augmentative alternative communication for Alzheimer's patients and families using SAVION. : ACM; 2011 Presented at: 4th International Conference on Pervasive Technologies Related to Assistive Environments; May 25-27, 2010; Crete, Greece p. 1-4 URL: https://doi.org/10.1145/2141622.2141677 [doi: 10.1145/2141622.2141677]
26. Berezina-Blackburn V, Oliszewski A, Cleaver D, Udakandage L. Virtual reality performance platform for learning about dementia. 2018 Presented at: 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing; November 3-7, 2018; New York, NY p. 153-156 URL: https://doi.org/10.1145/3272973.3274043 [doi: 10.1145/3272973.3274043]
27. Boger J, Hoey J, Poupart P, Boutilier C, Fernie G, Mihailidis A. A planning system based on Markov decision processes to guide people with dementia through activities of daily living. IEEE Trans Inf Technol Biomed 2006 Apr;10(2):323-333. [doi: 10.1109/titb.2006.864480] [Medline: 16617621]
28. Burleson W, Lozano C, Ravishankar V, Lee J, Mahoney D. An Assistive Technology System that Provides Personalized Dressing Support for People Living with Dementia: Capability Study. JMIR Med Inform 2018 May 01;6(2):e21 [FREE Full text] [doi: 10.2196/medinform.5587] [Medline: 29716885]
29. Chan J, Nejat G. Minimizing task-induced stress in cognitively stimulating activities using an intelligent socially assistive robot. : IEEE; 2011 Presented at: IEEE International Workshop on Robot and Human communicaton; July 31-August 3, 2011; Atlanta, GA URL: https://doi.org/10.1109/ROMAN.2011.6005275 [doi: 10.1109/roman.2011.6005275]
30. Chen W. MGuider: mobile guiding and tracking system in public transit system for individuals with cognitive impairments. New York, NY: Association for Computing Machinery; 2009 Presented at: ASSETS09: The 11th International ACM SIGACCESS Conference on Computers and Accessibility; October 2009; Pittsburgh, PA p. 271-272 URL: https://doi.org/10.1145/1639642.1639711 [doi: 10.1145/1639642.1639711]
31. Czarnuch S, Mihailidis A. The design of intelligent in-home assistive technologies: assessing the needs of older adults with dementia and their caregivers. Gerontechnology 2011;10(3):169-182 [FREE Full text] [doi: 10.4017/gt.2011.10.3.005.00]
32. de Jong M, Stara V, von Döllen V, Bolliger D, Heerink M. Users requirements in the design of a virtual agent for patients with dementia and their caregivers. 2018 Presented at: Goodtechs '18: Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good; November 28-30, 2018; Bologna, Italy p. 136-141 URL: https://doi.org/10.1145/3284869.3284899 [doi: 10.1145/3284869.3284899]
33. Edmeads J, Metatla O. Designing for reminiscence with people with dementia. 2019 Presented at: ACM CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, UK URL: https://doi.org/10.1145/3290607.3313059 [doi: 10.1145/3290607.3313059]
34. Hawkey K, Inkpen K, Rockwood K, McAllister M, Slonim J. Requirements gathering with Alzheimer's patients and caregivers. : ACM; 2005 Presented at: 7th International ACM SIGACCESS Conference on Computers and Accessibility; Oct 2005; Baltimore, MD URL: https://doi.org/10.1145/1090785.1090812 [doi: 10.1145/1090785.1090812]
35. Horwitz CM, Mueller M, Wiley D, Tentler A, Bocko M, Chen L, et al. Is home health technology adequate for proactive self-care? Methods Inf Med 2008;47(1):58-62 [FREE Full text] [doi: 10.3414/me9101] [Medline: 18213429]
36. Hwang A, Truong K, Mihailidis A. Using participatory design to determine the needs of informal caregivers for smart home user interfaces. 2012 Presented at: 6th International Conference on Pervasive Computing Technologies for Healthcare; May 21-24, 2012; San Diego, CA URL: https://doi.org/10.4108/icst.pervasivehealth.2012.248671 [doi: 10.4108/icst.pervasivehealth.2012.248671]
37. Hwang A, Truong K, Cameron J, Lindqvist E, Nygård L, Mihailidis A. Co-Designing Ambient Assisted Living (AAL) Environments: Unravelling the Situated Context of Informal Dementia Care. Biomed Res Int 2015;2015:720483 [FREE Full text] [doi: 10.1155/2015/720483] [Medline: 26161410]
38. Jean-Baptiste E, Mihailidis A. Analysis and comparison of two task models in a partially observable Markov decision process based assistive system. : IEEE; 2017 Presented at: 2017 IEEE 4th International Conference on Soft Computing & Machine Intelligence (ISCMI); Nov 23-24, 2017; Port Louis, Mauritius URL: https://doi.org/10.1109/ISCMI.2017.8279623 [doi: 10.1109/iscmi.2017.8279623]

39. Klein P, Uhlig M. Interactive Memories: technology-aided reminiscence therapy for people with dementia. : ACM; 2016 Presented at: 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments; June 2016; Corfu, Greece URL: https://doi.org/10.1145/2910674.2935838 [doi: 10.1145/2910674.2935838]
40. König A, Francis L, Joshi J, Robillard J, Hoey J. Qualitative study of affective identities in dementia patients for the design of cognitive assistive technologies. J Rehabil Assist Technol Eng 2017;4:2055668316685038 [FREE Full text] [doi: 10.1177/2055668316685038] [Medline: 31186921]
41. López J, Martin D, Moreno F. Acceptance of cognitive games through smart tv applications in patients with Parkinson's disease. : ACM; 2018 Presented at: Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference; June 2018; Corfu, Greece p. 428-433 URL: https://doi.org/10.1145/3197768.3201553 [doi: 10.1145/3197768.3201553]
42. Mihailidis A, Barbenel J, Fernie G. The efficacy of an intelligent cognitive orthosis to facilitate handwashing by persons with moderate to severe dementia. Neuropsychological Rehabilitation 2004 Mar;14(1-2):135-171 [FREE Full text] [doi: 10.1080/09602010343000156]
43. Mihailidis A, Boger J, Craig T, Hoey J. The COACH prompting system to assist older adults with dementia through handwashing: an efficacy study. BMC Geriatr 2008 Nov 07;8:28 [FREE Full text] [doi: 10.1186/1471-2318-8-28] [Medline: 18992135]
44. Mihailidis A, Fernie G, Barbenel J. The use of artificial intelligence in the design of an intelligent cognitive orthosis for people with dementia. Assist Technol 2001;13(1):23-39. [doi: 10.1080/10400435.2001.10132031] [Medline: 12212434]
45. Navarro R, Rodriguez MD, Favela J. Intervention Tailoring in Augmented Cognition Systems for Elders With Dementia. IEEE J Biomed Health Inform 2014 Jan;18(1):361-367 [FREE Full text] [doi: 10.1109/jbhi.2013.2267542]
46. Osamu T, Ryu T, Hayashida A. A smart system for home monitoring of people with cognitive impairment. 2014 Presented at: 2014 IEEE Canada International Humanitarian Technology Conference; June 1-4, 2014; Montreal, QC, Canada URL: https://doi.org/10.1109/IHTC.2014.7147550 [doi: 10.1109/ihtc.2014.7147550]
47. Rialle V, Ollivet C, Guigui C, Hervé C. What do family caregivers of Alzheimer's disease patients desire in smart home technologies? Contrasted results of a wide survey. Methods Inf Med 2008;47(1):63-69. [doi: 10.3414/ME9102] [Medline: 18213430]
48. Simão H, Guerreiro T. MATY: Designing an assistive robot for people with Alzheimer's. In: CHI EA '19: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems.: ACM; 2019 Presented at: 2019 CHI Conference on Human Factors in Computing Systems; May 2019; Glasgow, Scotland URL: https://doi.org/10.1145/3290607.3313016 [doi: 10.1145/3290607.3313016]
49. Wang R, Sudhama A, Begum M, Huq R, Mihailidis A. Robots to assist daily activities: views of older adults with Alzheimer's disease and their caregivers. Int Psychogeriatr 2017 Jan;29(1):67-79. [doi: 10.1017/S1041610216001435] [Medline: 27660047]
50. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
51. Burns P, Rohrich R, Chung K. The levels of evidence and their role in evidence-based medicine. Plast Reconstr Surg 2011 Jul;128(1):305-310 [FREE Full text] [doi: 10.1097/PRS.0b013e318219c171] [Medline: 21701348]

