JMIR Medical Informatics

Contents

Original Papers

Reviews

Original Paper

Weight Loss Trajectories After Bariatric Surgery for Obesity: Mathematical Model and Proof-of-Concept Study

Chloe Dimeglio[1], PhD; Guillaume Becouarn[2], MD; Philippe Topart[2], MD; Rodolphe Bodin[3], MD; Jean Christophe Buisson[4], DPhil; Patrick Ritz[1], DPhil, MD

[1]Centre Hospitalier Universitaire de Toulouse, Toulouse, France

[2]Clinique de l'Anjou, Angers, France

[3]Hôpital d'Instruction des Armées Robert Picqué, Service de Santé des Armées, Bordeaux, France

[4]École Nationale Supérieure d'Électrotechnique, d'Électronique, d'Informatique, d'Hydraulique et des Télécommunications, Toulouse, France

Corresponding Author:
Patrick Ritz, DPhil, MD
Centre Hospitalier Universitaire de Toulouse
Place J Pouilhes
Toulouse
France
Phone: 33 068227932
Email: patrick.ritz@wanadoo.fr

Abstract

Background: Obesity surgery has proven its effectiveness in weight loss. However, after a loss phase of about 12 to 18 months, between 20% and 40% of patients regain weight. Prediction of weight evolution is therefore useful for early detection of weight regain.

Objective: This proof-of-concept study aimed to analyze the postoperative weight trajectories and to identify "curve families" for early prediction of weight regain.

Methods: This was a monocentric retrospective study with calculation of the weight trajectory of patients having undergone gastric bypass surgery. Data on 795 patients after a 2-year follow-up allowed modeling of weight trajectories according to a hierarchical cluster analysis (HCA) tending to minimize the intragroup distance according to Ward. Clinical judgement was used to finalize the identification of clinically relevant representative trajectories. This modeling was validated on a group of 381 patients for whom the observed weight at 18 months was compared to the predicted weight.

Results: Two successive HCA produced 14 representative trajectories, distributed among 4 clinically relevant families: Of the 14 weight trajectories, 6 decreased systematically over time or decreased and then stagnated; 4 decreased, increased, and then decreased again; 2 decreased and then increased; and 2 stagnated at first and then began to decrease. A comparison of observed weight and that estimated by modeling made it possible to correctly classify 98% of persons with excess weight loss (EWL) >50% and more than 58% of persons with EWL between 25% and 50%. In the category of persons with EWL >50%, weight data over the first 6 months were adequate to correctly predict the observed result.

Conclusions: This modeling allowed correct classification of persons with EWL >50% and could identify early after surgery the patients with potentially less that optimal weight loss. Further studies are needed to validate this model in other populations, with other types of surgery, and with other medical-surgical teams.

Introduction

Obesity surgery has proved to be effective in the long term for weight loss and for remission or improvement of comorbidities associated with excess weight. In the long term, however, weight regain occurs in 20% to 40% of patients [1-7], and this is a major factor in the recurrence of comorbidities.

Many factors have been studied to predict weight regain but few have been found to be relevant [8]. Ritz et al [9] showed

that patients in failure at 2 years according to the Reinhold criteria [10] developed an unfavorable weight trajectory from the early postoperative period. Careful monitoring of early postoperative weight trajectories could thus be one way to achieve early identification of patients with weight regain. Ideally, self-monitoring of weight could empower patients to alert health care providers sufficiently early so that solutions can be studied.

Wise et al [11] have shown that an exponential decay can describe the weight path in the 3 months after gastric bypass. They used excess weight loss (EWL) and did not show a long-term 5-year prediction value of the classification. EWL calculations can be difficult for patients who are not used to this arithmetic. An easy-to-use application for computing weight paths using only weight would help patients in monitoring their weight loss.

This proof-of-concept study aimed to analyze the postoperative weight trajectories and identify "curve families" for early prediction of weight regain. We combined a method of classification of unsupervised curves and clinical expertise to define the reference weight trajectories after gastric bypass surgery.

Methods

Study Design

This was a monocentric retrospective study modeling the weight trajectories of patients having undergone gastric bypass surgery.

Ethics

All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Patients

Group 1 comprised 795 patients who had gastric bypass surgery between 2003 and 2012 with at least 2 years of follow-up in the same reference center, performed by two surgeons (GB and PT). They were included for modeling their weight trajectories. They were within the Haute Autorité de Santé criteria for obesity surgery [12]. The patients analyzed represented 63.00% (795/1262) of the initial patients. A total of 11.01% (139/1262) were excluded for redo surgery; 28.62% (320/1118) were excluded because they were lost to follow-up within the 2 years after surgery.

Group 2 comprised 381 patients with a follow-up of 5 years after gastric bypass. They were classified according to the Reinhold criteria and allowed us to assess predictive capability of the model when applied within the first 18 months to predict later weight.

Evaluation Criteria for Weight Loss

Weight was measured with patients wearing undergarments during postsurgery follow-up visits (1, 3, 6, 9, 12, and 18 months). The reference weight was that of the day before surgery. Certain data were missing. The EWL was calculated

based on a body mass index (BMI) of 25 kg/m^2, with equation EWL = (initial weight – actual weight)/(initial weight – 25* height^2); weight in kg and height in meters. The Reinhold [10] criteria were used to define failure (EWL after 18 months of less than 25%), success (EWL after 18 months exceeding 50%), and an intermediate result (EWL between 25% and 50%).

Calculations and Statistics

Calculations were performed using R version 3.0.2 (R Foundation for Statistical Computing) and MATLAB version 2010a (The MathWorks, Inc).

Objectives

The main objective was to establish weight trajectories representative of all trajectories in group 1 (795 patients). After imputation for missing values, all the initial trajectories were clustered into an optimal number y of trajectory groups. One representative trajectory was extracted for each group of trajectories. All the representative trajectories were merged into a smaller number of families according to the behavior of the trajectories evaluated by clinical expertise. The validity of this collection of representative trajectories was evaluated using the weight trajectories of group 2.

Management of Missing Data

The proportion of missing data was 9.4% at 1 month, 56.7% at 3 months, 51.7% at 6 months, 37.1% at 9 months, 35.6% at 12 months, and 38.7% at 18 months. The missing data were replaced by multiple imputation [13] in 5 iterations based on age, sex, presurgery weight, height, and the presence of diabetes.

Creation of a Catalog of Weight Trajectories Representative of the Base

To constitute trajectory groups that maximize the interclass distance and minimize the intraclass distance (the Ward method), successive hierarchical cluster analyses (HCA) without a priori on the number of classes that were used [14]. The number of classes retained (x) after the first HCA was that of preceding an overly heterogeneous association (while optimizing distance of Ward). Starting from this first classification, one representative curve per class was extracted using a criterion of representativity of the group, according to Dimeglio [15]. We thus obtain x representatives.

We performed as many additional HCA as necessary on this base of x representatives to continue reducing the number of classes to a clinically acceptable level, which we defined here as a maximum of 15 classes. Here y is the number of representative trajectories obtained at the last iteration. These y representatives constituted the reference base of weight trajectories.

Clinical Classification of Representative Trajectories

The final y representative curves were grouped into a number of families according to the behavior of weight trajectories evaluated by clinical expertise (gradual decrease, weight rebound, etc).

XSL•FO

RenderX

Validation of Representative Trajectories

For a given patient, their weight trajectory was brought closer

to one of the y representative trajectories by minimizing the distance between them [16]. Figure 1 gives a graphic illustration of the procedure.

Figure 1. Graphic illustration of the procedure.
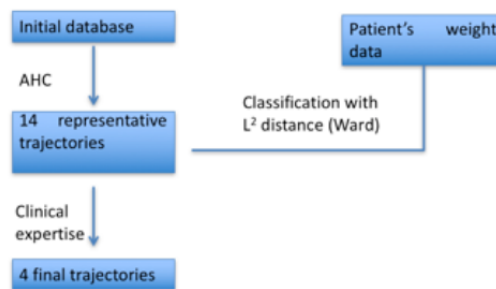
Application to the Weight Trajectories of Group 2

To determine the predictive validity of the reference base (the y representative trajectories), the procedure was applied to the 381 patients in group 2. Each patient was associated to a representative trajectory that predicted EWL at 18 months according to the Reinhold classification [10]. Similarly, the Reinhold classification was applied to the weight result observed

at 5 years. The two values obtained after these Reinhold classifications were compared. Figure 2 illustrates the procedure. To estimate the variability of our classification results, we chose to randomly extract 100 times 100 trajectories among the 381 patients in group 2 (bootstrap procedure). Thus, we compared 100 times the known identification of 100 patients to that obtained by bringing them closer to one of the representative trajectories.

Figure 2. Procedure for validation of trajectories in patients in group 2.
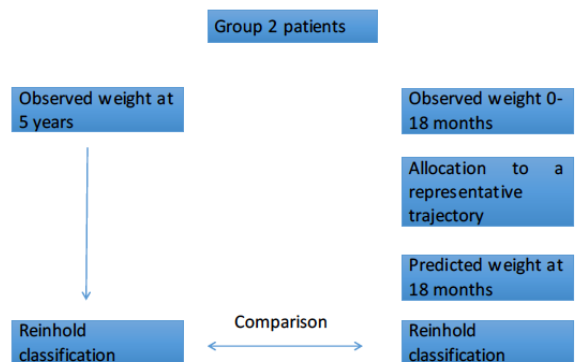
Results

Constitution of Weight Trajectories Representatives

A total of 795 patients were included in this study (141 men and 654 women). The variability of data did not allow us to summarize the information in a single representative trajectory

(Figure 3). Homogeneous groups of trajectories were necessary to summarize the entirety of the information. A total of 56 classes of representative trajectories came from the first HCA. After two successive HCA, the number of representative trajectories was reduced to 14 classes. Figure 4 shows the trajectories of the representative curve of each class.

XSL•FO

RenderX

Figure 3. Weight trajectories after surgery in group 1 (n=795).
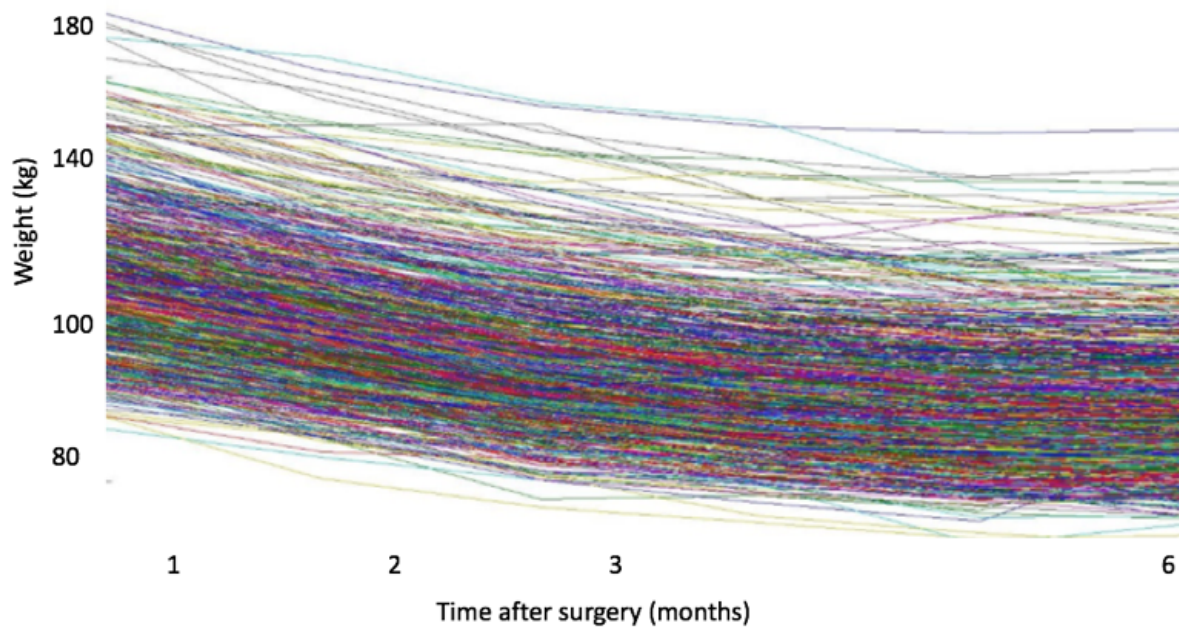
Figure 4. Weight trajectories after two hierarchical cluster analyses in group 1 (n=795).
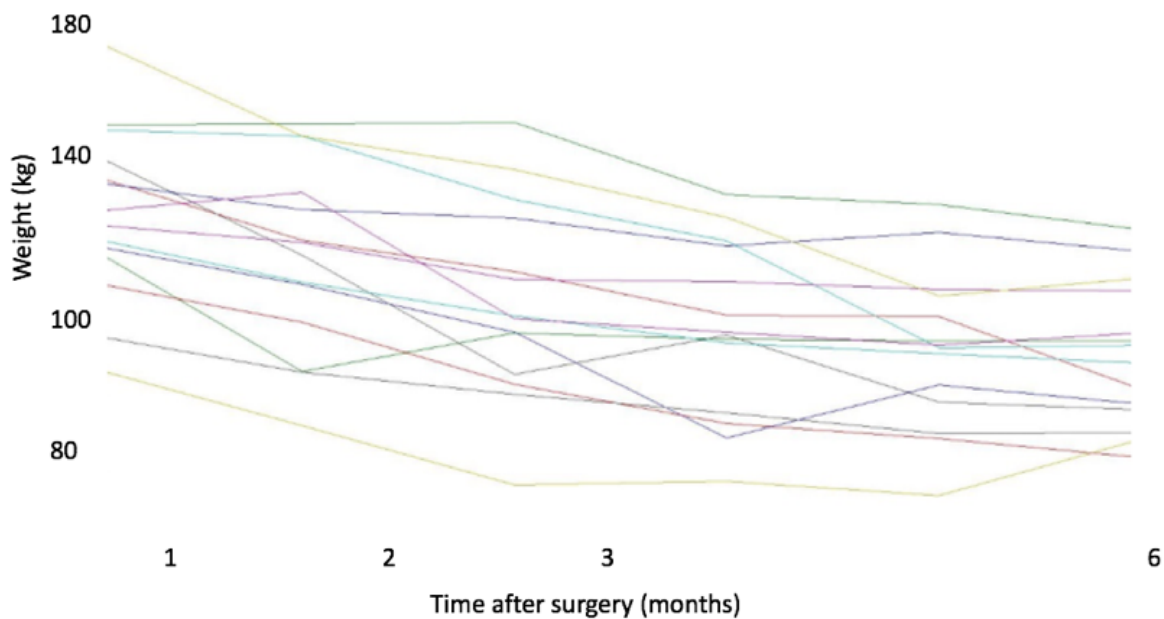
Clinical Classification of Representative Trajectories

Several of these representative trajectories had the same behavior (Figure 5):

- Weight decreased systematically over time or decreased and then remained stable (class 1). This family includes 6 of the 14 trajectories.

- Weight decreased, then increased, and then decreased again (class 2). This family includes 4 of the trajectories.
- Weight decreased and then increased (class 3). This family includes 2 of the 14 trajectories.
- Weight stagnated at first and then started to decrease (class 4). This family includes 2 of the 14 trajectories.

Preoperative factors (age, sex, presence of diabetes) did not modify the trajectory predictions.

Figure 5. The four families of weight trajectories.
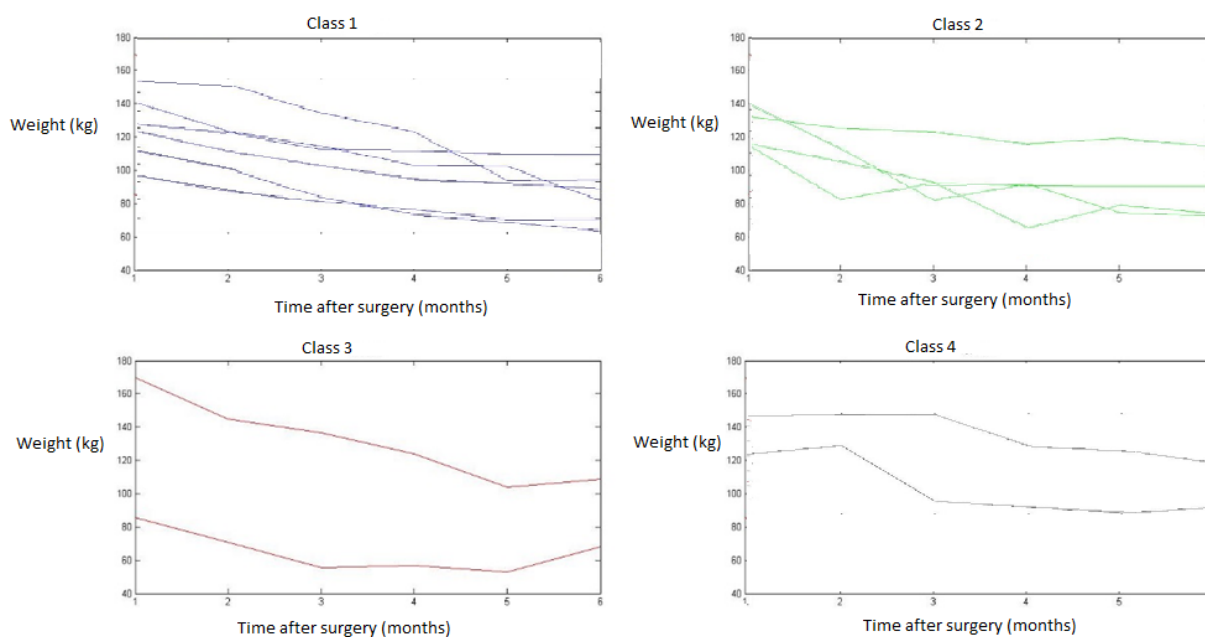
Application to the Weight Trajectories of Group 2

The Reinhold classification applied to weight observed at 18 months identified 87.4% (333/381) of patients in the category EWL >50% and 12.6% (78/411) of patients in the category EWL between 25% and 50%. There were no patients in the category EWL <25%.

The Reinhold classification was applied to the estimated weight at 18 months based on representative trajectories. Only one trajectory (7.1%) was classified EWL <25%. Six trajectories (42.9%) were classified EWL between 25% and 50%. Seven trajectories (50%) were classified EWL >50%.

The match between the Reinhold classification [10] based on observed or predicted weight is shown in Table 1. Classification with reference trajectories produced an overall rate of correct classification of more than 93%. It allowed correct classification of 97.6% (325/333) of persons with EWL >50% and more than 59% (46/78) of persons with EWL between 25% and 50%.

Table 2 shows the influence of the number of early weight values used in the prediction on the results at 5 years. The more the number of points increased, the more precise was the classification of the different profiles (best mean and lowest variability). For the persons who would be in the category EWL >50% at 18 months, the match rates increased with the number of values over time, but as of 6 months postsurgery, 91.6% (305/333) were well classified.

For the persons who would be in the category EWL between 25% and 50%, the match rate showed little increase over time (going from 47% to 56%). Therefore, between 44% and 53% of patients were poorly classified by the trajectories.

Table 1. Match between the Reinhold classification applied to observed and predicted weight of persons in group 2.

EWL calculated with observed weight		Predicted weight, mean (SD)		
		EWL[a] >50%	EWL 25%-50%	EWL <25%
Observed weight, mean (SD)	EWL >50%	97.6 (1.49)	2.37 (1.49)	0
Observed weight, mean (SD)	EWL 25%-50%	41.3 (12.8)	58.7 (12.8)	0

[a]EWL: excess weight loss.

Table 2. Influence of the number of early weight values used in the prediction of weight identification at 5 years. The mean numbers in the columns are the percentages of patients correctly classified.

EWL calculated with observed weight		Predicted weight with 6 months data mean (SD)		Predicted weight with 9 months data mean (SD)		Predicted weight with 12 months data mean (SD)	
		EWL[a] >50%	EWL 25%-50%	EWL >50%	EWL 25%-50%	EWL >50%	EWL 25%-50%
Observed weight, mean (SD)	EWL >50%	91.6 (2.96)	8.4 (2.96)	96 (1.86)	4 (1.86)	97.4 (1.6)	2.6 (1.6)
Observed weight, mean (SD)	EWL 25%-50%	52.7 (13.3)	47.3 (13.3)	46.3 (13.9)	53.7 (13.9)	43.7 (13.5)	56.3 (13.5)

[a]EWL: excess weight loss.

Discussion

Principal Findings

This study shows that it is possible to model weight loss over the 18-month period after a gastric bypass in representative trajectories. Applied to a group of patients at 5 years postsurgery, the representative trajectories correctly predicted more than 93% of the values. Postoperative data available for 1 to 6 months are sufficient to provide a satisfactory prediction of EWL >50%.

Strengths and Limitations

Unsupervised classification methods like the HCA that we used here makes it possible to define families of trajectories without a priori on the number of classes to obtain. This is a very commonly used method. In this study, it allowed us to determine families of weight trajectories only from their general appearance and without using additional variables, like in the research of Wise [11]. One of the advantages is that the classification can be updated as new weight data are added to the previous one. Also, we chose to use weight and not EWL (or the percentage of weight loss) to develop the trajectories because they are accompanied by arithmetic biases. For example, for a very high initial BMI, EWL will be less than for a lower BMI, despite an equivalent weight loss. We only used EWL to compare the value predicted by the trajectory to the observed value in group 2. The concept of predicting later weight is therefore valid from this data. The use of weight trajectories rather than trajectories of EWL or of percentages of weight loss was also preferred for conserving the maximum of behaviors within the trajectories.

The cohort came from the same center with two surgeons and a formalized protocol for the preparation and follow-up of patients. However, that limits the variability introduced by multiple teams using different preparation and follow-up protocols. This study did not analyze known factors of failure or of inadequate weight result like those identified by other authors [11]. The prediction is better for the patients who lose a lot of weight. This is explained by the initial imbalance between the three classes based on EWL in the database. Further analyses on more balanced bases are needed.

An early prediction of weight trajectory is clinically useful. A success prediction may empower postsurgical patients to carry on and reassure them, as many are anxious about regaining weight because of their past history of weight loss failures. This

does not mean that weight will no longer be monitored. This also suggests that if a person is not classified as a later success, every effort should be made to understand the reasons and implement early corrective strategies. In that sense, the factors classically associated with less than optimal weight loss can be used not as predictors (with little efficiency) but as targets for an early additive strategy. Despite the low power to predict future results in this category, the idea of an early alert is clinically useful. Validation studies on other cohorts are needed. We should also point out that the validation database was very imbalanced (87.4% of EWL >50% vs 12.6% with EWL between 25% and 50% and no patient in weight loss failure). It is possible that the classification of patients with average weight loss will improve with better-balanced validation database. A machine learning process added to the application would provide evolutive classifications as the amount of data increases [17-19].

Analysis of the Results

This study shows that a hierarchical cluster analysis allowed us to identify 4 profiles of weight trajectories associated with clinical expertise. The patients who were the most successful were those who lost weight regularly. The persons who lost the least had difficulties in the initial phase or had a secondary weight regain.

None of the preoperative parameters differed between the families of curves. This agrees with other published studies [8]. It seems difficult to predict weight loss after surgery with factors that do not consider the capacity for personal, psychological, or metabolic adaptation during the postsurgery period. The factors classically associated with poor weight loss results are postoperative (recurrence of depression, eating disorders, lack of follow-up, lack of arrangements for changes of eating habits, and physical activity) and usually occur late after surgery (24 months). The early identification of a trajectory is therefore an interesting feature of this model. Applied to a population of persons having undergone surgery and at 18 months after this surgery, this classification correctly identified more than 97% of EWL >50% and more than 58% of EWL between 25% and 50%.

Wise et al [11] recently published a comparable study. In a retrospective analysis of a cohort of gastric bypassed patients followed for 3 years, the authors demonstrated the exponential nature of the decrease of EWL and analyzed the factors influencing the trajectory, such as age. Our study used an original mathematical approach and adds to the study of Wise et al [11] because it validates the concept of trajectories in an

independent population at 5 years. Like Wise et al [11], we conducted the analysis using EWL without showing better results than with weight only.

Clinical Relevance

Given the prevalence of weight regain after obesity surgery, early identification of a nonoptimal weight trajectory would help to strengthen the second lines of treatment by reinforcing measures for behavioral change (self-care, physical activity, nutrition). This study proves the concept that a simple application could allow surgical patients to enter weight data themselves, evaluate the result, and eventually alert the health care provider.

Conflicts of Interest

None declared.

References

1. Magro DO, Geloneze B, Delfini R, Pareja BC, Callejas F, Pareja JC. Long-term weight regain after gastric bypass: a 5-year prospective study. Obes Surg 2008 Jun;18(6):648-651 [FREE Full text] [doi: 10.1007/s11695-007-9265-1] [Medline: 18392907]
2. Monaco-Ferreira DV, Leandro-Merhi VA. Weight regain 10 years after Roux-en-Y gastric bypass. Obes Surg 2017 Dec;27(5):1137-1144. [doi: 10.1007/s11695-016-2426-3] [Medline: 27798793]
3. Sjöström L, Lindroos A, Peltonen M, Torgerson J, Bouchard C, Carlsson B, et al. Lifestyle, diabetes, and cardiovascular risk factors 10 years after bariatric surgery. N Engl J Med 2004 Dec 23;351(26):2683-2693. [doi: 10.1056/NEJMoa035622] [Medline: 15616203]
4. Arterburn DE, Olsen MK, Smith VA, Livingston EH, Van Scoyoc L, Yancy WS, et al. Association between bariatric surgery and long-term survival. JAMA 2015 Jan 06;313(1):62-70. [doi: 10.1001/jama.2014.16968] [Medline: 25562267]
5. Sjöström L, Narbro K, Sjöström CD, Karason K, Larsson B, Wedel H, et al. Effects of bariatric surgery on mortality in Swedish obese subjects. N Engl J Med 2007 Aug 23;357(8):741-752. [doi: 10.1056/NEJMoa066254] [Medline: 17715408]
6. Sarwer DB, Steffen KJ. Quality of life, body image and sexual functioning in bariatric surgery patients. Eur Eat Disord Rev 2015 Nov;23(6):504-508. [doi: 10.1002/erv.2412] [Medline: 26608946]
7. Karmali S, Brar B, Shi X, Sharma AM, de Gara C, Birch DW. Weight recidivism post-bariatric surgery: a systematic review. Obes Surg 2013 Nov;23(11):1922-1933. [doi: 10.1007/s11695-013-1070-4] [Medline: 23996349]
8. Livhits M, Mercado C, Yermilov I, Parikh JA, Dutson E, Mehran A, et al. Patient behaviors associated with weight regain after laparoscopic gastric bypass. Obes Res Clin Pract 2011 Jul;5(3):e169-e266. [doi: 10.1016/j.orcp.2011.03.004] [Medline: 24331108]
9. Ritz P, Caiazzo R, Becouarn G, Arnalsteen L, Andrieu S, Topart P, et al. Early prediction of failure to lose weight after obesity surgery. Surg Obes Relat Dis 2013 Jan;9(1):118-121. [doi: 10.1016/j.soard.2011.10.022] [Medline: 22222302]
10. Reinhold RB. Critical analysis of long term weight loss following gastric bypass. Surg Gynecol Obstet 1982 Sep;155(3):385-394. [Medline: 7051382]
11. Wise ES, Felton J, Kligman MD. Exponential decay modeling can define parameters of weight loss trajectory after laparoscopic Roux-en-Y gastric bypass. Am J Surg 2018 Jul;216(1):120-123. [doi: 10.1016/j.amjsurg.2017.10.024] [Medline: 29089100]
12. Haute autorité de santé. 2009 Jun. [Obesité: prise en charge chirurgicale de l'adulte] URL: https://www.has-sante.fr/portail/jcms/c_765529/fr/obesite-prise-en-charge-chirurgicale-chez-l-adulte [accessed 2019-09-16]
13. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393 [FREE Full text] [Medline: 19564179]
14. Ward JH. Hierarchical grouping to optimize an objective function. J Am Statistical Assoc 1963 Mar;58(301):236-244. [doi: 10.1080/01621459.1963.10500845]
15. Dimeglio C, Gallón S, Loubes J, Maza E. A robust algorithm for template curve estimation based on manifold embedding. Computat Statistics Data Analysis 2014 Feb;70:373-386. [doi: 10.1016/j.csda.2013.09.030]
16. Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1979 Apr;PAMI-1(2):224-227. [doi: 10.1109/tpami.1979.4766909]
17. Baum A, Scarpa J, Bruzelius E, Tamler R, Basu S, Faghmous J. Targeting weight loss interventions to reduce cardiovascular complications of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the Look AHEAD trial. Lancet Diabetes Endocrinol 2017 Oct;5(10):808-815 [FREE Full text] [doi: 10.1016/S2213-8587(17)30176-6] [Medline: 28711469]

XSL•FO

RenderX

18. Triantafyllidis AK, Tsanas A. Applications of machine learning in real-life digital health interventions: review of the literature. J Med Internet Res 2019 Apr 05;21(4):e12286 [FREE Full text] [doi: 10.2196/12286] [Medline: 30950797]

19. Goldstein SP, Zhang F, Thomas JG, Butryn ML, Herbert JD, Forman EM. Application of machine learning to predict dietary lapses during weight loss. J Diabetes Sci Technol 2018 Sep;12(5):1045-1052 [FREE Full text] [doi: 10.1177/1932296818775757] [Medline: 29792067]

Abbreviations

BMI: body mass index
EWL: excess weight loss
HCA: hierarchical cluster analysis

XSL•FO
RenderX

Original Paper

Electronic Health Use in a Representative Sample of 18,497 Respondents in Norway (The Seventh Tromsø Study - Part 1): Population-Based Questionnaire Study

Rolf Wynn[1,2], MD, PhD; Sunday Oluwafemi Oyeyemi[3], MD, MSc; Andrius Budrionis[4], MSc, PhD; Luis Marco-Ruiz[4], MSc, PhD; Kassaye Yitbarek Yigzaw[4], MSc, PhD; Johan Gustav Bellika[1,4], MSc, PhD

[1]Department of Clinical Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway

[2]Division of Mental Health and Addictions, University Hospital of North Norway, Tromsø, Norway

[3]Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway

[4]Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway

Corresponding Author:
Johan Gustav Bellika, MSc, PhD
Department of Clinical Medicine
Faculty of Health Sciences
UiT The Arctic University of Norway
Postboks 6050 Langnes
Tromsø, 9037
Norway
Phone: 47 95748049
Email: johan.gustav.bellika@ehealthresearch.no

Abstract

Background: Electronic health (eHealth) services may help people obtain information and manage their health, and they are gaining attention as technology improves, and as traditional health services are placed under increasing strain. We present findings from the first representative, large-scale, population-based study of eHealth use in Norway.

Objective: The objectives of this study were to examine the use of eHealth in a population above 40 years of age, the predictors of eHealth use, and the predictors of taking action following the use of these eHealth services.

Methods: Data were collected through a questionnaire given to participants in the seventh survey of the Tromsø Study (Tromsø 7). The study involved a representative sample of the Norwegian population aged above 40 years old. A subset of the more extensive questionnaire was explicitly related to eHealth use. Data were analyzed using logistic regression analyses.

Results: Approximately half (52.7%; 9752/18,497) of the respondents had used some form of eHealth services during the last year. About 58% (5624/9698) of the participants who had responded to a question about taking some type of action based on information gained from using eHealth services had done so. The variables of being a woman (OR 1.58; 95% CI 1.47-1.68), of younger age (40-49 year age group: OR 4.28, 95% CI 3.63-5.04), with a higher education (tertiary/long: OR 3.77, 95% CI 3.40-4.19), and a higher income (>1 million kr [US $100,000]: OR 2.19, 95% CI 1.77-2.70) all positively predicted the use of eHealth services. Not living with a spouse (OR 1.14, 95% CI 1.04-1.25), having seen a general practitioner (GP) in the last year (OR 1.66, 95% CI 1.53-1.80), and having had some disease (such as heart disease, cancer, asthma, etc; OR 1.29, 95% CI 1.18-1.41) also positively predicted eHealth use. Self-rated health status did not significantly influence eHealth use. Taking some action following eHealth use was predicted with the variables of being a woman (OR 1.16, 95% CI 1.07-1.27), being younger (40-49 year age group: OR 1.72, 95% CI 1.34-2.22), having a higher education (tertiary/long: OR 1.65, 95% CI 1.42-1.92), having seen a GP in the last year (OR 1.58, 95% CI 1.41-1.77), and having ever had a disease (such as heart disease, cancer or asthma; OR 1.26, 95% CI 1.14-1.39).

Conclusions: eHealth appears to be an essential supplement to traditional health services for those aged above 40 years old, and especially so for the more resourceful. Being a woman, being younger, having higher education, having had a disease, and having seen a GP in the last year all positively predicted using the internet to get health information and taking some action based on this information.

XSL•FO
RenderX

KEYWORDS

adoption; digital health; eHealth; internet; Web search engine; health apps; social media; video service; population study; Tromsø study

Introduction

Online resources, including the use of search engines, social media, apps, and online video services, are becoming increasingly important for people in their everyday lives [1,2]. For example, 84% of US adults use the internet [3]. In the European Union (EU) in 2012, 73% of the population were online [4], and in Norway, in 2013, this was the case with 85% of the population [5]. When it comes to the specific use of social media, in 2016, a total of 68% of US adults used Facebook [6].

Electronic health (eHealth) is the use of Information and Communication Technology, such as the internet, to enable or improve health care [7]. While other eHealth activities, such as using health apps to make appointments and order prescriptions and using social media for communicating with health professionals, are becoming more popular, by far, the most frequent eHealth activity is finding information about health and illness on the internet [8-12]. Approximately 77% of online health searches start at a search engine [13]. In 2012, 59% of Americans had searched the internet for health information [12], and in 2013, 78% of Norwegians had used the internet for health purposes [8]. Prior studies have suggested that being young, being a woman, and being highly educated are central predictors of eHealth use [13]. A prior study found that 19% of smartphone users had at least one health app, and this was more frequent among women, those of a younger age, those with high incomes, and the college-educated [14]. One study found that 35.7% of all seniors and 89.1% of all those that were online used Facebook, Twitter, etc, to find and share health information [10]. Being a woman and highly educated predicted social media use for health information [10]. People with a primary health care provider, chronic disease, and of a younger age have been found to be more likely to use social media for health [11]. Thus, prior research has established that searching for health information is, overall, the most frequent eHealth activity. We also know that some demographic groups have been found to be more active online health-information searchers, but we lack updated information on this topic from a Nordic setting.

The aging population in many Western countries is likely to increase demands on health services. An increasing number of people with chronic illnesses are likely to stretch the capacity of health services further, and as many as 45% of US adults have one or more chronic illnesses [15]. eHealth may add to traditional services by engaging patients, helping patients to get information, empowering patients, increasing shared decision-making, and helping patients manage their health [16-21]. eHealth literacy is the competency required to use and make sense of eHealth tools and services [22], and it has been associated with younger age, a higher education, and having more devices [10]. eHealth has also been suggested as a means whereby shortages in health care staff and funding can be addressed (ie, the current crisis in the British National Health Service) [23]. Furthermore, eHealth, in combination with good eHealth literacy, might reduce the currently elevated number of unnecessary visits to doctors [24]. Thus, prior research has suggested that eHealth could help empower patients [25], and it has also been suggested that eHealth could help address resource shortages in the traditional health services, but the effect of eHealth use on traditional health service consumption has not been well established in the Nordic countries.

It remains unclear how eHealth influences traditional health service use in Norway, whether eHealth tools and services can replace traditional services or whether eHealth tools and services should be added to existing health service use [26]. Studies from other countries have found different results. One study found that 35% of US adults had gone online to figure out a medical condition and that 53% of these followed up with a visit to a medical professional [12]. Online self-diagnosing was more common among younger white adults, high earners, and the highly educated compared to others [12]. Other studies in the United States, Japan, and Taiwan have found that the use of the internet either increased or had no significant influence on traditional health care use [27-32]. For instance, Lee [27] found that increased internet use at survey wave one positively predicted health professional contact at survey wave two. Ybara and Suman [28] found that an increase in internet use increased the chance of the respondent visiting a physician. Hisieh et al [29], in a Taiwanese study controlling for a range of variables, including sex and age, found that increased internet use and chronic illness both predicted increased outpatient visits. Baker et al [31] found that female gender, younger age, higher education, and worse health were all variables that predicted increased internet use for health information searching but found that internet use had little influence on the use of traditional health professional consultations. Takahashi et al [32] found that younger age, higher education, and higher income were associated with an increase in searches for online health information but found no significant influence of online research on general practitioner (GP) visits. Interestingly, one Dutch study found a reduction in the use of traditional health services after the introduction of a high-quality, online health-information service [33]. More research is needed to explore how the use of eHealth will impact the use of traditional health services (such as GP visits) in Norway.

While prior studies have suggested that there are social divisions in the use of the internet for health purposes, many of these have been based on web-panels and other samples that might not have been fully representative of the general population [34]. In Europe, one major study included patients from seven countries [35], with a total of 8000 respondents. The study showed that Northern European countries, such as Denmark and Norway, had higher rates of eHealth use than Southern European countries, such as Portugal and Greece. Among those online, young adults, women, and the highly educated, as well as those who had visited a GP in the last year, and those who had a chronic illness more often searched online for health information. The study was carried out in 2005 and used random dialing within strata to sample respondents. A survey of 13,000

Europeans participating in an online panel found that younger adults, females, those living in larger households, those who had children or elderly family members, those with health problems, and those that were caring for others had a higher propensity towards intensive eHealth use [36]. Smaller studies from Europe of health information searching on the internet have suggested rates of searching are increasing across the continent [1,37,38].

In the United States, at least two larger studies have used representative samples: the Health Information National Trends Survey (HINTS) study [9,13,39,40] and the Pew Internet Study [3,6,12,14,15]. However, their findings might not be directly applicable to Western Europe, where services are organized differently and funded differently than in the United States. While the government, using taxation, funds nearly all health care in Nordic countries and many European countries (such as the United Kingdom), health care in the United States is typically based on different types of insurance. This means that while all citizens, regardless of their financial situation, have (at least in principle) equal access to health care in the Nordic model, access to and use of health services in the United States depends on an individual's insurance coverage. While little is known about the importance of traditional health service organization on eHealth use, one could speculate that eHealth could see less use in a model where traditional services are more or less free (ie, the Nordic model). A representative, population-based study in a Nordic country will give more reliable data about the use of eHealth services in the general population in a setting with a free-for-all (ie, tax-funded), government-operated health service, including data on the possibly increasing importance of more recent sources of health information, such as social media, apps, and online video services [41-44].

Nordic countries, especially the subarctic regions, are sparsely populated, and access to specialist health services may be limited in rural areas. eHealth services could be particularly relevant for stakeholders and policymakers in sparsely populated, rural, and remote areas [45,46]. It is essential for health providers and decision-makers in the health services to know how new media impacts information seeking about health and illness. Knowledge about the use of eHealth services, including health information consumption, may be used to modify and target health information to specific groups and to establish whether existing services are tailored to current needs [9]. Age, gender, educational level, health status, and others are variables that are likely to influence the use of eHealth services, as these factors may be of importance to questions that arise around the ability to use these services. In this light, it becomes central to establish which resources are available and are used, who utilizes eHealth services and who does not, and to what extent eHealth services are replacing traditional health services.