Abbreviations

ADRD: Alzheimer’s disease and related dementias
AI: artificial intelligence
CINAHL: Cumulative Index to Nursing and Allied Health Literature
MeSH: medical subject heading
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by C Lovis; submitted 10.02.20; peer-reviewed by X Yuan, C Kerssens, X Jing; comments to author 11.05.20; revised version received 20.05.20; accepted 21.06.20; published 20.08.20.

Please cite as:
Xie B, Tao C, Li J, Hilsabeck RC, Aguirre A
Artificial Intelligence for Caregivers of Persons With Alzheimer’s Disease and Related Dementias: Systematic Literature Review
JMIR Med Inform 2020;8(8):e18189
URL: http://medinform.jmir.org/2020/8/e18189/
doi:10.2196/18189
PMID:32663146

Original Paper

Nationwide Results of COVID-19 Contact Tracing in South Korea: Individual Participant Data From an Epidemiological Survey

Seung Won Lee[1*], MD, PhD; Woon Tak Yuh[2*], MD; Jee Myung Yang[3*], MD, PhD; Yoon-Sik Cho[1], PhD; In Kyung Yoo[4], MD, PhD; Hyun Yong Koh[5], MD, PhD; Dominic Marshall[6], MD; Donghwan Oh[7], MD; Eun Kyo Ha[8], MD; Man Yong Han[9], MD; Dong Keon Yon[9,10], MD

[1]Department of Data Science, Sejong University College of Software Convergence, Seoul, Republic of Korea
[2]Department of Neurosurgery, Seoul National University Hospital, Seoul, Republic of Korea
[3]Department of Ophthalmology, Asan Medical Center, Seoul, Republic of Korea
[4]Department of Gastroenterology, CHA Bundang Medical Center, Seongnam, Republic of Korea
[5]FM Kirby Neurobiology Center, Boston Children's Hospital, Harvard Medical School, Boston, MA, United States
[6]Critical Care Research Group, Nuffield Department of Clinical Neurosciences, Oxford, United Kingdom
[7]Department of Internal Medicine, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea
[8]Department of Pediatrics, Kangnam Sacred Heart Hospital, Hallym University College of Medicine, Seoul, Republic of Korea
[9]Department of Pediatrics, CHA Bundang Medical Center, CHA University School of Medicine, Seongnam, Republic of Korea
[10]Armed Force Medical Command, Republic of Korea Armed Forces, Seongnam, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Dong Keon Yon, MD
Armed Force Medical Command
Republic of Korea Armed Forces
81 Saemaeul-ro 177
Seongnam, 463-040
Republic of Korea
Phone: 82 2 6935 2476
Fax: 82 504 478 0201
Email: yonkkang@gmail.com

Abstract

Background: Evidence regarding the effectiveness of contact tracing of COVID-19 and the related social distancing is limited and inconclusive.

Objective: This study aims to investigate the epidemiological characteristics of SARS-CoV-2 transmission in South Korea and evaluate whether a social distancing campaign is effective in mitigating the spread of COVID-19.

Methods: We used contract tracing data to investigate the epidemic characteristics of SARS-CoV-2 transmission in South Korea and evaluate whether a social distancing campaign was effective in mitigating the spread of COVID-19. We calculated the mortality rate for COVID-19 by infection type (cluster vs noncluster) and tested whether new confirmed COVID-19 trends changed after a social distancing campaign.

Results: There were 2537 patients with confirmed COVID-19 who completed the epidemiologic survey: 1305 (51.4%) cluster cases and 1232 (48.6%) noncluster cases. The mortality rate was significantly higher in cluster cases linked to medical facilities (11/143, 7.70% vs 5/1232, 0.41%; adjusted percentage difference 7.99%; 95% CI 5.83 to 10.14) and long-term care facilities (19/221, 8.60% vs 5/1232, 0.41%; adjusted percentage difference 7.56%; 95% CI 5.66 to 9.47) than in noncluster cases. The change in trends of newly confirmed COVID-19 cases before and after the social distancing campaign was significantly negative in the entire cohort (adjusted trend difference –2.28; 95% CI –3.88 to –0.68) and the cluster infection group (adjusted trend difference –0.96; 95% CI –1.83 to –0.09).

Conclusions: In a nationwide contact tracing study in South Korea, COVID-19 linked to medical and long-term care facilities significantly increased the risk of mortality compared to noncluster COVID-19. A social distancing campaign decreased the spread of COVID-19 in South Korea and differentially affected cluster infections of SARS-CoV-2.

(JMIR Med Inform 2020;8(8):e20992) doi:10.2196/20992

KEYWORDS

COVID-19; contact tracing; coronavirus; South Korea; survey; health data; epidemiology; transmission

Introduction

The novel coronavirus that emerged in Wuhan, China, termed SARS-CoV-2, has caused a rapidly spreading outbreak of COVID-19 worldwide [1,2]. As of April 7, 2020, there were 1,279,722 human COVID-19 cases and 72,614 deaths worldwide [3], prompting public health interventions that mitigate transmission of the pandemic such as wearing face masks, practicing social distancing, and following home confinement recommendations. As China is a unitary one-party socialist republic with strong governmental control, entire cities in the Wuhan Province were locked down and underwent aggressive measures that brought the epidemic under control [1,4]. However, little is known about public health interventions in democratic countries.

The democratic republic of South Korea, one of the geographical neighbors of China, had the second highest number of COVID-19 cases until February 2020 [5]. However, with a well-organized testing program, contact tracing, strict case isolation, and public cooperation that included wearing masks and washing hands, Korea has emerged as a model country with exemplary public health interventions [6]. As of April 11, 2020, COVID-19 cases have dropped sharply, and only 30 new infections have been reported in South Korea since. Further, there have been no new infections in the Daegu Region, which had the highest proportion of COVID-19 cases (65% of South Korea's total number of cases) [3]. Therefore, epidemiological data and experience regarding the characteristics of SARS-CoV-2 transmission in Korea are valuable to find the right strategies to combat COVID-19.

Based on the experience with the Middle East respiratory syndrome (MERS) outbreak, South Korea has set up a novel monitoring system to collect information and manage patients with COVID-19 and their contacts by using GPD (cell phone location), card transaction logs, closed-circuit television (CCTV), and a history of medical facility use [7]. Using data acquired by this monitoring system, we investigated the epidemiological characteristics of SARS-CoV-2 transmission in South Korea and evaluated whether the social distancing campaign is effective in mitigating the spread of COVID-19.

Methods

Data Collection

Data were collected from individuals with laboratory-confirmed SARS-CoV-2 infection who subsequently completed the preliminary epidemiological surveillance conducted by each local government of South Korea (Seoul, Incheon, Sejong, Daegu, Gwangju, Ulsan, Busan, Gyeonggi-do, Gangwon-do, Chungcheongbuk-do [Chungbuk], Chungcheongnam-do [Chungnam], Gyeongsangbuk-do [Gyeongbuk], Gyeongsangnam-do [Gyeongnam], Jeollabuk-do [Jeonbuk], Jeollanam-do [Jeonnam], and Jeju) [8-12] and the Korea Centers for Disease Control and Prevention (KCDC) between January 19, 2020, and April 7, 2020. Epidemiological surveillance data were collected by epidemic intelligence service officers of each local government and the KCDC using the novel monitoring system that uses GPS (cell phone location), card transaction logs, CCTV, and a history medical facilities use. The study protocol was approved by the Institutional Review Board of Sejong University (SJU-HR-E-2020-003) and written informed consent was waived by the ethics commission, owing to the urgent need to collect data.

A cluster infection was defined as a group of similar COVID-19 cases that occurred in the same area during a short time interval. Nonclustered cases were patients with COVID-19 unrelated to any other patients with COVID-19 in time or place [13]. Laboratory confirmation of SARS-CoV-2 infection was defined as a positive result of real-time reverse transcriptase polymerase chain reaction assay of nasal or pharyngeal swabs, in agreement with the World Health Organization (WHO) guideline [14]. Information on age, sex, region of residence, and infection route was obtained for each participant. Death data as of April 7, 2020, were obtained by the KCDC.

Statistical Analysis

We set January 19, 2020, as the index date (epidemiologic day 1) and April 7, 2020, as epidemiologic day 80. The primary endpoint was the mortality risk among participants with noncluster infection and those with cluster infection. Analysis of covariance was used to calculate the adjusted mean difference and 95% CI after adjustment. The following factors were considered potential confounders: age (0-19 years, 20-39 years, 40-59 years, and 60 years or older), sex, diagnosis date, and region of residence (urban [Seoul, Incheon, Sejong, Daegu, Gwangju, Ulsan, and Busan] vs rural [Gyeonggi-do, Gangwon-do, Chungbuk, Chungnam, Gyeongbuk, Gyeongnam, Jeonbuk, Jeonnam, and Jeju]).

Our secondary endpoint was whether a social distancing campaign was effective in mitigating the spread of COVID-19. We divided the population into two distinct periods: before the social distancing campaign (January 19, 2020, to March 22, 2020) and after the social distancing campaign (March 23, 2020, to April 7, 2020). We tested whether trends in newly confirmed COVID-19 cases changed after the social distancing campaign compared with those before the campaign. We implemented interrupted time series analysis to detect a change of slope after the launch of the nationwide social distancing campaign. We introduced the following equation to compare the effect of the campaign, where:

Y_t is the newly infected person on day t; T is the number of days elapsed from the first confirmed infectious case; is the breakpoint day with the day when the nationwide social distancing campaign was launched (64); α_0 is the number of infected patients on the first day of the infection; α_1 is the slope of novel cases per day before the campaign; α_2 is the newly infected cases at the start of the campaign compared to α_0; α_3 is the difference in novel infection rate before and after launching the campaign. Therefore, $\alpha_1 + \alpha_3$ is the trend of the number of daily new infections after the onset of the campaign. X_{rt}, X_{at}, and X_{st} are vectors each containing region specificity, age distribution, and gender composition of the patients on day t, and β_1, β_2, and β_3 are the proportional coefficients of each covariate vector. DOW_t is the day of the week (eg, Saturday) on day t, γ is its coefficient, and e_t is an error term.

Network visualization was performed using Gephi version 0.9.2 [15]. The relative positions of nodes and edges were implemented by the Fruchterman-Reingold algorithm [16]. The algorithm would optimally draw the whole layout of the graph to cluster similar nodes and simplify the path of edges to express the transmission routes clearer. Next, we added the “Nooverlap” option to increase visibility further. The dot represented an individual and the line represented an individual tracing result. Larger dots represent clustered infections, where size was proportional to the number of infected individuals. Overseas influx and influx of community-acquired infections (Daegu and two cities in Gyeongbuk [Cheongdo and Gyeongsan]) connected 641 and 229 dots, respectively.

Each categorical value is reported as the number of patients (percentage). Statistical analyses were performed using SPSS version 25.0 (IBM Corp), and R software version 3.6.2 (R Foundation for Statistical Computing). A two-sided *P* value<.05 was considered statistically significant.

Patient and Public Involvement

No patients were directly involved in designing the research question or conducting the research. No patients were asked to interpret or write up the results. There are no plans to involve patients or relevant patient communities in dissemination at this moment.

Results

From January 19, 2020, to April 7, 2020, there were 10,046 patients with laboratory-confirmed COVID-19 in South Korea. Among the 10,046 patients, 7509 were excluded for the following reasons: epidemiological investigation was not possible due to community-level outbreaks (Daegu and two cities in Gyeongbuk [Cheongdo and Gyeongsan]; n=7493) or because the epidemiological investigation was incomplete (n=16). The final sample size was 2537 (1160 men and 1377 women; Figure 1).

Figure 1. Our study population in each region (number of our study population/number of total patients with confirmed COVID-19). Of 9550 patients with confirmed COVID-19, there were 2134 patients with confirmed COVID-19 who completed the epidemiological surveillance.