The seventh, population-based Tromsø Study included a questionnaire about the use of eHealth. In a series of four papers, we explore data from this questionnaire and how the use of eHealth related to a range of other variables that were measured in the Tromsø Study. In the first paper (this paper), we present our main findings regarding the characteristics of the participants and their use of eHealth. In the present study, we examined the use of eHealth in the population above 40 years of age, predictors of eHealth use, and predictors of participants acting following their use of eHealth services. In the second paper [47] we will present and discuss how having different illnesses influences the use of eHealth, in the third paper [48] we will examine outcomes of the use of eHealth, and in the fourth paper [49] we will study how eHealth consumption influences actual doctor visits.

Methods

The Seventh Survey of the Tromsø Study

The Tromsø Study is a population-based, longitudinal health study conducted by the University of Tromsø in cooperation with several other Norwegian public agencies [50]. Inhabitants of the municipality of Tromsø were invited to participate in the study. Tromsø is the major city of North Norway, with a population of about 75,000 inhabitants. The seventh survey of the Tromsø Study (Tromsø 7) was conducted in 2015-2016. All inhabitants from the age of 40 years old and older in the Tromsø municipality were mailed an invitation. A total of 21,083 subjects (10,009 men and 11,074 women) aged 40 years old and above attended, which was 65% of those invited to participate.

Questionnaire

As part of a more extensive questionnaire on health and illness (in total more than 300 questions), the participants completed a questionnaire with data about their use of different types of eHealth services. The following question was asked:

> *How often during the last year have you used the following Internet-services for information and advice on health and disease issues: Applications ('Apps') for smart phone or tablet?, Search engines (like Google)?, Social media (like Facebook)?, Video services (like YouTube)?*

For each item, it was possible to respond either "never," "once," "a few times," or "often." Those who responded that they had used at least one of the services were subsequently asked the following question:

> *If you during the last year have used Internet-services for information and advice on health and disease issues, based on the information you found on the Internet: Have you decided to go to the doctor?, Have you decided not to go the doctor?, Have you discussed the information with a doctor?, Have you changed your medication without consulting a doctor?, Have you been unsure whether the treatment you have received is correct?, Have you decided to seek out complementary or alternative treatment?, Have you made lifestyle changes?, Have you felt anxiety?, Have you felt reassured?, Have you felt more knowledgeable?, Have you felt more confused?*

For each of the items, it was possible to respond either "never," "once," "a few times," or "often."

The questions and their respective response options are also available online at the Tromsø Study website [51], and this was

the first time these types of questions on eHealth were included in the Tromsø Study. Participants could choose to complete the questionnaire on paper or online, with most completing the questionnaire at home. However, all participants were required to attend the study center in order to participate in the study. The questionnaire (of which the eHealth questions were a small part) was supplemented with a range of tests that required people to attend in person (ie, blood tests, body measurements, electrocardiograms, ultrasounds of various organs).

Study Sample

Variables obtained from the Tromsø 7 questionnaire included gender, age, education, occupation/work status, household income, whether the participant had seen a GP in the last year, assessment of own health, living status with a spouse, self-reported diseases, and use of the internet for finding health information. We excluded participants who had missing information on the use of the internet for health information searching (through search engines, social media, apps, or video services; n=384), and those with missing information on any of the other variables: gender, age, education, occupation, household income, GP consultation, assessment of own health, living status with spouse, and self-reported diseases (n=2202). The final analytical sample consisted of 18,497 participants (9138 men and 9359 women).

We also carried out separate analyses, including those who took health decisions (acted or not acted) following information gathering from online services (search engines, social media, apps, or video services). This subcohort included 9698 participants (4243 men and 5455 women), who had given information on these variables.

Assessment of the Use of the Internet for Health Information and Self-Reported Diseases

Information on the use of the internet for health and participants' self-reported diseases was taken from the Tromsø 7 questionnaires. Self-reported disease conditions included: high blood pressure, heart attack, heart failure, atrial fibrillation, angina, stroke, diabetes, kidney disease, bronchitis, asthma, cancer, rheumatoid arthritis, arthrosis, migraine, psychological problems, and chronic pain. The options on these questions were "no," or "yes," or "yes, previously."

The information on those (n=9698) who completed questions regarding the effect of using internet resources for health information or advice (through search engines, social media, health apps, or video services) was used in the subcohort analyses. The responses included in the present analyses were: if they had decided to visit (or not visit) the doctor, discussed information found online with a doctor, changed medication without consulting a doctor, if they had been unsure about whether the treatment they had received was correct, if they had made lifestyle changes, and if they had sought alternative or complementary treatment. The options were "never," "once," "a few times," or "often."

Statistical Analyses

We performed multivariable logistic regression analysis with the use of the internet for health information as the dichotomous dependent variable, and gender, age, education, occupation/work status, household income, GP consultation, assessment of own health, living status with spouse, and self-reported diseases as the independent variables. The use of the internet for health information was dichotomized into never/ever by grouping those who had never used any of the resources (search engines, social media, health apps, or video services) as never, and those who had used at least one of the resources for health advice as ever. Similarly, we grouped participants who never had any of the disease conditions as never, and those participants who previously or currently had at least one condition as ever. Age was grouped into four groups: 40-49, 50-59, 60-69, and 70 years old and older. Education was grouped into four groups: primary or partly secondary education (up to 10 years of school), upper secondary education (minimum of three years), short tertiary education (college or university for less than four years), and long tertiary education (college or university for four years or more). Occupation/work status was categorized into works full time, works part-time, unemployed, housekeeper, retired, student/in military service, on disability benefit or work assessment allowance, and on family income supplement. Household income in kr per annum: less than 250,000 (<US $25,000), 250,000-450,000 (US $25,000-$45,000), 451,000-750,000 (US $45,100-$75,000), 751,000-1,000,000 (US $75,100-$100,000), and more than 1,000,000 (>US $100,000). Living status with a spouse and consultation with the GP (during the last year) were either yes or no. Assessment of own health was either very bad, bad, neither good nor bad, good, or excellent.

We checked for possible interactions between education and occupation/work status, education and income, and occupation/work status and disease condition. We further explored the relationship between the use of the internet for health information and the independent variables stratified by disease conditions (never/ever).

All *P* values were considered statistically significant at a level of <.05, and all statistical tests were two-sided. We used Stata for Windows version 15.0 (StataCorp, College Station, Texas, United States) to conduct all statistical analyses.

Ethics

All participants gave written informed consent. The Regional Committee for Medical and Health Research Ethics approved the study (REK Nord, reference 2014/940).

Results

Participants' Characteristics

Regarding age, about 60% (11,036/18,497) of the participants were within the 40-59 years old age range. Only about 15% (2759/18,497) were 70 years old or older (see Multimedia Appendix 1). The male participants had a mean age of 57.5, while the female participants had a mean age of 56.9 years. Our study sample consisted of an approximately equal proportion of men (49.4%; 9138/18,497) and women (50.6%; 9359/18,497).

For education, occupation/work status, and income, about half of the participants had tertiary education while the other half had either primary or secondary school education. The

respondents were mostly in full time employment (60%; 11,188/18,497) or retired (21%; 3886/18,497). About half (51%; 9474/18,497) earned more than 750,000 kr (US $75,000) per annum, while less than 5% (890/18,497) earned 250,000 (US $25,000) or less.

A clear majority of the respondents (77.3%; 14,305/18,497) stated they were living with a spouse. As for health and psychological variables, most of the participants (80%; 14,781/18,497) had had at least one appointment with their GP during the last year, even though 70% (12,901/18,497) rated their health as excellent or good. About 73% (13,552/18,497) had had at least one of the diseases of interest in this study (see Multimedia Appendix 1).

The Use of Electronic Health Services

One of the main findings of this study was that 52.7% (9752/18,497) of the respondents in the last year had used at least one eHealth service (ie, search engine, social media, apps, or video services) to get information and advice about health and illness (see Multimedia Appendix 1). However, the odds of using the internet for health information decreased significantly with age, with senior citizens (70 years old or older) mostly at a disadvantage (OR 0.23, 95% CI 0.20-0.28) when compared to those in the age range of 40-49 years old (see Multimedia Appendix 1).

In the multivariable analyses, we found that women had 1.58 times the odds of using internet resources (at least one of these: search engine, social media, apps, or video services) for health information when compared to men (OR 1.58, 95% CI 1.47-1.68). Also, educational level and household income positively predicted the use of the internet for health information searching. Those who had a long tertiary education had 3.77 times the odds of using internet resources to look for health information compared to those who only had primary or partly secondary school education (OR 3.77, 95% CI 3.40-4.19). Similarly, those who earned the most were significantly at increased odds of using internet resources (OR 2.19, 95% CI 1.77-2.70) when compared to those who earned the least. Occupation or work status did not predict the use of internet resources for health information. However, those on disability benefits and other family welfare benefits had 1.71 times the odds of using the internet for health information when compared to those who worked full time (OR 1.71, 95% CI 1.05-2.78).

In regard to living with a spouse, those participants had 0.88 times the odds of using the internet for health information when compared to those who were not living with a spouse (OR 0.88, 95% CI 0.80-0.97). We also found that those who had consulted their GP in the last year had 1.66 times the odds of using internet resources for health information compared to those who had not consulted their GP. Similarly, those who had ever had at least one of the diseases of interest were at increased odds of using the internet for health information (OR 1.29, 95% CI 1.18-1.41). Intriguingly, assessment of own health did not predict the use of internet resources for health information searching (see Multimedia Appendix 1).

Taking Action After Obtaining Information

About 58% (5624/9696) of those who answered this question took some form of action after having obtained information about health and illness on the internet (see Multimedia Appendix 1). The action taken varied from deciding to see a doctor or not to see a doctor, discussing the information with a doctor, changing a medication without consulting a doctor, questioning previous treatment, deciding to seek alternative or complementary treatment, or changing lifestyle.

In the multivariable analyses of the subcohort (n=9698) who made health decisions following use of the internet for health information searching, we found that similar to the use of internet resources for health information, the odds of making health-related decisions following use decreased with age. Those aged 70 years old and above had nearly half the odds of making health-related decisions/actions when compared to those that were 40-49 years old (OR 0.58, 95% CI 0.45-0.75). Also, we found that women had 1.16 times the odds of making health-related decisions following the use of internet resources when compared to men (OR 1.16, 95% CI 1.07-1.27).

Regarding education and income, educational level positively predicted making health-related decisions following the use of internet resources for health information. Those with a long tertiary education had 1.65 times the odds of making health-related decisions following use when compared to those who had primary or partly secondary school education (OR 1.65, 95% CI 1.42-1.92). However, household income did not significantly predict health-related decision-making following the use of internet resources.

Unlike in the use of the internet for health information, not living with a spouse did not significantly predict health-related decision-making following the use of internet resources. Additionally, those who had consulted their GP in the last year (OR 1.58, 95% CI 1.41-1.77) and those who had had at least one of the diseases of interest (OR 1.26, 95% CI 1.41-1.39) had increased odds of taking health actions following internet resources use (see Multimedia Appendix 1). We found that occupation or work status and assessment of own health did not predict health-related decision-making following the use of internet resources, which is similar to our findings concerning the use of internet resources for health information.

Discussion

Use of Electronic Health and Predictors of Use

Overview

We found that approximately half of the respondents had used some form of eHealth during the last year. This figure is lower than what has been suggested in some prior studies [8,12,13]. However, as our study consisted of a sample of the population that is 40 years old and older, we did not include the younger generation that is likely to have a higher internet use. Moreover, our sample was from the general population and not restricted to internet users, which in part may explain the lower rate of eHealth use in our study [34].

Age

Younger age was a significant positive predictor of eHealth use, in line with the findings of several prior studies [8,13,17,35]. Unsurprisingly, the younger are more knowledgeable and comfortable with eHealth services, as more young people use online services [13]. This finding is also consistent with prior studies that have determined that younger users are more accurate and have more attention to detail when using eHealth resources [52,53]. Moreover, older adults may have barriers to technology use because of perceived complexity that may limit confidence and interest in engaging with the technology [54]. Internet use has been shown to be especially lower among those above 75 years of age [55]. Increased age has been related to lower rates of shared decision-making in traditional health services [56]. As the use of eHealth could be related to shared decision-making (ie, informed and empowered patients are more likely to be interested in shared decision-making), lower eHealth use in the oldest age groups might in part be explained by a lower rate of shared decision-making among the most senior.

Some prior studies have suggested higher rates of use in the older age groups than we did in the present study [26,39,57], but differences in the range of the age groups between studies make direct comparisons challenging. Another possible cause of differences in results regarding use could be related to whether participants are sampled from the general population or from groups of internet users (eHealth use is likely to be higher among internet users). It might be of concern that those who are in the age groups most in need of health services (ie, the elderly, who typically are iller than the younger generation) use these services less. Helping elderly patients find appropriate online information and better adapting the information to suit their needs (in terms of content, style, readability), might increase use [58]. However, this age-related difference might diminish as more older people are accessing the internet. In the EU in 2012, 42% of those in the 55-74 years old age group were regular users [4], and in 2015 a total of 58% of US senior citizens were online [3]. In summary, our finding that the younger use eHealth more often is supported by prior literature.

Gender

We also found that being a woman was a significant predictor of use, in line with previous findings [8,9,13,35,59]. This gender difference in eHealth use could be explained by the fact that women are more engaged in health care in general because they often serve as family caregivers, holders of health information, and care managers [60-62], and also have a higher use of social media [63], possibly because of gender roles. Thus, prior studies have suggested that women are more active eHealth users than men, and our findings strengthen this idea.

Education, Occupation/Work Status, and Income

We found that having a higher education positively predicted the use of eHealth. Higher education has also previously been shown to predict eHealth use [9,13,17,35], a finding that could be related to higher health literacy (ie, "the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions" [64]) and higher patient engagement (ie, to "promote and support active patient and public involvement in health and healthcare and to strengthen their influence on healthcare decisions, at both the individual and collective level" [65]) in health among the higher educated.

A review study has found that patients' engagement with digital health decreases with higher age and lower health literacy [66]. People working in some professions, such as those who routinely use the internet at work [67], may have more time to search for health information. Some prior studies have suggested that occupation and work status might be of importance [17,68], although the finding has not been consistent [8]. In our study, occupation and work status were, overall, not significant predictors. As our study has a large sample and consequently high statistical power, our finding may suggest that occupation and work status are less significant central predictors, at least in a Nordic setting. In line with a prior study [13], we found that increased household income also positively predicted eHealth use. In summary, while higher education and income positively predicted eHealth use in our study, we did not find a similar relationship for occupation and work status.

Living With a Spouse

While loneliness is known to increase the risk of death [69], living with a spouse reduces rates of illness and death from a range of illnesses [70]. These positive health effects have been associated with the support spouses offer each other [70]. An American study found that 39% of adults were caregivers, and many of these cared for their partner [71]. Those who search online for health information for others are more likely to live in households with others [72]. Drawing on some prior research [12,71], one could assume that spouses might have a higher use for eHealth because they might be searching for information relating to their partner. However, living with a spouse negatively predicted eHealth use in our study, possibly because spouses might either inform or comfort each other in such a way that the need for health information from other sources is reduced. Our finding that living with a spouse negatively predicted eHealth use thus stands somewhat in contrast to what some prior researchers have found.

Health and Psychological Variables

Health and psychological variables have, to varying degrees, been found to predict health-related internet use [17]. In our study, having seen a GP in the last year positively predicted eHealth use. Respondents who stated that they had (or previously had) an illness also used eHealth more, while self-reported health status was not a significant predictor. Poor or fair self-reported health or chronic illness has predicted health-related internet use in some other studies [8,31,35,73]. Fox and Duggan [12] found that living with a chronic disease had an independent negative effect on internet access in the United States. However, internet users with chronic illness were more likely to gather online information about medical problems. As a high percentage of Norwegians have internet access, one might consequently expect that having an illness would predict eHealth use, as we found in the present study. Thus, having an illness or having seen a GP positively predicted eHealth use in our study, as might be expected in a country with a very high internet access rate.

Taking Action Based on Online Information

Overview

The finding that about 6/10 acted based on information gained from using eHealth services suggests that health information on the internet plays a surprisingly important role in people's decision-making processes regarding their health care. The action taken included deciding to see a doctor or not to see a doctor. It is not surprising that health information may help people make such a decision, and many people probably search for health information online to get a basis for deciding whether they need professional help or not.

Other actions taken were discussing the information with a doctor, changing medication without talking with a doctor, deciding to see an alternative practitioner, or changing lifestyle. Prior studies have suggested that many patients obtain information from the internet that they want to discuss with their doctor [8,39,74], and sometimes such information may lead the patient to question the diagnosis or the treatment given by the doctor [74]. Lifestyle advice, such as advice relating to starting to exercise, stopping smoking, or dieting, is one of the most popular types of health-related information that people seek on the internet [8]. Many people use complementary and alternative medicine (CAM) and search for information related to CAM online [75,76]. In summary, we found that online health information was important for many in making decisions relating to their health.

Gender, Age, Education, and Health

Many of the same variables were of importance to acting on the information as to accessing it in the first place, and being a woman, being of younger age, having higher education, having seen a GP in the last year, and having ever had an illness all predicted taking some form of action. Searching for information and acting on this information are qualitatively different processes. However, both behaviors are determined by many of the same variables. Household income was not a predictor of acting on the information, possibly because health care is covered by national insurance in Norway. Thus, searching for information and acting on this information were predicted by mostly the same variables.

Electronic Health and Traditional Health Services

eHealth was associated with the use of traditional health services (ie, having seen the GP during the last year). It is possible that using online health information may increase traditional health service consumption. We know that health-related information on the internet, on social media, and video services may be wrong, misleading, or biased [17,77-79] and that this information may generate increased uncertainty or anxiety among users and result in a need for clarification and interpretation [27,29,39,80,81]. Doctors are still considered the most reliable source of information [82], and most (88%) Norwegians still favor seeing their GP face-to-face [8]. In paper 4 [49] in this series, we further explore the association between eHealth use and an increase in GP visits in Norway.

Reducing the Digital Divide

We have found that higher age, being male, and having lower education, not having an illness, and not having seen a GP in the last year were associated with a lower use of eHealth services. We do not know why some subgroups used the internet less for health purposes. We suggest that a lower degree of engagement in health, in general, might explain some of the differences in eHealth use. Furthermore, some may not access eHealth services because they are unaware of the service [83], because they find it difficult to use [84], because they find it irrelevant [85], or because they find it difficult to trust. Many websites with quality health information have low readability and may be difficult to understand for people with low literacy levels [86]. The fact that some groups use online health information and eHealth tools less often suggests that these services and tools need to be matched to the eHealth-use abilities of these underserved groups [84]. Suggesting appropriate sources of online information and using other types of health information, including traditional offline media, might be considered as a strategy for reaching those who use the internet less for health purposes [13].

Strengths and Limitations

This is the first representative, large-scale, population-based study of eHealth use in Norway. We have given a representative picture of the use of eHealth in a population 40 years old and older, predictors of eHealth use, and predictors of taking action following the use of eHealth services. There are important differences in the organization and funding of health care in the United States, Norway, and much of Europe. Despite these differences, lower age, female gender, higher educational level, and having a chronic illness seemed to predict increased eHealth use both in the United States and in Norway.

There are some central limitations to this study. There might be a self-selection bias because not everyone who was invited chose to participate. As this study was based on cross-sectional data obtained from questionnaires, there is a possibility of recall bias (ie, that participants either underestimated or overestimated their use of eHealth or their actions taken). However, the validity and reproducibility of self-reported (ie, recalled) findings from the Tromsø Study have been reported as quite high and of sufficient quality for epidemiological research [87,88]. Also, due to the cross-sectional design of the study, we are unable to establish causation. There is also a risk that there might be unmeasured confounding variables. We did not have variables on patients' trust in online information or on patients' literacy levels. One variable used was self-reported health, which has been shown to be influenced by socioeconomic class [89]. The questionnaire did not include more detailed questions about how people use and experience different eHealth services, and this issue is an important avenue for future large-scale studies.

Conclusions

About half of respondents used some form of eHealth in the last year, and about 6/10 of this half used the information to take some form of action. The use of eHealth was associated with the use of traditional health services. This study has provided new knowledge about the importance of the internet,

social media, apps, and online videos for health information and how this information impacts patients and the general public. While one might hope that eHealth services can benefit those most in need, the present study suggests that it is those with the most resources, the highly educated and well-off, that consume eHealth services the most. Being in poor health did not predict the use of online health information. Clinicians should be aware that many patients above 40 years of age use eHealth to find information about health and illness, and that they also often act on this information [26,39,57]. The provision

of high-quality eHealth services should, therefore, be a priority for clinicians and health service providers. Clinicians should be aware that the use of eHealth sometimes has important medical consequences, such as when patients decide not to visit their doctor or to stop taking their medication without consulting their doctor. Some groups of patients, such as the oldest and those with little education, appear to use eHealth less than other groups, possibly because the services are not adequately matched to their needs. Clinicians might consider recommending adapted online or paper-based information specifically for these groups.

Acknowledgments

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study tables.
[DOCX File , 28 KB - medinform_v8i3e13106_app1.docx]

References

1. Kummervold P, Wynn R. Health information accessed on the internet: the development in 5 European countries. Int J Telemed Appl 2012;2012:297416 [FREE Full text] [doi: 10.1155/2012/297416] [Medline: 23304133]
2. Wangberg S, Andreassen H, Kummervold P, Wynn R, Sørensen T. Use of the internet for health purposes: trends in Norway 2000-2010. Scand J Caring Sci 2009 Dec;23(4):691-696. [doi: 10.1111/j.1471-6712.2008.00662.x] [Medline: 19473317]
3. Perrin A, Duggan M. Pew Research Center. 2015. Americans´ Internet access 2000-2015. URL: http://www.pewinternet.org/2015/06/26/americans-internet-access-2000-2015/ [accessed 2019-12-02] [WebCite Cache ID 73J1bOrOr]
4. Seybert H. Eurostat, European Commission. 2012. Internet use in households and by individuals 2012. URL: https://ec.europa.eu/eurostat/web/products-statistics-in-focus/-/KS-SF-12-050?inheritRedirect=true [accessed 2018-10-20] [WebCite Cache ID 73J1tNXoT]
5. Statistics Norway. 2018. Norwegian Media Barometer 2013. URL: https://www.ssb.no/en/kultur-og-fritid/artikler-og-publikasjoner/norsk-mediebarometer-2013 [accessed 2018-10-20] [WebCite Cache ID 73J2J6khE]
6. Greenwood S, Perrin A, Duggan M. Pew Research Center. 2016. Social Media Update 2016. URL: http://www.pewinternet.org/2016/11/11/social-media-update-2016/ [accessed 2018-10-20] [WebCite Cache ID 73J2zYoIr]
7. Pagliari C, Sloan D, Gregor P, Sullivan F, Detmer D, Kahan JP, et al. What is eHealth (4): a scoping exercise to map the field. J Med Internet Res 2005 Mar 31;7(1):e9 [FREE Full text] [doi: 10.2196/jmir.7.1.e9] [Medline: 15829481]
8. Sørensen T, Andreassen H, Wangberg S. Prosjektrapport. Tromsø, Norway: Norwegian centre for e-health research; 2014. Ehelse i Norge 2013. URL: https://ehealthresearch.no/files/documents/Prosjektrapporter/NST-rapport_2014-02_e-helse_i_Norge_2013.pdf [accessed 2019-12-02] [WebCite Cache ID 73J38ARiT]
9. Atkinson NL, Saperstein SL, Pleis J. Using the internet for health-related activities: findings from a national probability sample. J Med Internet Res 2009 Feb 20;11(1):e4 [FREE Full text] [doi: 10.2196/jmir.1035] [Medline: 19275980]
10. Tennant B, Stellefson M, Dodd V, Chaney B, Chaney D, Paige S, et al. eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. J Med Internet Res 2015 Mar 17;17(3):e70 [FREE Full text] [doi: 10.2196/jmir.3992] [Medline: 25783036]
11. Thackeray R, Crookston BT, West JH. Correlates of health-related social media use among adults. J Med Internet Res 2013 Jan 30;15(1):e21 [FREE Full text] [doi: 10.2196/jmir.2297] [Medline: 23367505]
12. Fox S, Duggan M. Pew Research Center. 2013 Jan 15. Health online 2013. URL: http://www.pewinternet.org/2013/01/15/health-online-2013/ [accessed 2018-10-20] [WebCite Cache ID 73J3ImXkx]
13. Kontos E, Blake KD, Chou WS, Prestin A. Predictors of eHealth usage: insights on the digital divide from the Health Information National Trends Survey 2012. J Med Internet Res 2014 Jul 16;16(7):e172 [FREE Full text] [doi: 10.2196/jmir.3117] [Medline: 25048379]

14. Fox S, Duggan M. Pew Research Centre. 2012 Nov 8. Mobile Health 2012. URL: http://www.pewinternet.org/2012/11/08/mobile-health-2012/ [accessed 2018-10-20] [WebCite Cache ID 73J3T3jwr]

15. Fox S, Duggan M. Pew Research Center. 2013 Nov 26. The Diagnosis Difference. URL: http://www.pewinternet.org/2013/11/26/the-diagnosis-difference/ [accessed 2018-10-20] [WebCite Cache ID 73J3bPp5s]

16. Kreps GL, Neuhauser L. New directions in eHealth communication: opportunities and challenges. Patient Educ Couns 2010 Mar;78(3):329-336. [doi: 10.1016/j.pec.2010.01.013] [Medline: 20202779]

17. van Uden-Kraan CF, Drossaert CHC, Taal E, Smit WM, Moens HJB, Siesling S, et al. Health-related Internet use by patients with somatic diseases: frequency of use and characteristics of users. Inform Health Soc Care 2009 Jan;34(1):18-29. [doi: 10.1080/17538150902773272] [Medline: 19306196]

18. van Uden-Kraan C, Drossaert C, Taal E, Seydel E, van de Laar M. Participation in online patient support groups endorses patients' empowerment. Patient Educ Couns 2009 Jan;74(1):61-69. [doi: 10.1016/j.pec.2008.07.044] [Medline: 18778909]

19. Chewning B, Bylund CL, Shah B, Arora NK, Gueguen JA, Makoul G. Patient preferences for shared decisions: a systematic review. Patient Educ Couns 2012 Jan;86(1):9-18 [FREE Full text] [doi: 10.1016/j.pec.2011.02.004] [Medline: 21474265]

20. Santana S, Lausen B, Bujnowska-Fedak M, Chronaki CE, Prokosch H, Wynn R. Informed citizen and empowered citizen in health: results from an European survey. BMC Fam Pract 2011 Apr 16;12:20 [FREE Full text] [doi: 10.1186/1471-2296-12-20] [Medline: 21496309]

21. Friedman CP. A "fundamental theorem" of biomedical informatics. J Am Med Inform Assoc 2009;16(2):169-170 [FREE Full text] [doi: 10.1197/jamia.M3092] [Medline: 19074294]

22. Norman CD, Skinner HA. eHealth Literacy: Essential Skills for Consumer Health in a Networked World. J Med Internet Res 2006 Jun 16;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]

23. Iacobucci G. Stop rollout of online consultations until there is evidence of benefit to patients, say LMCs. BMJ 2017 Nov 10;359:j5247. [doi: 10.1136/bmj.j5247] [Medline: 29127095]

24. Campbell D. The Guardian. 2016 Nov 5. Millions of patients putting strain on NHS with minor ailments. URL: https://www.theguardian.com/society/2016/nov/05/millions-of-patients-putting-strain-on-nhs-with-minor-ailments [accessed 2018-10-20] [WebCite Cache ID 73J3mMS2r]

25. Risling T, Martinez J, Young J, Thorp-Froslie N. Evaluating Patient Empowerment in Association With eHealth Technology: Scoping Review. J Med Internet Res 2017 Sep 29;19(9):e329 [FREE Full text] [doi: 10.2196/jmir.7809] [Medline: 28963090]

26. Medlock S, Eslami S, Askari M, Sent D, de Rooij SE, Abu-Hanna A. The consequences of seniors seeking health information using the internet and other sources. Stud Health Technol Inform 2013;192:457-460. [Medline: 23920596]

27. Lee C. Does the internet displace health professionals? J Health Commun 2008;13(5):450-464. [doi: 10.1080/10810730802198839] [Medline: 18661387]

28. Ybarra ML, Suman M. Help seeking behavior and the Internet: a national survey. Int J Med Inform 2006 Jan;75(1):29-41. [doi: 10.1016/j.ijmedinf.2005.07.029] [Medline: 16129659]

29. Hsieh RW, Chen L, Chen T, Liang J, Lin T, Chen Y, et al. The Association Between Internet Use and Ambulatory Care-Seeking Behaviors in Taiwan: A Cross-Sectional Study. J Med Internet Res 2016 Dec 07;18(12):e319 [FREE Full text] [doi: 10.2196/jmir.5498] [Medline: 27927606]

30. Starcevic V, Berle D. Cyberchondria: towards a better understanding of excessive health-related Internet use. Expert Rev Neurother 2013 Feb;13(2):205-213. [doi: 10.1586/ern.12.162] [Medline: 23368807]

31. Baker L, Wagner TH, Singer S, Bundorf MK. Use of the Internet and e-mail for health care information: results from a national survey. JAMA 2003 May 14;289(18):2400-2406. [doi: 10.1001/jama.289.18.2400] [Medline: 12746364]

32. Takahashi Y, Ohura T, Ishizaki T, Okamoto S, Miki K, Naito M, et al. Internet use for health-related information via personal computers and cell phones in Japan: a cross-sectional population-based survey. J Med Internet Res 2011 Dec 14;13(4):e110 [FREE Full text] [doi: 10.2196/jmir.1796] [Medline: 22169526]

33. Spoelman WA, Bonten TN, de Waal MWM, Drenthen T, Smeele IJM, Nielen MMJ, et al. Effect of an evidence-based website on healthcare usage: an interrupted time-series study. BMJ Open 2016 Nov 09;6(11):e013166 [FREE Full text] [doi: 10.1136/bmjopen-2016-013166] [Medline: 28186945]

34. Hays RD, Liu H, Kapteyn A. Use of Internet panels to conduct surveys. Behav Res Methods 2015 Sep;47(3):685-690 [FREE Full text] [doi: 10.3758/s13428-015-0617-9] [Medline: 26170052]

35. Andreassen HK, Bujnowska-Fedak MM, Chronaki CE, Dumitru RC, Pudule I, Santana S, et al. European citizens' use of E-health services: a study of seven countries. BMC Public Health 2007 Apr 10;7:53 [FREE Full text] [doi: 10.1186/1471-2458-7-53] [Medline: 17425798]

36. Torrent-Sellens J, Díaz-Chao Á, Soler-Ramos I, Saigí-Rubió F. Modelling and Predicting eHealth Usage in Europe: A Multidimensional Approach From an Online Survey of 13,000 European Union Internet Users. J Med Internet Res 2016 Jul 22;18(7):e188 [FREE Full text] [doi: 10.2196/jmir.5605] [Medline: 27450189]

37. Rasmussen J, Rohde I, Ravn BL, Sørensen T, Wynn R. Developments in the population's use of the internet for health-related purposes [in Danish]. Ugeskr Laeger 2009 Mar 16;171(12):1007-1011. [Medline: 19284923]

38. Wynn R, Pettersen G, Rosenvinge J, Bergvik S, Friborg O. Nearly All Students and Patients Use the Internet to Search for Health Information: Results from a Scandinavian Study. In: Stud Health Technol Inform. 2014 Presented at: MIE 2014; August 31-September 3; Istanbul, Turkey p. 1244.

39. Chung JE. Patient-provider discussion of online health information: results from the 2007 Health Information National Trends Survey (HINTS). J Health Commun 2013;18(6):627-648. [doi: 10.1080/10810730.2012.743628] [Medline: 23590202]

40. Shen H, Xu J, Wang Y. Applying Situational Theory of Problem Solving in Cancer Information Seeking: A Cross-Sectional Analysis of 2014 HINTS Survey. J Health Commun 2019 Mar;24(2):165-173. [doi: 10.1080/10810730.2019.1587111] [Medline: 30849294]

41. Gabarron E, Årsand E, Wynn R. Social Media Use in Interventions for Diabetes: Rapid Evidence-Based Review. J Med Internet Res 2018 Aug 10;20(8):e10303 [FREE Full text] [doi: 10.2196/10303] [Medline: 30097421]

42. Gabarron E, Dorronzoro E, Bradway M, Rivera-Romero O, Wynn R, Årsand E. Preferences and interests of diabetes social media users regarding a health-promotion intervention. PPA 2018 Nov;Volume 12:2499-2506. [doi: 10.2147/ppa.s184369]

43. Gabarron E, Wynn R. Use of social media for sexual health promotion: a scoping review. Glob Health Action 2016;9:32193 [FREE Full text] [doi: 10.3402/gha.v9.32193] [Medline: 27649758]

44. Gabarron E, Serrano JA, Wynn R, Armayones M. Avatars using computer/smartphone mediated communication and social networking in prevention of sexually transmitted diseases among North-Norwegian youngsters. BMC Med Inform Decis Mak 2012 Oct 30;12:120 [FREE Full text] [doi: 10.1186/1472-6947-12-120] [Medline: 23110684]

45. Acharibasam JW, Wynn R. Telemental Health in Low- and Middle-Income Countries: A Systematic Review. Int J Telemed Appl 2018;2018:9602821 [FREE Full text] [doi: 10.1155/2018/9602821] [Medline: 30519259]

46. Bradford NK, Caffery LJ, Smith AC. Telehealth services in rural and remote Australia: a systematic review of models of care and factors influencing success and sustainability. Rural Remote Health 2016;16(4):4268 [FREE Full text] [Medline: 27817199]

47. Marco-Ruiz L, Wynn R, Budrionis A, Yigzaw K, Oyeyemi S, Bellika J. Impact of Illness on Electronic Health Use (The Seventh Tromsø Study - Part 2): Population-Based Questionnaire Study. J Med Internet Res 2020 Mar 05;22(3):e13116 [FREE Full text] [doi: 10.2196/13116]

48. Budrionis A, Wynn R, Ruiz L, Yigzaw K, Bergvik S, Oyeyemi S, et al. Impact of the Use of Electronic Health Tools on the Psychological and Emotional Well-Being of Electronic Health Service Users (The Seventh Tromsø Study - Part 3): Population-Based Questionnaire Study. J Med Internet Res 2020 Mar 05;22(3):e13118 [FREE Full text] [doi: 10.2196/13118]

49. Yigzaw K, Ruiz L, Budrionis A, Oyeyemi S, Fagerlund A, Wynn R, et al. The Association Between Health Information Seeking on the Internet and Doctor Visits (The Seventh Tromsø Study - Part 4): Population-Based Questionnaire Study. J Med Internet Res 2020 Mar 05;22(3):e13120 [FREE Full text] [doi: 10.2196/13120]

50. Jacobsen BK, Eggen AE, Mathiesen EB, Wilsgaard T, Njølstad I. Cohort profile: the Tromso Study. Int J Epidemiol 2012 Aug;41(4):961-967 [FREE Full text] [doi: 10.1093/ije/dyr049] [Medline: 21422063]

51. The Tromsø Study. Tromsø, Norway: UIT-The Arctic University of Norway; 2019. Welcome to research data from the Tromsø Study. URL: http://tromsoundersokelsen.uit.no/tromso/ [accessed 2019-12-02] [WebCite Cache ID 78JiCzDqV]

52. Marco-Ruiz L, Bønes E, de la Asunción E, Gabarron E, Aviles-Solis JC, Lee E, et al. Combining multivariate statistics and the think-aloud protocol to assess Human-Computer Interaction barriers in symptom checkers. J Biomed Inform 2017 Oct;74:104-122 [FREE Full text] [doi: 10.1016/j.jbi.2017.09.002] [Medline: 28893671]

53. Luger TM, Houston TK, Suls J. Older adult experience of online diagnosis: results from a scenario-based think-aloud protocol. J Med Internet Res 2014 Jan 16;16(1):e16 [FREE Full text] [doi: 10.2196/jmir.2924] [Medline: 24434479]

54. Hill R, Betts LR, Gardner SE. Older adults' experiences and perceptions of digital technology: (Dis)empowerment, wellbeing, and inclusion. Computers in Human Behavior 2015 Jul;48:415-423. [doi: 10.1016/j.chb.2015.01.062]

55. Zickuhr K, Madden M. Pew Research Center. 2012 Jun 6. Older adults and internet use. URL: https://www.pewresearch.org/internet/2012/06/06/older-adults-and-internet-use/ [accessed 2019-12-02]

56. Haesebaert J, Adekpedjou R, Croteau J, Robitaille H, Légaré F. Shared decision-making experienced by Canadians facing health care decisions: a Web-based survey. CMAJ Open 2019;7(2):E210-E216 [FREE Full text] [doi: 10.9778/cmajo.20180202] [Medline: 30948650]

57. Johnson E, Johnson C S. Internet Use and Access to Health Information among Canadians: Are the Elderly on the Sidelines? J Gerontol Geriatr Res 2016;05(06). [doi: 10.4172/2167-7182.1000367]

58. Nguyen MH, Smets EMA, Bol N, Loos EF, Van Weert JCM. How Tailoring the Mode of Information Presentation Influences Younger and Older Adults' Satisfaction with Health Websites. J Health Commun 2018 Jan;23(2):170-180. [doi: 10.1080/10810730.2017.1421729] [Medline: 29345531]

59. Wangberg SC, Sørensen T, Andreassen HK. Using the Internet to Support Exercise and Diet: A Stratified Norwegian Survey. Med 2 0 2015 Aug 26;4(2):e3 [FREE Full text] [doi: 10.2196/med20.4116] [Medline: 26310277]

60. Etingen B, Miskevics S, Malhiot A, LaVela SL. Patient Engagement in VA Health Care: Does Gender Play a Role? Defence and Peace Economics 2018 May 16:1-10. [doi: 10.1080/10242694.2018.1465676]

61. National Alliance for Caregiving. 2009 Nov. Caregiving in the US 2009. URL: https://www.caregiving.org/pdf/research/Caregiving_in_the_US_2009_full_report.pdf [accessed 2019-12-02]

62. Friedemann M, Buckwalter K. Family Caregiver Role and Burden Related to Gender and Family Relationships. J Fam Nurs 2014 Aug;20(3):313-336 [FREE Full text] [doi: 10.1177/1074840714532715] [Medline: 24777069]

63. Ramanadhan S, Viswanath K. Health and the information nonseeker: a profile. Health Commun 2006;20(2):131-139 [FREE Full text] [doi: 10.1207/s15327027hc2002_4] [Medline: 16965250]

XSL•FO
RenderX

64. Healthypeople.gov. Washington DC: Office of Disease Prevention and Health Promotion; 2019 Dec 03. Social Determinants of Health. URL: https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health [accessed 2019-12-02]

65. Coulter A. Engaging Patients In Healthcare. Berkshire, England: Open University Press; 2019.

66. Irizarry T, DeVito Dabbs A, Curran CR. Patient Portals and Patient Engagement: A State of the Science Review. J Med Internet Res 2015 Jun 23;17(6):e148 [FREE Full text] [doi: 10.2196/jmir.4255] [Medline: 26104044]