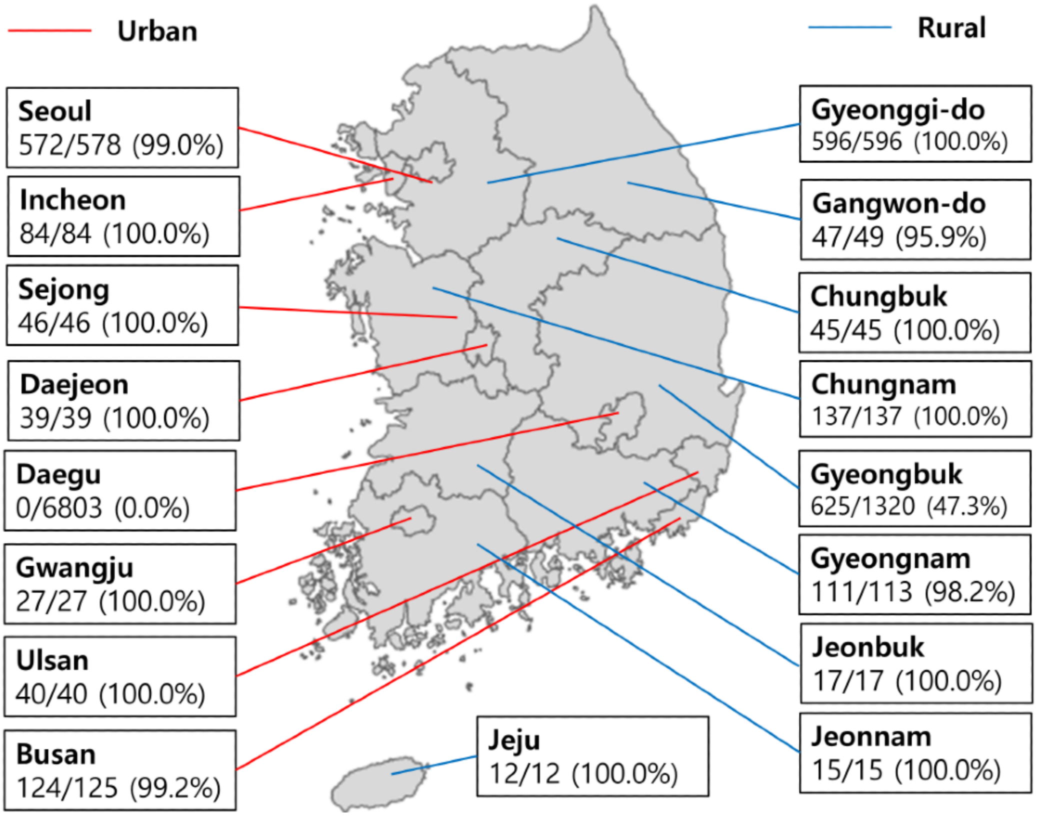


Table 1 shows the demographic characteristics of the participants. There were 1305 cluster cases (51.4%) and 1232 noncluster cases (48.6%; Figure 2). Cluster cases were linked to medical facilities (n=143, 5.6%), long-term care facilities (n=221, 8.7%), religious facilities (n=486, 19.2%), and other locations (n=455, 17.9%), which included military units, dance studios, karaoke bars, internet cafés, public transport, prisons, and the workplaces of each patient. Noncluster cases were linked to the overseas influx (n=641, 25.3%), influx in community-infection outbreak areas (n=229, 9.0%), and sporadic cases (n=362, 14.3%). Figure 3 and Multimedia Appendix 1 show the infection spread network visualization of COVID-19.

Table 1. Demographic characteristics of patients with confirmed COVID-19 in South Korea.

Characteristic	Entire cohort, n (%)	Cluster and contact cases, n (%)				Noncluster cases[a], n (%)
		Linked to medical facilities	Linked to long-term care facilities	Linked to religious facilities	Others[b]	
Patients	2537 (100)	143 (5.6)	221 (8.7)	486 (19.2)	455 (17.9)	1232 (48.6)
Age (years)						
0-19	151 (6.0)	3 (2.1)	2 (0.9)	33 (6.8)	35 (7.7)	78 (6.3)
20-39	974 (38.4)	28 (19.6)	15 (6.8)	196 (40.3)	123 (27.0)	612 (49.7)
40-59	805 (31.7)	44 (30.8)	39 (17.6)	162 (33.3)	240 (52.7)	320 (26.0)
≥60	607 (23.9)	68 (47.6)	165 (74.7)	95 (19.5)	57 (12.5)	222 (18.0)
Sex						
Male	1160 (45.7)	49 (34.3)	68 (30.8)	228 (46.9)	176 (38.7)	639 (51.9)
Female	1377 (54.3)	94 (65.7)	153 (69.2)	258 (53.1)	279 (61.3)	593 (48.1)
Region of residence						
Urban	934 (36.8)	25 (17.5)	8 (3.6)	146 (30.0)	210 (46.2)	545 (44.2)
Rural	1603 (63.2)	118 (82.5)	213 (96.4)	340 (70.0)	245 (53.8)	687 (55.8)
Died						
No	2500 (98.5)	132 (92.3)	202 (91.4)	485 (99.8)	454 (99.8)	1227 (99.6)
Yes	37 (1.5)	11 (7.7)	19 (8.6)	1 (0.2)	1 (0.2)	5 (0.4)

[a] Noncluster cases were linked to overseas influx (641/2537, 25.3%), influx for community-infection outbreak areas (229/2537, 9.0%), and sporadic cases (362/2537, 14.3%).

[b] Other facilities included military units, dance studios, karaoke, internet cafés, public transport, prisons, and workplaces of each patient.

Figure 2. Number of infections based on infection type (cluster and contact cases vs noncluster cases) in South Korea from January 19, 2020, to April 7, 2020.

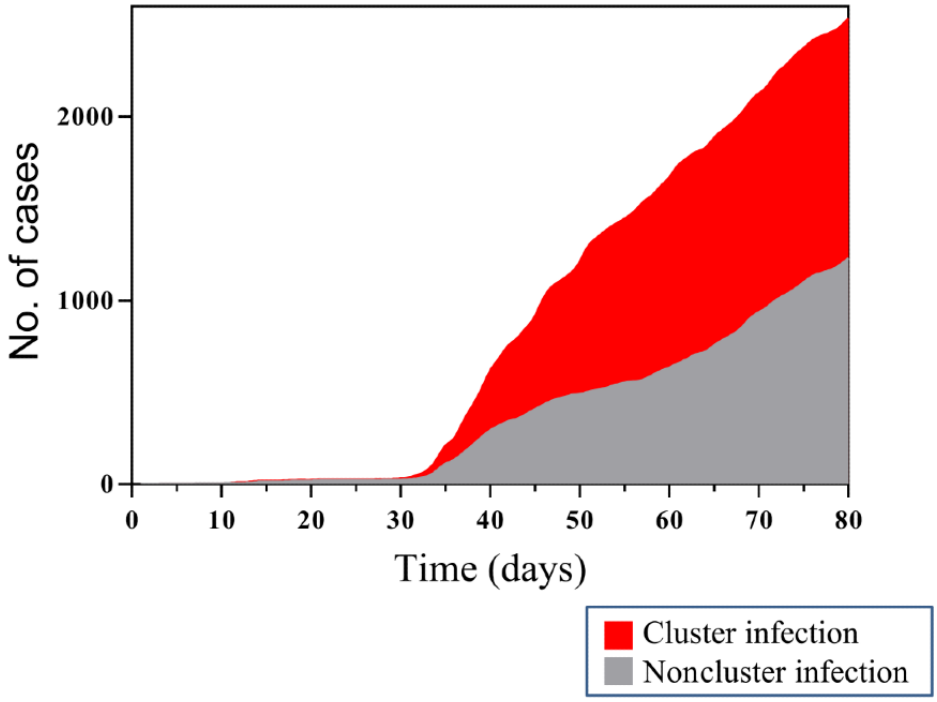


Figure 3. Infection spread network visualization of COVID-19 in South Korea from January 19, 2020, to April 7, 2020. Each dot represents an individual, and each line represents an individual's tracing results. Overseas influx and influx of community-acquired infections (Daegu and two cities in Gyeongbuk [Cheongdo and Gyeongsan]) are shown by 641 and 229 connected dots, respectively. CA: community-acquired.

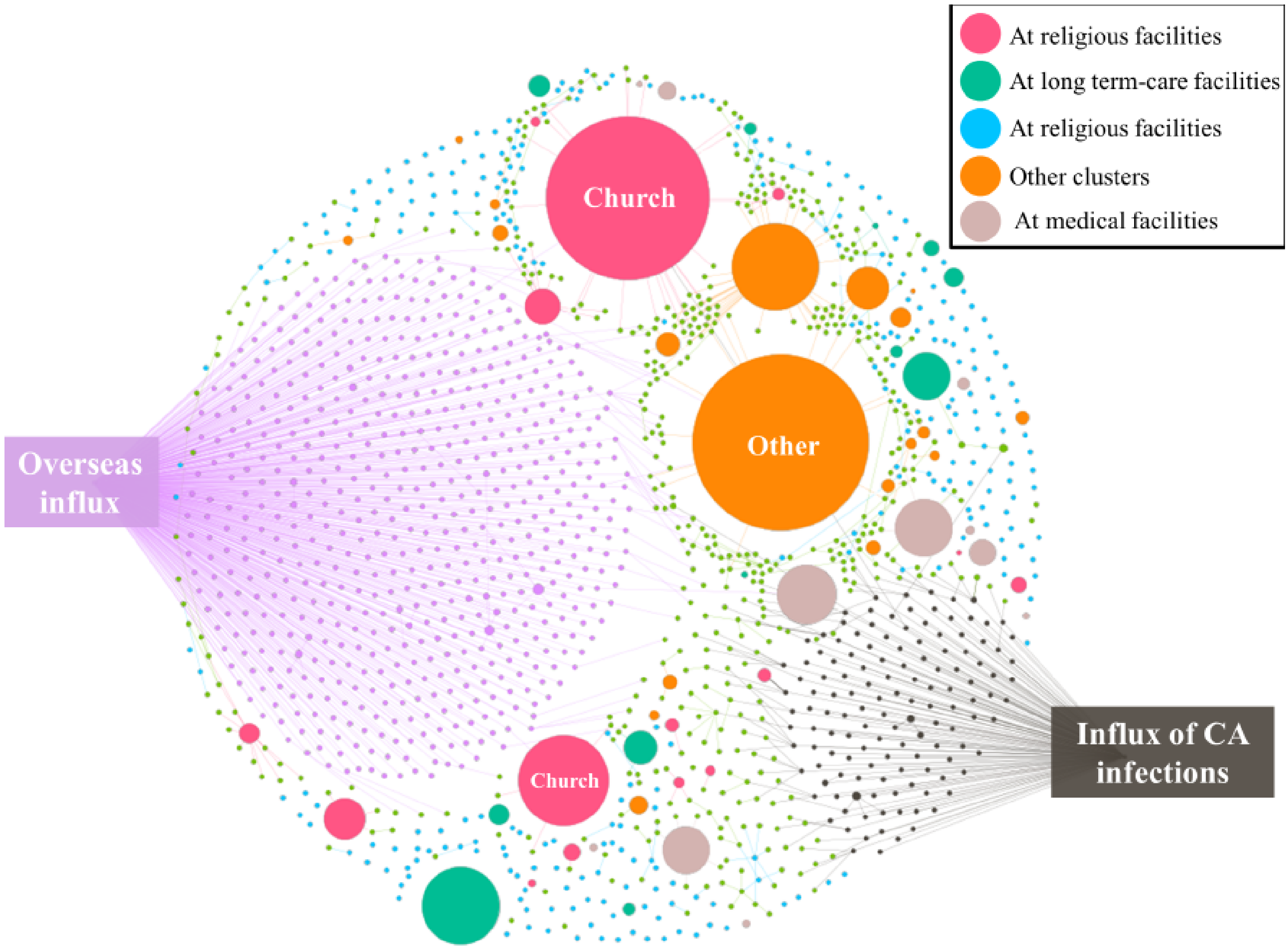


Table 2 indicates the mortality rate of COVID-19 according to the infection route. The multivariable regression analysis showed that the mortality was significantly higher in cluster cases linked to medical facilities (11/143, 7.70% vs 5/1232, 0.41%; adjusted percentage difference 7.99%; 95% CI 5.83 to 10.14) and long-term care facilities (19/221, 8.60% vs 5/1232, 0.41%; adjusted percentage difference 7.56%; 95% CI 5.66 to 9.47) than in noncluster cases.

Table 2. Mortality rate for COVID-19 according to the infection route in South Korea (n=2134).[a]

Cases	Mortality percentage (95% CI)	Adjusted difference (95% CI)	*P* value
Noncluster cases	0.41 (–0.25 to 1.06)	Reference	
Cluster and their contact cases			
Linked to medical facilities	7.70 (5.78 to 9.61)	7.99 (5.83 to 10.14)	<.001
Linked to long-term care facilities	8.60 (7.06 to 10.14)	7.56 (5.66 to 9.47)	<.001
Linked to religious facilities	0.21 (–0.83 to 1.24)	–0.14 (–1.40 to 1.13)	.88
Others	0.22 (–0.85 to 1.29)	–0.14 (–1.42 to 1.15)	.88

[a] Risk factors were adjusted by age (0-19 years, 20-39 years, 40-59 years, and 60 years or older), sex, diagnosis date, and region of residence (urban [Seoul, Incheon, Sejong, Daegu, Gwangju, Ulsan, and Busan] vs rural [Gyeonggi-do, Gangwon-do, Chungbuk, Chungnam, Gyeongbuk, Gyeongnam, Jeonbuk, Jeonnam, and Jeju]).