67. Vitak J, Crouse J, LaRose R. Personal Internet use at work: Understanding cyberslacking. Computers in Human Behavior 2011 Sep;27(5):1751-1759. [doi: 10.1016/j.chb.2011.03.002]

68. Bass SB, Ruzek SB, Gordon TF, Fleisher L, McKeown-Conn N, Moore D. Relationship of Internet health information use with patient behavior and self-efficacy: experiences of newly diagnosed cancer patients who contact the National Cancer Institute's Cancer Information Service. J Health Commun 2006 Mar;11(2):219-236. [doi: 10.1080/10810730500526794] [Medline: 16537289]

69. Hawkley LC, Cacioppo JT. Loneliness matters: a theoretical and empirical review of consequences and mechanisms. Ann Behav Med 2010 Oct;40(2):218-227 [FREE Full text] [doi: 10.1007/s12160-010-9210-8] [Medline: 20652462]

70. Kiecolt-Glaser JK, Newton TL. Marriage and health: His and hers. Psychological Bulletin 2001;127(4):472-503. [doi: 10.1037//0033-2909.127.4.472]

71. Fox S, Duggan M, Purcell K. Pew Research Center. 2013 Jun 20. Family Caregivers are Wired for Health. URL: http://www.pewinternet.org/2013/06/20/family-caregivers-are-wired-for-health/ [accessed 2019-12-02] [WebCite Cache ID 73J41FmNQ]

72. Cutrona SL, Mazor KM, Vieux SN, Luger TM, Volkman JE, Finney Rutten LJ. Health information-seeking on behalf of others: characteristics of "surrogate seekers". J Cancer Educ 2015 Mar;30(1):12-19 [FREE Full text] [doi: 10.1007/s13187-014-0701-3] [Medline: 24989816]

73. Wagner TH, Baker LC, Bundorf MK, Singer S. Use of the Internet for health information by the chronically ill. Prev Chronic Dis 2004 Oct;1(4):A13 [FREE Full text] [Medline: 15670445]

74. Tan SS, Goonawardene N. Internet Health Information Seeking and the Patient-Physician Relationship: A Systematic Review. J Med Internet Res 2017 Jan 19;19(1):e9 [FREE Full text] [doi: 10.2196/jmir.5729] [Medline: 28104579]

75. Fox S, Purcell K. Pew Research Centre. 2010. Chronic Disease and the Internet. URL: http://www.pewinternet.org/2010/03/24/chronic-disease-and-the-internet/ [accessed 2019-12-02] [WebCite Cache ID 73J4EAQGh]

76. National Center for Complementary and Integrative Health. 2007. The Use of Complementary and Alternative Medicine in the United States. URL: https://nccih.nih.gov/research/statistics/2007/camsurvey_fs1.htm#use [accessed 2018-10-20] [WebCite Cache ID 73J4PlbO0]

77. Gabarron E, Serrano JA, Wynn R, Lau AYS. Tweet content related to sexually transmitted diseases: no joking matter. J Med Internet Res 2014 Oct 06;16(10):e228 [FREE Full text] [doi: 10.2196/jmir.3259] [Medline: 25289463]

78. Oyeyemi SO, Gabarron E, Wynn R. Ebola, Twitter, and misinformation: a dangerous combination? BMJ 2014 Oct 14;349:g6178. [doi: 10.1136/bmj.g6178] [Medline: 25315514]

79. Chen Y, Chen L, Kao Y, Chu T, Huang T, Ko W. The over-optimistic portrayal of life-supporting treatments in newspapers and on the Internet: a cross-sectional study using extra-corporeal membrane oxygenation as an example. BMC Med Ethics 2014 Aug 01;15:59 [FREE Full text] [doi: 10.1186/1472-6939-15-59] [Medline: 25081491]

80. Slev VN, Mistiaen P, Pasman HRW, Verdonck-de Leeuw IM, van Uden-Kraan CF, Francke AL. Effects of eHealth for patients and informal caregivers confronted with cancer: A meta-review. Int J Med Inform 2016 Mar;87:54-67. [doi: 10.1016/j.ijmedinf.2015.12.013] [Medline: 26806712]

81. Rising CJ, Bol N, Kreps GL. Age-Related Use and Perceptions of eHealth in Men With Prostate Cancer: A Web-Based Survey. JMIR Cancer 2015 Jun 25;1(1):e6 [FREE Full text] [doi: 10.2196/cancer.4178] [Medline: 28410165]

82. van de Poll-Franse LV, van Eenbergen MCHJ. Internet use by cancer survivors: current use and future wishes. Support Care Cancer 2008 Oct;16(10):1189-1195. [doi: 10.1007/s00520-008-0419-z] [Medline: 18293014]

83. Jung M, Loria K. Acceptance of Swedish e-health services. J Multidiscip Healthc 2010 Nov 16;3:55-63 [FREE Full text] [doi: 10.2147/JMDH.S9159] [Medline: 21289860]

84. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan 27;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]

85. Zickuhr K. Pew Research Center. 2013 Sep 25. Who's not online and why. URL: http://www.pewinternet.org/2013/09/25/whos-not-online-and-why/ [accessed 2019-12-02] [WebCite Cache ID 73J4eW0Gx]

86. McInnes N, Haglund BJA. Readability of online health information: implications for health literacy. Inform Health Soc Care 2011 Dec;36(4):173-189. [doi: 10.3109/17538157.2010.542529] [Medline: 21332302]

87. Lundblad MW, Jacobsen BK. The reproducibility of self-reported age at menarche: The Tromsø Study. BMC Womens Health 2017 Aug 22;17(1):62 [FREE Full text] [doi: 10.1186/s12905-017-0420-0] [Medline: 28830397]

88. Engstad T, Bonaa KH, Viitanen M. Validity of self-reported stroke : The Tromso Study. Stroke 2000 Jul;31(7):1602-1607. [doi: 10.1161/01.str.31.7.1602] [Medline: 10884460]

XSL•FO

RenderX

89.   Delpierre C, Lauwers-Cances V, Datta GD, Lang T, Berkman L. Using self-rated health for analysing social inequalities in health: a risk for underestimating the gap between socioeconomic groups? J Epidemiol Community Health 2009 Jun;63(6):426-432. [doi: 10.1136/jech.2008.080085] [Medline: 19218251]

Abbreviations

CAM: complementary and alternative medicine
eHealth: electronic health
EU: European Union
GP: general practitioner
HINTS: Health Information National Trends Survey
Tromsø 7: seventh survey of the Tromsø Study

XSL•FO
RenderX

Original Paper

Towards an Adoption Framework for Patient Access to Electronic Health Records: Systematic Literature Mapping Study

Hugo J T van Mens[1,2], MSc; Ruben D Duijm[1], MSc; Remko Nienhuis[2], MSc; Nicolette F de Keizer[1], PhD; Ronald Cornet[1], PhD

[1]Department of Medical Informatics, Amsterdam Public Health, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, Netherlands

[2]Department of Research & Development, ChipSoft, Amsterdam, Netherlands

Corresponding Author:
Hugo J T van Mens, MSc
Department of Medical Informatics
Amsterdam Public Health
Amsterdam University Medical Centers, University of Amsterdam
Meibergdreef 9, PO Box 22700
Amsterdam, 1100 DE
Netherlands
Phone: 31 20 566 51 88
Email: h.j.vanmens@amsterdamumc.nl

Abstract

Background: Patient access to electronic health records (EHRs) is associated with increased patient engagement and health care quality outcomes. However, the adoption of patient portals and personal health records (PHRs) that facilitate this access is impeded by barriers. The Clinical Adoption Framework (CAF) has been developed to analyze EHR adoption, but this framework does not consider the patient as an end-user.

Objective: We aim to extend the scope of the CAF to patient access to EHRs, develop guidance documentation for the application of the CAF, and assess the interrater reliability.

Methods: We systematically reviewed existing systematic reviews on patients' access to EHRs and PHRs. Results of each review were mapped to one of the 43 CAF categories. Categories were iteratively adapted when needed. We measured the interrater reliability with Cohen's unweighted kappa and statistics regarding the agreement among reviewers on mapping quotes of the reviews to different CAF categories.

Results: We further defined the framework's inclusion and exclusion criteria for 33 of the 43 CAF categories and achieved a moderate agreement among the raters, which varied between categories.

Conclusions: In the reviews, categories about people, organization, system quality, system use, and the net benefits of system use were addressed more often than those about international and regional information and communication technology infrastructures, standards, politics, incentive programs, and social trends. Categories that were addressed less might have been underdefined in this study. The guidance documentation we developed can be applied to systematic literature reviews and implementation studies, patient and informal caregiver access to EHRs, and the adoption of PHRs.

Introduction

Patient access to electronic health records (EHRs) is becoming increasingly common and is even a legal right in many countries. EHRs have been associated with increased patient engagement and improved health care quality outcomes [1-8]. However, there are also barriers to patients' access to EHRs. For example, some patients have difficulties logging in to patient portals and personal health records (PHRs), which facilitate access, due to complicated security procedures [1-8]. A framework is needed to assess the determinants and outcomes of PHR and EHR adoption that facilitates this access. This framework should

XSL·FO
RenderX

consider patients and informal caregivers as users rather than health care providers alone. This framework would enable the comparison and aggregation of evidence, and provide an overview of any important factors involved, which can then be used as a guide in implementations and health care policies as well as to address the gaps in knowledge.

"The Clinical Adoption Framework (CAF) is a general evaluation framework to assess the success of health information system (HIS) adoption in healthcare organisations" [9,10]. PHRs and EHRs are types of HISs, and thus this framework is also applicable to them. "As shown in Figure 1, it addresses the micro level, which encompasses the dimensions of quality, use and net benefits of the HIS; the meso level, consisting of the dimensions people, organisation and implementation; and the macro level, incorporating the dimensions healthcare standards, legislation, policy and governance, funding and incentives, and

societal, political and economic trends" [10]. Within each dimension, several categories were distinguished, for example "01. Functionality", "02. Performance", and "03. Security" are categories of the dimension "HIS quality" at the micro level. "It is hence an integrated framework that covers a wide range of aspects involved in HIS adoption" [10]. The CAF was developed and validated through consultation with health information technology professionals, comparisons with other survey instruments, and a meta-review of 50 systematic reviews on HIS implementation [11]. Categories, dimensions, and levels of the CAF were originally described by Lau, Price, and Kashevjee [9]. Throughout the categories, dimensions, and levels there are feedback loops, which are indicated by the arrows in Figure 1, that resembles the interplay between the factors and nondeterministic characteristics of HIS adoption and the outcomes of HIS use [9,11]. The CAF was applied in over 30 studies [12-18].

Figure 1. Clinical Adoption Framework with levels, dimensions, and categories. HIS: health information system. Originally published in [10, 18].
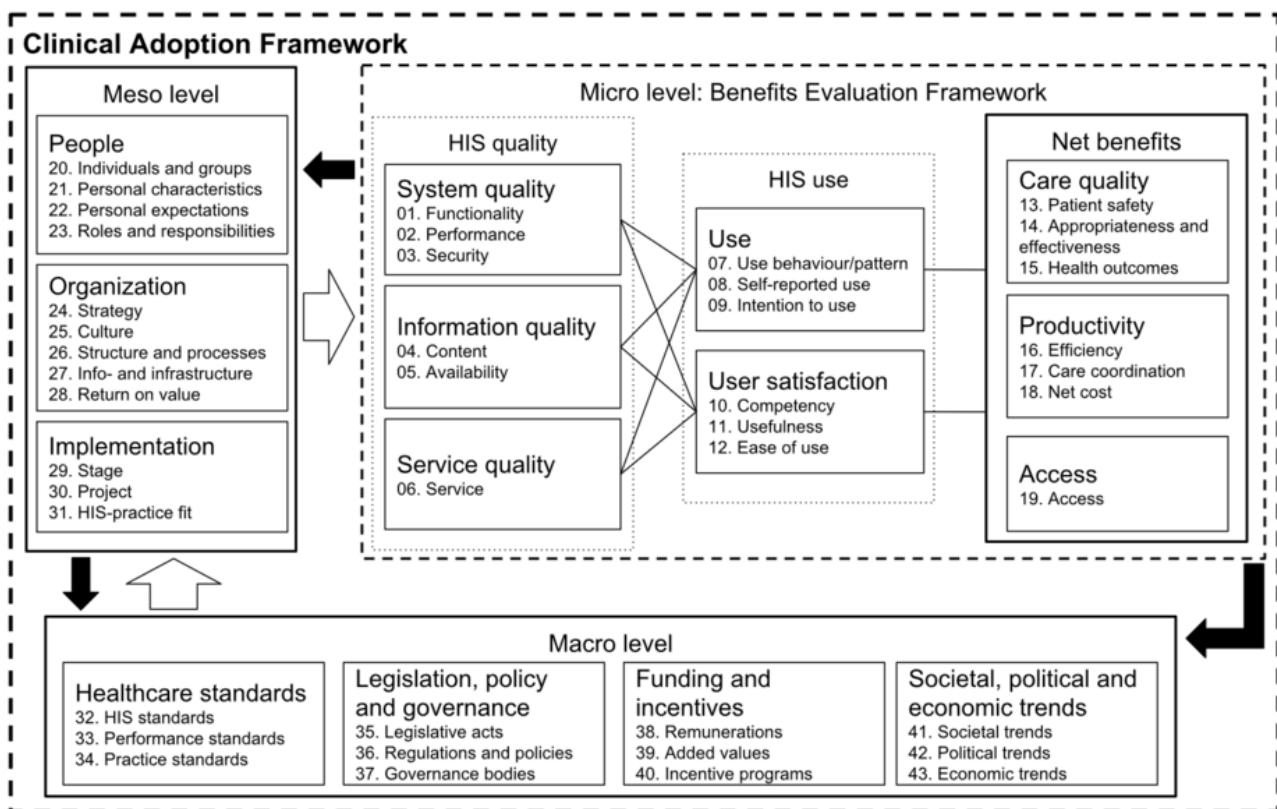
The CAF is a complex framework consisting of 43 categories that belong to 15 dimensions (illustrated as small boxes in Figure 1), which are further separated into the 3 previously mentioned micro, meso, and macro levels [9]. The CAF was considered difficult to apply as there was no guidance documentation with explicit descriptions and rules regarding its use [11]. Consequently, studies [15-18] that have applied the CAF differed in their interpretations and applications. Furthermore, HIS adoption increasingly involves sharing medical data with patients and informal caregivers. Therefore, patients and caregivers should also be considered when understanding successful HIS implementation, because they might value different factors than health care providers. This patient and

caregiver perspective was not explicitly taken into account during the development of the CAF.

The primary objective of this study was to extend the CAF to make it useful for evaluating patients' access to EHRs and the adoption of PHRs. The second objective was to improve the consistent application of the CAF in literature and implementation studies. For this purpose, we aimed to assess the interrater reliability of applying the framework, discuss which areas of the CAF could be improved, and develop guidance documentation.

Methods

We systematically reviewed existing systematic review papers on determinants and outcomes of patients' access to their personal health data. Results from each review paper were mapped to categories in the CAF, which was adapted when needed to reach consensus. The protocol for this review study was developed using the first 6 review papers [19-24], which were the most recent publications at the beginning of this review study. We used 13 subsequent review papers [25-37] in this study to refine the CAF and to assess the interrater reliability. The review protocol was registered at PROSPERO under CRD42018084542 [38]. We then reported the results of adapting the CAF, including its reliability, to make it suitable for an evaluation of the adoption of PHRs and patients' access to EHRs. The results of the review study on the determinants and outcomes of patients' access to medical records were reported separately [10].

To improve the CAF and its definitions, one reviewer (HM) extracted quotes from the literature that described determinants and outcomes for the adoption of EHRs and PHRs, and another reviewer (RD) verified these extracted quotes. The two reviewers independently mapped the extracted quotes. The interrater reliability for the agreement on the mapping was calculated with Cohen's unweighted kappa [39,40]. Each quote was mapped to two CAF categories: one for the determinant of the quote and the other for the outcome. Within each category, the quotes were classified into metrics by thematic analysis, as illustrated in Figure 2. The metrics and categorizations were iteratively revised to ensure consistent and meaningful categories for summarizing results, which was similar to the process described in Bassi, Lau, and Lesperance [15]. The mapping to two categories is visualized in Figure 2. For example, in the quote "Online record access and service users tended to be slightly older (t-test, $P<0.001$)" [31], the determinant metric could be "Age" and the outcome metric "Adoption". Age would be classified as "21. Personal characteristics", under the dimension "People" at the meso level, while adoption would be classified as "07. Use behaviour/pattern", under the dimension "Use" at the micro level. For the sake of the review, we added the category "44. Other" to denote when a quote could not be classified using the CAF.

Figure 2. Example of how a quote is mapped to a determinant metric and outcome metric. CAF: Clinical Adoption Framework.
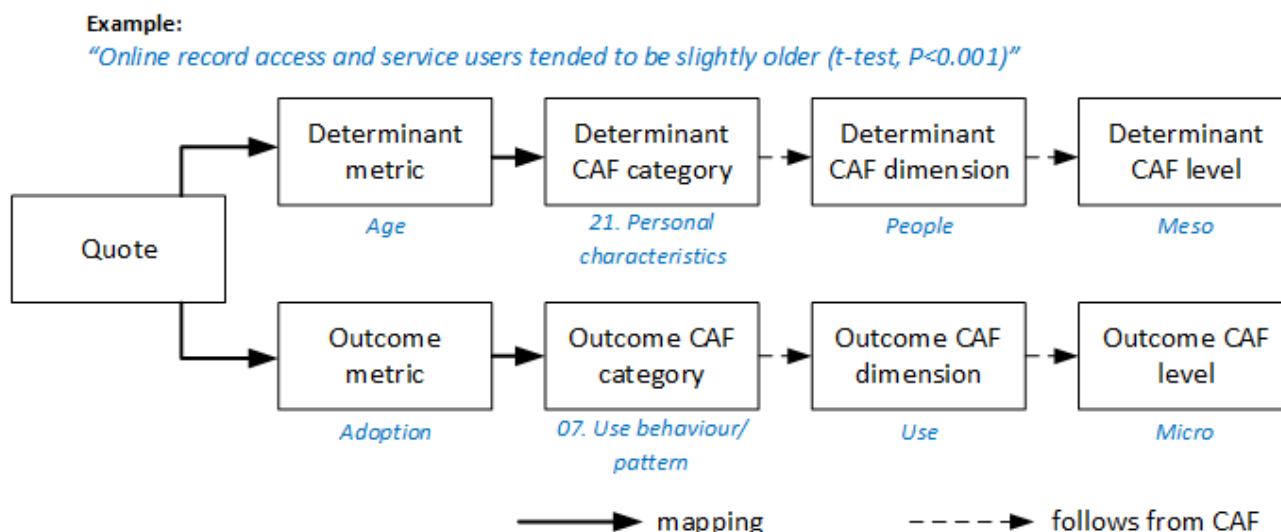
The results of this mapping and the differences in quote interpretation and CAF categorization were discussed among the two reviewers to achieve consensus. When necessary to achieve consensus, the definitions of the CAF were adapted and extended with inclusion and exclusion criteria to make them clearer. We presented the number of definitions for categories that were introduced, extended, or unchanged in each level. For agreements and disagreements between reviewers on mapping quotes to categories, we calculated the number of times each unique combination was agreed or disagreed upon (ie, number of times there was agreement on one certain category or disagreement between two specific categories). We counted the number of quotes that were classified into each category by a reviewer as well as how many quotes could not be mapped to the CAF. The level of agreement between reviewers on mapping quotes to each category indicated how ambiguous or well-defined the category was. This process resulted in defined categories of the CAF with inclusion and exclusion criteria and a list of metrics that we distinguished. Statistical analysis was carried out in R version 3.5.0 (R Foundation for Statistical Computing, Vienna, Austria) with RStudio 1.1.453 (RStudio Inc, Boston, MA). The R script can be found in Multimedia Appendix 1.

Results

In this section, we first list the definitions that were unchanged, extended, or introduced. Second, we discuss the interrater reliability and the spread of mapping quotes to CAF categories.

Adaptation of CAF Categories and Found Metrics

Definitions were introduced to the CAF for the 19 micro level categories, because they were missing in the original publication of the CAF. For example, the category "01. Functionality" of the dimension "System quality" was defined with the inclusion criteria "Actual or missing features/functionalities of the HIS

and their quality" and the exclusion criteria "If adoption or use of the HIS in general, without a particular functionality, then choose 07. Use behaviour/pattern". Thus, the exclusion criteria were made explicit for when a quote must be classified in another category. For the 24 meso- and macro level categories, the definitions from Lau, Price, and Kashevjee [9] were used, either unchanged (9 categories) or extended (15 categories), to cover cases of patient and informal caregiver use and disambiguate the categories with refinements and exclusion criteria. For example, the definition for the category "21. Personal characteristics" of the dimension "People" was

extended with "socio-economic status, ethnicity, computer skills, (health) literacy, health status" and "Behaviour". These are factors that were found to be important for the adoption of a HIS by patients and caregivers, and were not included in the original CAF category definition. Table 1 shows the number of categories for each level and how many were introduced, extended, or unchanged. Table 2 shows the categories that were changed and provides an example for each level. The resulting definitions for disambiguation in each category are listed in Table A in Multimedia Appendix 2. The metrics of each category can be found in Table B in Multimedia Appendix 2.

Table 1. The number of categories with introduced, extended, and unchanged definitions per level.

Level	Introduced	Extended	Unchanged	Total
Micro	19	0	0	19
Meso	0	9	3	12
Macro	0	6	6	12

Table 2. Categories where inclusion and exclusion criteria were added.

Level	Categories changed	Example [additions in brackets]
Micro	All categories from "01. Functionality" to "19. Access"	Inclusion criteria introduced for "01. Functionality":
		[Actual or missing features/functionalities of the HIS[a] and their quality.]
		Exclusion criteria introduced for "01. Functionality":
		[For adoption or use of the HIS in general, not a particular functionality, use category "07. Use behaviour/pattern"]
Meso	• "20. Individuals and groups" • "21. Personal characteristics" • "22. Personal expectations" • "23. Roles and responsibilities" • "25. Culture" • "27. Info- and infrastructure" • "28. Return on value" • "30. Project" • "31. HIS-practice fit"	Inclusion criteria extended for "21. Personal characteristics": "Degree to which an individual's characteristics, such as age, gender, education, [socio-economic status, ethnicity, computer skills, (health) literacy, health status,] experience and expertise can affect the adoption of an HIS" [9]. [Behaviour].
Macro	• "35. Legislative acts" • "36. Regulations and policies" • "39. Added values" • "41. Societal trends" • "42. Political trends" • "43. Economic trends"	Definition extended with exclusion criteria for "35. Legislative acts": [For privacy concerns use category "22. Personal expectation."]

[a]HIS: health information system.

Interrater Reliability and Spread

From the 13 reviews [25-37], we extracted 624 quotes. Each of the 624 quotes were mapped twice (ie, to a determinant and an outcome category) resulting in 1248 mappings. We achieved a percentage agreement of 67.0% (418) and a kappa of 0.58 for the determinant category, and a percentage agreement of 62.5% (390) and a kappa of 0.55 for the outcome category. As shown in Table C and Table D in Multimedia Appendix 2, the three categories that were least ambiguous, based on their high agreement scores, were "16. Efficiency", "21. Personal characteristics", and "13. Patient safety". In contrast, categories "09. Intention to use", "04. Content", and "30. Project" showed low agreement scores. Some disagreements between two

categories occurred more often than others. For example, a feature relating to secure messaging or access to medical records was interpreted by one reviewer as "'01. Functionality" and the other reviewer as "'07. Use behaviour/pattern" 94 times. This happened for instance with the quote "Patients experienced easier communication and interactive discussion with their physician after reading the medical file" [37]. There was one quote that did not fit into any one of the categories: "Two articles proposed achieving data exchange by setting up (Regional) Health Information Exchanges that can standardize data and facilitate exchange among different organizations." [24]. This result referred to infrastructure that exists outside of an organization to facilitate data exchange between organizations and would, therefore, fall into the macro level.

Discussion

Principal Findings

The definitions of the CAF categories were extended to be applicable to patient access to EHRs and the adoption of PHRs. This was achieved by adding factors that were found in the reviewed literature on patient access to EHRs, but were not present in the CAF yet, as was illustrated in the example of "21. Personal characteristics" in the results section. In addition, we developed guidance documentation in the form of inclusion criteria, exclusion criteria, and a list of metrics found. The interrater reliability of the reviewers applying the adapted CAF was moderate. However, we found the CAF to be a highly suitable and comprehensive framework to address patients' access to EHRs, as we could achieve consensus on the mappings through discussion, and almost all results could be categorized in the CAF. The original content for the definitions of the CAF were unchanged and only extended with additional inclusion and exclusion criteria for disambiguation and for the application to patients' access to EHRs. The number of agreements and disagreements and percentage of agreements varied among the CAF categories, just like the number of quotes that were mapped to each category. Some categories were not found at all in the reviews, especially those on the macro level.

Strengths and Weaknesses of the Study

We showed how the CAF can be applied to studies evaluating patient access to EHRs and PHRs. Despite many publications on the application of the CAF, we are the first, to our knowledge, to provide measures on the interrater reliability. However, the unweighted Cohen's kappa does not consider that categories actually reflect an order and results within each review are all correlated and come from the same study. Nonetheless, the moderate agreement indicates that the extended CAF is applicable in a consistent way. Because this study was a systematic review of systematic reviews, we have not investigated how to apply these results in primary implementation studies. The categories that were mapped to a lesser extent might have been underdefined, especially those at the macro level. It is possible that these categories may not have been reported in the literature, but also the literature may not have addressed the topics from those categories, or those categories could have been reported in other types of literature such as in policy, law, or gray literature, rather than scientific medical literature. Those categories with relatively high disagreement should also be further evaluated and redefined. Furthermore, the CAF could be used in studies to present their results in a more structured and standardized way. This will improve the ability to compare the results of different studies.

Results in Relation to Other Studies

The variability in the application of the CAF categories found in previous studies [15-18] can be explained by ambiguities that were addressed by the inclusion and exclusion criteria of this study. In addition, we found that mapping to a determinant and an outcome CAF category, instead of only one, decreased some of the ambiguity. Only one result, concerning regional information exchange, could not be mapped in the original CAF. This shows that overall the CAF is sufficiently comprehensive. However, we believe that the infrastructure that is available in the environment of an organization forms a missing category in the framework. This category could be introduced in the framework at the macro level to incorporate regional information and communication technologies (ICT) infrastructure, which might be more advanced in some regions than in others.

Implications of the Study

This adapted framework can be used in other reviews and in implementation studies of HISs, especially when the HIS has patients and informal caregivers as users. The definitions and metrics provided will still be of value to implementation studies by pointing out several aspects and metrics that have to be considered when carrying out HIS implementations. Furthermore, the results of this study fulfill part of the need for more guidance documentation when applying the CAF [11]. Our definitions with inclusion and exclusion criteria as well as the metrics found may contribute to a more consistent application of the framework. We recommend addressing specific relationships between determinants and outcomes using this framework, as we did by mapping quotes from the literature to two CAF categories.

Conclusions

The scope of the CAF was extended to the adoption of PHRs, in addition to EHRs, by health care providers, patients, and informal caregivers. Further definitions and inclusion and exclusion criteria disambiguate and guide the application of each category. We found moderate interrater reliability in applying the framework and variance among the categories in the framework. Future research should address the application of the CAF in primary implementation studies and studies focusing on macro level topics such as international and regional ICT infrastructures, standards, politics, incentive programs, and social trends.

Authors' Contributions

All authors were involved in the systematic review. RC, NK, and RN reviewed two of the first three reviews and contributed to the revision of the CAF definitions. RD and HM reviewed all reviews and continued revising the CAF iteratively. HM prepared the text and carried out the analyses with feedback from RC, NK, and RN. RC, NK, RN, and RD critically reviewed the manuscript. All authors agreed to submit the manuscript.

XSL·FO

RenderX

software vendor that develops the health information system HiX. HiX Care Portal is a web application that includes a patient portal. RC and NK declare that they have no known conflicts of interest and are not affiliated with any vendor, patient portal, or personal health record.

Multimedia Appendix 1
R script with syntax for analysis.
[TXT File , 14 KB - medinform_v8i3e15150_app1.txt]

Multimedia Appendix 2
Supplementary tables.
[DOCX File , 61 KB - medinform_v8i3e15150_app2.docx]

References

1. Ross SE, Lin C. The effects of promoting patient access to medical records: a review. J Am Med Inform Assoc 2003;10(2):129-138 [FREE Full text] [doi: 10.1197/jamia.m1147] [Medline: 12595402]
2. Ammenwerth E, Schnell-Inderst P, Hoerbst A. The impact of electronic patient portals on patient care: a systematic review of controlled trials. J Med Internet Res 2012 Nov 26;14(6):e162 [FREE Full text] [doi: 10.2196/jmir.2238] [Medline: 23183044]
3. Prey JE, Polubriaginof F, Kuperman GJ, Tiase V, Collins SA, Vawdrey DK. International perspectives on sharing clinical data with patients. Int J Med Inform 2016 Feb;86:135-141. [doi: 10.1016/j.ijmedinf.2015.11.007] [Medline: 26631956]
4. Beard L, Schein R, Morra D, Wilson K, Keelan J. The challenges in making electronic health records accessible to patients. J Am Med Inform Assoc 2012;19(1):116-120 [FREE Full text] [doi: 10.1136/amiajnl-2011-000261] [Medline: 22120207]
5. Saparova D. Motivating, influencing, and persuading patients through personal health records: a scoping review. Perspect Health Inf Manag 2012;9:1f [FREE Full text] [Medline: 22783154]
6. Goldzweig C, Orshansky G, Paige N, Towfigh AA, Haggstrom D, Miake-Lye I, et al. Electronic patient portals: evidence on health outcomes, satisfaction, efficiency, and attitudes: a systematic review. Ann Intern Med 2013 Nov 19;159(10):677-687. [doi: 10.7326/0003-4819-159-10-201311190-00006] [Medline: 24247673]
7. Farrelly S, Brown GE, Flach C, Barley E, Laugharne R, Henderson C. User-held personalised information for routine care of people with severe mental illness. Cochrane Database Syst Rev 2013 Oct 05(10):CD001711. [doi: 10.1002/14651858.CD001711.pub2] [Medline: 24096715]
8. Goldzweig CL, Orshansky G, Paige NM, Towfigh AA, Haggstrom DA, Miake-Lye I, et al. Electronic patient portals: evidence on health outcomes, satisfaction, efficiency, and attitudes: a systematic review. Ann Intern Med 2013 Nov 19;159(10):677-687. [doi: 10.7326/0003-4819-159-10-201311190-00006] [Medline: 24247673]
9. Lau F, Price M, Keshavjee K. From benefits evaluation to clinical adoption: making sense of health information system success in Canada. Healthc Q 2011;14(1):39-45. [doi: 10.12927/hcq.2011.22157] [Medline: 21301238]
10. van Mens HJT, Duijm RD, Nienhuis R, de Keizer NF, Cornet R. Determinants and outcomes of patient access to medical records: Systematic review of systematic reviews. Int J Med Inform 2019 Sep;129:226-233. [doi: 10.1016/j.ijmedinf.2019.05.014] [Medline: 31445260]
11. Craven CK, Doebbeling B, Furniss D, Holden RJ, Lau F, Novak LL. Evidence-based health informatics frameworks for applied use. Stud Health Technol Inform 2016;222:77-89. [Medline: 27198094]
12. Lau F, Kuziemsky C. Handbook of eHealth evaluation: an evidence-based approach. Victoria: University of Victoria; 2016.
13. Ahmadi H, Darvishi M, Nilashi M, Almaee A, Ibrahim O, Zolghadri A. Evaluating the critical factors for electronic medical record adoption using fuzzy approaches. International Journal of Innovation and Scientific Research 2014;9(2):268-284.
14. Ahmadi H, Salahshour Rad M, Nilashi M, Ibrahim O, Almaee A. Ranking the Micro Level Critical Factors of Electronic Medical Records Adoption Using Topsis Method. HIIJ 2013 Nov 30;2(4):19-32. [doi: 10.5121/hiij.2013.2402]
15. Bassi J, Lau F, Lesperance M. Perceived impact of electronic medical records in physician office practices: a review of survey-based research. Interact J Med Res 2012 Jul 28;1(2):e3 [FREE Full text] [doi: 10.2196/ijmr.2113] [Medline: 23611832]
16. Gagnon M, Payne-Gagnon JA, Sicotte C, Langué-Dubé JA, Motulsky A. Connecting primary care clinics and community pharmacies through a nationwide electronic prescribing network: A qualitative study. J Innov Health Inform 2015 Oct 05;22(3):359-367 [FREE Full text] [doi: 10.14236/jhi.v22i3.168] [Medline: 26577426]
17. Lau F, Partridge C, Randhawa G, Bowen M. Applying the clinical adoption framework to evaluate the impact of an ambulatory electronic medical record. Stud Health Technol Inform 2013;183:15-20. [Medline: 23388247]
18. Lau F, Price M, Boyd J, Partridge C, Bell H, Raworth R. Impact of electronic medical record on physician practice in office settings: a systematic review. BMC Med Inform Decis Mak 2012 Feb 24;12:10 [FREE Full text] [doi: 10.1186/1472-6947-12-10] [Medline: 22364529]

XSL•FO
RenderX

19. Rathert C, Mittler JN, Banerjee S, McDaniel J. Patient-centered communication in the era of electronic health records: What does the evidence say? Patient Educ Couns 2017 Jan;100(1):50-64. [doi: 10.1016/j.pec.2016.07.031] [Medline: 27477917]

20. Thompson MJ, Reilly JD, Valdez RS. Work system barriers to patient, provider, and caregiver use of personal health records: A systematic review. Appl Ergon 2016 May;54:218-242. [doi: 10.1016/j.apergo.2015.10.010] [Medline: 26851482]

21. Turner K, Klaman SL, Shea CM. Personal health records for people living with HIV: a review. AIDS Care 2016 Sep;28(9):1181-1187. [doi: 10.1080/09540121.2016.1153594] [Medline: 26917113]

22. Bush RA, Connelly CD, Fuller M, Pérez A. Implementation of the Integrated Electronic Patient Portal in the Pediatric Population: A Systematic Review. Telemed J E Health 2016 Feb;22(2):144-152 [FREE Full text] [doi: 10.1089/tmj.2015.0033] [Medline: 26258289]

23. Irizarry T, DeVito Dabbs A, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015 Jun 23;17(6):e148 [FREE Full text] [doi: 10.2196/jmir.4255] [Medline: 26104044]

24. Otte-Trojel T, de Bont A, Rundall TG, van de Klundert J. What do we know about developing patient portals? a systematic literature review. J Am Med Inform Assoc 2016 Apr;23(e1):e162-e168 [FREE Full text] [doi: 10.1093/jamia/ocv114] [Medline: 26335985]

25. Coughlin SS, Prochaska JJ, Williams LB, Besenyi GM, Heboyan V, Goggans DS, et al. Patient web portals, disease management, and primary prevention. Risk Manag Healthc Policy 2017;10:33-40 [FREE Full text] [doi: 10.2147/RMHP.S130431] [Medline: 28435342]

26. Hemsley B, Rollo M, Georgiou A, Balandin S, Hill S. The health literacy demands of electronic personal health records (e-PHRs): An integrative review to inform future inclusive research. Patient Educ Couns 2018 Jan;101(1):2-15. [doi: 10.1016/j.pec.2017.07.010] [Medline: 28864052]

27. Kelly MM, Coller RJ, Hoonakker PL. Inpatient portals for hospitalized patients and caregivers: a systematic review. J Hosp Med 2018 Jun 01;13(6):405-412 [FREE Full text] [doi: 10.12788/jhm.2894] [Medline: 29261819]

28. Kneale L, Demiris G. Lack of diversity in personal health record evaluations with older adult participants: a systematic review of literature. J Innov Health Inform 2017 Jan 15;23(4):881 [FREE Full text] [doi: 10.14236/jhi.v23i4.881] [Medline: 28346131]

29. Kruse CS, Argueta DA, Lopez L, Nair A. Patient and provider attitudes toward the use of patient portals for the management of chronic disease: a systematic review. J Med Internet Res 2015 Feb 20;17(2):e40 [FREE Full text] [doi: 10.2196/jmir.3703] [Medline: 25707035]

30. Kruse CS, Bolton K, Freriks G. The effect of patient portals on quality outcomes and its implications to meaningful use: a systematic review. J Med Internet Res 2015 Feb 10;17(2):e44 [FREE Full text] [doi: 10.2196/jmir.3171] [Medline: 25669240]

31. Mold F, de Lusignan S, Sheikh A, Majeed A, Wyatt JC, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic review in primary care. Br J Gen Pract 2015 Mar;65(632):e141-e151 [FREE Full text] [doi: 10.3399/bjgp15X683941] [Medline: 25733435]

32. Powell KR. Patient-perceived facilitators of and barriers to electronic portal use: a systematic review. Comput Inform Nurs 2017 Nov;35(11):565-573. [doi: 10.1097/CIN.0000000000000377] [Medline: 28723832]

33. Price M, Bellwood P, Kitson N, Davies I, Weber J, Lau F. Conditions potentially sensitive to a personal health record (PHR) intervention, a systematic review. BMC Med Inform Decis Mak 2015 Apr 18;15:32 [FREE Full text] [doi: 10.1186/s12911-015-0159-1] [Medline: 25927384]

34. Sakaguchi-Tang DK, Bosold AL, Choi YK, Turner AM. Patient portal use and experience among older adults: systematic review. JMIR Med Inform 2017 Oct 16;5(4):e38 [FREE Full text] [doi: 10.2196/medinform.8092] [Medline: 29038093]

35. Sartain SA, Stressing S, Prieto J. Patients' views on the effectiveness of patient-held records: a systematic review and thematic synthesis of qualitative studies. Health Expect 2015 Dec;18(6):2666-2677 [FREE Full text] [doi: 10.1111/hex.12240] [Medline: 25059439]

36. Showell C. Barriers to the use of personal health records by patients: a structured review. PeerJ 2017;5:e3268 [FREE Full text] [doi: 10.7717/peerj.3268] [Medline: 28462058]

37. Vermeir P, Degroote S, Vandijck D, Van Tiggelen H, Peleman R, Verhaeghe R, et al. The patient perspective on the effects of medical record accessibility: a systematic review. Acta Clin Belg 2017 Jun;72(3):186-194. [doi: 10.1080/17843286.2016.1275375] [Medline: 28056665]

38. van Mens HJT, Nienhuis R, Duijm RD, Cornet R, de Keizer NF. National Institute for Health Research.: PROSPERO; 2018. An overview of determinants and outcomes of providing patients access to and control over their own medical data: a systematic review of reviews URL: http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42018084542

39. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 2016 Jul 02;20(1):37-46. [doi: 10.1177/001316446002000104]

40. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968 Oct;70(4):213-220. [doi: 10.1037/h0026256] [Medline: 19673146]

XSL·FO
RenderX

Abbreviations

CAF: Clinical Adoption Framework
EHR: electronic health record
HIS: health information system
ICT: information and communication technologies
PHR: personal health record.

XSL•FO
RenderX

<u>Review</u>

Clinical Text Data in Machine Learning: Systematic Review

Irena Spasic[1], PhD; Goran Nenadic[2], PhD

[1]School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom
[2]Department of Computer Science, University of Manchester, Manchester, United Kingdom

Corresponding Author:
Irena Spasic, PhD
School of Computer Science and Informatics
Cardiff University
5 The Parade
Cardiff, CF24 3AA
United Kingdom
Phone: 44 02920870320
Email: spasici@cardiff.ac.uk

Abstract

Background: Clinical narratives represent the main form of communication within health care, providing a personalized account of patient history and assessments, and offering rich information for clinical decision making. Natural language processing (NLP) has repeatedly demonstrated its feasibility to unlock evidence buried in clinical narratives. Machine learning can facilitate rapid development of NLP tools by leveraging large amounts of text data.

Objective: The main aim of this study was to provide systematic evidence on the properties of text data used to train machine learning approaches to clinical NLP. We also investigated the types of NLP tasks that have been supported by machine learning and how they can be applied in clinical practice.

Methods: Our methodology was based on the guidelines for performing systematic reviews. In August 2018, we used PubMed, a multifaceted interface, to perform a literature search against MEDLINE. We identified 110 relevant studies and extracted information about text data used to support machine learning, NLP tasks supported, and their clinical applications. The data properties considered included their size, provenance, collection methods, annotation, and any relevant statistics.

Results: The majority of datasets used to train machine learning models included only hundreds or thousands of documents. Only 10 studies used tens of thousands of documents, with a handful of studies utilizing more. Relatively small datasets were utilized for training even when much larger datasets were available. The main reason for such poor data utilization is the annotation bottleneck faced by supervised machine learning algorithms. Active learning was explored to iteratively sample a subset of data for manual annotation as a strategy for minimizing the annotation effort while maximizing the predictive performance of the model. Supervised learning was successfully used where clinical codes integrated with free-text notes into electronic health records were utilized as class labels. Similarly, distant supervision was used to utilize an existing knowledge base to automatically annotate raw text. Where manual annotation was unavoidable, crowdsourcing was explored, but it remains unsuitable because of the sensitive nature of data considered. Besides the small volume, training data were typically sourced from a small number of institutions, thus offering no hard evidence about the transferability of machine learning models. The majority of studies focused on text classification. Most commonly, the classification results were used to support phenotyping, prognosis, care improvement, resource management, and surveillance.

Conclusions: We identified the data annotation bottleneck as one of the key obstacles to machine learning approaches in clinical NLP. Active learning and distant supervision were explored as a way of saving the annotation efforts. Future research in this field would benefit from alternatives such as data augmentation and transfer learning, or unsupervised learning, which do not require data annotation.

KEYWORDS

XSL•FO
RenderX

Introduction

Clinical narratives represent the main form of communication within health care. In comparison with generically coded elements of electronic health records (EHRs), the narrative notes provide a more detailed and personalized account of patient history and assessments, offering a better context for clinical decision making [1]. Natural language processing (NLP) is a subfield of artificial intelligence that studies the ways in which the analysis and synthesis of information expressed in a natural language can be automated. It has repeatedly demonstrated its feasibility to unlock evidence buried in clinical narratives, making it available for large-scale analysis down the stream [2]. Traditionally, rule-based approaches were commonly used to unlock evidence of specific types [3]. Their development requires some form of direct interaction with clinical experts to convert their knowledge, often tacit, into a set of explicit pattern-matching rules.

Machine learning has long been hailed as a silver bullet solution for the knowledge elicitation bottleneck, the main argument being that the task of annotating the data manually is easier than that of eliciting the knowledge [4]. Nonetheless, the amount of data required to train a machine learning model may require as much time to annotate as the knowledge elicitation itself [5]. Much like the law of energy conservation, it seems that the knowledge required to inform the creation of an accurate computational model is simply transferred from one form to another. Instead of explicit knowledge in the form of rules, machine learning is based on implicit knowledge in the form of annotations and their distribution, with the time involved in their acquisition remaining virtually constant.

Another problem associated with the machine learning approach is the availability of clinical narratives given the sensitive nature of health data and privacy concerns [6]. These problems (ie, unavailability of manually annotated data) may result in the lack of representativeness of the training data and consequently substandard performance of the corresponding machine learning models. For these reasons, the main aim of this review was to provide systematic evidence on the properties of data used to train machine learning approaches to clinical NLP. In addition,

we investigate the types of NLP tasks that have been supported by machine learning and how they can be applied in clinical practice.

The remainder of the paper is organized as follows. We start by explaining the methodology of this systematic review in detail. We then discuss the main findings of the review. Finally, we conclude the review by outlining future research directions in this field.

Methods

Overview

On the basis of the guidelines for performing systematic reviews described by Kitchenham [7], our methodology is structured around the following steps. First, research questions (RQs) were used to define the scope, depth, and the overall aim of the review. Next, a search strategy was designed to identify all studies that are relevant to the RQs in an efficient and reproducible manner. In addition, inclusion and exclusion criteria were defined to refine the scope. A critical appraisal of the included studies was conducted to ensure that the findings of the review are valid. During data extraction, the relevant information was identified from the included studies and semistructured to facilitate the synthesis of evidence and support the findings of the review.

Research Questions

The overarching topic of this review is concerned with the properties of text data used to enable machine learning approaches to clinical NLP. The main aim of the review was to answer the RQs given in Table 1. RQ1 aims at describing the properties of data that are relevant for interpreting the performance of machine learning. These properties include size, provenance, heterogeneity, annotations, and others. Here, heterogeneity refers to content, structure, and clinical domains. RQ2 classifies the problems addressed by machine learning in the context of NLP into different types of computational tasks. Finally, RQ3 focuses on the ways in which NLP based on machine learning can be applied to tackle practical problems encountered in clinical practice.

Table 1. Research questions.

ID	RQ[a]
RQ1	What are the key properties of data used to train and evaluate machine learning models?
RQ2	What types of NLP[b] tasks have been supported by machine learning?
RQ3	How can NLP based on machine learning be applied in clinical practice?

[a]RQ: research question.

[b]NLP: natural language processing.

Search Strategy

We used PubMed as a search engine to retrieve relevant documents from the MEDLINE database of 28 million citations from life sciences and biomedical literature, which are indexed by Medical Subject Headings (MeSH). MeSH is a hierarchically organized controlled vocabulary used for manually indexing

articles in MEDLINE in a uniform and consistent manner to facilitate their retrieval. We derived a list of search terms to describe the topic of this review: *machine learning, deep learning, text, natural language, clinical, health, health care,* and *patient*. Here, machine learning and deep learning are used to retrieve articles that employ this methodology. Note that MeSH includes the term *machine learning*, thus making it

XSL•FO

RenderX

unnecessary to include specific machine learning techniques such as *support vector machines* or *conditional random fields* into the search query. The following 2 search terms, *text* and *natural language*, refer to the relevant type of input into the learning methods. The final 4 terms were used to refer to clinical applications. Owing to the broad nature and common use of the last 6 terms, their mentions were restricted to titles and abstracts only. In an attempt to prevent retrieval of nonoriginal studies and NLP applications developed to support systematic reviews, we negated the terms *literature*, *bibliometric*, and *systematic review*. Finally, to focus on the emerging application of machine learning, we restricted the search to the period from January 1, 2015. The search was performed on August 8, 2018. The search terms were combined into a PubMed query as follows:

(("machine learning"[All Fields] OR "deep learning"[All Fields]) AND (text[Title/Abstract] OR "natural language" [Title/Abstract]) AND (clinical[Title/Abstract] OR health [Title/Abstract] OR healthcare[Title/Abstract] OR patient [Title/Abstract]) NOT (literature[Title/Abstract] OR bibliometric [Title/Abstract] OR "systematic review"[Title/Abstract]) AND ("2015/01/01"[PDat] : "2018/08/08"[PDat])

We identified 389 candidate articles according to the described search strategy. The results were further screened against the selection criteria.

Selection Criteria

The scope of this systematic review was formally defined by the inclusion and exclusion criteria given in Textboxes 1 and 2, respectively. Having screened the retrieved articles against the inclusion and exclusion criteria, a total of 149 articles were retained for further processing.

Textbox 1. Inclusion criteria.

1. The study has to use natural language processing.

2. Machine learning has to be used to support such processing.

3. Input text has to be routinely collected within health care boundaries.

4. Input text has to be written or dictated.

5. The article has to be peer reviewed.

6. The full text has to be freely available online for academic use.

Textbox 2. Exclusion criteria.

1. Articles written in a language other than English.

2. Natural language processing of a language other than English.

3. Natural language processing of spoken language.

Given the interdisciplinary nature of articles considered for this review, we encountered a wide diversity of venues in which they were published. Not surprisingly, some studies put an emphasis on the clinical aspects but neglected to describe the computational aspects of the study in sufficient detail to support its reproducibility. To be included in this review, articles needed to provide sufficient information to support answering RQs defined in Table 1. In other words, they needed to describe the datasets used; define the NLP problem clearly; describe the features used to support NLP; state the machine learning methods used and, where appropriate, their parameters; and provide a formal evaluation of the results. A total of 39 studies were found not to match these criteria. This further reduced the number of selected articles to 110 [8-117]. Figure 1 summarizes the outcomes of the 4 major stages in the systematic literature review.

Figure 1. Flow diagram of the literature review process.
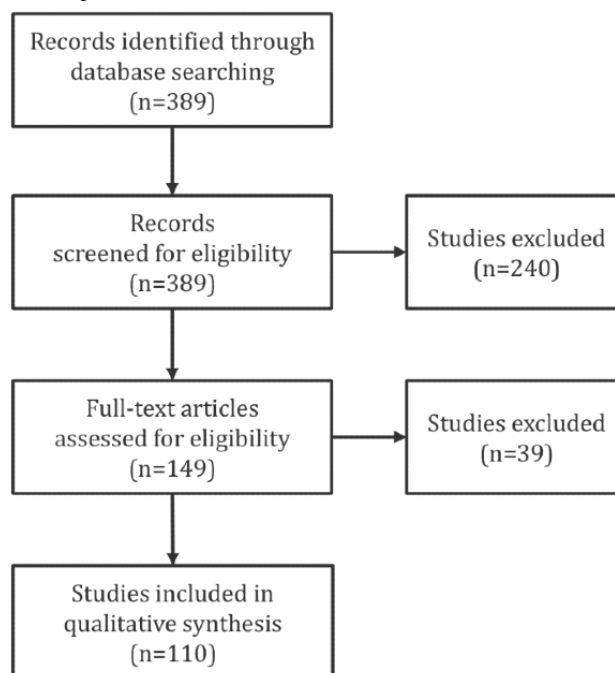
Data Extraction

We explored the selected studies to extract data that contribute to answering the RQs given in Table 1. Data were extracted from the full text of articles under the following headings: data, task, clinical domain, and clinical application. The data properties considered included their size, provenance, collection methods, annotation, and any relevant statistics. The task was defined as a subfield of NLP (eg, text classification, information extraction (IE), named entity recognition (NER), and word sense disambiguation [WSD]). This was supplemented with task-specific information; for example, for NER, we also specified the type of named entities considered. Clinically relevant information was extracted to identify the potential for practical applications. The extracted data were then used to facilitate a narrative synthesis of the main findings.

Results

The first step in developing a machine learning model is to collect data relevant to the problem at hand. Ultimately, the model's performance will depend on the properties of such a dataset. We summarized these properties, including data size, key data sources, training annotations, and types of clinical documents considered.

Size

Among other factors, the performance of machine learning models and the significance of test results depend on the size of the dataset used for training and testing, respectively. In this section, we examine the size of datasets used in the studies included in this review. Owing to large variations in data sizes, we used a logarithmic scale to fit this information into the chart shown in Figure 2, which stratifies the datasets according to their order of magnitude. Some studies used as few as 40 documents [48] and as few as 15 patients [28]. The vast majority of datasets have the cardinality in the range of hundreds or thousands. Only 10 studies used tens of thousands of documents, with a handful of studies utilizing more than that despite the fact that machine learning approaches are data hungry in the sense that their performance is strongly correlated with the amount of training data available.

Relatively small datasets were utilized even when much larger datasets were available. Figure 3 demonstrates data utilization on a logarithmic scale, with some studies utilizing as little as 0.002% of available data [44] and as much as 11.88% [11]. Specific examples illustrate this issue: 500 from 188,843 [32], 300 from 4025 [59], 62 from 6343 [25], 323 from 16,000 [24], 1188 from 10,000 [11], 1610 from 52,746 [39], 1004 from 96,303 [112], 1058 from 376,487 [34], 10,000 from 103,564 sentences [36], less than 12,000 out of 137,522+28,159 [101], 562 from 2.5 million [44], 8288 from 2,977,739 [13], 6174 from 2.6 million [113], 3467 from 8,168,330 [68], and 2159 from 24 million [19].

Figure 2. Distribution of data size on a logarithmic scale.
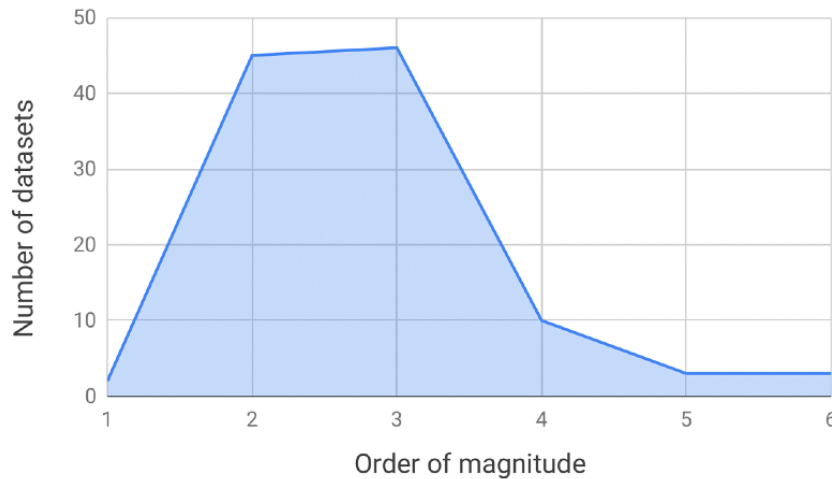
Figure 3. Data utilization on a logarithmic scale.
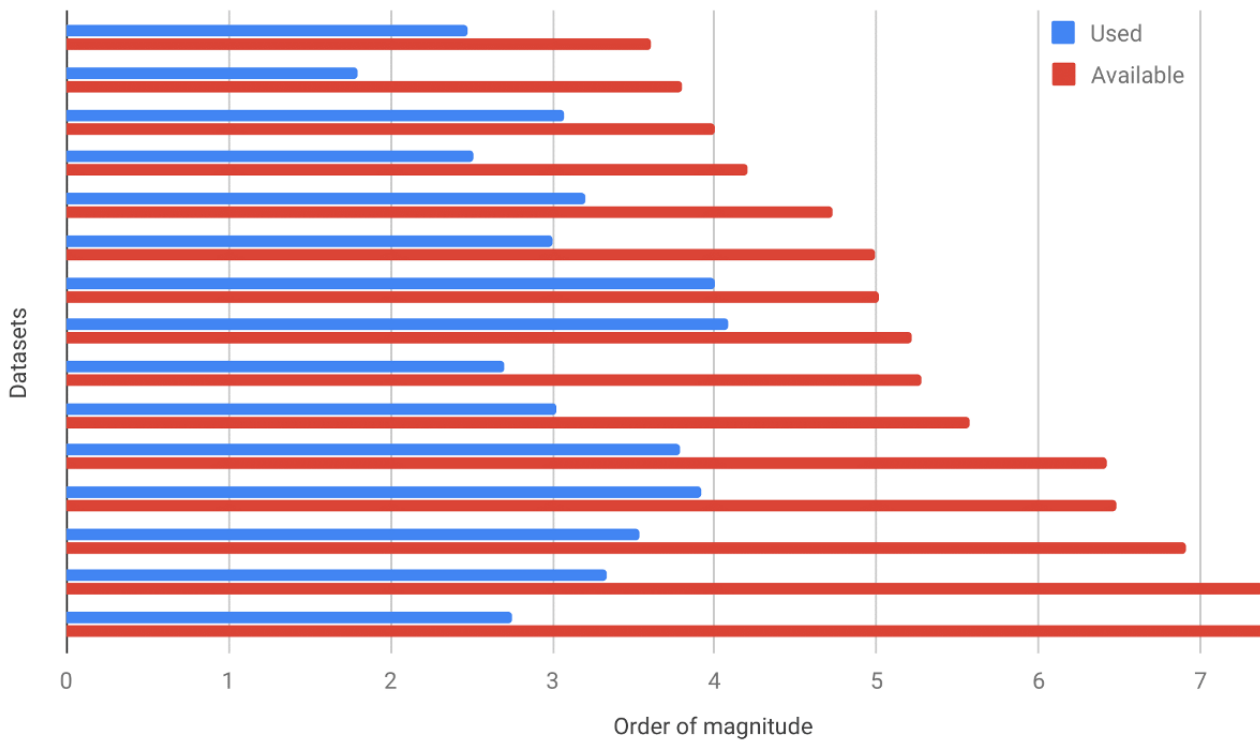
Annotation

The main reason for such poor data utilization is the annotation bottleneck faced by supervised machine learning algorithms, which require training data to be annotated to generalize them into predictive mathematical models. Compiling manually annotated corpora is both labor-intensive and error prone. The fact that annotations are task-specific means that the training data rarely get to be recycled. The labor and time limitations imposed on individual studies will naturally be correlated with the volume of manually annotated training data. Active learning aims to address the annotation bottleneck by involving human experts in the machine learning process in an attempt to improve performance with relatively small annotation effort [20,54,100]. An active learning algorithm can iteratively sample a subset of data for manual annotation, depending on the current predictive performance. Sampling strategies can be based on a disagreement between different predictive models or different measures of uncertainty, density, and expectation of a single predictive model. Such sampling depends on the quality of a predictive model and may not be efficient when retraining the model lasts relatively long. Alternatively, diversity measures can be used to prioritize annotation. For instance, pair-wise cosine similarity was used to compare sentences and prioritize those least similar to annotated sentences for annotation [20]. However, this may lead to the selection of outliers, whose presence in the training data can result in a degradation of predictive models. By considering representativeness and informativeness, outliers are less likely to be selected, thus leading to better coverage of the data characteristics and, consequently better predictive models. Here, the average similarity between a sentence and all other sentences indicates

how representative it is [54]. The higher the similarity, the more representative the sentence is.

In principle, supervised learning approaches are convenient when labels are readily available. For instance, EHRs combine different types of data elements from unstructured data such as free text and images to structured data (ie, discrete elements such as numbers, dates, and codes) from controlled medical terminologies [118]. In the studies included in this systematic review, larger datasets (ie, those ranging from tens of thousands to millions, see Figure 2), were used mostly in cases where existing structured data were utilized as labels. For instance, in relation to hospitalization, readily available information about events such as in-hospital death [102], discharge [90], readmission [9], and emergency department visits [37] was used to train models to predict future events of this type well in advance to inform an appropriate course of action. Similarly, in relation to diagnostics, both prior (eg, imaging protocol [17,94]) and posterior (eg, test result [69]) information was utilized for supervision. International Classification of Diseases (ICD) diagnosis codes were used to train predictive models from historical data to identify patients at risk [16,22,50] or to facilitate disease surveillance [76]. Similarly, supervised models trained with ICD procedure codes otherwise used for billing can be used for cost optimization but also improving the quality of care [81]. Indeed, all of these examples have clear applications in care improvement and resource management. In some other cases (eg, classification of clinical notes into medical subdomains [103]), the utility of such information remains unclear.

Some types of learning problems such as WSD lend themselves well to semiautomated labeling based on greedy matching. Not surprisingly, the corresponding methods were tested on large datasets [33,105]. Similarly, using the concept of distant supervision, which utilizes an existing knowledge base to automatically annotate raw text, as much as 9.5 million clinical notes were annotated with adverse drug events [99]. Where manual annotation was unavoidable, crowdsourcing was explored. This approach is suitable for patient-facing problems such as readability of medical documents [116], where lay annotators are indeed ideally suited for the annotation task. The concept of crowdsourcing was explored for problems that require medical expertise [24]. Even though the interannotator agreement among crowdsourced workers was found to be much lower than that of medical experts, with Krippendorff alpha coefficient over .7, it was still found to be good agreement beyond chance. However, privacy constraints do undermine the feasibility of crowdsourcing in the context of clinical narratives.

Provenance

Besides the small volume of training data, another issue that might affect the performance of machine learning methods trained on such data is their provenance. The structure and style of clinical narratives may vary greatly between institutions [119]. Therefore, when the provenance of data is confined to a small number of contributing institutions, the data may not be representative. This, in turn, may lead to overfitting, a modeling error that occurs when a complex model adapts to the idiosyncrasies of the training data and fails to generalize the underlying properties of the problem. Unfortunately, the majority of studies reviewed here were limited to the authors' host institutions [8,10,12,15,17,22,24,25,28,30-33,35,40, 41,44,66,70,76,79,84-86,89,90,94,95,99,105,106,111,113]. Rarely are such datasets freely accessible to the community. A notable exception is the Medical Information Mart for Intensive Care (MIMIC) [120], a freely accessible critical care database that stores a wide range of clinical narratives, including radiology reports [87], clinical notes [102] and discharge summaries [16,39]. Although it is a single-site dataset, some consolation may be found in the sheer volume of data. More importantly, its public availability allows for rigorous and detailed direct comparison of competing approaches, a rare commodity in clinical NLP.

Only 9 studies used data from 2 institutions [36,47,50,56,61,100,103,109,112]. Three studies used data from 3 institutions [45,71,87]. A handful of studies managed to obtain data from multiple sources: 5 [38], 6 [73], 18 [19], and 28 [37]. The Veterans Health Administration (VHA) [121,122], as the largest integrated health care system in the United States, provides centralized access to data from multiple institutions, enhancing the credibility of results achieved on such data [13,14,29,34,55,68,72,77,97].

Availability

Most datasets used in the included studies originated from a few institutions, thus offering no hard evidence about the transferability of machine learning models. Knowing that the format and style of clinical notes may vary substantially across institutions [119], it is not uncommon to observe a significant drop in performance when training a model in one institution and testing it in another [33,61,75,105,109]. This remains an ongoing concern for the clinical NLP community, where the confidentiality of data involved requires a careful balance between accessibility and privacy protection. In this section, we discuss wider availability of data that provide opportunities for secondary uses, including research. In this context, the NLP community challenges play an important role in providing access to clinical data to a wider pool of researchers and establishing benchmarks for future comparisons. Not surprisingly, many studies reviewed here have been enabled by the datasets shared in community challenges, which are described in Table 2.

Table 2. Datasets used in clinical natural language processing community challenges.

Dataset	Provenance	Documents	Size[a]	Annotations	Studies
2010 i2b2/VA [123]	PHC[b], BIDMC[c], UPMC[d]	Discharge summaries, progress reports	871	Medical problems, treatments, tests, and relations among them	[20,49,64,67,96,104]
2011 i2b2/VA [124]	PHC, BIDMC, UPMC, Mayo[e]	Discharge summaries, progress reports, radiology reports, pathology reports, other reports	978+164	Coreference chains for the problem, person, test, result, treatment, anatomical site, disease or syndrome, sign or symptom, etc	[63]
2012 i2b2 [125]	PHC, BIDMC	Discharge summaries	310	Clinical events, temporal expressions, temporal relations	[64]
2013 ShARe/CLEF eHealth [126]	BIDMC	Discharge summaries, electrocardiogram reports, echocardiogram reports, radiology reports	300	Disorders, acronyms, and abbreviations	[54,57,88,98,114]
2014 i2b2/UTHealth [127,128]	PHC	Longitudinal medical records	1304	Protected health information; risk factors for heart disease	[18,21,26,52,62,64,80,82,91,107,108]
2015 SemEval/THYME [129]	Mayo	Clinical notes, pathology reports	600	Times, events, and temporal relations among them	[60]
2016 CEGS N-GRID [130,131]	PHC	Psychiatric intake records	1000	Protected health information; symptom severity	[23,27,42,53,58,65,78,83,92]

[a]Size is expressed as the number of documents.

[b]Partners Health Care (PHC) is a nonprofit hospital and physician network that includes Brigham and Women's Hospital and Massachusetts General Hospital.

[c]Beth Israel Deaconess Medical Center (BIDMC) is a teaching hospital of Harvard Medical School. Both organizations are based in Boston, Massachusetts, United States.

[d]The University of Pittsburgh Medical Center (UPMC) is a global nonprofit health enterprise that integrates over 35 hospitals, 600 clinical locations, and a health insurance division.

[e]The Mayo Clinic is a nonprofit academic medical center based in Rochester, Minnesota, which focuses on integrated clinical practice, education, and research. The clinic specializes in treating difficult cases through tertiary care.

Similarly, MIMIC dataset represents a key driver of open research in clinical NLP. It is notable for being the only freely accessible critical care database of its kind [120]. Data analysis is unrestricted once a data use agreement is accepted, enabling clinical research and education internationally. The open nature of the data supports the reproducibility of findings and enables continual research advances. MIMIC is a large, single-center database that stores deidentified, comprehensive clinical information relating to patients admitted to critical care units at the Beth Israel Deaconess Medical Centre in Boston, Massachusetts, United States, a large tertiary care hospital. Its content, which spans more than a decade, integrates different types of data (see Table 3). Of interest to this systematic review are free-text data, which include various types of notes and reports. Their integration with coded data offers an opportunity to circumvent manual annotation of data for supervised learning and evaluation purposes. For instance, Berndorfer and Henriksson [16] used a large dataset of 59,531 discharge summaries with at least one assigned ICD diagnosis code to automate the process of diagnosis coding. However, in many cases, accurate classification of medical conditions exists only in clinical narratives. Therefore, it may be necessary to annotate relevant phrases in the free text to train classification models. For instance, Gehrmann et al [39] manually annotated 1610 discharge summaries from MIMIC to automatically learn which phrases are relevant for 10 patient phenotypes considered. Similarly, Tahmasebi et al [87] manually annotated 860 radiology reports from MIMIC and 2 other institutions to evaluate an unsupervised approach to detecting and normalizing anatomical phrases.

Table 3. Description of clinical data types in the Medical Information Mart for Intensive Care.

Type	Description
Billing	Coded data recorded primarily for billing and administrative purposes.
Descriptive information	Demographic information, admission and discharge times, and dates of death.
Dictionaries	Look-up tables for cross-referencing identifiers (eg, codes) with associated definitions.
Interventions	Procedures such as dialysis, imaging studies, and placement of lines.
Laboratory measurements	Blood chemistry, hematology, urine analysis, and microbiology test results.
Medications	Administration records of intravenous medications and medication orders.
Notes	Free-text notes such as provider progress notes and hospital discharge summaries.
Physiologic information	Nurse-verified vital signs, approximately hourly (eg, heart rate, blood pressure, and respiratory rate).
Reports	Free-text reports of electrocardiogram and imaging studies (x-ray, computed tomography, ultrasound, and magnetic resonance imaging).

In addition to openness, an important driver of advancing state of the art in clinical NLP is an ability to access a wide range of data sources, many of which may not be compatible with national or organization-wide standards. As the largest integrated health care system in the United States, which provides care at 1243 health care facilities, including 172 medical centers and 1062 outpatient sites of care of varying complexity, the VHA [121,122] has the potential to address this challenge. The VHA offers veterans (ie, those who served in the active military, naval, or air service and who were discharged or released under conditions other than dishonorable) a wide range of inpatient, outpatient, mental health, rehabilitation, and long-term care services, which are all linked by an EHR platform. The construction of the VHA's information infrastructure, the Veterans Information Systems Technology Architecture (VistA), began in 1982 and became operational in 1985. VistA integrates multiple applications seamlessly that are accessible via a graphical user interface, the Computerized Patient Record System, first launched in 1997. Designed primarily to support clinical care delivery rather than billing, the system has been used since 2004 to document all routine clinical activities currently storing more than 16 billion clinical entries.

On average, 1 million free-text notes (eg, progress notes and discharge summaries), 1.2 million provider-entered electronic orders, 2.8 million images (radiologic studies, electrocardiograms, and photographs), and 1 million vital signs were stored in VistA daily. Such proliferation of data quickly outgrew the original plans for storage capacity, network bandwidth, support staff, and information technology budget, leading to the construction of the Corporate Data Warehouse (CDW) in 2006. The new repository for patient-level data aggregated from across the VHA's national health delivery system also hosts data from the legacy system, each featuring its own data rules, definitions, and structures. Given the slow process of normalizing these idiosyncrasies to a common standard and the rapidly increasing volume of data, the CDW allowed selective streaming of data from VistA and structuring them pragmatically in a way that minimizes redundancy. The CDW stores comprehensive patient-level data, which are used primarily to support health care delivery, but their unprecedented richness and volume provide a great opportunity for secondary uses such as quality improvement and research. To facilitate

such uses, the VHA has partitioned a section of the CDW for use by health services and informatics investigators, who can access these data in secure workspaces within the VHA's firewall. The VHA is developing mechanisms to fully deidentify data extracts so that they can be shared outside of the VHA.

Similar to MIMIC, integration of structured (coded) and unstructured (free-text) data offers an opportunity to circumvent manual annotation of data for supervised learning and evaluation purposes. In this manner, Ben-Ari et al [14] utilized postoperative notes of 32,636 patients by cross-referencing them to prescription data. However, most studies still rely on manual annotation of information that is not well documented in structured data. For example, Bates et al [13] manually annotated 8288 radiology reports as *fall* or *not fall* at the document level. Similarly, Maguen et al [68] annotated 3467 randomly selected psychotherapy notes with respect to the use of evidence-based psychotherapy. Patterson et al [77] manually annotated 2000 colonoscopy procedure notes with an indication, which included screening, nonscreening, noncolonoscopy, and unknown. Walsh et al [97] annotated 3900 snippets of text referring to axial spondyloarthritis in a corpus sampled from 500 million clinical notes and 120 million radiology notes. Divita et al [29] sampled 948 records from 164 preselected document types and annotated them manually to identify 5819 positively asserted symptoms within the documents. Kim et al [55] annotated a corpus of 1465 echocardiography reports, radiology reports, and other note types from multiple medical centers sampled at random for mentions and assessments of left ventricular ejection fraction. Fodeh et al [34] sampled 1058 clinical notes of 101 types and manually annotated fine-grained information about pain assessments, which included not only pain mention but also its features such as intensity, quality, site, and etiology. Meystre et al [72] sampled a cohort of 1083 patients and annotated their clinical notes of more than 10 preselected types with information regarding congestive heart failure treatment performance measures. These in-document annotations were summarized at the clinical note and patient level for binary classification of patients as meeting the treatment performance measure or not. These studies illustrate the extent of manual annotation effort involved in developing machine learning approaches to clinical NLP. Unfortunately, manual annotations remain underexploited

because the fruits of such labor are rarely shared outside the original teams of investigators.

Types of Narratives

The vast majority of studies focused on a single type of clinical narrative. This may be driven by a specific clinical application. For instance, Mai and Krauthammer [69] focused exclusively on free-text test orders to predict whether a patient would test positive for a particular virus in a quest to reduce viral testing volumes. To support service improvement, Elmessiry et al [30] focused solely on patient complaints. Similarly, applications related to patient safety focused on relevant documents such as adverse event reports [15], patient safety event reports [35], and incident reports [101].

Not surprisingly, most clinical applications of NLP focus on diagnosis and prognosis as they are central to medicine. Clinicians and health policymakers need to make predictions about the diagnosis and disease prognosis to support their decision making. These 2 applications focus primarily on various types of reports. For instance, electroencephalography reports were used to study epilepsy [41,70], whereas echocardiography reports were used to extract information relevant to cardiovascular medicine [55]. Most studies explored radiology reports [13,24,43,45,85,87,110,111]. They typically focus on a single imaging modality such as computer tomography [11,48,71,106,112] or magnetic resonance imaging (MRI) [17,47,94]. Such a segregated approach may be warranted by the intrinsic differences in the types of images produced, which may be reflected in the types of information discussed in the corresponding reports. For instance, MRI better differentiates between soft tissues than x-ray imaging does. Therefore, their respective reports may focus on different types of anatomical structures and their pathologies. This implies that machine learning models trained on one type of report may not be transferrable to another.

Nonetheless, aggregating findings from multiple imaging modalities [19,46,73] and other types of examination may increase diagnostic accuracy, especially when planning surgical treatments. In particular, pathology and radiology form the core of cancer diagnosis, leading to an initiative to integrate pathology and radiology studies to support making correct diagnoses and appropriate patient management and treatment decisions [132]. In this context, Bahl et al [10] combined mammographic reports, image-guided core needle biopsy reports, and surgical pathologic reports to avoid unnecessary surgical excisions. An important data source that supports this type of integration is RadBank, a database that links radiology and pathology reports [133]. It contains more than 2 million reports and allows full-text search by patient history, findings, and diagnosis by radiology and pathology. Still, the majority of studies focused on pathology reports alone [8,22,38,66,75,76]. Combinations of different report types were mostly used in enabling studies that focused on NLP tasks without a specific clinical application in mind (eg, NER approaches trained on electrocardiography, echocardiography, and radiology reports) [54,57,88,98,114].

Heterogeneity across different types of reports, including cardiac catheterization procedure reports, coronary angiographic reports

together with integrated reports that combine history and physical report, discharge summary, outpatient clinic notes, outpatient clinic letter, and inpatient discharge medication report retrieved from the Emory Cardiovascular Biobank [134] was utilized to train robust machine learning models [115]. Different subsets drawn from clinical notes, admission notes, discharge summaries, progress reports, radiology reports, allergy entries, and free-text medication orders are typically used to support fundamental NLP applications such as spell-checking [56]; coreference resolution [63]; WSD [100], including that of abbreviations [105]; and NER [20,64]. Finally, colonoscopy reports were used to explore the feasibility of NLP in a clinical setting [77,93].

Discharge summaries are used as the primary communication means between hospitals and primary care and, as such, are essential for ensuring patient safety and continuity of care. Their content and structure may vary greatly between institutions and clinicians [135]. Typical components include dates of admission and discharge, reason for hospitalization, significant findings from history and examination, significant laboratory findings, significant radiological findings, significant findings from other tests, list of procedures performed, procedure report findings, stress test report findings, pathology report findings, discharge diagnosis, condition at discharge, discharge medications, follow‐up issues, pending test results, and information provided to patients and family. Practically, discharge summaries may be viewed as amalgamations of different types of clinical narratives, some of which we discussed previously. Although this may make their processing more challenging, any algorithms trained on discharge summaries are more likely to be applicable across a wider range of clinical narratives. Discharge summaries tend to provide the most informative accounts of patient phenotypes and have been used to automate cohort selection [39]. This also makes them well suited for training and testing NER approaches [59,96,104], extraction of relationships between them [49,67], or predicting diagnoses [16].

Other types of clinical narratives considered include physician notes [84], progress notes [25,40,90], EHR notes [74,81,116], surgical notes [14,79], and emergency department notes [50,109]. Unspecified type of clinical notes [102] were used mostly for classification [9,12,31,61,86,95,103,113], WSD [33], and disambiguation and IE [36,51,99].

Psychiatric notes were used mainly in an NLP community challenge to extract protected health information and symptom severity [23,27,42,53,58,65,78,83,92]. These narratives are key enablers of mental health informatics as the fine-grained context of actionable information does not readily lend itself to predefined coding schemes. Other types of documents used to support mental health applications include psychotherapy notes [68], event and correspondence notes [32], progress notes [40], and those in general clinical context including admission notes and discharge summaries [117].

Longitudinal EHRs were mainly used in NLP community challenges [18,21,26,52,62,80,82,91,107,108]. In practical applications, cumulative patient profiles were used to predict frequent emergency department visits [37]. Longitudinal records

consisting of encounter and clinical notes were used to determine whether a candidate problem is genuine or not [28]. Similarly, encounter notes were used to determine whether a specific dermatological problem was definite, probable, or negative [44].

Clinical Applications

This section focuses on the clinical applications of NLP approaches based on machine learning. We mapped 21 clinical

applications against 7 NLP tasks (see Figure 4). It should be noted that we excluded a total of 39 studies that did not provide sufficient information to support answering RQs defined in Table 1. These studies may have described their own clinical applications, which are not discussed in this section.

Figure 4. Clinical applications underpinned by natural language processing tasks.

	Classification	Clustering	Coreference resolution	Information extraction	Named entity recognition	Ranking	Word sense disambiguation	Total
Care improvement	8			2				10
Comparative effectiveness	1							1
Data management	2			1	1			4
Diagnosis	2							2
Efficiency	1							1
Enabling	7		2	4	14		3	30
Interactive NLP	1			1				2
Knowledge acquisition				1				1
Patient literacy						1		1
Pharmacovigilance	3			1				4
Phenotyping	13							13
Prognosis	13			3				16
Quality	2							2
Referral	1							1
Resource management	8							8
Risk prediction	1							1
Safety	4							4
Service improvement	1							1
Surveillance	5			1	1			7
Triage	1	2		1	1			5
Unclear	1							1
Total	**75**	**2**	**2**	**15**	**17**	**1**	**3**	

Not surprisingly, the vast majority of studies focused on the task of text classification, which naturally lends itself to supervised machine learning. Most commonly, the classification results were used to support phenotyping, prognosis, care improvement, resource management, and surveillance.

EHR-based phenotyping approaches leverage data collected routinely in the course of health care delivery to identify cohorts of individuals that share certain clinical characteristics, events, and service patterns [136]. Their data can then be used for the secondary purposes of observational and interventional studies, prospective recruitment into clinical trials, health services research, public health surveillance, and comparative effectiveness research. Standardized computable phenotypes can enable large-scale studies while ensuring reliability and reproducibility. For instance, historical trial patient enrollment decisions were used to demonstrate the potential of NLP to increase trial screening efficiency by 450% and reduce workload associated with patient cohort identification by 90% [137]. Different types of events identified from EHRs include falls [13] and long bone fractures [43]. Most often, EHR phenotyping focused on a single medical condition, eg, axial spondyloarthritis [97], hypertension [89], systemic lupus erythematosus [95], dermatitis [44], obesity [61], celiac disease [22], epilepsy [41], autism [84], or psychiatric problems in general [40]. Two studies differentiated between multiple disorders. Tran and Kavuluru

[92] focused on 11 mental disorders including attention-deficit hyperactivity disorder, anxiety, bipolar disorder, dementia, depression, eating disorder, grief, obsessive compulsive spectrum disorder, psychosis, and posttraumatic stress disorder. Gehrmann et al [39] focused on a less homogeneous list of 10 disorders including advanced cancer, advanced heart disease, advanced lung disease, chronic neurologic dystrophies, chronic pain, alcohol abuse, substance abuse, obesity, psychiatric disorders, and depression.

In terms of prognosis, text classification results were used to predict 3-month survival [12], the likelihood of intracranial hemorrhage [11] and the development of coronary artery disease [18,26,52,62,80,82,91,107,108] or prognosis based on cancer staging [75].

At the other end of the spectrum from text classification were lower-level tasks such as coreference resolution [63,110] and WSD [33,100,105], which were not associated with any particular clinical application. However, their importance lies in enabling other higher-level NLP tasks. Similarly, as a subtask of IE, NER can be used to support structuring text into predefined templates, whose slots need to be filled with named entities of relevant types. The majority of NER studies were related to NLP community challenges such as those described in studies by Uzuner et al [123], Suominen et al [126], and Stubbs et al [131]. They focused on entities such as medical

problems, tests, and treatments [20,49,67,96,104]; disorders [54,57,88,98,114]; and protected health information [27,58,65].