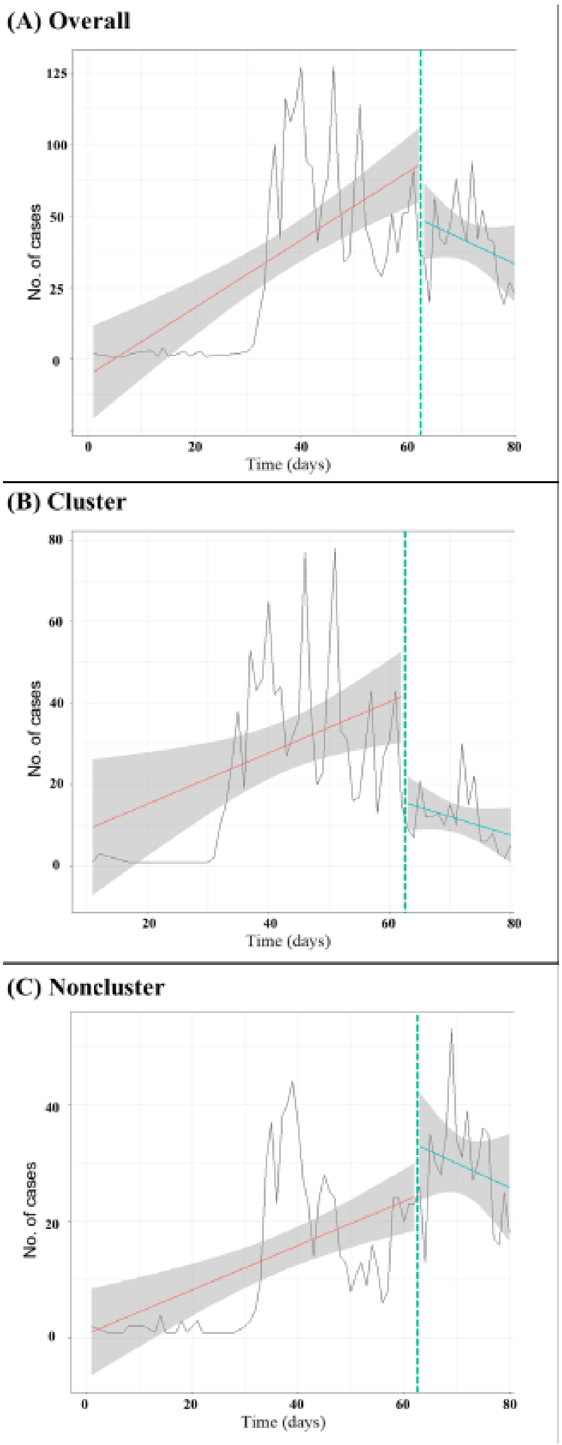
Table 3 and Figure 4 show the trend in newly confirmed COVID-19 cases after the social distancing campaign by infection route. The trend was significantly negative in the overall population (adjusted trend difference –2.28; 95% CI –3.88 to –0.68) and the cluster infection group (adjusted trend difference, –0.96; 95% CI –1.83 to –0.09).

Table 3. New confirmed COVID-19 cases trends before and after a social distancing campaign in South Korea.[a]

Groups	Trend before the social distancing campaign (95% CI)	Trend after the social distancing campaign (95% CI)	Trend difference (95% CI)	*P* value
Overall	1.11 (0.62 to 1.59)	–1.18 (–2.70 to 0.34)	–2.28 (–3.88 to –0.68)	.005
Cluster	0.43 (–0.10 to 0.96)	–0.53 (–1.22 to 0.17)	–0.96 (–1.83 to –0.09)	.03
Noncluster	0.35 (0.16 to 0.54)	–0.34 (–1.34 to 0.67)	–0.69 (–1.71 to 0.33)	.19

[a]Risk factors were adjusted by age, sex, and region of residence.

Figure 4. Number of new confirmed COVID-19 cases over the study period. The dashed vertical line at March 22, 2020, indicates the launch of the social distancing campaign. The solid red (before the social distancing campaign) and blue (after the social distancing campaign) lines represent the linear trends of new confirmed COVID-19 cases. Shaded areas represent 95% CIs for the linear trends.



Discussion

Principal Findings

To our knowledge, this is the first study to investigate the results of nationwide contact tracing of patients with COVID-19 and examine whether a social distancing campaign is effective in mitigating the spread of COVID-19. Cases of cluster infection and their contacts, which accounted for 51.4% (1305/2537) of the cases in this study, were linked to medical facilities, long-term care facilities, religious facilities, and other locations (military units, dance studios, karaoke bars, internet cafés, public transport, prisons, and workplaces of each patient). Moreover, COVID-19 linked to medical and long-term care facilities significantly increased the risk of mortality compared to noncluster COVID-19. Our study also showed that the social distancing campaign decreased the spread of COVID-19 in South Korea and differentially affected cluster infections of SARS-CoV-2. Therefore, strategies for the prevention of cluster infection of SARS-CoV-2 should be personalized and comprehensive, and multidisciplinary strategies to prevent COVID-19 should be developed. In particular, special attention should be paid to prevent cluster infections of SARS-CoV-2, especially in medical and long-term care facilities.

The pandemic spread of COVID-19 is exponentially escalating [17,18]. Cases of COVID-19 grew by several thousand each day in China in late January and early February, and took 2-3 days to double from 1000 to 2000 outside of China [1,17,19]. The velocity of the SARS-CoV-2 spread is substantially higher than that of the coronaviruses causing severe acute respiratory syndrome (SARS) and MERS (48 days for the first 1000 people to be diagnosed with COVID-19 compared to 130 days for SARS and 903 days for MERS) [2,20]. Aside from the characteristics of the virus itself, we investigated the epidemiological aspects of SARS-CoV-2 transmission in South Korea using contact tracing of confirmed cases and analyzed factors that may accelerate infection and death. We found three significant factors. First, cluster cases accounted for the highest portion of SARS-CoV-2–positive cases; second, overseas influx was significantly involved; and third, the majority of cases were confined to a specific area (Daegu Region).

An in-depth analysis of clustered cases revealed that a higher proportion of confirmed COVID-19 cases were related to religious, long-term care, and medical facilities. Cases from medical and long-term care facilities had a high mortality rate (11/143, 7.70% and 19/221, 8.60%, respectively) due to a higher proportion of vulnerable people including older adults and patients who are chronically ill present among these cases. These facilities are typically crowded with people in enclosed rooms, which create favorable conditions for transmission of respiratory diseases [21,22]. South Korea has the highest number of nursing hospitals (long-term care hospitals: 27.35 per 1000 people aged≥65 years) and the longest average length of hospital stay (average 18.5 days) of all Organisation for Economic Co-operation and Development countries [23]. Therefore, more care with strict regulation and quarantine programs should be applied to these kinds of facilities to avoid massive clusters of infection.

The enforced social distancing campaign was introduced by the Korean government on March 22, 2020. Our data support the enforced social distancing campaign as a highly effective method for preventing clustered infections. Our analyses demonstrated a significant reduction in clustered SARS-CoV-2 infections (adjusted trend difference –0.96; 95% CI –1.83 to –0.09) after the launch of the nationwide campaign. Since SARS-CoV-2 is transmitted via respiratory droplets [24,25], the purpose of the campaign was to keep a minimum distance to avoid transmission while maintaining personal hygiene. A droplet will fall under gravity or evaporate within 2 meters of the infected individual; therefore, staying 2 meters, or approximately three steps, away from other individuals will theoretically prevent droplet-induced transmission [26]. In addition to keeping personal distance, enforced social distancing includes following basic guidelines at work, religious facilities, sports and entertainment facilities, and other high-risk facilities, such as refraining from going outdoors when experiencing respiratory symptoms; having online gatherings instead of personal meetings; keeping a distance and avoiding talking when you eat; using personal belongings instead of sharing items; and keeping hand sanitizer available at entrances of buildings, elevators, and stairways.