Unlike NER, the more complex task of IE found a wider variety of clinical applications, the most prominent of which include prognosis and care improvement. For instance, cancer stage is one of the most important prognostic parameters in cancer, but this information is typically recorded in clinical narratives, which means that medical abstractors have to read through large volumes of text to extract such information. Given the importance and laboriousness of this task, it is not a coincidence that all IE approaches with prognosis as the most obvious clinical application focused on cancer staging [8,38,111]. Another IE approach related to cancer focused on extraction of symptoms experienced by patients during chemotherapy [36]. Rather than prognosis, this information can be used to improve patient care through modifying treatments and recognizing and managing symptoms. Similarly, extraction of information about assessments and medications can be used to improve management and outpatient treatment of patients suffering from chronic heart failure [72].

Triage is a process for sorting patients into groups based on their need for or likely benefit from medical treatment. Clustering, which is the task of grouping objects in a way that objects within a cluster are more similar to one another than to those in other clusters, can, therefore, naturally be applied to triage patients. Clustering was used to identify latent groups of lymphoma patients from their pathology reports [66]. Another study confirmed that automatically generated clusters of radiology reports coincided with major topics in radiology investigations [46]. Surprisingly, triage was not found to be a common clinical application of NLP and was largely associated with a single author [45-48].

Summary

In this review, we examined the key properties of data used to train and evaluate machine learning models. We found that the size of the training dataset tends to be relatively small. For instance, the vast majority of studies included only hundreds or thousands of documents. Relatively small proportions were utilized for training even when much larger datasets were available. Beside their volume being small, training data were typically sourced from few institutions. In addition to the NLP community challenges such as i2b2, ShARe/CLEF eHealth, and CEGS N-GRID, most commonly used data sources were MIMIC and VHA. The vast majority of studies focused on a single type of clinical narratives, which ranged from imaging reports to hospital discharge summaries. Most often, training data were used to support the tasks of text classification, IE, and NER. Only a handful of studies focused on tasks such as clustering, ranking, coreference resolution, and WSD. Most commonly, the classification results were used to support clinical applications such as phenotyping, prognosis, care improvement, resource management, and surveillance. The remaining NLP tasks did not have clear clinical applications. In fact, the majority were used to enable other higher-level NLP tasks.

Discussion

The use of text data in health informatics applications present quite a few challenges, the main ones being the preservation of patient privacy and the annotation bottleneck. Consequently, the training datasets become inflicted with problems typically associated with an unrepresentative sample. In other words, they may not reflect the distribution of characteristics of the target problem. In machine learning, such bias may lead to overfitting, a modeling error that occurs when a complex model adapts to idiosyncrasies of the training data and fails to generalize the underlying properties of the problem.

Unfortunately, most datasets used in the included studies originated from few institutions, thus offering no hard evidence about the generalizability and transferability of machine learning models. With the format and style of clinical notes varying substantially across institutions [119], a significant drop in performance was observed when training a model in one institution and testing it in another [33,61,75,105,109]. In this context, NLP community challenges play an important role in providing access to clinical data to a wider pool of researchers and establishing benchmarks for future comparisons. Not surprisingly, many studies included in this systematic review were enabled by the datasets shared in NLP community challenges. Unfortunately, relying on these challenges to provide clinical text data to NLP researchers seems like putting a Band-Aid on a proverbial bullet wound. Alternative opportunities have presented themselves in the form of synthetic health data, which contain the health records of realistic albeit not real patients. For instance, Synthea, the original open source synthetic health data software, can be used to simulate disease progression and the corresponding medical care to produce risk-free health care records at scale [138]. As synthetic data are not associated with any privacy concerns, crowdsourcing remains an option for their annotation though it may still require medical expertise, which remains an expensive commodity.

In terms of data annotation, lessons can be learned from other fields such as computer vision and speech processing, which have similarly been plagued by the lack of annotated data. They use data augmentation techniques to diversify data available for training of machine learning models without actually collecting any new data [139]. Similar techniques are now increasingly used to augment text data in a quest to improve generalization performance of the corresponding machine learning models [140-143]. Alternatively, transfer learning can be applied to leverage knowledge (features, parameters, etc) acquired in one domain and/or task with sufficient training data to support learning in another, which has got significantly less training data, thereby reducing expensive data annotation efforts [144,145]. In some cases, manual data annotation can be avoided altogether by applying the concept of distant supervision, which relies on an existing knowledge base to annotate text data automatically [146].

Some problems (eg, in-hospital death [102], discharge [90], readmission [9], and emergency department visits [37]), where labels are readily available, lend themselves naturally to supervised learning approaches. For instances, EHRs combine

free-text data with codes from controlled medical terminologies, which can be utilized as class labels [118]. These codes were used to train predictive models from historical data to identify patients at risk [16,22,50], facilitate disease surveillance [76], or optimize the cost and quality of care [81]. For other problems, where data have to be annotated manually from scratch, insisting on supervised learning is very much like trying to fit a square peg through a round hole, leaving unsupervised approaches such as topic modeling largely underexplored even though they may be better fit for purpose for clinical applications such as EHR phenotyping, patient triage, care, and service improvement.

In summary, we identified the data annotation bottleneck as one of the key obstacles to machine learning approaches in clinical NLP. Active learning has been explored as a way of using the annotation efforts in a more strategic manner. However, the clinical NLP community could benefit from using alternatives such as data augmentation, transfer learning, and distant supervision. Ultimately, unsupervised learning avoids the need for data annotation altogether and, therefore, should be used more frequently to support clinical NLP.

Acknowledgments

Conflicts of Interest

None declared.

References

1. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011;18(2):181-186 [FREE Full text] [doi: 10.1136/jamia.2010.007237] [Medline: 21233086]

2. Spasić I, Uzuner O, Zhou L. Emerging clinical applications of text analytics. Int J Med Inform 2020 Feb;134:103974. [doi: 10.1016/j.ijmedinf.2019.103974] [Medline: 31630961]

3. Spasić I, Livsey J, Keane JA, Nenadić G. Text mining of cancer-related information: review of current status and future directions. Int J Med Inform 2014 Sep;83(9):605-623 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.06.009] [Medline: 25008281]

4. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Second Edition (Morgan Kaufmann Series in Data Management Systems). Burlington, Massachusetts, USA: Morgan Kaufmann; 2008.

5. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. IEEE Intell Syst 2009 Mar;24(2):8-12. [doi: 10.1109/mis.2009.36]

6. Berman JJ. Confidentiality issues for medical data miners. Artif Intell Med 2002;26(1-2):25-36. [doi: 10.1016/s0933-3657(02)00050-7] [Medline: 12234715]

7. Kitchenham B. Department of Computer Science: Keele University. 2004. Procedures for Performing Systematic Reviews URL: http://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf [accessed 2020-03-24]

8. AAlAbdulsalam A, Garvin J, Redd A, Carter M, Sweeny C, Meystre S. Automated extraction and classification of cancer stage mentions from unstructured text fields in a central cancer registry. AMIA Jt Summits Transl Sci Proc 2018;2017:16-25 [FREE Full text] [Medline: 29888032]

9. Agarwal A, Baechle C, Behara R, Zhu X. A natural language processing framework for assessing hospital readmissions for patients with COPD. IEEE J Biomed Health Inform 2018 Mar;22(2):588-596. [doi: 10.1109/JBHI.2017.2684121] [Medline: 28328520]

10. Bahl M, Barzilay R, Yedidia AB, Locascio NJ, Yu L, Lehman CD. High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. Radiology 2018 Mar;286(3):810-818. [doi: 10.1148/radiol.2017170549] [Medline: 29039725]

11. Banerjee I, Madhavan S, Goldman R, Rubin D. Intelligent word embeddings of free-text radiology reports. AMIA Annu Symp Proc 2017;2017:411-420 [FREE Full text] [Medline: 29854105]

12. Banerjee I, Gensheimer MF, Wood DJ, Henry S, Aggarwal S, Chang DT, et al. Probabilistic prognostic estimates of survival in metastatic cancer patients (PPES-Met) utilizing free-text clinical narratives. Sci Rep 2018 Jul 3;8(1):10037 [FREE Full text] [doi: 10.1038/s41598-018-27946-5] [Medline: 29968730]

13. Bates J, Fodeh SJ, Brandt CA, Womack JA. Classification of radiology reports for falls in an HIV study cohort. J Am Med Inform Assoc 2016 Apr;23(e1):e113-e117 [FREE Full text] [doi: 10.1093/jamia/ocv155] [Medline: 26567329]

14. Ben-Ari A, Chansky H, Rozet I. Preoperative opioid use is associated with early revision after total knee arthroplasty: a study of male patients treated in the veterans affairs system. J Bone Joint Surg Am 2017 Jan 4;99(1):1-9. [doi: 10.2106/JBJS.16.00167] [Medline: 28060227]

15. Benin AL, Fodeh SJ, Lee K, Koss M, Miller P, Brandt C. Electronic approaches to making sense of the text in the adverse event reporting system. J Healthc Risk Manag 2016 Aug;36(2):10-20. [doi: 10.1002/jhrm.21237] [Medline: 27547874]

16.   Berndorfer S, Henriksson A. Automated diagnosis coding with combined text representations. Stud Health Technol Inform 2017;235:201-205. [doi: 10.3233/978-1-61499-753-5-201] [Medline: 28423783]

17.   Brown A, Marotta T. Using machine learning for sequence-level automated MRI protocol selection in neuroradiology. J Am Med Inform Assoc 2018 May 1;25(5):568-571. [doi: 10.1093/jamia/ocx125] [Medline: 29092082]

18.   Buchan K, Filannino M, Uzuner O. Automatic prediction of coronary artery disease from clinical narratives. J Biomed Inform 2017 Aug;72:23-32 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.019] [Medline: 28663072]

19.   Castro SM, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S, Bekhuis T, et al. Automated annotation and classification of BI-RADS assessment from radiology reports. J Biomed Inform 2017 May;69:177-187 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.011] [Medline: 28428140]

20.   Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. A study of active learning methods for named entity recognition in clinical text. J Biomed Inform 2015 Dec;58:11-18 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.010] [Medline: 26385377]

21.   Chen Q, Li H, Tang B, Wang X, Liu X, Liu Z, et al. An automatic system to identify heart disease risk factors in clinical texts over time. J Biomed Inform 2015 Dec;58(Suppl):S158-S163 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.002] [Medline: 26362344]

22.   Chen W, Huang Y, Boyle B, Lin S. The utility of including pathology reports in improving the computational identification of patients. J Pathol Inform 2016;7:46 [FREE Full text] [doi: 10.4103/2153-3539.194838] [Medline: 27994938]

23.   Clark C, Wellner B, Davis R, Aberdeen J, Hirschman L. Automatic classification of RDoC positive valence severity with a neural network. J Biomed Inform 2017 Nov;75S:S120-S128 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.005] [Medline: 28694118]

24.   Cocos A, Qian T, Callison-Burch C, Masino AJ. Crowd control: effectively utilizing unscreened crowd workers for biomedical data annotation. J Biomed Inform 2017 May;69:86-92 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.003] [Medline: 28389234]

25.   Cohen KB, Glass B, Greiner HM, Holland-Bouley K, Standridge S, Arya R, et al. Methodological issues in predicting pediatric epilepsy surgery candidates through natural language processing and machine learning. Biomed Inform Insights 2016;8:11-18 [FREE Full text] [doi: 10.4137/BII.S38308] [Medline: 27257386]

26.   Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Combining knowledge- and data-driven methods for de-identification of clinical narratives. J Biomed Inform 2015 Dec;58(Suppl):S53-S59 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.029] [Medline: 26210359]

27.   Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Learning to identify protected health information by integrating knowledge- and data-driven algorithms: a case study on psychiatric evaluation notes. J Biomed Inform 2017 Nov;75S:S28-S33 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.005] [Medline: 28602908]

28.   Devarakonda MV, Mehta N, Tsou C, Liang JJ, Nowacki AS, Jelovsek JE. Automated problem list generation and physicians perspective from a pilot study. Int J Med Inform 2017 Sep;105:121-129 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.05.015] [Medline: 28750905]

29.   Divita G, Luo G, Tran LT, Workman TE, Gundlapalli AV, Samore MH. General symptom extraction from VA electronic medical notes. Stud Health Technol Inform 2017;245:356-360. [Medline: 29295115]

30.   Elmessiry A, Cooper WO, Catron TF, Karrass J, Zhang Z, Singh MP. Triaging patient complaints: Monte Carlo cross-validation of six machine learning classifiers. JMIR Med Inform 2017 Jul 31;5(3):e19 [FREE Full text] [doi: 10.2196/medinform.7140] [Medline: 28760726]

31.   Fan Y, Zhang R. Using natural language processing methods to classify use status of dietary supplements in clinical notes. BMC Med Inform Decis Mak 2018 Jul 23;18(Suppl 2):51 [FREE Full text] [doi: 10.1186/s12911-018-0626-6] [Medline: 30066648]

32.   Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Sci Rep 2018 May 9;8(1):7426 [FREE Full text] [doi: 10.1038/s41598-018-25773-2] [Medline: 29743531]

33.   Finley G, Pakhomov S, McEwan R, Melton G. Towards comprehensive clinical abbreviation disambiguation using machine-labeled training data. AMIA Annu Symp Proc 2016;2016:560-569 [FREE Full text] [Medline: 28269852]

34.   Fodeh SJ, Finch D, Bouayad L, Luther SL, Ling H, Kerns RD, et al. Classifying clinical notes with pain assessment using machine learning. Med Biol Eng Comput 2018 Jul;56(7):1285-1292 [FREE Full text] [doi: 10.1007/s11517-017-1772-1] [Medline: 29280092]

35.   Fong A, Harriott N, Walters DM, Foley H, Morrissey R, Ratwani RR. Integrating natural language processing expertise with patient safety event review committees to improve the analysis of medication events. Int J Med Inform 2017 Aug;104:120-125. [doi: 10.1016/j.ijmedinf.2017.05.005] [Medline: 28529113]

36.   Forsyth AW, Barzilay R, Hughes KS, Lui D, Lorenz KA, Enzinger A, et al. Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. J Pain Symptom Manage 2018 Jun;55(6):1492-1499. [doi: 10.1016/j.jpainsymman.2018.02.016] [Medline: 29496537]

37.   Frost DW, Vembu S, Wang J, Tu K, Morris Q, Abrams HB. Using the electronic medical record to identify patients at high risk for frequent emergency department visits and high system costs. Am J Med 2017 May;130(5):601.e17-601.e22. [doi: 10.1016/j.amjmed.2016.12.008] [Medline: 28065773]

38. Gao S, Young M, Qiu J, Yoon H, Christian J, Fearn P, et al. Hierarchical attention networks for information extraction from cancer pathology reports. J Am Med Inform Assoc 2018 Mar 1;25(3):321-330. [doi: 10.1093/jamia/ocx131] [Medline: 29155996]

39. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS One 2018;13(2):e0192360 [FREE Full text] [doi: 10.1371/journal.pone.0192360] [Medline: 29447188]

40. Geraci J, Wilansky P, de Luca V, Roy A, Kennedy JL, Strauss J. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. Evid Based Ment Health 2017 Aug;20(3):83-87 [FREE Full text] [doi: 10.1136/eb-2017-102688] [Medline: 28739578]

41. Goodwin T, Harabagiu S. Multi-modal patient cohort identification from EEG report and signal data. AMIA Annu Symp Proc 2016;2016:1794-1803 [FREE Full text] [Medline: 28269938]

42. Goodwin TR, Maldonado R, Harabagiu SM. Automatic recognition of symptom severity from psychiatric evaluation records. J Biomed Inform 2017 Nov;75S:S71-S84 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.020] [Medline: 28576748]

43. Grundmeier RW, Masino A, Casper T, Dean J, Bell J, Enriquez R, Pediatric Emergency Care Applied Research Network. Identification of long bone fractures in radiology reports using natural language processing to support healthcare quality improvement. Appl Clin Inform 2016 Nov 9;7(4):1051-1068 [FREE Full text] [doi: 10.4338/ACI-2016-08-RA-0129] [Medline: 27826610]

44. Gustafson E, Pacheco J, Wehbe F, Silverberg J, Thompson W. A machine learning algorithm for identifying atopic dermatitis in adults from electronic health records. IEEE Int Conf Healthc Inform 2017 Aug;2017:83-90 [FREE Full text] [doi: 10.1109/ICHI.2017.31] [Medline: 29104964]

45. Hassanpour S, Langlotz CP. Information extraction from multi-institutional radiology reports. Artif Intell Med 2016 Jan;66:29-39 [FREE Full text] [doi: 10.1016/j.artmed.2015.09.007] [Medline: 26481140]

46. Hassanpour S, Langlotz CP. Unsupervised topic modeling in a large free text radiology report repository. J Digit Imaging 2016 Feb;29(1):59-62 [FREE Full text] [doi: 10.1007/s10278-015-9823-3] [Medline: 26353748]

47. Hassanpour S, Langlotz CP, Amrhein TJ, Befera NT, Lungren MP. Performance of a machine learning classifier of knee mri reports in two large academic radiology practices: a tool to estimate diagnostic yield. AJR Am J Roentgenol 2017 Apr;208(4):750-753. [doi: 10.2214/AJR.16.16128] [Medline: 28140627]

48. Hassanpour S, Bay G, Langlotz CP. Characterization of change and significance for clinical findings in radiology reports through natural language processing. J Digit Imaging 2017 Jun;30(3):314-322 [FREE Full text] [doi: 10.1007/s10278-016-9931-8] [Medline: 28050714]

49. He B, Guan Y, Dai R. Classifying medical relations in clinical text via convolutional neural networks. Artif Intell Med 2019 Jan;93:43-49. [doi: 10.1016/j.artmed.2018.05.001] [Medline: 29778673]

50. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLoS One 2017;12(4):e0174708 [FREE Full text] [doi: 10.1371/journal.pone.0174708] [Medline: 28384212]

51. Jagannatha A, Yu H. Bidirectional RNN for medical event detection in electronic health records. Proc Conf 2016 Jun;2016:473-482 [FREE Full text] [doi: 10.18653/v1/n16-1056] [Medline: 27885364]

52. Jonnagaddala J, Liaw S, Ray P, Kumar M, Dai H, Hsu C. Identification and progression of heart disease risk factors in diabetic patients from longitudinal electronic health records. Biomed Res Int 2015;2015:636371 [FREE Full text] [doi: 10.1155/2015/636371] [Medline: 26380290]

53. Karystianis G, Nevado AJ, Kim C, Dehghan A, Keane JA, Nenadic G. Automatic mining of symptom severity from psychiatric evaluation notes. Int J Methods Psychiatr Res 2018 Mar;27(1) [FREE Full text] [doi: 10.1002/mpr.1602] [Medline: 29271009]

54. Kholghi M, Sitbon L, Zuccon G, Nguyen A. Active learning: a step towards automating medical concept extraction. J Am Med Inform Assoc 2016 Mar;23(2):289-296. [doi: 10.1093/jamia/ocv069] [Medline: 26253132]

55. Kim Y, Garvin J, Goldstein M, Meystre S. Classification of contextual use of left ventricular ejection fraction assessments. Stud Health Technol Inform 2015;216:599-603 [FREE Full text] [doi: 10.3233/978-1-61499-564-7-599] [Medline: 26262121]

56. Lai KH, Topaz M, Goss FR, Zhou L. Automated misspelling detection and correction in clinical free-text records. J Biomed Inform 2015 Jun;55:188-195 [FREE Full text] [doi: 10.1016/j.jbi.2015.04.008] [Medline: 25917057]

57. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. J Biomed Inform 2015 Oct;57:28-37 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.010] [Medline: 26187250]

58. Lee H, Wu Y, Zhang Y, Xu J, Xu H, Roberts K. A hybrid approach to automatic de-identification of psychiatric notes. J Biomed Inform 2017 Nov;75S:S19-S27 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.006] [Medline: 28602904]

59. Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. BMC Med Inform Decis Mak 2015 May 6;15:37 [FREE Full text] [doi: 10.1186/s12911-015-0160-8] [Medline: 25943550]

60. Lin C, Dligach D, Miller TA, Bethard S, Savova GK. Multilayered temporal modeling for the clinical domain. J Am Med Inform Assoc 2016 Mar;23(2):387-395 [FREE Full text] [doi: 10.1093/jamia/ocv113] [Medline: 26521301]

61. Lingren T, Thaker V, Brady C, Namjou B, Kennebeck S, Bickel J, et al. Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers. Appl Clin Inform 2016 Jul 20;7(3):693-706 [FREE Full text] [doi: 10.4338/ACI-2016-01-RA-0015] [Medline: 27452794]

62. Liu Z, Chen Y, Tang B, Wang X, Chen Q, Li H, et al. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. J Biomed Inform 2015 Dec;58(Suppl):S47-S52 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.009] [Medline: 26122526]

63. Liu S, Liu H, Chaudhary V, Li D. An infinite mixture model for coreference resolution in clinical notes. AMIA Jt Summits Transl Sci Proc 2016;2016:428-437 [FREE Full text] [Medline: 27595047]

64. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. BMC Med Inform Decis Mak 2017 Jul 5;17(Suppl 2):67 [FREE Full text] [doi: 10.1186/s12911-017-0468-7] [Medline: 28699566]

65. Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. J Biomed Inform 2017 Nov;75S:S34-S42 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.023] [Medline: 28579533]

66. Luo Y, Xin Y, Hochberg E, Joshi R, Uzuner O, Szolovits P. Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. J Am Med Inform Assoc 2015 Sep;22(5):1009-1019 [FREE Full text] [doi: 10.1093/jamia/ocv016] [Medline: 25862765]

67. Luo Y, Cheng Y, Uzuner O, Szolovits P, Starren J. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. J Am Med Inform Assoc 2018 Jan 1;25(1):93-98 [FREE Full text] [doi: 10.1093/jamia/ocx090] [Medline: 29025149]

68. Maguen S, Madden E, Patterson OV, DuVall SL, Goldstein LA, Burkman K, et al. Measuring use of evidence based psychotherapy for posttraumatic stress disorder in a large national healthcare system. Adm Policy Ment Health 2018 Jul;45(4):519-529. [doi: 10.1007/s10488-018-0850-5] [Medline: 29450781]

69. Mai M, Krauthammer M. Controlling testing volume for respiratory viruses using machine learning and text mining. AMIA Annu Symp Proc 2016;2016:1910-1919 [FREE Full text] [Medline: 28269950]

70. Maldonado R, Goodwin T, Skinner M, Harabagiu S. Deep learning meets biomedical ontologies: knowledge embeddings for epilepsy. AMIA Annu Symp Proc 2017;2017:1233-1242 [FREE Full text] [Medline: 29854192]

71. Martinez D, Ananda-Rajah MR, Suominen H, Slavin MA, Thursky KA, Cavedon L. Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans. J Biomed Inform 2015 Feb;53:251-260 [FREE Full text] [doi: 10.1016/j.jbi.2014.11.009] [Medline: 25460203]

72. Meystre SM, Kim Y, Gobbel GT, Matheny ME, Redd A, Bray BE, et al. Congestive heart failure information extraction framework for automated treatment performance measures assessment. J Am Med Inform Assoc 2017 Apr 1;24(e1):e40-e46. [doi: 10.1093/jamia/ocw097] [Medline: 27413122]

73. Meystre S, Gouripeddi R, Tieder J, Simmons J, Srivastava R, Shah S. Enhancing comparative effectiveness research with automated pediatric pneumonia detection in a multi-institutional clinical repository: a PHIS+ pilot study. J Med Internet Res 2017 May 15;19(5):e162 [FREE Full text] [doi: 10.2196/jmir.6887] [Medline: 28506958]

74. Munkhdalai T, Liu F, Yu H. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: classical learning versus deep learning. JMIR Public Health Surveill 2018 Apr 25;4(2):e29 [FREE Full text] [doi: 10.2196/publichealth.9361] [Medline: 29695376]

75. Napolitano G, Marshall A, Hamilton P, Gavin AT. Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. Artif Intell Med 2016 Jun;70:77-83. [doi: 10.1016/j.artmed.2016.06.001] [Medline: 27431038]

76. Osborne JD, Wyatt M, Westfall AO, Willig J, Bethard S, Gordon G. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. J Am Med Inform Assoc 2016 Nov;23(6):1077-1084 [FREE Full text] [doi: 10.1093/jamia/ocw006] [Medline: 27026618]

77. Patterson O, Forbush T, Saini S, Moser SE, du Vall SL. Classifying the indication for colonoscopy procedures: a comparison of NLP approaches in a diverse national healthcare system. Stud Health Technol Inform 2015;216:614-618. [Medline: 26262124]

78. Posada JD, Barda AJ, Shi L, Xue D, Ruiz V, Kuan P, et al. Predictive modeling for classification of positive valence system symptom severity from initial psychiatric evaluation records. J Biomed Inform 2017 Nov;75S:S94-104 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.019] [Medline: 28571784]

79. Rastegar-Mojarad M, Sohn S, Wang L, Shen F, Bleeker TC, Cliby WA, et al. Need of informatics in designing interoperable clinical registries. Int J Med Inform 2017 Dec;108:78-84 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.10.004] [Medline: 29132635]

80. Roberts K, Shooshan SE, Rodriguez L, Abhyankar S, Kilicoglu H, Demner-Fushman D. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. J Biomed Inform 2015 Dec;58(Suppl):S111-S119 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.010] [Medline: 26122527]

81. Roysden N, Wright A. Predicting health care utilization after behavioral health referral using natural language processing and machine learning. AMIA Annu Symp Proc 2015;2015:2063-2072 [FREE Full text] [Medline: 26958306]

82. Sabra S, Mahmood Malik K, Alobaidi M. Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives. Comput Biol Med 2018 Mar 1;94:1-10 [FREE Full text] [doi: 10.1016/j.compbiomed.2017.12.026] [Medline: 29353160]

83. Scheurwegs E, Sushil M, Tulkens S, Daelemans W, Luyckx K. Counting trees in random forests: predicting symptom severity in psychiatric intake reports. J Biomed Inform 2017 Nov;75S:S112-S119 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.007] [Medline: 28602906]

84. Schuler A, Liu V, Wan J, Callahan A, Udell M, Stark D, et al. Discovering patient phenotypes using generalized low rank models. Pac Symp Biocomput 2016;21:144-155 [FREE Full text] [doi: 10.1142/9789814749411_0014] [Medline: 26776181]

85. Sevenster M, Buurman J, Liu P, Peters J, Chang P. Natural language processing techniques for extracting and categorizing finding measurements in narrative radiology reports. Appl Clin Inform 2015;6(3):600-610 [FREE Full text] [doi: 10.4338/ACI-2014-11-RA-0110] [Medline: 26448801]

86. Sohn S, Larson DW, Habermann EB, Naessens JM, Alabbad JY, Liu H. Detection of clinically important colorectal surgical site infection using Bayesian network. J Surg Res 2017 Mar;209:168-173 [FREE Full text] [doi: 10.1016/j.jss.2016.09.058] [Medline: 28032554]

87. Tahmasebi AM, Zhu H, Mankovich G, Prinsen P, Klassen P, Pilato S, et al. Automatic normalization of anatomical phrases in radiology reports using unsupervised learning. J Digit Imaging 2019 Feb;32(1):6-18 [FREE Full text] [doi: 10.1007/s10278-018-0116-5] [Medline: 30076490]

88. Tang B, Chen Q, Wang X, Wu Y, Zhang Y, Jiang M, et al. Recognizing disjoint clinical concepts in clinical text using machine learning-based methods. AMIA Annu Symp Proc 2015;2015:1184-1193 [FREE Full text] [doi: 10.1186/1472-6947-13-S1-S1] [Medline: 26958258]

89. Teixeira PL, Wei W, Cronin RM, Mo H, van Houten JP, Carroll RJ, et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. J Am Med Inform Assoc 2017 Jan;24(1):162-171 [FREE Full text] [doi: 10.1093/jamia/ocw071] [Medline: 27497800]

90. Temple MW, Lehmann C, Fabbri D. Natural language processing for cohort discovery in a discharge prediction model for the neonatal ICU. Appl Clin Inform 2016;7(1):101-115 [FREE Full text] [doi: 10.4338/ACI-2015-09-RA-0114] [Medline: 27081410]

91. Torii M, Fan J, Yang W, Lee T, Wiley MT, Zisook DS, et al. Risk factor detection for heart disease by applying text analytics in electronic medical records. J Biomed Inform 2015 Dec;58(Suppl):S164-S170 [FREE Full text] [doi: 10.1016/j.jbi.2015.08.011] [Medline: 26279500]

92. Tran T, Kavuluru R. Predicting mental conditions based on 'history of present illness' in psychiatric notes with deep neural networks. J Biomed Inform 2017 Nov;75S:S138-S148 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.010] [Medline: 28606869]

93. Trivedi G, Pham P, Chapman W, Hwa R, Wiebe J, Hochheiser H. NLPReViz: an interactive tool for natural language processing on clinical text. J Am Med Inform Assoc 2018 Jan 1;25(1):81-87 [FREE Full text] [doi: 10.1093/jamia/ocx070] [Medline: 29016825]

94. Trivedi H, Mesterhazy J, Laguna B, Vu T, Sohn JH. Automatic determination of the need for intravenous contrast in musculoskeletal MRI examinations using IBM Watson's natural language processing algorithm. J Digit Imaging 2018 Apr;31(2):245-251 [FREE Full text] [doi: 10.1007/s10278-017-0021-3] [Medline: 28924815]

95. Turner CA, Jacobs AD, Marques CK, Oates JC, Kamen DL, Anderson PE, et al. Word2Vec inversion and traditional text classifiers for phenotyping lupus. BMC Med Inform Decis Mak 2017 Aug 22;17(1):126 [FREE Full text] [doi: 10.1186/s12911-017-0518-1] [Medline: 28830409]

96. Unanue IJ, Borzeshi EZ, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. J Biomed Inform 2017 Dec;76:102-109 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.007] [Medline: 29146561]

97. Walsh JA, Shao Y, Leng J, He T, Teng C, Redd D, et al. Identifying axial spondyloarthritis in electronic medical records of US veterans. Arthritis Care Res (Hoboken) 2017 Sep;69(9):1414-1420 [FREE Full text] [doi: 10.1002/acr.23140] [Medline: 27813310]

98. Wang C, Akella R. A hybrid approach to extracting disorder mentions from clinical notes. AMIA Jt Summits Transl Sci Proc 2015;2015:183-187 [FREE Full text] [Medline: 26306265]

99. Wang G, Jung K, Winnenburg R, Shah NH. A method for systematic discovery of adverse drug events from clinical notes. J Am Med Inform Assoc 2015 Nov;22(6):1196-1204 [FREE Full text] [doi: 10.1093/jamia/ocv102] [Medline: 26232442]

100. Wang Y, Zheng K, Xu H, Mei Q. Clinical word sense disambiguation with interactive search and classification. AMIA Annu Symp Proc 2016;2016:2062-2071 [FREE Full text] [Medline: 28269966]

101. Wang Y, Coiera E, Runciman W, Magrabi F. Using multiclass classification to automate the identification of patient safety incident reports by type and severity. BMC Med Inform Decis Mak 2017 Jun 12;17(1):84 [FREE Full text] [doi: 10.1186/s12911-017-0483-8] [Medline: 28606174]

102. Weissman GE, Hubbard RA, Ungar LH, Harhay MO, Greene CS, Himes BE, et al. Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay. Crit Care Med 2018 Jul;46(7):1125-1132 [FREE Full text] [doi: 10.1097/CCM.0000000000003148] [Medline: 29629986]

XSL•FO

RenderX

103.   Weng WH, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak 2017 Dec 1;17(1):155 [FREE Full text] [doi: 10.1186/s12911-017-0556-8] [Medline: 29191207]

104.   Wu Y, Jiang M, Xu J, Zhi D, Xu H. Clinical named entity recognition using deep learning models. AMIA Annu Symp Proc 2017;2017:1812-1819 [FREE Full text] [Medline: 29854252]

105.   Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Wang L, et al. A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD). J Am Med Inform Assoc 2017 Apr 1;24(e1):e79-e86. [doi: 10.1093/jamia/ocw109] [Medline: 27539197]

106.   Yadav K, Sarioglu E, Choi H, Cartwright WB, Hinds PS, Chamberlain JM. Automated outcome classification of computed tomography imaging reports for pediatric traumatic brain injury. Acad Emerg Med 2016 Feb;23(2):171-178 [FREE Full text] [doi: 10.1111/acem.12859] [Medline: 26766600]

107.   Yang H, Garibaldi JM. A hybrid model for automatic identification of risk factors for heart disease. J Biomed Inform 2015 Dec;58(Suppl):S171-S182 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.006] [Medline: 26375492]

108.   Yang H, Garibaldi JM. Automatic detection of protected health information from clinic narratives. J Biomed Inform 2015 Dec;58(Suppl):S30-S38 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.015] [Medline: 26231070]

109.   Ye Y, Wagner MM, Cooper GF, Ferraro JP, Su H, Gesteland PH, et al. A study of the transferability of influenza case detection systems between two large healthcare systems. PLoS One 2017;12(4):e0174970 [FREE Full text] [doi: 10.1371/journal.pone.0174970] [Medline: 28380048]

110.   Yim W, Kwan SW, Yetisgen M. Tumor reference resolution and characteristic extraction in radiology reports for liver cancer stage prediction. J Biomed Inform 2016 Dec;64:179-191 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.005] [Medline: 27729234]

111.   Yim W, Denman T, Kwan S, Yetisgen M. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. AMIA Jt Summits Transl Sci Proc 2016;2016:455-464. [doi: 10.1148/radiographics.21.1.g01ja18237] [Medline: 27570686]

112.   Zech J, Pain M, Titano J, Badgeley M, Schefflein J, Su A, et al. Natural language-based machine learning models for the annotation of clinical radiology reports. Radiology 2018 May;287(2):570-580. [doi: 10.1148/radiol.2018171093] [Medline: 29381109]

113.   Zhang R, Ma S, Shanahan L, Munroe J, Horn S, Speedie S. Discovering and identifying New York heart association classification from electronic health records. BMC Med Inform Decis Mak 2018 Jul 23;18(Suppl 2):48 [FREE Full text] [doi: 10.1186/s12911-018-0625-7] [Medline: 30066653]

114.   Zhang E, Thurier Q, Boyle L. Improving clinical named-entity recognition with transfer learning. Stud Health Technol Inform 2018;252:182-187. [doi: 10.1007/978-3-319-96893-3_20] [Medline: 30040703]

115.   Zheng S, Lu JJ, Ghasemzadeh N, Hayek SS, Quyyumi AA, Wang F. Effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies. JMIR Med Inform 2017 May 9;5(2):e12 [FREE Full text] [doi: 10.2196/medinform.7235] [Medline: 28487265]

116.   Zheng J, Yu H. Assessing the readability of medical documents: a ranking approach. JMIR Med Inform 2018 Mar 23;6(1):e17 [FREE Full text] [doi: 10.2196/medinform.8611] [Medline: 29572199]

117.   Zhou L, Baughman AW, Lei VJ, Lai KH, Navathe AS, Chang F, et al. Identifying patients with depression using free-text clinical documents. Stud Health Technol Inform 2015;216:629-633. [doi: 10.3233/978-1-61499-564-7-629] [Medline: 26262127]

118.   Knake LA, Ahuja M, McDonald EL, Ryckman KK, Weathers N, Burstain T, et al. Quality of EHR data extractions for studies of preterm birth in a tertiary care center: guidelines for obtaining reliable data. BMC Pediatr 2016 Apr 29;16:59 [FREE Full text] [doi: 10.1186/s12887-016-0592-z] [Medline: 27130217]

119.   Sohn S, Wang Y, Wi C, Krusemark E, Ryu E, Ali M, et al. Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. J Am Med Inform Assoc 2018 Mar 1;25(3):353-359. [doi: 10.1093/jamia/ocx138] [Medline: 29202185]

120.   Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]

121.   Brown SH, Lincoln MJ, Groen PJ, Kolodner RM. VistA--US Department of Veterans Affairs national-scale HIS. Int J Med Inform 2003 Mar;69(2-3):135-156. [doi: 10.1016/s1386-5056(02)00131-4] [Medline: 12810119]

122.   Fihn SD, Francis J, Clancy C, Nielson C, Nelson K, Rumsfeld J, et al. Insights from advanced analytics at the veterans health administration. Health Aff (Millwood) 2014 Jul;33(7):1203-1211. [doi: 10.1377/hlthaff.2014.0054] [Medline: 25006147]

123.   Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]

124.   Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc 2012;19(5):786-791 [FREE Full text] [doi: 10.1136/amiajnl-2011-000784] [Medline: 22366294]

XSL·FO
RenderX

125. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013;20(5):806-813 [FREE Full text] [doi: 10.1136/amiajnl-2013-001628] [Medline: 23564629]

126. Suominen H, Salanterä S, Velupillai S, Chapman W, Savova G, Elhadad N. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages. 2013 Presented at: CLEF'13; September 23-36, 2013; Valencia, Spain p. 212-231. [doi: 10.1007/978-3-642-40802-1_24]

127. Stubbs A, Uzuner O. Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. J Biomed Inform 2015 Dec;58(Suppl):S20-S29 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.020] [Medline: 26319540]

128. Stubbs A, Uzuner O. Annotating risk factors for heart disease in clinical narratives for diabetic patients. J Biomed Inform 2015 Dec;58(Suppl):S78-S91 [FREE Full text] [doi: 10.1016/j.jbi.2015.05.009] [Medline: 26004790]

129. Bethard S, Derczynski L, Savova G, Pustejovsky J, Verhagen M. SemEval-2015 Task 6: Clinical TempEval. In: Proceedings of the 9th International Workshop on Semantic Evaluation. 2015 Presented at: SemEval'15; June, 2015; Denver, Colorado p. 806-814. [doi: 10.18653/v1/s15-2136]

130. Filannino M, Stubbs A, Uzuner O. Symptom severity prediction from neuropsychiatric clinical records: overview of 2016 CEGS N-GRID shared tasks Track 2. J Biomed Inform 2017 Nov;75S:S62-S70 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.017] [Medline: 28455151]

131. Stubbs A, Filannino M, Uzuner O. De-identification of psychiatric intake records: overview of 2016 CEGS N-GRID shared tasks Track 1. J Biomed Inform 2017 Nov;75S:S4-18 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.011] [Medline: 28614702]

132. Sorace J, Aberle DR, Elimam D, Lawvere S, Tawfik O, Wallace WD. Integrating pathology and radiology disciplines: an emerging opportunity? BMC Med 2012 Sep 5;10:100 [FREE Full text] [doi: 10.1186/1741-7015-10-100] [Medline: 22950414]

133. Rubin DL, Desser TS. A data warehouse for integrating radiologic and pathologic data. J Am Coll Radiol 2008 Mar;5(3):210-217. [doi: 10.1016/j.jacr.2007.09.004] [Medline: 18312970]

134. Ko YA, Hayek S, Sandesara P, Tahhan AS, Quyyumi A. Cohort profile: the emory cardiovascular biobank (EmCAB). BMJ Open 2017 Dec 29;7(12):e018753 [FREE Full text] [doi: 10.1136/bmjopen-2017-018753] [Medline: 29288185]

135. O'Leary KJ, Liebovitz DM, Feinglass J, Liss DT, Evans DB, Kulkarni N, et al. Creating a better discharge summary: improvement in quality and timeliness using an electronic discharge summary. J Hosp Med 2009 Apr;4(4):219-225. [doi: 10.1002/jhm.425] [Medline: 19267397]

136. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH health care systems collaboratory. J Am Med Inform Assoc 2013 Dec;20(e2):e226-e231 [FREE Full text] [doi: 10.1136/amiajnl-2013-001926] [Medline: 23956018]

137. Ni Y, Kennebeck S, Dexheimer J, McAneney C, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc 2015 Jan;22(1):166-178 [FREE Full text] [doi: 10.1136/amiajnl-2014-002887] [Medline: 25030032]

138. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc 2018 Mar 1;25(3):230-238. [doi: 10.1093/jamia/ocx079] [Medline: 29025144]

139. Ho D, Liang E, Chen X, Stoica I, Abbeel P. Population Based Augmentation: Efficient Learning of Augmentation Policy Schedules. In: Proceedings of the International Conference on Machine Learning. 2019 Presented at: CML'19; June 10-15, 2019; Long Beach, USA p. 2731-2741 URL: http://proceedings.mlr.press/v97/ho19b/ho19b.pdf

140. Li Y, Cohn T, Baldwin T. Robust Training under Linguistic Adversity. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017 Presented at: ACL'17; April 3-7, 2017; Valencia, Spain p. 21-27 URL: https://www.aclweb.org/anthology/E17-2004/ [doi: 10.18653/v1/e17-2004]

141. Xie Z, Wang SI, Li J, Levy D, Nie A, Jurafsky D, et al. Data Noising as Smoothing in Neural Network Language Models. In: Proceedings of the 5th International Conference on Learning Representations. 2017 Presented at: CLR'17; April 24-26, 2017; Toulon, France URL: https://nlp.stanford.edu/pubs/xie2017noising.pdf

142. Kobayashi S. Contextual Augmentation: Data Augmentation by Words With Paradigmatic Relations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018 Presented at: ACL'18; 2018; New Orleans, Louisiana, USA p. 452-457 URL: https://www.aclweb.org/anthology/N18-2072/ [doi: 10.18653/v1/n18-2072]

143. Wei J, Zou K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019 Presented at: EMNLP-IJCNLP; 2019; Hong Kong, China p. 6382-6388 URL: https://www.aclweb.org/anthology/D19-1670/ [doi: 10.18653/v1/d19-1670]

144. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010 Oct;22(10):1345-1359 [FREE Full text] [doi: 10.1109/tkde.2009.191]

145. Do C, Ng A. Transfer learning for text classification. Adv Neural Inf Process Syst 2006;18:299-306 [FREE Full text]

146. Mintz M, Bills S, Snow R, Jurafsky D. Distant Supervision for Relation Extraction Without Labeled Data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural

Language Processing of the AFNLP: Volume 2. 2009 Presented at: ACL'09; 2009; Suntec, Singapore p. 1003-1011 URL: https://www.aclweb.org/anthology/P09-1113/ [doi: 10.3115/1690219.1690287]

Abbreviations

CDW: Corporate Data Warehouse
EHR: electronic health record
ICD: International Classification of Diseases
IE: information extraction
MeSH: Medical Subject Headings
MIMIC: Medical Information Mart for Intensive Care
MRI: magnetic resonance imaging
NER: named entity recognition
NLP: natural language processing
RQs: research question
VHA: Veterans Health Administration
VistA: Veterans Information Systems Technology Architecture
WSD: word sense disambiguation

XSL·FO
RenderX

Original Paper

Discrepancies in Written Versus Calculated Durations in Opioid Prescriptions: Pre-Post Study

Benjamin H Slovis[1,2], MA, MD; John Kairys[1,3], MD; Bracken Babula[1,4], MD; Melanie Girondo[5], RN, MSN; Cara Martino[5], RN, MSN; Lindsey M Roke[5], PharmD; Jeffrey Riggio[1,4], MD

[1]Office of the Chief Medical Information Officer, Thomas Jefferson University, Philadelphia, PA, United States

[2]Department of Emergency Medicine, Thomas Jefferson University, Philadelphia, PA, United States

[3]Department of Surgery, Thomas Jefferson University, Philadelphia, PA, United States

[4]Department of Medicine, Thomas Jefferson University, Philadelphia, PA, United States

[5]Information Services and Technology, Thomas Jefferson University, Philadelphia, PA, United States

Corresponding Author:
Benjamin H Slovis, MA, MD
Office of the Chief Medical Information Officer
Thomas Jefferson University
833 Chestnut St, 10th Floor
Philadelphia, PA, 19107
United States
Phone: 1 215 955 7975
Email: Benjamin.Slovis@jefferson.edu

Abstract

Background: The United States is in the midst of an opioid epidemic. Long-term use of opioid medications is associated with an increased risk of dependence. The US Centers for Disease Control and Prevention makes specific recommendations regarding opioid prescribing, including that prescription quantities should not exceed the intended duration of treatment.

Objective: The purpose of this study was to determine if opioid prescription quantities written at our institution exceed intended duration of treatment and whether enhancements to our electronic health record system improved any discrepancies.

Methods: We examined the opioid prescriptions written at our institution for a 22-month period. We examined the duration of treatment documented in the prescription itself and calculated a duration based on the quantity of tablets and doses per day. We determined whether requiring documentation of the prescription duration affected these outcomes.

Results: We reviewed 72,314 opioid prescriptions, of which 16.96% had a calculated duration that was greater than what was documented in the prescription. Making the duration a required field significantly reduced this discrepancy (17.95% vs 16.21%, $P<.001$) but did not eliminate it.

Conclusions: Health information technology vendors should develop tools that, by default, accurately represent prescription durations and/or modify doses and quantities dispensed based on provider-entered durations. This would potentially reduce unintended prolonged opioid use and reduce the potential for long-term dependence.

Introduction

The United States is in the midst of an opioid epidemic [1,2]. The rate of death from opioids is increasing; more than 67% of the 70,237 overdose deaths in the United States in 2017 were attributed to opioid drugs [3]. Drug overdoses could soon become the number one cause of death of Americans under the age of 50 years [4]. Opioid overdoses are a public health crisis associated with significant medical costs and resource utilization [5-11]. Chronic opioid use is often preceded by treatment of acute pain [12], and long-term prescription opioids are associated with progression to illicit opioid abuse [13]. Despite the clear correlation between prescription opioid use and mortality, opioids continue to be routinely prescribed [2,14,15]. While the United States makes up only 4.3% of the global

population [16], Americans consume more than 80% of the world's supply of opioid medications [17].

In the Commonwealth of Pennsylvania, there were 5456 drug-related overdoses in 2017 [18], the third-highest of any state for drug-related deaths in the nation [19]. In 2016, 60% of all Pennsylvania counties had prescribing rates higher than the national average. In 2017, even though there was an overall decline in opioid prescriptions, enough oxycodone and hydrocodone were dispensed to provide every Pennsylvanian with 32 dosage units of these drugs [18].

Reduction in the number of opioid prescriptions, opioid doses, and duration of treatment are critical to curbing the opioid epidemic. State prescription drug monitoring programs (PDMPs) allow physicians to assess a patient's controlled-substance prescription history [20]. Chronic opioid use and risk of death have been associated with higher prescription doses and longer durations [12,21]. After the third day of use in an opioid-naïve individual, the likelihood of chronic opioid use increases daily. This increase is most drastic on days 5 and 31 [22]. Therefore, the US Centers for Disease Control and Prevention (CDC) has authored guidelines on opioid dosing and duration for acute and chronic pain [23,24]. Specifically, they recommend a duration of 3 days or less for acute pain, and advise that more than 7 days is unlikely to be required. The CDC also recommends that prescriptions should be of no greater quantity than what is required for the duration of treatment [23]. Consequently, opioid prescriptions with unintentionally prolonged durations of treatment may result in increased risk of chronic opioid use and associated morbidity and mortality [25].

Health information technology can play a critical role in improving quality of care while improving guideline adherence and decreasing errors [26]. Electronic health records (EHRs) have been associated with advanced quality of patient care [27-29], and computerized provider order entry (CPOE) is associated with improved efficiency and safety [30]. CPOE can improve medication ordering through clinical decision support systems in both active (ie, a pop-up warning of drug-drug interactions) and passive (ie, a dose-appropriate default prescription setting) approaches [31].

Some health information technology interventions that require providers to review aspects of current or prior opioid prescriptions have been demonstrated to affect opioid prescribing patterns. For instance, PDMP mandates that require the review of controlled substance prescription history prior to and during opioid prescribing have demonstrated a reduction in opioid prescribing rates [32]. Additionally, while some hospitals have attempted to control opioid prescribing patterns through prescription presets of a specific number of tablets [33,34], others have demonstrated both reductions and increases in tablets dispensed when requiring manual entry of number of tablets to dispense in a prescription [35,36].

Our current vendor-based EHR's CPOE ordering screen has four fields for opioid prescription entry: (1) dose (ie, number of tablets), (2) frequency (eg, three times a day), (3) duration (ie, number of doses or days), and (4) quantity (ie, the number of tablets dispensed). Previously, the system did not auto-calculate duration from dose, frequency, and quantity or

vice versa. Instead, it alerted the provider within the CPOE ordering screen and recommended updating the duration or quantity. This means a provider could write a prescription with a duration of 3 days and a quantity of 200 tablets with only a soft alert. Additionally, prior to August 2018, our institution did not require a duration value to be entered into our CPOE, meaning this field could be left blank. In August 2018, we introduced a number of interventions via modifications to our EHR's opioid prescription settings: (1) duration was set to be a required field, (2) a *quick button* for 3 days' duration was added to coincide with CDC guidelines for acute pain, and (3) tablet quantity for all opioid orders was preset to 10.

The purpose of this research is to examine if there are differences between the duration of treatment as written in an opioid prescription versus the duration associated with the dosing frequency and number of pills dispensed. That is to say, we examined how long the number of dispensed pills at the prescribed frequency would last (ie, the *calculated duration*) and compared that to the duration documented by the prescriber (ie, the *written duration*) for opioids ordered by providers at our institution. Additionally, we examined whether the requirement of a value in the duration field—from the interventions described above—had any effect on accurately representing the *calculated duration* of the prescription. We take special interest in those durations that are calculated to be longer than what was written in the duration field, as these directly contradict the CDC prescribing guidelines [23] and may increase risk of the associated negative effects of prolonged opioid use.

Methods

Data Acquisition

We queried the EHR system of the Center City division of our health care system, which includes an urban academic tertiary care center, an urban academic community hospital, and multiple ambulatory clinics. We examined data generated by Epic (Epic Systems Corporation) via a third-party analytics software, Qlik Sense (QlikTech International), to develop a list of all outpatient opioid prescriptions, including discharge medications, written over 22 months from October 2017 to July 2019. This includes an 11-month preintervention period and an 11-month postintervention period.

We extracted a number of variables, including the quantity of tablets, the dosage units (ie, mg), the route of administration (ie, oral vs buccal), the *discrete dose* (ie, 15 mg vs 10-15 mg, based on the number of tablets per dose), the written duration, and the frequency of administration.

For the purposes of this study, we limited the route of administration to *oral* and excluded all nontablet formulations. Finally, we excluded the medications buprenorphine and methadone, as these are routinely used for the management of opioid use disorder.

Calculated Duration

In order to compute the calculated duration of each prescription, we took each of the unique possible frequencies in the system (ie, twice a day or every 4 hours) and mapped these to the

equivalent number of administrations per 24-hour day. For pro re nata (PRN) medications, we selected the maximum frequency possible. Therefore, "twice a day PRN" resulted in two daily administrations and "every other day" resulted in a daily administration of 0.5.

Next, we took all doses that contained ranges for their *discrete dose* (ie, 20-30 mg) and isolated the maximum dose possible per administration (ie, 30 mg). We calculated the number of tablets per administration by dividing the maximum dose by the dose per tablet. We calculated the total number of possible doses by dividing the quantity dispensed by number of tablets per administration. Finally, we computed the calculated duration, measured in days, by dividing the total number of doses by the number of administrations per 24 hours. The choice of the maximum dose per administration means that our calculated duration is the shortest possible when holding the other variables constant.

Statistical Analysis

We first converted all written durations to be represented in units of *days*. We then examined the proportion of prescriptions that did not have a written duration to determine if our intervention had an effect on the documentation of this field.

We also examined the proportion of prescriptions for which we could not compute a calculated duration to ensure the preintervention and postintervention periods were similar. Finally, we compared the number of prescriptions written with ranges (ie, 20-30 mg) to see if our intervention had an effect on this category of prescriptions.

To compare what was documented (ie, written duration) to what was dispensed (ie, calculated duration) for each prescription, we then generated our study cohort by excluding any prescription that did not hold a value for both of these fields. Also, since we computed the calculated duration in units of days, we excluded all written durations not also documented in units of days.

We examined the written duration of each prescription and compared it to the calculated duration. We categorized each relationship as a written duration equal to, greater than, or less than the calculated duration. We examined whether there were any changes in the relationships between written duration and calculated duration before and after our interventions, with specific interest regarding the requirement to document a written duration. See Table 1 for examples of each field and how we coded relationships.

Table 1. Examples of each field extracted from the electronic health record system, with written duration and calculated duration demonstrated, along with their relationship.

Drug	Prescription instructions	Dispense quantity	Written duration	Administrations per day	Maximum dose per administration	Calculated duration[a]	Relationship
Morphine: 15 mg tablets	Take 1-2 tablets as needed twice a day for 3 days	12	3 days	2	30 mg (2 tablets)	3 days	Written duration is equal to calculated duration
Morphine: 15 mg tablets	Take 1-2 tablets as needed twice a day for 3 days	8	3 days	2	30 mg (2 tablets)	2 days	Written duration is longer than calculated duration
Morphine: 15 mg tablets	Take 1-2 tablets as needed twice a day for 3 days	20	3 days	2	30 mg (2 tablets)	5 days	Written duration is shorter than calculated duration

[a]Calculated duration=((quantity of tablets × drug dose in mg) / maximum dose per administration) / administrations per day.

Data cleaning and calculations were performed in Qlik Sense. Statistical analysis was performed in the statistical software R, version 3.3.2 (The R Foundation). Chi-square analysis was used for categorical values. The *t* test was used for parametric data and the Wilcoxon rank-sum test was used for nonparametric data.

Results

General Results

There were 92,462 unique opioid prescriptions that met initial inclusion criteria during our study period for 30,426 individual patients. There was a mean number of prescriptions per month of 4202.82 (SD 204.32) and a median number of prescriptions per person per month of 1 (IQR 1-2).

In the preintervention period, there were 47,131 prescriptions for 16,863 individual patients; in the postintervention period, there were 45,331 prescriptions for 17,483 individual patients. There was no significant difference (*P*=.06) between mean

number of prescriptions per month in the preintervention period (mean 4284.64 [SD 200.68]) versus the postintervention period (mean 4121.00 [SD 180.74]). The median number of 1 prescription per person per month did not change (IQR 1-2, *P*=.37).

There was a statistically significant decrease in the proportion of prescriptions with no written duration documented postintervention (33.54%, 95% CI 33.12-33.97, vs 9.45%, 95% CI 9.19-9.72, *P*<.001). Evaluation of the remaining 9.45% of prescriptions without a written duration appear to be due to refills of prior prescriptions, which were exempt from the new documentation requirement.

There was a small but significant difference in the proportion of prescriptions in which we were unable to compute the calculated duration when comparing the preperiod and the postperiod (6.01%, 95% CI 5.80-6.23, vs 3.61%, 95% CI 3.44-3.79, *P*<.001).

There was no significant difference pre- and postintervention in the number of prescriptions that contained ranged doses (ie, 20-30 mg) (4.42%, 95% CI 4.23-4.61, vs 4.66%, 95% CI 4.47-4.86, *P*=.08).

There were 72,364 out of 92,462 (78.26%) total prescriptions that had a written duration documented. Of these, there were 2632 (3.64%) prescriptions whose written duration units were converted from other duration units to *days*, while the rest were already written in this unit of measurement. Out of 72,364 prescriptions, 50 (0.07%) were excluded because the calculated duration could not be computed. This resulted in 72,314 out of 92,462 (78.21%) total prescriptions meeting the inclusion criteria for our comparison: out of 72,314 prescriptions, 31,300 (43.28%) were in the preintervention period and 41,014 (56.72%) were in the postintervention period. Figure 1 demonstrates the inclusion and exclusion criteria and separate cohorts for comparison.

Figure 1. Breakdown of cohorts based on inclusion and exclusion criteria and separation into preperiod and postperiod cohorts.
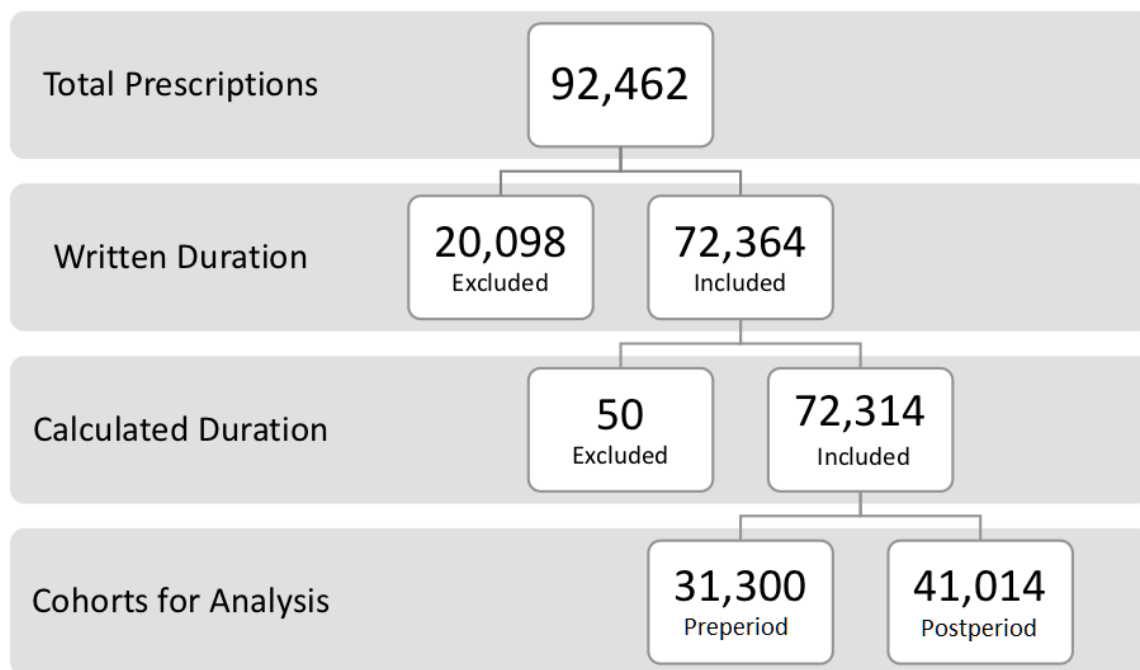
Comparison of Written and Calculated Durations

Of the 72,314 prescriptions, 41.97% (30,353 total, 95% CI 41.61-42.33) had calculated durations that were equivalent to their written durations, 41.06% (29,694 total, 95% CI 40.70-41.42) had calculated durations that were less than their written durations, and 16.96% (12,267 total, 95% CI 16.69-17.24) had a calculated duration that was greater than the written duration.

Requiring a written duration to be documented significantly improved the number of calculated durations that were equal to their corresponding written duration, from 38.86% (12,163/31,300, 38.86%, 95% CI 38.32-39.40) to 44.35% (18,190/41,014, 44.35%, 95% CI 43.87-44.83, *P*<.001). Additionally, requiring a written duration resulted in a reduction in prescriptions where the calculated duration was longer than the written duration (5617/31,300, 17.95%, 95% CI 17.52-18.37, vs 6650/41,014, 16.21%, 95% CI 15.86-16.57, *P*<.001). Changes in percentages of each relationship between written duration and calculated duration, pre- and postintervention, are presented in Table 2.

Table 2. Totals and percentages of each cohort and their associated relationships for pre- and postintervention.

Relationship	Preperiod (N=31,300), n (%)	Postperiod (N=41,014), n (%)	*P* value	Total (N=72,314), n (%)
Calculated duration equal to written duration	12,163 (38.86)	18,190 (44.35)	<.001	30,353 (41.97)
Calculated duration less than written duration	13,520 (43.19)	16,174 (39.44)	<.001	29,694 (41.06)
Calculated duration greater than written duration	5617 (17.95)	6650 (16.21)	<.001	12,267 (16.96)

Discussion

Principal Findings

The opioid epidemic is a major concern in the United States, specifically in the Commonwealth of Pennsylvania [1,18] where the governor has issued a state emergency disaster declaration since 2018 [37]. Our hospitals and clinics are located in the heart of a major urban area within the state, in a county that hosts the highest estimated frequency of overdose deaths [38] and neighbors a district considered to be the epicenter for illicit opioid trade in the region [39].

In order to aid in many state-, city-, and hospital-level initiatives, our clinical informatics group has recently employed solutions to combat the opioid epidemic. These include simplifying PDMP queries, modifications of prescription settings in the emergency department, and analytics tools to track opioid prescribing throughout the institution [40,41].

Reduction in opioid prescriptions is a key step in combating this epidemic, and the duration of prescriptions has a significant effect on a patient's risk of becoming a chronic opioid user [12,22].

In order to study our upcoming planned interventions, we realized we needed a reliable way to calculate durations of opioid prescriptions. Given the prior lack of requirement to document a written duration in our EHR system, we wanted to explore whether this was an accurate representation of a prescription's true duration and whether making this a requirement improved overall representation of the duration of a prescription as measured by the calculated duration.

Overall, written duration was *equal to* the calculated duration (41.97%) or *less than* the calculated duration (41.06%) the majority of the time. However, 16.96% of the time our prescriptions had a longer calculated duration than the written duration, which contradicts CDC prescribing guidelines. We are concerned that unintentionally prolonged opioid prescription durations could be contributing to an increased risk of opioid dependence. We must take this opportunity to stress that we do not imply that these prescriptions were intentionally written to break guidelines or to prolong a patient's duration of opioid therapy. We are simply stating that we observed a difference in these two values and we believe this requires attention. One solution to this issue could be modifications to the EHR system.

In order to provide peak levels of patient safety and end-user satisfaction, EHRs require continuous review and optimization. A lack of a required field allowed prescription durations to not match what the provider had documented, or likely intended. This is apparent in that requiring documentation of this field reduced those prescriptions without written durations, improved the percentage of written durations that accurately represented their corresponding calculated durations, and reduced the number of calculated durations that exceeded written durations. As discussed in our results, we explored the remaining 9.45% of prescriptions that continued to lack a written duration after our intervention and determined that they were likely due to renewals of prescriptions written prior to our modifications, where the system allows the field previously left blank to remain. Additionally, our intervention did yield a small but significant reduction in prescriptions where we were unable to compute a calculated duration (6.01% vs 3.61%). We attribute this to the addition of a quick button, as described in our methods.

Many studies have demonstrated how modifications to CPOE can affect opioid prescribing, though most of the literature focuses on tablet counts and not duration of treatment. Delgado et al demonstrated that when transitioning from an EHR that had no preset dispense quantity to an EHR that required a preset of 10 tablets, the median number of oxycodone 5 mg/ acetaminophen 325 mg tablets dispensed by two emergency departments decreased from 11.3 to 10.0 and from 12.6 to 10.9, respectively [33]. Similarly, Chiu et al demonstrated that when opioids were prescribed at discharge for outpatient surgery, modification of the default pill count from 30 tablets to 12 tablets reduced the median number of tablets dispensed from 30 to 20 tablets [34]. However, they did not examine prescription duration, noting that duration guidelines are usually "far longer than most patients will need," and modifications of number of tablets dispensed will have a more profound affect [34]. While we believe reduction of tablets dispensed is critical to combating the opioid epidemic, at this time, CDC guidelines consider prescription duration to be an important measurement in reducing opioid therapy and combating the epidemic [23,24].

Conversely, other studies have demonstrated results from entirely removing presets for opioid tablets dispensed. Santistevan et al showed that in an emergency department setting, removal of the default of 20 tablets for hydrocodone and oxycodone, and the requirement of the prescriber to enter a number of tablets to dispense, reduced the median number of tablets prescribed from 20 to 15 [35]. The authors concluded that EHR presets may hinder providers' ability to prescribe opioids to appropriately fit the variability in patient care (ie, more painful clinical conditions need more tablets) and that prescriptions written at the provider's discretion may be more appropriate. Contrary to these results, Zwank et al demonstrated an increase in mean number of tablets from 15.31 to 15.77 after removing their 15-tablet preset and requiring manual entry of tablets dispensed [36].

Crothers et al also examined the effects of transitioning between EHR systems, from a homegrown system to a vendor-based system [42]. Their prior EHR system auto-calculated the maximum number of *dispense units* for PRN opioid prescriptions based on dose, frequency, and duration documented. During their implementation, they removed this functionality and instead developed preset dispense quantities, with outpatient prescriptions for clinics and inpatient discharges defaulted to 30 tablets for oxycodone and *null* for hydrocodone. This resulted in a decrease of 1.4 dispense units overall and 3.9 dispense units for inpatient discharge [42].

While these studies have examined the effect of changes in the requirement to document the number of tablets dispensed [33-36], to our knowledge we are the first to examine the requirement of documenting a duration field and comparing the duration intended by a prescriber (ie, written duration) to the

calculated duration based on the number of tablets dispensed and instructions provided.

Our results confirm that simple documentation of an opioid prescription's duration is not sufficient. When a duration field is not directly linked to other elements of the prescription and there is no automated calculation of duration from dose, frequency, and quantity—or vice versa—inaccuracies remain. While it is interesting that Crothers et al demonstrated that removal of such functionality reduced opioid dispensing [42], this was performed during a transition of EHR systems and with the addition of prescription presets, which likely influenced their results. Given the importance of the duration of an opioid prescription in the associated risk of long-term use, as well as further potential for overdose, it is vital to have data that accurately represents this value. Additionally, we have demonstrated that retrospective computation of a calculated duration is a viable alternative when written durations are not available for the study of EHR prescribing data.

In order to improve compliance to CDC opioid prescribing guidelines, EHR system vendors should consider rapidly developing tools that, by default, accurately represent prescription durations and/or modify doses and quantities dispensed based on provider-entered durations. As described in the Introduction, this calculation already exists in our system, but no hard stop or passive updating of prescription fields exists. If a provider intends to prescribe one tablet, three times a day for 3 days, the system should automatically set the quantity at nine tablets; if the provider reduces the quantity to six tablets, the duration of the prescription should automatically be reduced to 2 days.

Limitations

Our study was performed at a single, urban academic institution and its associated ambulatory clinics; therefore, our results may not represent the majority of the country's hospitals. However, given that our institution is located in the state with the third-highest rate of drug overdose, our results may be valid when compared to other states of similar rates of overdose.

Additionally, as discussed in our methods, we intentionally biased our process of computing the calculated duration to make our resulting durations as short as possible by assuming individual doses were the higher of the possible range (ie, if a possible dose was 5-10 mg, we assumed the dose was 10 mg, thereby meaning that the quantity was used faster and the duration was shorter). This means that our calculations may underrepresent the number of prescriptions where the calculated duration was longer than the written duration if patients were to ration their medications and take lower doses over a longer period of time. Further, while our main intervention was making the duration field a requirement, we also added quick-action buttons and modified some prescription settings, which may have influenced our postperiod results, though we expect this to have been minimal. Finally, we used EHR data to represent durations of medications. While we believe it is important for our prescriptions to accurately represent the intended therapy and to abide by CDC guidelines, we did not determine whether individual patients took their prescribed medication as written, nor did we examine whether each prescription was filled at a pharmacy.

Conclusions

Accurate documentation of an opioid prescription's duration is critical, both for patient safety and for secondary use in analysis of the status of the opioid epidemic, as well as for evaluating interventions implemented to combat this public health crisis. Our study demonstrates that more than 17% of prescriptions written at our institution had durations documented in the EHRs that were shorter than durations calculated via the dose, frequency, and quantity of tablets prescribed. Requiring documentation of the duration field in a prescription improved these errors statistically but, clinically, a large number of prescriptions continued to not match the calculated duration. EHR vendors should invest in research and development to create functions that automatically calculate and fill values of the opioid prescription to ensure prescriptions are accurately represented, while physicians and hospitals should invest in informatics initiatives to study and improve provider-prescribing practices.

Conflicts of Interest

None declared.

References

1. The Opioid Epidemic by the Numbers. Washington, DC: US Department of Health & Human Services (HHS); 2019 Jan. URL: https://www.hhs.gov/opioids/sites/default/files/2019-09/opioids-infographic.pdf [accessed 2020-01-27]
2. Rudd RA, Seth P, David F, Scholl L. Increases in drug and opioid-involved overdose deaths: United States, 2010-2015. MMWR Morb Mortal Wkly Rep 2016 Dec 30;65(50-51):1445-1452 [FREE Full text] [doi: 10.15585/mmwr.mm655051e1] [Medline: 28033313]
3. Scholl L, Seth P, Kariisa M, Wilson N, Baldwin G. Drug and opioid-involved overdose deaths: United States, 2013-2017. MMWR Morb Mortal Wkly Rep 2018 Jan 04;67(5152):1419-1427 [FREE Full text] [doi: 10.15585/mmwr.mm675152e1] [Medline: 30605448]

XSL•FO

RenderX

4.   Katz J. The New York Times. 2017. Drug deaths in America are rising faster than ever URL: https://www.nytimes.com/interactive/2017/06/05/upshot/opioid-epidemic-drug-overdose-deaths-are-rising-faster-than-ever.html [accessed 2017-08-19]

5.   Kuehn BM. Alarming nonfatal overdose rates found for opioids, sedatives, and tranquilizers. JAMA 2010 May 26;303(20):2020-2021. [doi: 10.1001/jama.2010.641] [Medline: 20501919]

6.   Coben JH, Davis SM, Furbee PM, Sikora RD, Tillotson RD, Bossarte RM. Hospitalizations for poisoning by prescription opioids, sedatives, and tranquilizers. Am J Prev Med 2010 May;38(5):517-524. [doi: 10.1016/j.amepre.2010.01.022] [Medline: 20409500]

7.   Dunn KM, Saunders KW, Rutter CM, Banta-Green CJ, Merrill JO, Sullivan MD, et al. Opioid prescriptions for chronic pain and overdose: A cohort study. Ann Intern Med 2010 Jan 19;152(2):85-92 [FREE Full text] [doi: 10.7326/0003-4819-152-2-201001190-00006] [Medline: 20083827]

8.   Coffin PO, Galea S, Ahern J, Leon AC, Vlahov D, Tardiff K. Opiates, cocaine and alcohol combinations in accidental drug overdose deaths in New York City, 1990-98. Addiction 2003 Jun;98(6):739-747. [doi: 10.1046/j.1360-0443.2003.00376.x] [Medline: 12780362]

9.   Darke S. Oxycodone poisoning: Not just the 'usual suspects'. Addiction 2011 Jun;106(6):1035-1036. [doi: 10.1111/j.1360-0443.2011.03395.x] [Medline: 21564365]

10. Häkkinen M, Launiainen T, Vuori E, Ojanperä I. Comparison of fatal poisonings by prescription opioids. Forensic Sci Int 2012 Oct 10;222(1-3):327-331. [doi: 10.1016/j.forsciint.2012.07.011] [Medline: 22884575]

11. Hall AJ, Logan JE, Toblin RL, Kaplan JA, Kraner JC, Bixler D, et al. Patterns of abuse among unintentional pharmaceutical overdose fatalities. JAMA 2008 Dec 10;300(22):2613-2620. [doi: 10.1001/jama.2008.802] [Medline: 19066381]

12. Edlund MJ, Martin BC, Russo JE, DeVries A, Braden JB, Sullivan MD. The role of opioid prescription in incident opioid abuse and dependence among individuals with chronic noncancer pain: The role of opioid prescription. Clin J Pain 2014 Jul;30(7):557-564 [FREE Full text] [doi: 10.1097/AJP.0000000000000021] [Medline: 24281273]

13. Jones CM. Heroin use and heroin use risk behaviors among nonmedical users of prescription opioid pain relievers: United States, 2002-2004 and 2008-2010. Drug Alcohol Depend 2013 Sep 01;132(1-2):95-100. [doi: 10.1016/j.drugalcdep.2013.01.007] [Medline: 23410617]

14. Gaither JR, Leventhal JM, Ryan SA, Camenga DR. National trends in hospitalizations for opioid poisonings among children and adolescents, 1997 to 2012. JAMA Pediatr 2016 Dec 01;170(12):1195-1201. [doi: 10.1001/jamapediatrics.2016.2154] [Medline: 27802492]

15. Nuckols TK, Anderson L, Popescu I, Diamant AL, Doyle B, Di Capua P, et al. Opioid prescribing: A systematic review and critical appraisal of guidelines for chronic pain. Ann Intern Med 2014 Jan 07;160(1):38-47 [FREE Full text] [doi: 10.7326/0003-4819-160-1-201401070-00732] [Medline: 24217469]

16. United States Census Bureau. 2020. US and world population clock URL: https://www.census.gov/popclock/ [accessed 2020-01-27]

17. Gusovsky D. CNBC. 2016 Apr 27. Americans consume vast majority of the world's opioids URL: https://www.cnbc.com/2016/04/27/americans-consume-almost-all-of-the-global-opioid-supply.html [accessed 2019-09-04]

18. DEA Philadelphia Division, University of Pittsburgh. The Opioid Threat in Pennsylvania. Arlington, VA: United States Drug Enforcement Administration; 2018 Sep. URL: https://www.dea.gov/sites/default/files/2018-10/PA%20Opioid%20Report%20Final%20FINAL.pdf [accessed 2019-09-04]

19. Centers for Disease Control and Prevention. 2019 Jan 10. Drug overdose mortality by state URL: https://www.cdc.gov/nchs/pressroom/sosmap/drug_poisoning_mortality/drug_poisoning.htm [accessed 2019-06-26]

20. Griggs CA, Weiner SG, Feldman JA. Prescription drug monitoring programs: Examining limitations and future approaches. West J Emerg Med 2015 Jan;16(1):67-70 [FREE Full text] [doi: 10.5811/westjem.2014.10.24197] [Medline: 25671011]

21. Bohnert AS, Valenstein M, Bair MJ, Ganoczy D, McCarthy JF, Ilgen MA, et al. Association between opioid prescribing patterns and opioid overdose-related deaths. JAMA 2011 Apr 06;305(13):1315-1321. [doi: 10.1001/jama.2011.370] [Medline: 21467284]

22. Shah A, Hayes CJ, Martin BC. Characteristics of initial prescription episodes and likelihood of long-term opioid use: United States, 2006-2015. MMWR Morb Mortal Wkly Rep 2017 Mar 17;66(10):265-269 [FREE Full text] [doi: 10.15585/mmwr.mm6610a1] [Medline: 28301454]

23. Dowell D, Haegerich TM, Chou R. CDC guideline for prescribing opioids for chronic pain: United States, 2016. JAMA 2016 Apr 19;315(15):1624-1645 [FREE Full text] [doi: 10.1001/jama.2016.1464] [Medline: 26977696]

24. Dowell D, Haegerich TM, Chou R. CDC guideline for prescribing opioids for chronic pain: United States, 2016. MMWR Recomm Rep 2016 Mar 18;65(1):1-49. [doi: 10.15585/mmwr.rr6501e1] [Medline: 26987082]

25. Han B, Compton WM, Blanco C, Crane E, Lee J, Jones CM. Prescription opioid use, misuse, and use disorders in US adults: 2015 National Survey on Drug Use and Health. Ann Intern Med 2017 Sep 05;167(5):293-301. [doi: 10.7326/M17-0865] [Medline: 28761945]

26. Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, et al. Systematic review: Impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med 2006 May 16;144(10):742-752. [doi: 10.7326/0003-4819-144-10-200605160-00125] [Medline: 16702590]

27.   Erstad TL. Analyzing computer based patient records: A review of literature. J Healthc Inf Manag 2003;17(4):51-57. [Medline: 14558372]

28.   Menachemi N, Brooks RG. Reviewing the benefits and costs of electronic health records and associated patient safety technologies. J Med Syst 2006 Jun;30(3):159-168. [doi: 10.1007/s10916-005-7988-x] [Medline: 16848129]

29.   Wang SJ, Middleton B, Prosser LA, Bardon CG, Spurr CD, Carchidi PJ, et al. A cost-benefit analysis of electronic medical records in primary care. Am J Med 2003 Apr 01;114(5):397-403. [doi: 10.1016/s0002-9343(03)00057-3] [Medline: 12714130]

30.   Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011;4:47-55 [FREE Full text] [doi: 10.2147/RMHP.S12985] [Medline: 22312227]

31.   Tolley CL, Slight SP, Husband AK, Watson N, Bates DW. Improving medication-related clinical decision support. Am J Health Syst Pharm 2018 Feb 15;75(4):239-246. [doi: 10.2146/ajhp160830] [Medline: 29436470]

32.   Wen H, Hockenberry JM, Jeng PJ, Bao Y. Prescription drug monitoring program mandates: Impact on opioid prescribing and related hospital use. Health Aff (Millwood) 2019 Sep;38(9):1550-1556. [doi: 10.1377/hlthaff.2019.00103] [Medline: 31479368]

33.   Delgado MK, Shofer FS, Patel MS, Halpern S, Edwards C, Meisel ZF, et al. Association between electronic medical record implementation of default opioid prescription quantities and prescribing behavior in two emergency departments. J Gen Intern Med 2018 Apr;33(4):409-411 [FREE Full text] [doi: 10.1007/s11606-017-4286-5] [Medline: 29340937]

34.   Chiu AS, Jean RA, Hoag JR, Freedman-Weiss M, Healy JM, Pei KY. Association of lowering default pill counts in electronic medical record systems with postoperative opioid prescribing. JAMA Surg 2018 Nov 01;153(11):1012-1019 [FREE Full text] [doi: 10.1001/jamasurg.2018.2083] [Medline: 30027289]

35.   Santistevan JR, Sharp BR, Hamedani AG, Fruhan S, Lee AW, Patterson BW. By default: The effect of prepopulated prescription quantities on opioid prescribing in the emergency department. West J Emerg Med 2018 Mar;19(2):392-397 [FREE Full text] [doi: 10.5811/westjem.2017.10.33798] [Medline: 29560071]

36.   Zwank MD, Kennedy SM, Stuck LH, Gordon BD. Removing default dispense quantity from opioid prescriptions in the electronic medical record. Am J Emerg Med 2017 Oct;35(10):1567-1569. [doi: 10.1016/j.ajem.2017.04.002] [Medline: 28416266]

37.   Wolf T. Proclamation: Amendment to Proclamation of Disaster Emergency. Harrisburg, PA: Commonwealth of Pennsylvania, Office of the Governor; 2019 Mar 20. URL: https://www.governor.pa.gov/wp-content/uploads/2019/03/20190320-opioid-disaster-emergency-renual.pdf [accessed 2019-09-04]

38.   OpenDataPA, Commonwealth of Pennsylvania. 2019 Jul 19. Estimated accidental and undetermined drug overdose deaths CY 2017-2018: County Health URL: https://data.pa.gov/Opioid-Related/Estimated-Accidental-and-Undetermined-Drug-Overdos/azzc-q64m [accessed 2020-01-27]

39.   Percy J. The New York Times Magazine. 2018 Oct 10. Trapped by the 'Walmart of Heroin' URL: https://www.nytimes.com/2018/10/10/magazine/kensington-heroin-opioid-philadelphia.html [accessed 2019-09-04]

40.   Slovis B, London K, Randolph F, Aini M, Mammen P, Martino C, et al. The effect of implementing electronic health record default prescribing preferences on opioid prescriptions written in the emergency department. Ann Emerg Med 2018 Oct;72(4, Supplement):S133 [FREE Full text] [doi: 10.1016/j.annemergmed.2018.08.341]

41.   Martino C. Upshot. 2018 Jul 09. Fighting the opioid epidemic with the right people and the right tools URL: https://upshotstories.com/stories/fighting-the-opioid-epidemic-with-the-right-people-and-the-right-tools [accessed 2019-09-04]

42.   Crothers G, Edwards DA, Ehrenfeld JM, Woo E, McCluggage L, Lobo B. Evaluating the impact of auto-calculation settings on opioid prescribing at an academic medical center. Jt Comm J Qual Patient Saf 2019 Jun;45(6):416-422. [doi: 10.1016/j.jcjq.2019.02.010] [Medline: 30935884]

Abbreviations

CDC: US Centers for Disease Control and Prevention
CPOE: computerized provider order entry
EHR: electronic health record
PDMP: prescription drug monitoring program
PRN: pro re nata

Original Paper

Insurance Customers' Expectations for Sharing Health Data: Qualitative Survey Study

Casandra Grundstrom[1], BTECH, MSc; Olli Korhonen[1], MSc; Karin Väyrynen[1], PhD; Minna Isomursu[1], PhD

Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu, Finland

Corresponding Author:
Casandra Grundstrom, BTECH, MSc
Faculty of Information Technology and Electrical Engineering
University of Oulu
Pentti Kaiteran katu 1
PO Box 4500
Oulu, FI-90014
Finland
Phone: 358 0503536025
Email: casandra.grundstrom@gmail.com

Abstract

Background: Insurance organizations are essential stakeholders in health care ecosystems. For addressing future health care needs, insurance companies require access to health data to deliver preventative and proactive digital health services to customers. However, extant research is limited in examining the conditions that incentivize health data sharing.