It is interesting to note that the overseas influx had a significant role in the spread of the virus in South Korea. Recently, many countries have imposed government-issued international travel restrictions [27]. Although restricting travel may be useful in the early stage of the outbreak, it may be less successful once the outbreak is widespread [28]. Therefore, banning visitors from China or other COVID-19 high-risk countries to reduce the risk of reintroduction of the virus might be effective in countries that are at the early stage of the COVID-19 outbreak. However, for countries with a high incidence of COVID-19, an alternative strategy must be applied to mitigate SARS-CoV-2 transmission.

Policy Implications

As the nature of COVID-19 is subclinical in some individuals, isolating early detected confirmed cases before transmission can occur is difficult [29]. Therefore, substantial effort should be made to prevent the virus from spreading by developing effective public health policy. First, public health policy should advise against social gatherings such as mass conferences, sporting events, musical concerts, and religious meetings. Instead, working remotely, online conferences, and online religious services should be encouraged. Second, strict screening and quarantine should be applied to those entering or leaving a region. Routine screening for SARS-CoV-2 and self-isolation should be required of visitors from areas of high incidence of COVID-19. Third, individuals should be advised against travel to regions of high COVID-19 incidence. Surveys of medical or long-term care facility visitors should be routinely conducted to screen for a history of visits to areas of high COVID-19 incidence. In addition, testing for COVID-19 should be required for patients and residents as well as staff and visitors in medical and long-term care facilities to prevent the introduction of COVID-19 in those facilities.

Strengths and Limitations

First, as previously mentioned, one of the strengths of our study is that novel individual contact tracing data acquired by the KCDC and each local government in South Korea was used. By tracing individual data, we could categorize the source and characteristics of the transmission. Additionally, most other countries have not performed epidemiological surveys that include contact tracing; South Korea is thus far the only country to conduct epidemiological surveys with contact tracing. Therefore, we were able to identify the spread dynamics of COVID-19. Second, our study has a clear time point when a nationwide social distancing campaign was launched. Therefore, we could compare the trends of transmission before and after the campaign and evaluate the effectiveness of the public health intervention. Nonetheless, our study has some limitations. First, our data did not contain clinical information because we could not link hospital data to the epidemiological survey expeditiously. Second, we are still developing epidemiological surveys that include information on socioeconomic status (personal occupation and income) and time to development of COVID-19–related symptoms; hence, we were unable to analyze the time to symptom onset or socioeconomic status. Third, although the WHO stated that contact tracing includes the process of identifying, assessing, and managing people who have been exposed to a disease to prevent onward transmission [30], we only had tracing from confirmed cases; tracing for exposure remains for future study. Finally, epidemiological surveillance was not possible in some regions due to community-level outbreaks (Daegu and two cities in Gyeongbuk [Cheongdo and Gyeongsan]). Therefore, data from those regions were excluded.

Conclusion

In this study, we investigated the nationwide contact tracing results of patients with COVID-19 and whether the social distancing campaign was effective in mitigating the spread of COVID-19. COVID-19 linked to medical and long-term care facilities significantly increased the risk of mortality compared with noncluster COVID-19. Moreover, our study shows that the social distancing campaign decreased the spread of COVID-19 in South Korea and differentially affected cluster infections of SARS-CoV-2. Therefore, our data may support driving public health policies in other countries and help normalize and restore social activities while minimizing the risk of transmission. Further cooperative global epidemic studies and updates are warranted to drive the best policy to control the transmission of SARS-CoV-2.

Acknowledgments

This work was supported by the Institute of Information and Communications Technology Planning & Evaluation grant funded by the Korea government (MSIT); No.2020-0-00512, Data Refinement and Improvement through Data Quality Evaluation. The funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Authors' Contributions

DKY had full access to all of the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. All authors approved the final version before submission. The study concept and design was done by SWL, WTY, and DKY. Acquisition, analysis, or interpretation of data was done by SWL, WTY, and DKY. Drafting of the manuscript was done by JMY and DKY. Critical revision of the manuscript for important intellectual content was done by Y-SC, IKY, HYK, DM, DO, EKH, MYH, and DKY. Statistical analysis was done by SWL, Y-SC, and DKY. Study supervision was done by DKY. DKY is guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The dynamic infection spread network video of the COVID-19 in South Korea. The bottom bar represents the elapsed day after the first infection has occurred. The video includes the first 80 days of spread. Each line represents the spread occurrence between persons, and each dot represents the infected individual. Larger dots represent clustered infections, with dynamically increasing size, which means the number of infections in that cluster as time proceeds.

[MP4 File (MP4 Video), 6720 KB - medinform_v8i8e20992_app1.mp4]

References

1. Pan A, Liu L, Wang C, Guo H, Hao X, Wang Q, et al. Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China. JAMA 2020 Apr 10 [FREE Full text] [doi: 10.1001/jama.2020.6130] [Medline: 32275295]
2. Park SE. Epidemiology, virology, and clinical features of severe acute respiratory syndrome -coronavirus-2 (SARS-CoV-2; Coronavirus Disease-19). Clin Exp Pediatr 2020 Apr;63(4):119-124 [FREE Full text] [doi: 10.3345/cep.2020.00493] [Medline: 32252141]