Objective: This study aimed to (1) identify the expectations of insurance customers when sharing health data, (2) determine the perceived intrinsic value of health data, and (3) explore the conditions that aid in incentivizing health data sharing in the relationship between an insurance organization and its customer.

Methods: A Web-based survey was distributed to randomly selected customers from a Finnish insurance organization through email. A single open-text answer was used for a qualitative data analysis through inductive coding, followed by a thematic analysis. Furthermore, the 4 constructs of commitment, power, reciprocity, and trust from the social exchange theory (SET) were applied as a framework.

Results: From the 5000 customers invited to participate, we received 452 surveys (response rate: 9.0%). Customer characteristics were found to reflect customer demographics. Of the 452 surveys, 48 (10.6%) open-text responses were skipped by the customer, 57 (12.6%) customers had no expectations from sharing health data, and 44 (9.7%) customers preferred to abstain from a data sharing relationship. Using the SET framework, we found that customers expected different conditions to be fulfilled by their insurance provider based on the commitment, power, reciprocity, and trust constructs. Of the 452 customers who completed the surveys, 64 (14.2%) customers required that the insurance organization meets their data treatment expectations (commitment). Overall, 4.9% (22/452) of customers were concerned about their health data being used against them to profile their health, to increase insurance prices, or to deny health insurance claims (power). A total of 28.5% (129/452) of customers expected some form of benefit, such as personalized digital health services, and 29.9% (135/452) of customers expected finance-related compensation (reciprocity). Furthermore, 7.5% (34/452) of customers expected some form of empathy from the insurance organization through enhanced transparency or an emotional connection (trust).

Conclusions: To aid in the design and development of digital health services, insurance organizations need to address the customers' expectations when sharing their health data. We established the expectations of customers in the social exchange of health data and explored the perceived values of data as intangible goods. Actions by the insurance organization should aim to increase trust through a culture of transparency, commitment to treat health data in a prescribed manner, provide reciprocal benefits through digital health services that customers deem valuable, and assuage fears of health data being used to prevent providing insurance coverage or increase costs.

KEYWORDS

XSL•FO

RenderX

Introduction

Background

The paradigm shift toward person-centric health care promotes the engagement of individuals in their care process, providing pervasive digital interventions for preventative and proactive health and wellness [1,2]. At the core of this paradigm shift are health data, as data usage is the cornerstone for health care and pivotal for the success of the health data economy [3]. Individuals are being empowered with health data sharing capabilities through mechanisms such as interoperability and portability, which are slated to facilitate a wide range of financial, academic, societal, and personal benefits for health and wellness [4]. For example, patient-led health data sharing on the digital platform, *PatientsLikeMe*, has led to better outcomes in areas such as symptom management and medication adherence [5]. Similarly, monitoring platforms for home-based self-measurements can support the decision-making processes of health care professionals using real-time patient-generated health data [6]. The mechanisms for health data sharing are enacted in the General Data Protection Regulation (GDPR) throughout Europe and implicate data processors and controllers to facilitate desirable outcomes such as access to health data [7]. Consequently, traditional business models are being disrupted and transformed to sustain organizations, now that the power to control health data has been shifted toward data subjects such as patients, individuals, or customers [8-11].

Extant literature has mostly focused on the barriers of health data sharing in the public setting [4], emphasizing personalization, care improvement, and the intended use of health data as conditions that facilitate sharing [12-14]. Objections to health data sharing with insurance organizations are much more prevalent in the existing literature. Nearly half of the participants in a study in the United States, including doctors and patients, were found to strongly object to disclosing health data because of concerns about discrimination. The participants expected their shared health data to be used as leverage to impede medical care or insurance coverage [15]. In the Nordic countries especially, there is only limited empirical research on customers sharing health data in the private insurance sector [16]. In particular, there is a lack of case study research wherein relevant stakeholders are studied holistically to develop insights into the motivations for and perceptions about sharing health data [17]. Studies that are available indicate similar findings; the majority of individuals are unwilling to share their health data with insurance organizations. Approximately 57% of customers from a Finnish insurance organization indicated an unwillingness to share their health data [18]. Further conflating the matter, institutional barriers within insurance organizations can obstruct obtaining a holistic understanding of the customers' willingness to share health data [19]. A study in Canada reported that an overwhelming 79% of physicians and 67% of patients do not want private insurance organizations to have access to health data, even if anonymized or used for research [20], depicting private insurance organizations to be more untrustworthy than the pharmaceutical industry or the government for sharing health data.

Severe distrust is reflected across other health data sharing studies as well, especially when insurance organizations are suspected of profiting from selling or using health data—an unintended consequence of health data sharing [21-23]. When customers perceive that organizations are only using their data for self-interest value creation, customers will not be motivated to share their data and ultimately lose trust in the organization [24]. The social exchange theory (SET) presumes that to get something, you must give something of equivalent perceived value in return [25]. In the context of the insurance industry, this would mean that for customers to share their health data, the insurance organization would need to offer the customers something in return. What this *something* is, however, is unclear.

Objectives

As previous research indicates, there exists a high reluctance of individuals to share health data across the insurance industry. However, at the same time, a need for health data sharing to drive digital health services is increasing as part of the data economy, marking a clear research gap. In response to this gap, we asked the following question: What would customers expect in return for sharing their health data?

We aimed to address this gap by investigating the expectations of insurance customers for sharing their health data. Furthermore, we sought to garner insight into the customers' perceived values that are intrinsic to their expectations when sharing health data. Finally, we explored the conditions that the insurance organizations should facilitate to incentivize customers' health data sharing, all to aid in the design and development of proactive digital health services.

Methods

Case Study

The insurance organization chosen for this case study is one of the largest in Finland. For anonymity purposes, the moniker *Omega* is used. In Finland, health care is mostly decentralized and has three main avenues of distribution: primary public health care; occupational health care, which all employers are obligated to provide; and private services, where voluntary private insurance exists [26]. Omega offers both occupational health care and private health care services, including a Web-based virtual hospital. As business models are transforming in response to an abundance of data [8], Omega is currently making strategic movements away from traditional insurance models to a more proactive one. To be able to provide more proactive health services that help prevent illness or injury to the customer, Omega requires access to customers' health data [27,28]. The control of how any type of data is shared or managed affects the organizations that rely on it. Understanding the conditions under which customers will share data is vital for driving the shift in preventative health care provision. This change makes Omega highly suitable for this study to determine customers' expectations in sharing health data.

Survey Design and Development

The information provided about the survey follows the *Checklist for Reporting Results of Internet E-Surveys* [29]. The survey was designed iteratively. It included a variety of survey utilities

that were both quantitative and qualitative in nature. No randomization of question order was included. There were 5 screens (ie, 4 subject areas and informed consent) and a total of 23 questions across 4 key subject areas: customer characteristics (5 questions), health data (7 questions), value creation (6 questions), and social media (5 questions). The survey was pilot tested with 4 Finnish testers through Web-based testing and feedback and 2 cognitive walkthroughs in English and Swedish. The feedback was summarized and used to iterate the next version of the survey. The content and clarity of questions were improved to ensure uniform meaning across the 3 languages (ie, Finnish, Swedish, and English) in which the survey was available. Native speakers were involved when translating the survey into all 3 languages. A check for completeness was added, as were nonresponse options, the ability to review answers by moving back and forth through the survey, and a progress bar to indicate the percentage of completion at the bottom of the screen.

Data Collection

A probability-based sampling technique was used within the population of Omega's customers as the response rate in previously distributed surveys varied based on topic and was therefore unpredictable. A simple random sampling approach was chosen to provide a high degree of sample probability and minimize probabilistic sampling method biases [30]. The survey itself was randomly allocated to 5000 customers with the only constraints being that respondents had to be Omega's existing customers aged older than 18 years (for consent purposes) and had to have email addresses (so they may be reached). A minimum of 385 responses from Omega's customer base of approximately 1.3 million was required to accommodate our goal of acquiring a 95% CI (descriptive statistics of the same survey [18]). A total of 452 viable responses were received from 5000 customers (response rate: 9.0%) selected as a predetermined sample size. The 452 responses provided us with a rich dataset for a qualitative analysis of free-text answers. In this paper, we did not use the data for the quantitative analysis. The survey was sent on January 30, 2018, and was live for 6 weeks (until March 15, 2018). The time frame was determined by monitoring the number of responses received. Most of the responses were received within the first 15 days.

Data Analysis

For the purposes of this paper, only the customers' characteristics and qualitative health data responses have been discussed. More detailed descriptive statistics about Omega can be found in another paper [18]. A single open-text question from the health data section was analyzed. The question was "If you were to share your health data with Omega, what would you expect in return?" The *expectation* of the customer was asked in a neutral manner so that the customer was not biased toward negative or positive expectations. The output of the text

was intended to be analyzed qualitatively to avoid the pitfall of quasi-content of free text in surveys [31]. A total of 452 translated quotes were extracted from the survey and sequenced into a Microsoft Excel spreadsheet. With the exception of one response, which was in Swedish, all open-text answers were in Finnish and translated to English for analysis. During the analysis, a native Finnish speaker (second author) actively participated to allow for nuances and social contexts to be discussed. Each quote was inductively coded by the first and second authors. No text recognition or automated frequency software was used; all analyses were performed manually to ensure coding familiarity. Once the codes were saturated, similar open-text answers were grouped together through content-driven analysis and then built into themes. Finally, the SET framework was applied (detailed in the following section).

Social Exchange Theory Framework

With classic origins, the SET is a prominent concept used across a variety of disciplines [32] and has been used in more modern information and digital settings, such as knowledge-sharing power plays in interdisciplinary collaborations [33]. Here, we applied Emerson's [25] definition of SET, which is built on the work by Blau in sociology [34] and Homans in social psychology [35]. Emerson [25] advocates that SET is in fact not a theory but rather a framework to gather relevant assumptions in the context of structural functionalism where exchange occurs upon the contingency of perceived values. Other disciplines have previously borrowed SET to generate new theoretical insights [35,36]. However, few studies have connected intangible values such as data with social exchanges between organizations and their customers [37]. The SET framework used in our analysis was adopted from a study by Wu et al [38], which synthesized SET issues to develop a model for partner prerequisites for information sharing in supply chains. They determined that for partners to be willing to share information, the constructs of commitment, power, reciprocity, and trust need to be established. This framework was suitable for adoption into our context as data are entangled with supply and value chains in organizations. More specifically, in the case of an insurance organization strategizing to develop more proactive health services, access to their customers' health data is necessitated, making health data an integral part of their supply chain. Definitions of the 4 SET constructs are summarized in Table 1. Columns 1 and 2 of Table 1 show the SET construct as well as a summary of the definition of the respective construct based on the study by Wu et al [38]. Furthermore, they are contextualized to elucidate how the framework was used in the analysis of the survey data and the delineation of the customer expectation. Column 3 describes how the SET constructs were present in our data. Column 4 gives an example of how the expectations identified in the empirical data were interpreted and can be *read* in the context of the SET constructs.

Table 1. Summary of the social exchange theory framework definitions, application for data analysis, and expectation delineation.

Social exchange theory construct	Summarized definition [38]	Meaning in data analysis	Explanation of expectation
Commitment	Commitment is the inclination to achieve a shared purpose with the ongoing expectation to conditionally maintain the relationship	Activities for health data processing that motivate the formation and stabilization of the relationship between Omega and its customers	Customers expect Omega to commit to certain conditions when sharing their health data
Power	Power is the expectation that the partner with control over the desired resources influences the behavior of the other dependent partner	How Omega uses health data as part of its decision-making process for their customers regarding health-related outcomes	Customers expect Omega to use its power to act in either a positive or a negative manner when health data are being shared
Reciprocity	Reciprocity is the expectation of mutually beneficial outcomes between partners	Entering or perpetuating a relationship that offers perceived advantages for both Omega and its customers when using health data	Customers expect Omega to reciprocate with some form of a benefit in exchange for sharing their health data
Trust	Trust is the ability to establish equal confidence between partners with the expectation that the partners will act in each other's best interest	Indications of what behaviors would increase the customers' confidence in Omega when providing Omega with their health data	To be able to trust Omega with shared data, customers expect Omega to behave in a certain manner

Ethical Considerations

No ethical declaration was required as the survey data used by the researchers were anonymous, no customer under the age of 18 years was asked to participate, and no identifying personal data were requested, only attitudes and opinions. Full disclosure text on the purpose of the study, the affiliation with a Horizon 2020 project, and the expected length of time to take the survey (between 5 and 10 min) were made available to all customers during the commencement of the survey in 3 languages.

Results

Customer Characteristics

Of the 5000 customers who were sent the survey, 452 surveys (response rate: 9.0%) were completed, and all were determined eligible. Table 2 presents the customer characteristics. The results of the customer characteristics were compared with internal demographic models at Omega from January 2018, and the survey sample was confirmed to be representative of Omega's customer base. In addition to customer characteristics are the 3 result categories that did not fit within the SET framework, as the answers in the survey were skipped, customers reported having no expectations, or customers were unwilling to enter the exchange relationship. First, 10.6% (48/452) of customers intentionally skipped the open-text question in the survey, which was mandatory. This was typically denoted with repeated punctuation such as "?????" Second, 12.6% (57 of 452 customers) of the total responses included some variation of "I can't really say." These responses indicated that the customers had no expectations when sharing health

data. Finally, 9.7% (44/452) of customers indicated an unwillingness to share health data, thus abstaining from the exchange relationship. Customers similarly stated, "I wouldn't share my health data" (Customer 254). Their unwillingness to share their data was grounded in the opinion that Omega had no business with their health data or that their health data were private. Furthermore, customers emphasized that their health data were not for sale and that there is no motivator or incentive that would influence their willingness to share health data.

We identified several expectations that customers had in relation to sharing their health data with Omega. Table 3 summarizes the expectations resulting from our analysis, displaying the SET construct (column 1 and first-order), the larger themes that we grouped the expectations into (second-order), the actual expectations (third-order), and the percentages of answers in which this expectation was expressed (column 2). It is possible for an answer to be coded in more than one of the constructs as the responses typically included more than one theme. For example, a passage may contain a desire for something in return for sharing health data and a fear of not receiving compensation for it. This means that the percentages provided in the results (Table 3) are not summative. Customers' responses that expressed an unwillingness to share health data (44/452, 9.7%) and thus an unwillingness to enter the social exchange overall could not be classified with SET and are not included in Table 3. Furthermore, customers' responses that expressed no expectations when sharing health data (57/452, 12.6%) were not included in Table 3.

Next, we describe our findings in detail.

XSL•FO
RenderX

Table 2. Customer characteristics representing the number of customers (N=452) and their reported gender, age, and highest level of education. Customers disengaged from the exchange relationship are also included.

Characteristics	Customers, n (%)
Gender	
Female	227 (50.2)
Male	224 (49.6)
Other	1 (0.2)
Age range (years)	
18-24	11 (2.4)
25-34	43 (9.5)
35-44	87 (19.2)
45-54	99 (21.9)
55-64	104 (23.0)
65-74	86 (19.0)
≥75	22 (4.9)
Highest level of education	
Primary or comprehensive school	32 (7.1)
High school or vocational school	145 (32.1)
Some college credit, no degree	30 (6.6)
Bachelor's degree	146 (32.3)
Master's degree	98 (21.7)
Doctoral degree	1 (0.2)
Customers disengaged from the exchange relationship	
Skipped survey responses	48 (10.6)
No expectations when sharing	57 (12.6)
Unwillingness to share	44 (9.7)
Total customers	149 (33.0)

Table 3. Expectations of insurance customers for sharing health data (N=452).

Social exchange theory construct, theme, expectation	Customers, n (%)
Commitment	
Requirements for data treatment	
Access and control	7 (1.5)
Security	16 (3.5)
Privacy	10 (2.2)
Use	31 (6.9)
Total customers	64 (14.2)
Power	
Negative consequences	
Policy	4 (0.9)
Profiling	18 (4.0)
Total customers	22 (4.9)
Reciprocity	
Compensation	
Discounts	128 (28.3)
Tangible goods	7 (1.5)
Total customers	135 (29.9)
Benefits	
Customer experience	14 (3.1)
General advantages	52 (11.5)
Personalization	63 (13.9)
Total customers	129 (28.5)
Trust	
Empathy	
Compassion	8 (1.8)
Confidence	13 (2.9)
Transparency	13 (2.9)
Total customers	34 (7.5)

Commitment

Commitment is the inclination to maintain the exchange relationship on certain conditions. Customers expected Omega to meet certain requirements for treating their health data. These requirements were access and control, security, privacy, and use.

The Requirement of Access and Control

Customers expected Omega to facilitate access and control to their health data. A total of 1.5% (7/452) of customers conveyed that they require access to their health data in some form and that access to their health data should be strictly regulated and only available to pertinent persons. Customers also valued the presence of accountable actions when access to health data occurs by tracking and logging why the data were accessed and who accessed them; this data log should also be made available to the customers upon request: "I would see who accesses my information" (Customer 396). Having access control to health data is also expected by customers to manage their health data to perform tasks such as correcting any misinformation, managing what information is shared, and deleting information if desired.

The Requirement of Security

Customers expected Omega to act as a data guardian to ensure the security of their health data. Overall, 3.5% (16/452) of the customers specifically mentioned that they would expect Omega to store and secure their data against external attacks. As part of the requirements for data treatment, customers expect technical and legal measures to be in place for storing, processing, and controlling their health data: "Taking care of information security" (Customer 234).

The Requirement of Privacy

A total of 2.2% (10/452) of customers expected Omega to act with a high degree of discretion, much like the belief in those who swear by the Hippocratic oath, that is, only those who are authorized to engage with the data will do so and will also conduct themselves with the level of confidentiality seen in health care professionals. This is illustrated by the following quote: "Similar restrictions and confidentiality obligations in relation to accessing the data, that exists in the healthcare sector..." (Customer 195).

The Requirement of Use

Customers expected Omega to provide details about the use of their health data in the insurance organization. Overall, 6.9% (31/452) of customers expressed that knowing the use or purpose of the data is an important condition for sharing their health data. The sentiment of being informed was echoed repeatedly: "Complete information about what the data will be used for" (Customer 102). This signifies that there is a current lack of or no understanding of what Omega would use health data for. Furthermore, 7 of the 31 customers reporting on their data use expectations explicitly were concerned about the *dark side* of use, that is, they expected assurances that their data were not going to be sold or used by external parties and that the terms of use would not be altered without their consent: "A hundred per cent certainty that the information will only be used for means that have been agreed on together, and Omega won't unilaterally expand the [means of] use" (Customer 201).

Power

Power is the level of influence one partner has by controlling access to desired resources. Omega has the capacity to make positive or negative decisions that affect the health and well-being of customers and to deny or limit financial support. Customers expected that sharing their health data would lead to Omega abusing their power in the form of misleading policy and health profiling.

The Negative Consequence of Policy

Customers expected Omega to purposefully make policy documentation difficult to understand in layman's terms. A total of 0.9% (4/452) of customers expected that Omega would intentionally mislead their customers through unclear policy practices: "You can't expect anything from the kind of insurance company that cheats their customers with nonexistent insurance policies" (Customer 62). All 4 customers were fearful that Omega would leverage their power through fine print in their policies to avoid compensating their customers in the case of an accident or illness.

The Negative Consequence of Profiling

Some customers expected Omega to use the health data they shared against them by profiling their health. However, a surprisingly small number of responses had negative expectations when sharing their health data with Omega. Overall, 18 of 452 customers (response rate: 4.0%) thought that sharing their health data would ultimately lead to Omega using these data to charge the customers more money, reduce compensation, and utilize the health data to exert their power

over the customer: "I wouldn't be ready to share my data because I believe that the issue would always be flipped around to be the customer's fault by using their health data, if compensation was required" (Customer 127). No positively oriented responses were made by customers.

Reciprocity

Reciprocity is the facilitation of mutually beneficial outcomes for partners in an exchange. Reciprocity is connected to the perceived values of each party. Customers provided extensive and detailed expectations for getting something in exchange for sharing their health data, which indicates that they perceive their health data to be valuable enough to merit something in return. Reciprocity is split into 2 themes: compensation, which is finance-related rewards for sharing health data, and benefits, which provide the customer with an advantage they deem worthy.

Discounts as Compensation

Customers expected discounts in exchange for sharing their health data, placing most of their perceived value on decreasing the overall insurance costs. A total of 128 of 452 customers (response rate: 28.3%) expected a form of financial compensation for sharing their health data through discounted insurance payments or services: "Lower insurance payments" (Customer 367). Despite the total number of responses falling under this theme (135/452, 29.9%), there is not much variation between financial compensation expectations. Customers generally indicated that shared health data provide value to Omega and expected to pay less in return. However, some of the customers emphasized that this discount should be substantial, significant, or in the form of "a whopping discount on insurance policies" (Customer 35).

Tangible Goods as Compensation

A total of 1.5% (7/452) of customers expected compensation in the form of tangible goods. Furthermore, 3 customers specifically requested a fitness device for monitoring purposes: "A device with which activity, exercise, and heartbeat is monitored" (Customer 382). The remaining 4 customers expected outright money in exchange as their privacy would be decreased or because Omega would be making additional profits: "Money and a lot of it..." (Customer 369).

Benefits of a Better Customer Experience

A total of 3.1% (14/452) of customers expected Omega to be able to provide a better customer experience if it has additional information (in the form of their health data). The majority of answers were vague as to what better customer services meant and how it would be implemented. However, some did specify what more meaningful services meant for them. Emphasized among them was the desire for standardization when it came to decisions about compensation for insurance claims, so that it would not matter which insurance clerk made the decision as it would always be the same decision, as highlighted in the following quote: "I would expect equality in processing, so that decisions wouldn't change based on who handles them" (Customer 36). This was also coupled with the desire to have a relationship with an employee to create a smoother and more familiar customer experience.

Benefits in the Form of General Advantages

Customers expected general advantages when sharing health data with Omega in a variety of forms. A total of 11.5% (52 of 452 responses) of the open-text answers alluded to an advantage in a vague form. This was typically characterized by a demand for "Perks" (Customer 153) or "Bonuses!!" (Customer 317). The advantage could also take the form of services where customers repeated the expectations for "good service" (Customers 133, 181, 259, and 386). However, there was no indication of what would specifically improve the services offered by Omega for these customers. Furthermore, 4 customers suggested that loyalty benefits would encourage the sharing of their health data with Omega: "The benefits for regular customers should be significantly better" (Customer 99).

Benefits of Personalization

One of the largest saturated results from this survey was that of personalized health services. A total of 13.9% (63/452) of customers expected some form of tailored help with their health in the form of personalized digital health services. The majority of these customers desired help with living a healthier lifestyle or improving their overall health through these customized services, as stated by a customer, "A proactive take on what [kind of things] one should watch out for" (Customer 21). However, many customers took it a step beyond proactive digital health services by also requesting that their insurance plans reflect this personalization as well, such as "[personalized] services and insurance policies 'designed' for me" (Customer 69). Customers also expected that having good health or living a healthy lifestyle would reduce their risk category. Supplemented by health data as proof of their health status and shared with Omega, the overall cost of their health services should decrease as they represent a decreased risk for expense.

Trust

Trust is the ability to establish confidence between partners. Customers expected Omega to show empathy by being compassionate, confident, and transparent with them.

Empathy Through Compassion

Customers expected compassion from Omega as they believed that sharing health data makes the relationship more human. In total, 1.8% (8/452) of customers described scenarios where they wished Omega would be compassionate or more understanding of their customers. Echoed in the open-text answers are expectations that during the turmoil of illness, Omega would strive to ensure that the insurance process is a source of support, not a burden. The following quote illustrates the expectations for emotional support in difficult life situations: "When falling seriously ill, one would get treatment, and wouldn't need to wrestle with the insurance company about finances" (Customer 314).

Empathy Through Confidence

Customers expected Omega to show confidence in its customers. To share health data, customers will be giving Omega the authority to process and control their data. Overall, 2.9% (13/452) of customers expected that sharing health data is aligned with absolute trust between themselves and Omega and that trust should be mutual. Therefore, Omega needs to conduct itself in a manner the customers deem trustworthy because by sharing health data, customers are establishing themselves as trustworthy. This was highlighted in the comments as "unconditional trust" (Customer 27). No specific mechanism is mentioned for what actions would enable mutual trust in practice.

Empathy Through Transparency

Customers expected transparency from Omega about its actions and intentions with health data. Similar to the conditions for data treatment (use), 2.9% (13/452) of customers wanted Omega to be more transparent about use. This was not just specific to health data but rather the organization's processes in general. They expected Omega to provide fair and open treatment to its customers as well as provide clear insurance policies that convey simple and meaningful information: "Fair, transparent, and egalitarian treatment" (Customer 284). Only 2 customers (Customers 195 and 431) suggested how Omega might enact this transparency, both suggesting external validation: "Transparency in the handling. Some outside body to evaluate the insurance company's compensation verdicts" (Customer 431). This indicated a certain level of distrust in Omega, despite Customer 431 purchasing multiple insurance services from Omega.

Discussion

Principal Findings

In this research, we asked what insurance customers would expect in return for sharing their health data. Our contribution was three-fold. First, we identified concrete customer expectations. Second, we determined the perceived values that are intrinsic to health data. Third, we explored the conditions that aid in incentivizing health data sharing in the relationship between an insurance organization and its customer. Our findings contribute to research on health data sharing to aid in the design and development of proactive digital health services [12,13]. Next, we discuss each of our contributions.

First, with regard to the expectations of customers, our findings showed that the majority of customers saw their health data to be valuable and thus wanted something in exchange. We classified these expectations with the help of SET and identified that customers expect the organization to be committed to fulfill certain requirements for data treatment (eg, access and control of health data), to provide the customer with some form of compensation or benefits (eg, personalization of health services) in reciprocity, and to show empathy toward the customers to increase the customers' trust (eg, by being transparent about the use of health data). The majority of Omega's customers expected reciprocal advantages in exchange for sharing their health data, hinging on the contingency that they would receive personalized digital health services that would help to prevent illness in their lifetime. This finding is congruent with the quantitative data of this survey presented in another paper [18]. Customers also expect digital health services to provide personalized and proactive interventions, a trade-off between the concerns of information privacy and the perceived value of digital health services [39]. A surprisingly small number of

customers (22/452, 4.9%) expected a negative outcome when sharing their health data either through profiling or using policy to assuage insurance claims, a consequence of Omega's perceived power.

Second, our results had some indication of the perceived value that is intrinsic to health data. Our findings showed great diversity in the value that customers ascribe to their health data. On the one end of the continuum are those who said that they want nothing in return for sharing their health data, thus ascribing a very low value to health data. On the other end are those who expressed that they would not share their health data with the company under any circumstances, thus ascribing a very high value to the data. In between these extremes, almost a third of the total expectations detailed a form of benefit or compensation as a reciprocal condition in the exchange relationship. Some customers, for example, expected money or tangible goods for sharing their health data. Interestingly, this gives health data a specific value, such as the price of a fitness tracker. For example, the approximate price of the base Fitbit Flex available is US $99, whereas the Fitbit Charge HR is slightly more expensive at US $149; the Apple watch varies greatly in price starting from approximately US $349 [40]. Perhaps the most significant finding was that health data are perceived to enable a shift of power from the insurance company to the customer. The understanding that the insurance company has power over the customer was present in several forms, including how insurance companies make decisions about compensating customer claims. This decision is made based on the information that is available to the insurance company through the claim process and is supported by Finnish legislation [19]. As power and data are intrinsically linked, data can be understood to be a form of currency, and customers understand that sharing their health data requires an exchange of power for this currency. To balance the inequality of power, conditions such as data treatment help shift the asymmetry of power away from the dominance of the insurance industry and more toward the customers. However, this requires strategic movements by the insurance organization to ascribe the customers' value in business models by incentivizing data sharing [9]. As made evident in our analysis, the customer (unconsciously or consciously) is willing to share health data if certain conditions are met to equalize the power in the relationship.

Insurance organizations, similar to most health care stakeholders, require access to health data [9]. As our third contribution, and based on the identified expectations, we were able to ascertain certain conditions that should be met by the insurance organization to incentivize health data sharing and correct negative perceptions held by the customers. Our results showed a certain lack of trust as some customers are skeptical about the trustworthiness of specific processes and policies in insurance contexts. Customers assume that complicated legal jargon and fine print writing in contract agreements will impede them from receiving health insurance coverage. Complex insurance policies act as barriers to a customer's access to health coverage, such as medication or surgery, potentially harming rather than helping customers [41]. Through transparent actions and by establishing a culture of transparency, customers could better understand how their health data will be used [42]. Increased transparency should also focus on decreasing the fear of health data being used to prevent insurance coverage provision or to increase insurance costs for individuals based on their health data.

Previous research shows that patients' access to their Web-based health data empowers them to make more informed decisions and supports a more proactive role in their health [43-45]. Primarily because of legal actions, such as the GDPR, and empowering movements, such as MyData [7,10], individuals now have more digital power than ever before and control how their health data are shared. Our findings indicated that customers are also aware of their rights, as 14.2% (64/452) of customers expressed certain data treatment requirements. Thus, the organization should make explicit how it will treat the customers' health data, who will have access to the data, how the data will be used, and what measures will be taken to protect the data from unwanted access. In addition, some customers either had no expectations when sharing health data (57/452, 12.6%) or preferred to not engage in an exchange relationship with Omega (44/452, 9.7%). This is interesting when comparing our findings with a Canadian primary care practice context, which reported a much higher degree of unwillingness to share data with private insurance companies (79% of physicians and 67% of patients) [20]. However, passive exchange partners can still form relationships without engaging in health data sharing beyond what is necessary, such as initial screening for health insurance. For Omega and the health data economy, this suggests that some individuals will be passive or even indifferent participants, some of whom would need additional stimulation or education to meaningfully be incentivized to engage in digital health solutions [3,12,46].

Limitations

There are some limitations to this study that are worth mentioning. First, a major criticism of SET is that it is most useful in describing post hoc patterns but has limited utility in pinpointing specific a priori predictions [47]. However, we used the SET framework as a way to develop an understanding of how customers perceive value in health data, providing conditions for insurance organizations to make future strategic movements for health data sharing, not predicting how the customer will act. Second, although our survey provided a rich source of data, the response rate could have been higher, given more time and resources to send follow-up fill-in reminders or prompts through other means. Third, despite our best efforts to verify that only the intended 5000 customers who fit our inclusion criteria (ie, an active customer, >18 years of age, and with an email address) participated in the survey, a relatively small number may not have been active customers. The final limitation, also related to our survey, is the response bias, as those who were more willing to fill in the survey may also be more willing to engage with Omega in other areas such as sharing health data, meaning that those who contributed represent a higher percentage of people who might wish to participate in sharing health data. However, as our survey results captured a diverse range of responses, in our view, Omega's customers are well represented.

Conclusions and Future Research

Data sharing is the foundation on which the health data economy can be formed to be mutually beneficial for all health care stakeholders [3]. Our survey found that the majority of customers in an insurance organization are open to exchanging their health data under certain conditions. However, it is apparent that no single offering or exchange for customers can apply as a *one-size-fits-all* solution. Personalization of research streams should aim to cover the scope of need of all customers in digital health services to aid in design and development.

In the case of reciprocal benefits to customers, there is the potential to perpetuate the exchange of health data under the right conditions. The ongoing collection and sharing of data from activity monitors provided by insurance organizations positively affect willingness to share data when customers perceive the benefits of sharing to have a positive impact on their health and wellness [48]. Organizational research should focus on the development of trust between the organization and its customers to improve proactive digital health services so that those services would provide more value to the customers. The power possessed by insurance organizations highlights the negative side of data sharing, and acts of transparency to increase trust could help alleviate this negative valence. Fortuitously, Omega is aware that transparency is a concern [19]. Future research should aim to understand transparent actions and how they can be implemented in a manner that shifts the balance of power for mutually beneficial outcomes.

Authors' Contributions

CG conceived and designed the study; carried out data collection, data analysis, interpretation of findings; and drafted as well as edited the manuscript. OK participated in the data analysis and interpretation of findings and modified the manuscript. KV was involved with study design, interpretation of the results, and modification of the manuscript and provided critical insights. MI participated in the study design and modification of the manuscript and provided critical insights. All authors approved the final version of the manuscript for submission.

Conflicts of Interest

None declared.

References

1. Ekman I, Swedberg K, Taft C, Lindseth A, Norberg A, Brink E, et al. Person-centered care--ready for prime time. Eur J Cardiovasc Nurs 2011 Dec;10(4):248-251. [doi: 10.1016/j.ejcnurse.2011.06.008] [Medline: 21764386]
2. Prilleltensky I. Promoting well-being: time for a paradigm shift in health and human services1. Scand J Public Health Suppl 2005 Oct;66:53-60. [doi: 10.1080/14034950510033381] [Medline: 16214724]
3. Tang C, Plasek JM, Bates DW. Rethinking data sharing at the dawn of a health data economy: a viewpoint. J Med Internet Res 2018 Nov 22;20(11):e11519 [FREE Full text] [doi: 10.2196/11519] [Medline: 30467103]
4. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health 2014 Nov 5;14:1144 [FREE Full text] [doi: 10.1186/1471-2458-14-1144] [Medline: 25377061]
5. Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, et al. Sharing health data for better outcomes on PatientsLikeMe. J Med Internet Res 2010 Jun 14;12(2):e19 [FREE Full text] [doi: 10.2196/jmir.1549] [Medline: 20542858]
6. Korhonen O, Väyrynen K, Isomursu M. Analyzing the role of a telemedicine system in health service personalization. In: Proceedings of the 27th International Conference on Information Systems Development. 2018 Presented at: ISD'18; August 22-24, 2018; Lund, Sweden URL: http://jultika.oulu.fi/files/nbnfi-fe2018102438642.pdf
7. EUR-Lex. 2016. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN [accessed 2019-09-03]
8. Vial G. Understanding digital transformation: a review and a research agenda. J Strateg Inf Syst 2019;28(2):118-144. [doi: 10.1016/j.jsis.2019.01.003]
9. Huhtala T, Pikkarainen M, Saraniemi S. Transformation of the business model in an occupational health care company embedded in an emerging personal data ecosystem: a case study in Finland. Int J Econ Manag Eng 2015;9(10) [FREE Full text] [doi: 10.5281/zenodo.1340252]

10. Iivari M, Pikkarainen M, Koivumäki T. How MyData is Transforming the Business Models for Health Insurance Companies. In: Proceedings of the Working Conference on Virtual Enterprises. Cham: Springer; 2017 Presented at: PRO-VE'17; September 18-20, 2017; Vicenza, Italy p. 323-332. [doi: 10.1007/978-3-319-65151-4_30]

11. Kemppainen L, Koivumäki T, Pikkarainen M, Poikola A. Emerging revenue models for personal data platform operators: when individuals are in control of their data. J Bus Model 2018;6(3):79-105 [FREE Full text] [doi: 10.5278/ojs.jbm.v6i3.2053]

12. Abdelhamid M, Gaia J, Sanders GL. Putting the focus back on the patient: How privacy concerns affect personal health information sharing intentions. J Med Internet Res 2017 Sep 13;19(9):e169 [FREE Full text] [doi: 10.2196/jmir.6877] [Medline: 28903895]

13. Kim J, Kim H, Bell E, Bath T, Paul P, Pham A, et al. Patient perspectives about decisions to share medical data and biospecimens for research. JAMA Netw Open 2019 Aug 2;2(8):e199550 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9550] [Medline: 31433479]

14. Weitzman ER, Kaci L, Mandl KD. Sharing medical data for health research: the early personal health record experience. J Med Internet Res 2010 May 25;12(2):e14 [FREE Full text] [doi: 10.2196/jmir.1356] [Medline: 20501431]

15. Weitzman ER, Kelemen S, Kaci L, Mandl KD. Willingness to share personal health record data for care improvement and public health: a survey of experienced personal health record users. BMC Med Inform Decis Mak 2012 May 22;12:39. [doi: 10.1186/1472-6947-12-39] [Medline: 22616619]

16. Grundstrom C, Väyrynen K, Isomursu M. Dimensions of Accessibility and Interoperability for Electronic Health Records in the Nordic Countries: A Qualitative Evidence Synthesis of Facilitators and Barriers. In: Proceedings of the 22nd Pacific Asia Conference on Information Systems.: AIS Electronic Library (AISeL); 2018 Presented at: PACIS'18; June 26-30, 2018; Yokohama, Japan URL: http://jultika.oulu.fi/files/nbnfi-fe2019090927416.pdf

17. Karampela M, Ouhbi S, Isomursu M. Accessing and Sharing Electronic Personal Health Data. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies. 2019 Presented at: BIOSTEC'19; February 22-24, 2018; Prague, Czech Republic p. 182-189 URL: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0007247301820189 [doi: 10.5220/0007247301820189]

18. Persson M, Grundstrom C, Giunti G. Customer Attitudes Towards Participation and Health Data Sharing in the Digital Transformation of Finnish Insurance. In: Proceedings of the 32nd Bled eConference Humanizing Technology for a Sustainable Society. 2019 Presented at: Bled eConference'19; June 16 - 19, 2019; Bled, Slovenia URL: http://jultika.oulu.fi/files/nbnfi-fe2019090627073.pdf [doi: 10.18690/978-961-286-280-0.43]

19. Grundstrom C, Väyrynen K, Persson M, Isomursu M. Health Data Access Barriers in a Finnish Insurance Company: A Case Study. In: Proceedings of the 27th International Conference on Information Systems Development. 2018 Presented at: ISD'18; August 22-24, 2018; Lund, Sweden URL: http://jultika.oulu.fi/files/nbnfi-fe2019090927454.pdf

20. Perera G, Holbrook A, Thabane L, Foster G, Willison DJ. Views on health information sharing and privacy from primary care practices using electronic medical records. Int J Med Inform 2011 Feb;80(2):94-101. [doi: 10.1016/j.ijmedinf.2010.11.005] [Medline: 21167771]

21. Ancker JS, Witteman HO, Hafeez B, Provencher T, van de Graaf M, Wei E. The invisible work of personal health information management among people with multiple chronic conditions: qualitative interview study among patients and providers. J Med Internet Res 2015 Jun 4;17(6):e137 [FREE Full text] [doi: 10.2196/jmir.4381] [Medline: 26043709]

22. Spencer K, Sanders C, Whitley EA, Lund D, Kaye J, Dixon WG. Patient perspectives on sharing anonymized personal health data using a digital system for dynamic consent and research feedback: a qualitative study. J Med Internet Res 2016 Apr 15;18(4):e66 [FREE Full text] [doi: 10.2196/jmir.5011] [Medline: 27083521]

23. Krahe M, Milligan E, Reilly S. Personal health information in research: perceived risk, trustworthiness and opinions from patients attending a tertiary healthcare facility. J Biomed Inform 2019 Jul;95:103222. [doi: 10.1016/j.jbi.2019.103222] [Medline: 31176040]

24. Saarijärvi H, Grönroos C, Kuusela H. Reverse use of customer data: implications for service-based business models. J Serv Mark 2014;28(7):529-537. [doi: 10.1108/JSM-05-2013-0111]

25. Emerson RM. Social exchange theory. Annu Rev Sociol 1976;2(1):335-362. [doi: 10.1146/annurev.so.02.080176.002003]

26. Tynkkynen L, Chydenius M, Saloranta A, Keskimäki I. Expanding choice of primary care in Finland: much debate but little change so far. Health Policy 2016 Mar;120(3):227-234 [FREE Full text] [doi: 10.1016/j.healthpol.2016.01.015] [Medline: 26819142]

27. Koivumäki T, Pekkarinen S, Lappi M, Väisänen J, Juntunen J, Pikkarainen M. Consumer adoption of future mydata-based preventive eHealth services: an acceptance model and survey study. J Med Internet Res 2017 Dec 22;19(12):e429 [FREE Full text] [doi: 10.2196/jmir.7821] [Medline: 29273574]

28. Pikkarainen MA, Pekkarinen S, Koivumäki T, Huhtala TT. Data as a driver for shaping the practices of a preventive healthcare service delivery network. J Innov Manag 2018;6(1):55-79. [doi: 10.24840/2183-0606_006.001_0005]

29. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]

30. Fricker Jr RD. Sampling methods for online surveys. In: Fielding NG, Lee RM, Blank G, editors. The SAGE Handbook of Online Research Methods. London: SAGE Publications Ltd; 2008:195-216.