3. Coronavirus disease 2019 (COVID-19) situation report – 78. World Health Organization. 2020 Apr 07. URL: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200407-sitrep-78-covid-19.pdf?sfvrsn=bc43e1b_2
4. Choi S, Kim HW, Kang J, Kim DH, Cho EY. Epidemiology and clinical features of coronavirus disease 2019 in children. Clin Exp Pediatr 2020 Apr;63(4):125-132 [FREE Full text] [doi: 10.3345/cep.2020.00535] [Medline: 32252139]
5. Korean Society of Infectious Diseases, Korean Society of Pediatric Infectious Diseases, Korean Society of Epidemiology, Korean Society for Antimicrobial Therapy, Korean Society for Healthcare-associated Infection Control and Prevention, Korea Centers for Disease Control and Prevention. Report on the epidemiological features of coronavirus disease 2019 (COVID-19) outbreak in the Republic of Korea from January 19 to March 2, 2020. J Korean Med Sci 2020 Mar 16;35(10):e112 [FREE Full text] [doi: 10.3346/jkms.2020.35.e112] [Medline: 32174069]
6. Normile D. Coronavirus cases have dropped sharply in South Korea. What's the secret to its success? Science 2020 Mar 18. [doi: 10.1126/science.abb7566]
7. COVID-19 National Emergency Response Center, Epidemiology & Case Management Team, Korea Centers for Disease Control & Prevention. Contact transmission of COVID-19 in South Korea: novel investigation techniques for tracing contacts. Osong Public Health Res Perspect 2020 Feb;11(1):60-63 [FREE Full text] [doi: 10.24171/j.phrp.2020.11.1.09] [Medline: 32149043]
8. Lee SW, Yon DK, James CC, Lee S, Koh HY, Sheen YH, et al. Short-term effects of multiple outdoor environmental factors on risk of asthma exacerbations: age-stratified time-series analysis. J Allergy Clin Immunol 2019 Dec;144(6):1542-1550.e1. [doi: 10.1016/j.jaci.2019.08.037] [Medline: 31536730]
9. Koh HY, Kim TH, Sheen YH, Lee SW, An J, Kim MA, et al. Serum heavy metal levels are associated with asthma, allergic rhinitis, atopic dermatitis, allergic multimorbidity, and airflow obstruction. J Allergy Clin Immunol Pract 2019;7(8):2912-2915.e2. [doi: 10.1016/j.jaip.2019.05.015] [Medline: 31129074]
10. Ha J, Lee SW, Yon DK. Ten-Year trends and prevalence of asthma, allergic rhinitis, and atopic dermatitis among the Korean population, 2008-2017. Clin Exp Pediatr 2020 Jul;63(7):278-283 [FREE Full text] [doi: 10.3345/cep.2019.01291] [Medline: 32023407]
11. Woo A, Lee SW, Koh HY, Kim MA, Han MY, Yon DK. Incidence of cancer after asthma development: 2 independent population-based cohort studies. J Allergy Clin Immunol 2020 May 15. [doi: 10.1016/j.jaci.2020.04.041] [Medline: 32417133]
12. Kong SG. Current use of safety restraint systems and front seats in Korean children based on the 2008-2015 Korea National Health and Nutrition Examination Survey. Korean J Pediatr 2018 Dec;61(12):381-386 [FREE Full text] [doi: 10.3345/kjp.2018.06604] [Medline: 30304902]
13. Lee C, Tsai H, Kunin CM, Lee SS, Wu K, Chen Y. Emergence of sporadic non-clustered cases of hospital-associated listeriosis among immunocompromised adults in southern Taiwan from 1992 to 2013: effect of precipitating immunosuppressive agents. BMC Infect Dis 2014 Mar 19;14:145 [FREE Full text] [doi: 10.1186/1471-2334-14-145] [Medline: 24641498]
14. Grasselli G, Zangrillo A, Zanella A, Antonelli M, Cabrini L, Castelli A, COVID-19 Lombardy ICU Network, et al. Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2 admitted to ICUs of the Lombardy Region, Italy. JAMA 2020 Apr 06 [FREE Full text] [doi: 10.1001/jama.2020.5394] [Medline: 32250385]
15. Gephi. URL: https://gephi.org/
16. Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Softw: Pract Exper 1991 Nov;21(11):1129-1164. [doi: 10.1002/spe.4380211102]
17. Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, China Medical Treatment Expert Group for Covid-19. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020 Apr 30;382(18):1708-1720 [FREE Full text] [doi: 10.1056/NEJMoa2002032] [Medline: 32109013]
18. Shah S, Meenakshisundaram R, Senthilkumaran S, Thirumalaikolundusubramanian P. COVID-19 in children: reasons for uneventful clinical course. Clin Exp Pediatr 2020 Jul;63(7):237-238 [FREE Full text] [doi: 10.3345/cep.2020.00801] [Medline: 32664708]
19. Lee K, Rhim J, Kang J. Immunopathogenesis of COVID-19 and early immunomodulators. Clin Exp Pediatr 2020 Jul;63(7):239-250 [FREE Full text] [doi: 10.3345/cep.2020.00759] [Medline: 32664709]
20. Callaway E, Cyranoski D, Mallapaty S, Stoye E, Tollefson J. The coronavirus pandemic in five powerful charts. Nature 2020 Mar;579(7800):482-483. [doi: 10.1038/d41586-020-00758-2] [Medline: 32203366]
21. McMichael TM, Currie DW, Clark S, Pogosjans S, Kay M, Schwartz NG, Public Health–Seattle and King County, EvergreenHealth, and CDC COVID-19 Investigation Team. Epidemiology of Covid-19 in a long-term care facility in King County, Washington. N Engl J Med 2020 May 21;382(21):2005-2011 [FREE Full text] [doi: 10.1056/NEJMoa2005412] [Medline: 32220208]
22. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of SARS-CoV-2 in the Icelandic population. N Engl J Med 2020 Jun 11;382(24):2302-2315 [FREE Full text] [doi: 10.1056/NEJMoa2006100] [Medline: 32289214]
23. OECD health statistics 2014 - frequently requested data. OECD. URL: https://www.oecd.org/els/health-systems/oecd-health-statistics-2014-frequently-requested-data.htm

24. Chan J, Yuan S, Kok K, To K, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet 2020 Feb;395(10223):514-523. [doi: 10.1016/s0140-6736(20)30154-9]
25. Anfinrud P, Stadnytskyi V, Bax CE, Bax A. Visualizing speech-generated oral fluid droplets with laser light scattering. N Engl J Med 2020 May 21;382(21):2061-2063 [FREE Full text] [doi: 10.1056/NEJMc2007800] [Medline: 32294341]
26. Xie X, Li Y, Chwang ATY, Ho PL, Seto WH. How far droplets can move in indoor environments--revisiting the Wells evaporation-falling curve. Indoor Air 2007 Jun;17(3):211-225. [doi: 10.1111/j.1600-0668.2007.00469.x] [Medline: 17542834]
27. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science 2020 Apr 24;368(6489):395-400 [FREE Full text] [doi: 10.1126/science.aba9757] [Medline: 32144116]
28. Kraemer MUG, Yang C, Gutierrez B, Wu C, Klein B, Pigott DM, Open COVID-19 Data Working Group, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science 2020 May 01;368(6490):493-497 [FREE Full text] [doi: 10.1126/science.abb4218] [Medline: 32213647]
29. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. N Engl J Med 2020 Mar 26;382(13):1199-1207. [doi: 10.1056/nejmoa2001316]
30. Ethical considerations to guide the use of digital proximity tracking technologies for COVID-19 contact tracing. World Health Organization. 2020 May 28. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-Ethics_Contact_tracing_apps-2020.1

Abbreviations

CCTV: closed-circuit television
KCDC: Korea Centers for Disease Control and Prevention
MERS: Middle East respiratory syndrome
SARS: severe acute respiratory syndrome
WHO: World Health Organization

Edited by G Eysenbach; submitted 03.06.20; peer-reviewed by A Ahmed, A Alfadl, P Akande, S Olowookere; comments to author 22.06.20; revised version received 04.07.20; accepted 24.07.20; published 25.08.20.

Please cite as:
Lee SW, Yuh WT, Yang JM, Cho YS, Yoo IK, Koh HY, Marshall D, Oh D, Ha EK, Han MY, Yon DK
Nationwide Results of COVID-19 Contact Tracing in South Korea: Individual Participant Data From an Epidemiological Survey
JMIR Med Inform 2020;8(8):e20992
URL: http://medinform.jmir.org/2020/8/e20992/
doi:10.2196/20992
PMID:32784189

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org