XSL•FO
RenderX

31. O'Cathain A, Thomas KJ. 'Any other comments?' Open questions on questionnaires - a bane or a bonus to research? BMC Med Res Methodol 2004 Nov 8;4:25 [FREE Full text] [doi: 10.1186/1471-2288-4-25] [Medline: 15533249]

32. Cropanzano R, Mitchell MS. Social exchange theory: an interdisciplinary review. J Manag 2005;31(6):874-900. [doi: 10.1177/0149206305279602]

33. Park JG, Lee J. Trust and Dependence Model of Knowledge Sharing in Systems Development. In: Proceedings of the 38th International Conference on Information Systems. 2017 Presented at: ICIS'17; December 10-13, 2017; Seoul, Korea.

34. Blau PM. Exchange and Power in Social Life. New York, NY: John Wiley and Sons; 1964.

35. Homans GC. Social behavior as exchange. Am J Sociol 1958 May;63(6):597-606. [doi: 10.1086/222355]

36. Hall H. Borrowed theory: applying exchange theories in information science research. Libr Inf Sci Res 2003;25(3):287-306. [doi: 10.1016/S0740-8188(03)00031-8]

37. Hall H, Widén-Wulff G. Social exchange, social capital information sharing in online environments: lessons from three case studies. In: Proceedings of the international conference for the celebration of the 20th anniversary of Information Studies. 2008 Presented at: USE'08; June 23-25, 2008; Oulu, Finland p. 73-86 URL: https://www.napier.ac.uk/~/media/worktribe/output-233352/social-exchange-social-capital-and-information-sharing-in-online-environments-lessons.pdf

38. Wu I, Chuang CH, Hsu CH. Information sharing and collaborative behaviors in enabling supply chain performance: a social exchange perspective. Int J Prod Econ 2014;148:122-132. [doi: 10.1016/j.ijpe.2013.09.016]

39. Li T, Unger T. Willing to pay for quality personalization? Trade-off between quality and privacy. Eur J Inf Syst 2012;21(6):621-642. [doi: 10.1057/ejis.2012.13]

40. Angulo G, Brogan D, Martini A, Wang J, Clevenger L. Health Features of Activity Trackers: Motivation, Goal Achievement, and Usability. In: The Conference Proceeding of the Michael L. Gargano 14th Annual Research Day. Health Features of Activity Trackers: Motivation, Goal Achievement, and Usability. Conf Proceeding Michael L Gargano 14th Annu Res Day. Seidenberg School of CSIS, Pace University; 2016 Presented at: CSIS'16; May 6th, 2016; Pace University, New York p. A5-1-A5-8 URL: http://csis.pace.edu/~ctappert/srd2016/2016PDF/a5.pdf

41. Simacek KF, Ko JJ, Moreton D, Varga S, Johnson K, Katic BJ. The impact of disease-modifying therapy access barriers on people with multiple sclerosis: mixed-methods study. J Med Internet Res 2018 Oct 30;20(10):e11168 [FREE Full text] [doi: 10.2196/11168] [Medline: 30377144]

42. Smith RJ, Grande D, Merchant RM. Transforming scientific inquiry: Tapping into digital data by building a culture of transparency and consent. Acad Med 2016 Apr;91(4):469-472 [FREE Full text] [doi: 10.1097/ACM.0000000000001022] [Medline: 26630607]

43. Denneson LM, Cromer R, Williams HB, Pisciotta M, Dobscha SK. A qualitative analysis of how online access to mental health notes is changing clinician perceptions of power and the therapeutic relationship. J Med Internet Res 2017 Jun 14;19(6):e208 [FREE Full text] [doi: 10.2196/jmir.6915] [Medline: 28615152]

44. Bergevi C, Lendahls L, Crang-Svalenius E, Oscarsson MG. From passive passenger to participating co-pilot - Pregnant women's expectations of being able to access their online journal from antenatal care. Sex Reprod Healthc 2018 Mar;15:35-39. [doi: 10.1016/j.srhc.2017.11.009] [Medline: 29389499]

45. Hägglund M, Scandurra I. Patients' online access to electronic health records: current status and experiences from the implementation in Sweden. Stud Health Technol Inform 2017;245:723-727. [Medline: 29295193]

46. Tully MP, Bozentko K, Clement S, Hunn A, Hassan L, Norris R, et al. Investigating the extent to which patients should control access to patient records for research: a deliberative process using citizens' juries. J Med Internet Res 2018 Mar 28;20(3):e112 [FREE Full text] [doi: 10.2196/jmir.7763] [Medline: 29592847]

47. Cropanzano R, Anthony EL, Daniels SR, Hall AV. Social exchange theory: a critical review with theoretical remedies. Acad Manag Annals 2017;11(1):479-516. [doi: 10.5465/annals.2015.0099]

48. Puri A, Kim B, Nguyen O, Stolee P, Tung J, Lee J. User acceptance of wrist-worn activity trackers among community-dwelling older adults: mixed method study. JMIR Mhealth Uhealth 2017 Nov 15;5(11):e173 [FREE Full text] [doi: 10.2196/mhealth.8211] [Medline: 29141837]

Abbreviations

GDPR: General Data Protection Regulation
SET: social exchange theory

XSL•FO
RenderX

<u>Original Paper</u>

Performance of Computer-Aided Diagnosis in Ultrasonography for Detection of Breast Lesions Less and More Than 2 cm: Prospective Comparative Study

Liang Yongping[1*], MD; Ping Zhou[1], MD; Zhang Juan[1*], MD; Zhao Yongfeng[1], MD; Wengang Liu[1], MD; Yifan Shi[1], MD

The Third Xiangya Hospital, Central South University, Changsha, Hunan, China
[*]these authors contributed equally

Corresponding Author:
Ping Zhou, MD
The Third Xiangya Hospital
Central South University
138 Tongzipo Road, Hexiyuelu District
Changsha, Hunan, 410013
China
Phone: 86 13974809881
Fax: 86 88618403
Email: zhouping1000@hotmail.com

Abstract

Background: Computer-aided diagnosis (CAD) is used as an aid tool by radiologists on breast lesion diagnosis in ultrasonography. Previous studies demonstrated that CAD can improve the diagnosis performance of radiologists. However, the optimal use of CAD on breast lesions according to size (below or above 2 cm) has not been assessed.

Objective: The aim of this study was to compare the performance of different radiologists using CAD to detect breast tumors less and more than 2 cm in size.

Methods: We prospectively enrolled 261 consecutive patients (mean age 43 years; age range 17-70 years), including 398 lesions (148 lesions>2 cm, 79 malignant and 69 benign; 250 lesions≤2 cm, 71 malignant and 179 benign) with breast mass as the prominent symptom. One novice radiologist with 1 year of ultrasonography experience and one experienced radiologist with 5 years of ultrasonography experience were each assigned to read the ultrasonography images without CAD, and then again at a second reading while applying the CAD S-Detect. We then compared the diagnostic performance of the readers in the two readings (without and combined with CAD) with breast imaging. The McNemar test for paired data was used for statistical analysis.

Results: For the novice reader, the area under the receiver operating characteristic curve (AUC) improved from 0.74 (95% CI 0.67-0.82) from the without-CAD mode to 0.88 (95% CI 0.83-0.93; *P*<.001) at the combined-CAD mode in lesions≤2 cm. For the experienced reader, the AUC improved from 0.84 (95% CI 0.77-0.90) to 0.90 (95% CI 0.86-0.94; *P*=.002). In lesions>2 cm, the AUC moderately decreased from 0.81 to 0.80 (novice reader) and from 0.90 to 0.82 (experienced reader). The sensitivity of the novice and experienced reader in lesions≤2 cm improved from 61.97% and 73.23% at the without-CAD mode to 90.14% and 97.18% (both *P*<.001) at the combined-CAD mode, respectively.

Conclusions: S-Detect is a feasible diagnostic tool that can improve the sensitivity for both novice and experienced readers, while also improving the negative predictive value and AUC for lesions≤2 cm, demonstrating important application value in the clinical diagnosis of breast cancer.

Trial Registration: Chinese Clinical Trial Registry ChiCTR1800019649; http://www.chictr.org.cn/showprojen.aspx?proj=33094

KEYWORDS

XSL·FO
RenderX

Introduction

Breast cancer is one of the most common cancers in women, and the second leading cause of cancer-related mortality worldwide [1,2]. Early diagnosis of breast cancer can increase the treatment options and survival rate of patients [3], in which breast ultrasound plays an important role in detecting breast cancer. Operator experience-dependence remains the main limitation in ultrasound-based diagnosis [4,5]. S-detect is a recently developed computer-assisted diagnosis (CAD) system for breast cancer, which is based on the Breast Imaging Reporting and Data System (BI-RADS) lexicon and classification [6]. Many studies have reported that S-detect has potential to become a novel diagnostic tool for radiologists [7-10]. However, no study has evaluated the diagnosis performance of CAD in breast lesions with respect to size (less and more than 2 cm). Therefore, the purpose of this study was to compare the performance of detecting breast cancer using CAD between radiologists with different levels of experience for lesions greater and less than 2 cm in size.

Methods

Patient Selection

We prospectively enrolled 261 patients who presented with a total of 398 lesions from November 2018 to May 2019. All patients underwent ultrasound before surgery. The mean age of the examined patients was 43.11 (SD 12.55) years (range 17-70 years). The diameter of lesions ranged from 0.26 to 9.50 cm, with a mean diameter of 1.92 (SD 1.26) cm. All 398 lesions were examined after surgery to confirm their pathological type. This prospective study was approved by the Institutional Review Board of Third Xiangya Hospital. Informed consent was obtained from all patients.

The inclusion criteria were follows: patients aged 17-70 years with breast tumor requiring surgery. The exclusion criteria were a history of neoadjuvant chemotherapy or endocrine therapy before surgery, lesions punctured by core-needle biopsy or Mammotome System, breast equipped with a prosthesis, lesions unclear as displayed by ultrasound, and patients unwilling to take part in the study.

Ultrasound Image Acquisition

All images were obtained with an RS80A ultrasound system (Samsung Medison Co Ltd, Seoul, Korea) with a 5-13–MHz bandwidth linear transducer. All ultrasound examinations were performed by an independent radiologist with 3 years of experience. Typical images of the tumor in longitudinal and transverse planes were stored in the ultrasound system.

Computer-Assisted Diagnostic System

Our CAD system (S-Detect) extracts features using an integration of artificial neural network classifiers internally installed in the ultrasound equipment (RS80A). The sensitivity of the instrument can be adjusted, with greater sensitivity yielding a higher potential rate of false-positive findings. We chose the default setting. To test the reproducibility of CAD marks with the same image, we randomly selected 20 of 398

(5.0%) examinations, which were sent through the CAD system three times, and the results showed that the markings were consistent in all images.

In S-Detect, the cursor on the center of the lesion was identified, and a region of interest was drawn along the border of the mass automatically by the ultrasound system. The ultrasound features of the lesion were analyzed according to the BI-RADS lexicon, and the final assessment classifications were automatically performed by the ultrasound system. If the borderline was considered inaccurate in any area of the tumor, it was manually edited to achieve the optimum fitness. In the S-Detect system, the final assessment classification was divided into "possibly benign" or "possibly malignant."

Diagnostic Criteria

According to the fifth version of BI-RADS, the radiologists classified the lesion from BI-RADS category 3 to BI-RADS category 5. BI-RADS category 4 was further subdivided into category 4A, 4B, and 4C. Category 3 is considered probably benign (<2% likelihood of malignancy) and categories 4A, 4B, 4C range from low to high suspicion (2-10%, 10-50%, 50-95% likelihood of malignancy, respectively). Category 5 indicates a high malignancy rate (>95% likelihood of malignancy). The malignant signs in breast ultrasound imaging included irregular shape, antiparallel orientation, noncircumscribed margin, microcalcification, acoustic halo, posterior shadowing, and abnormalities of the surrounding tissue. No definitive malignant sign is assigned to category 3; one, two, and three malignant signs are assigned to category 4A, 4B, and 4C, respectively; and more than four malignant signs is assigned to category 5. Accordingly, category 3 and 4A lesions were regarded as benign, and category 4B, 4C, and 5 lesions were regarded as malignant [11,12].

For assessments of the combination of ultrasound and the CAD system, we took longitudinal and transverse planes of the tumor for CAD. If one plane indicated "possibly malignant," it was considered a positive outcome, and the BI-RADS category diagnosis was increased by one level (ie, 3 to 4A, 4A to 4B, 4B to 4C, 4C to 5). If both planes indicated "possibly benign," it was considered a negative outcome, and the BI-RADS category diagnosis was decreased by one level (ie, 5 to 4C, 4C to 4B, 4B to 4A, 4A to 3) [13].

Readers, Reading Modes, and Training

Two readers were involved in the study: a novice reader with 1 year of ultrasound experience and an experienced reader with 5 years of ultrasound experience. Both readers were trained on the reading procedures with 20 ultrasound images that were not part of the study set, 10 of which were read in without-CAD mode. The other 10 images were assessed in combined-CAD mode, in which the readers first read the ultrasound images without CAD and then combined the indications of CAD marks to make the final decision.

Both readers reviewed every examination at each reading mode independently and were blinded to any information about the patients, including age, manifestation of symptoms, and previous radiology reports. The readers were asked to read for at least 2

hours a day to simulate the typical process of batch reading in such examinations.

Statistical Analysis

Statistical evaluation was performed using SPSS software (SPSS for Windows 19.0, SPSS Inc, Chicago, IL, USA). Taking the pathology results as the gold standard, we analyzed the diagnostic sensitivity, specificity, and area under the receiving operating characteristic curve values (AUCs) in without-CAD mode and combined-CAD mode [14]. The combined-CAD mode and without-CAD mode diagnostic parameters were compared using the McNemar test (sensitivity, specificity, positive predictive value [PPV], negative predictive value [NPV], accuracy) for match-paired data. We used the Hanley and McNeil method to analyze the differences between pairs of AUCs. For all statistical tests, $P<.05$ was considered to indicate statistical significance.

Results

Basic Characteristics of Lesions

Patient and lesion characteristics on the basis of lesion size are summarized in Table 1. Of the 398 breast lesions in the 261 patients included in this study, 250 (62.8%) were ≤2 cm and 148 (37.2%) were >2 cm. The mean sizes for all lesions, malignant lesions, and benign lesions at ultrasound were similar and close to 2 cm, with benign lesions being the smallest (1.73 cm) and malignant lesions being the largest (2.22 cm).

Table 1. Characteristics of patients and lesions.

Characteristic	All lesions (n=398)	Lesions≤2 cm (n=250)	Lesions>2 cm (n=148)
Patient age (years)			
Mean (SD)	43.10 (12.57)	43.62 (11.875)	42.22 (13.66)
Median (range)	45 (17-70)	45.0 (17-70)	44.5 (17-70)
Size of all lesions (cm)			
Mean (SD)	1.92 (1.26)	1.1629 (0.42)	3.1876 (1.19)
Median (range)	1.6 (0.26-9.5)	1.1 (0.26-2.0)	2.8 (2.1-9.5)
Size of malignant lesions (cm)			
Mean (SD)	2.22 (1.08)	1.331 (0.42)	3.02 (0.82)
Median (range)	2.11 (0.26-6.2)	1.30 (0.26-2.0)	2.8 (2.1-6.2)
Size of benign lesions (cm)			
Mean (SD)	1.73 (1.33)	1.10 (0.40)	3.38 (1.50)
Median (range)	1.3 (0.4-9.5)	1.0 (0.4-2.0)	2.9 (2.1-9.5)
Histologic type of malignant lesions, n (%)			
Total	150 (37.7)	71 (47.3)	79 (52.7)
Intraductal carcinoma in situ	5 (3.3)	5 (7.0)	0 (0.0)
Invasive lobular carcinoma	11 (7.3)	10 (14.1)	1 (1.3)
Mucinous adenocarcinoma	2 (1.3)	2 (2.8)	0 (0.0)
Medullary carcinoma	2 (1.3)	1 (1.4)	1 (1.3)
Invasive ductal carcinoma	130 (86.7)	53 (74.6)	77 (97.5)
Histological type of benign lesions, n (%)			
Total	248 (62.3)	179 (72.2)	69 (27.8)
Intraductal papilloma	29 (11.7)	29 (16.2)	0 (0.0)
Granulomatous mastitis	5 (2.0)	1 (0.6)	4 (5.8)
Fibroma	171 (69.0)	110 (61.5)	61 (88.4)
Hyperplasia-induced lesions	42 (16.9)	38 (21.2)	4 (5.8)
Scar tissue	1 (0.4)	1 (0.6)	0 (0.0)

Reader Performance

In all lesions, the AUCs of the reading improved at combined-CAD mode compared to those of the without-CAD mode for both the novice and experienced reader (Table 2, Figure 1). For the novice reader, the improvement in AUCs was significant between the without-CAD and combined-CAD modes ($Z=4.90$, $P<.001$), whereas there was no significant difference in AUCs between modes for the experienced reader ($Z=1.06$, $P=.29$).

In subgroup analysis, for lesions≤2 cm, the AUCs of the reading improved significantly in combined-CAD mode for both the novice and experienced readers. However, in lesions>2 cm, there were no significant differences in AUCs between two reading modes for both the novice and experienced readers (Table 2).

When a BI-RADS category 4A threshold was used, the sensitivity and NPV improved at the combined-CAD mode compared with that at the without-CAD mode for both the novice reader and experienced reader in all lesions and subgroup analyses (Table 2). However, in lesions≤2 cm, there were no

significant differences between without-CAD and combined-CAD modes for the novice reader with respect to specificity, PPV, and accuracy. By contrast, significant differences were observed for the experienced reader in specificity and PPV, whereas there was no significant difference in accuracy. In lesions>2 cm, there was a significant decrease in specificity and a significant increase in NPV between without-CAD and combined-CAD modes for both readers, and there was a significant decrease in PPV for only the experienced reader. There was a moderate reduction in accuracy between the without-CAD and combined-CAD modes for both readers, and in PPV for the novice reader (Table 2).

Table 2. Diagnostic performance of the readers in two reading modes with a Breast Imaging Reporting and Data System Category 4A threshold.

| Lesions	Novice Reader					Expert Reader				
	Without CAD[a]		Combined with CAD		*P* value	Without CAD		Combined with CAD		*P* value
	+	−	+	−		+	−	+	−	
Lesions≤2 cm										
Pathology										
+	44	22	64	23		52	9	69	33	
−	27	157	7	156		19	170	2	146	
Sensitivity[b]	61.97		90.14		<.001	73.24		97.18		<.001
Specificity[b]	87.71		87.15		.83	94.97		81.56		.004
PPV[c]	66.66		73.56		.22	85.25		67.65		.005
NPV[d]	85.33		95.71		.008	89.95		98.65		.005
Accuracy	80.40		88.00		.12	88.8		86		.52
AUC[e] (95% CI)	0.74 (0.67-0.82)		0.88 (0.83-0.93)		<.001	0.84 (0.77-0.90)		0.90 (0.86-0.94)		.002
Lesions>2 cm										
Pathology										
+	61	11	79	28		67	4	79	23	
−	18	58	0	41		12	65	0	46	
Sensitivity	77.22		100		<.001	86.67		100		<.001
Specificity	84.06		59.42		<.001	96.72		66.66		<.001
PPV	84.72		73.83		.05	96.30		77.45		<.001
NPV	76.32		100		<.001	88.06		100		<.001
Accuracy	80.41		81.08		.86	91.74		84.46		.13
AUC (95% CI)	0.81 (0.73-0.88)		0.80 (0.72-0.87)		.81	0.90 (0.84-0.95)		0.83 (0.76-0.91)		.03
All lesions										
Pathology										
+	105	33	143	51		119	13	148	52	
−	45	215	7	197		31	235	2	96	
Sensitivity	70		95.33		<.001	79.33		98.66		<.001
Specificity	86.69		79.43		.13	94.75		79.03		.001
PPV	76.08		73.71		.74	90.15		74.00		.003
NPV	82.69		96.57		.001	88.34		98.98		.002
Accuracy	80.40		85.43		.35	88.94		86.43		.52
AUC (95% CI)	0.78 (0.73-0.83)		0.87 (0.84-0.91)		<.001	0.87 (0.83-0.91)		0.89 (0.85-0.92)		.29

[a]CAD: computer-aided diagnosis.

[b]Breast Imaging Reporting and Data System assessment categories 4B, 4C, and 5 were considered positive for cancer for the calculation of sensitivity and specificity.

[c]PPV: positive predictive value.

[d]NPV: negative predictive value.

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

Figure 1. Comparison of receiver operating characteristic curves in lesions <2 cm, lesions >2 cm, and lesions in all size.
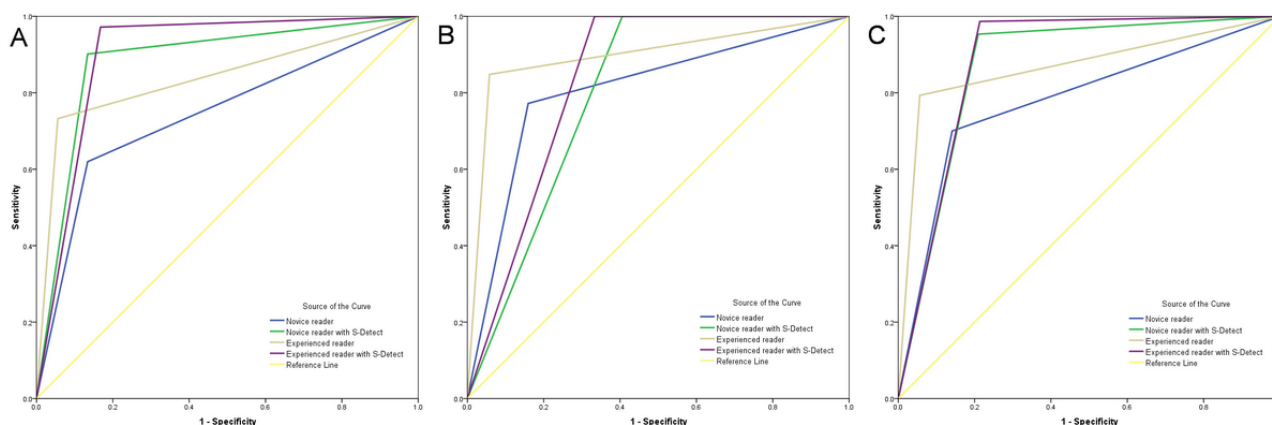
Management of Diagnostic Feature Decision Change

At the combined-CAD mode, the management of diagnostic feature decision changes compared with the without-CAD mode was 164/398 (41.2%) of all ultrasound images for the two readers. For the novice reader, with CAD support, 38/150 (25.3%) of the malignant lesions (20 lesions≤2 cm, 18 lesions>2 cm) were correctly upgraded from category 4A to 4B, whereas none (0%) of the malignant lesions was incorrectly changed from 4B to 4A. In addition, 18/248 (7.3%) benign lesions (12 lesions≤2 cm, 6 lesions>2 cm) were correctly downgraded from category 4B to 4A; however, 36/248 (14.5%) benign lesions (11 lesions≤2 cm, 24 lesions>2 cm) were incorrectly changed from category 4A to 4B.

For the experienced reader, 30/150 (20.0%) malignant lesions (18 lesions≤2 cm, 12 lesions>2 cm) were correctly changed from category 4A to 4B, and only 1 of 79 (1%) malignant lesions (≤2 cm) was incorrectly changed from category 4B to 4A. In addition, 4/248 (1.6%) benign lesions (3 lesions≤2 cm, 1 lesion>2 cm) were correctly downgraded from category 4B to 4A, whereas 38/248 (15.3%) benign lesions (3 lesions≤2 cm, 20 lesions>2 cm) were incorrectly changed from category 4A to 4B.

Discussion

Principal Findings

CAD systems have been recently applied to improve diagnostic performance in breast ultrasonography. S-Detect is a CAD system based on a neural network learning algorithm [7], which applies a novel feature extraction technique and vector machine classifier that categorizes breast masses into benign or malignant depending on the suggested feature based on the BI-RADS lexicon [15]. Choi et al [10] recently reported that both experienced and inexperienced readers had significantly higher specificity and AUCs in reading ultrasounds in combination with S-Detect, and the inexperienced reader also showed significant improvement in sensitivity. However, the diagnosis of breast lesions of different sizes is one of the most difficult challenges in clinical practice [16,17]. Radiologists with different levels of experience typically perform breast ultrasound, and thus the usefulness of S-Detect may be different according to experience. For example, radiologists with less

experience may have a greater benefit in using S-Detect for the diagnosis of small breast lesions.

In our study, when combining ultrasound reading with S-Detect, both the experienced and novice readers showed significantly higher sensitivity and NPV compared to those obtained without S-Detect, which is in line with the findings of the previous studies for CAD systems mentioned above. In addition, 38/150 (25.3%) and 30/150 (20.0%) breast cancers initially assessed as category 4A by the novice and experienced readers were categorized as probably malignant by S-Detect, regardless of size. Combining the results of S-Detect led to significant improvements in AUCs for both readers in lesions<2 cm. However, in lesions>2 cm, the combination of S-Detect did not confer improvements in accuracy and AUC for either reader.

Our results suggest that S-Detect could be used as an additional tool with breast ultrasound regardless of the experience of the reader, and may help to reduce the misdiagnosis ratio of early-stage breast cancer. Although the sensitivity, NPV, and AUCs were improved, there was no significant improvement in the accuracy of the readers when using S-Detect compared to that obtained by the ultrasound reading alone. This may be due to the fact that both readers already showed high AUC values with ultrasound alone, and therefore there was minimal room for improvement.

Strengths and Prospects

Our results showed that readers with less experience may benefit more by using S-Detect in detection of smaller breast lesions. Several studies have reported the application of different types of CAD to breast ultrasound [6,18,19]. Overall, these studies showed that the CAD systems promoted the diagnostic performance of breast ultrasound, especially specificity and accuracy. Shen et al [18] argued that CAD systems could be helpful in evaluating fuzzy category 4 lesions. Wang et al [19] suggested that combining CAD with ultrasound was more helpful for inexperienced radiologists than for experienced radiologists owing to greater improvement in the diagnostic performances observed in the inexperienced group. In our study, the sensitivity, NPV, and AUCs of both readers were improved, supporting the idea that S-Detect can reliably provide a second view that can be referred to by readers. High sensitivity is a remarkable superiority of S-Detect, and similar results were reported in some previous studies [20,21]. Compared to these

previous studies, there was a relatively smaller proportion of benign lesions in our study and the mean size of lesions in our study was larger. In addition, all patients had a breast mass as the prominent symptom, which may explain the different results. Moreover, since S-Detect provides the final assessment in a dichotomized form of possibly benign and possibly malignant, this factor may have also affected the accuracy of readers in the combined-CAD mode.

This result is encouraging for clinical breast cancer screening, as breast cancer is a highly aggressive disease with multiple pathological subtypes, including those associated with higher rates of metastases and poorer survival rates [22]. Thus, it is important to detect cancer early to reduce the mortality rate [23]. In addition, S-Detect is a user-friendly and concise program that is integrated in an ultrasound machine to enable obtaining a terse result for radiologists immediately during real-time ultrasonography, which can easily be applied to routine work. However, it is not recommended to apply CAD alone or use it as a replacement for a radiologist in the diagnosis of breast lesions, especially for tumors>2 cm, which is consistent with the results of Kim et al [13]. As one example from this study, a fibroadenoma lesion with a size of 2.94×1.76 cm (Figure 2A) that showed an unclear margin and a large lobulated shape was misdiagnosed as malignant by S-Detect, and was inversely excluded by the radiologist after combining the results with information on the patient's history. In another example, a lesion of invasive ductal carcinoma with a size of 3.09×1.36 cm (Figure 2B) showing a clear border and microcalcification was classified

as BI-RADS category 4B by conventional ultrasound, whereas S-Detect diagnosed this lesion as benign. Further investigation along with technical progress are anticipated to lead to the development of a more sophisticated algorithm using the multiple-planes assessment BI-RADS ultrasonographic categories.

Likewise, ultrasound scanning is a real-time and multi-angled imaging method, which can observe the lesion from different planes to collect the imaging features such as the internal situation, relation of the lesion with surrounding tissues, and the blood supply model, along with patient history and other available information. Therefore, more image data and clinical information can be obtained with ultrasound than with CAD. Consequently, in lesions≤2 cm, the combination of S-Detect and ultrasound allows for the weaknesses of each method to be counteracted by the strengths of the other, which could assist both novice and experienced readers in making a more accurate final diagnosis. As one example from this study, an invasive ductal carcinoma lesion with a size of 1.75×1.56 cm (Figure 3A) that showed an unclear margin, irregular shape, and microcalcification was correctly diagnosed as malignant by S-Detect and was classified as BI-RADS category 4C by both readers. In another example, a lesion of fibroadenoma with a size of 1.58×1.10 cm (Figure 3B) showing a clear border and regular shape was classified as BI-RADS category 3 by conventional ultrasound and was correctly diagnosed as benign by S-Detect.

Figure 2. A malignant and benign lesion > 2 cm misdiagnosed by S-Detect.
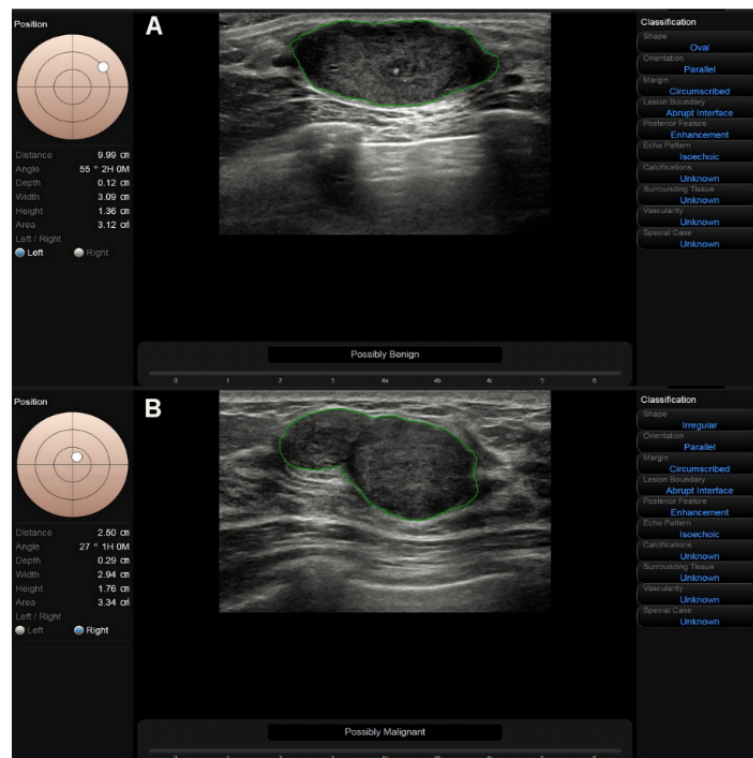
Figure 3. A malignant and benign lesion < 2 cm diagnosed correctly by S-Detect.
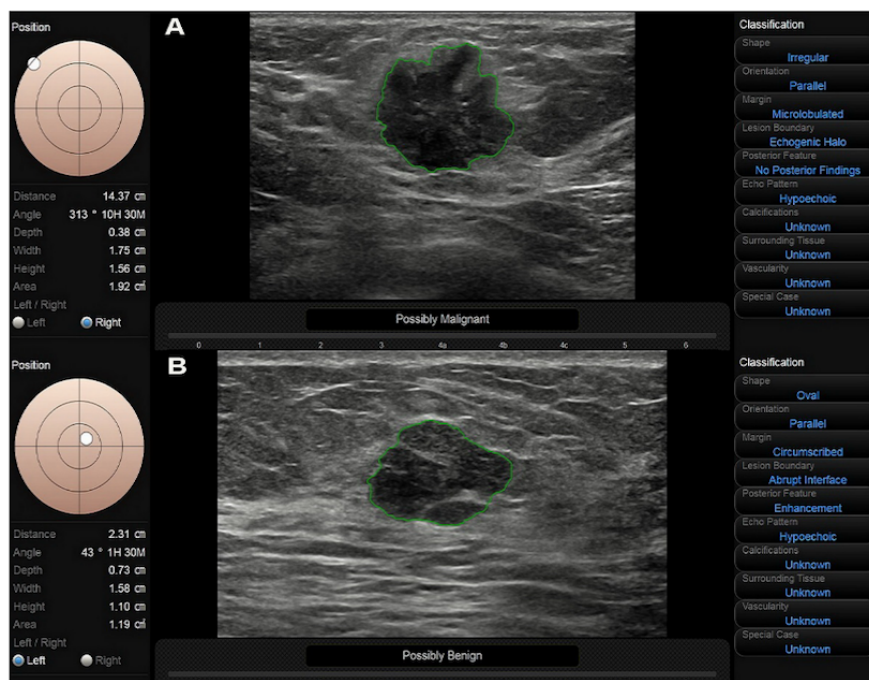
Limitations

There are several limitations of this study. First, a relatively small number of cases were included (N=261). Second, the presentation of calcifications was not included in the analysis, owing to the limited ability of S-Detect to detect microcalcifications [24]. Third, some small nodules (around 1 cm) classified as BI-RADS category 3 that were not surgically removed were excluded from the study, which may have affected the results. Fourth, only two representative still images of breast masses stored for analysis were chosen during image analysis by the radiologists and S-Detect, which may have caused variability in selecting images of representative planes. Fifth, the criterion of the size of lesions for comparison was set to 2 cm; thus, further studies using other stratifications with a larger number of samples may be warranted. Sixth, both readers had relatively minimal experience as breast imagers. In China, the specialty of breast imaging is somewhat new, and the staff in this field tend to be younger compared with staff of other imaging specialties. Therefore, these factors may have had a slight influence on our results.

Conclusion

In conclusion, S-Detect is a clinically feasible diagnostic tool that can improve the sensitivity of breast ultrasonography, in addition to improving the NPV and AUC for lesions≤2 cm, with important application value in the clinical diagnosis of breast cancer.

Acknowledgments

Conflicts of Interest

None declared.

References

1. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. CA Cancer J Clin 2017 Jan;67(1):7-30. [doi: 10.3322/caac.21387] [Medline: 28055103]
2. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. CA Cancer J Clin 2016;66(2):115-132. [doi: 10.3322/caac.21338] [Medline: 26808342]
3. Brinkley D, Haybrittle JL. The curability of breast cancer. Lancet 1975 Jul 19;2(7925):95-97. [doi: 10.1016/s0140-6736(75)90003-3] [Medline: 49738]
4. Takahashi R, Kajikawa Y. Computer-aided diagnosis: A survey with bibliometric analysis. Int J Med Inform 2017 May;101:58-67. [doi: 10.1016/j.ijmedinf.2017.02.004] [Medline: 28347448]
5. Cho E, Kim E, Song MK, Yoon JH. Application of Computer-Aided Diagnosis on Breast Ultrasonography: Evaluation of Diagnostic Performances and Agreement of Radiologists According to Different Levels of Experience. J Ultrasound Med 2018 Jan;37(1):209-216. [doi: 10.1002/jum.14332] [Medline: 28762552]

6.   Kim K, Song MK, Kim E, Yoon JH. Clinical application of S-Detect to breast masses on ultrasonography: a study evaluating
     the diagnostic performance and agreement with a dedicated breast radiologist. Ultrasonography 2017 Jan;36(1):3-9. [doi:
     10.14366/usg.16012] [Medline: 27184656]

7.   Komeda Y, Handa H, Watanabe T, Nomura T, Kitahashi M, Sakurai T, et al. Computer-Aided Diagnosis Based on
     Convolutional Neural Network System for Colorectal Polyp Classification: Preliminary Experience. Oncology 2017;93
     Suppl 1:30-34. [doi: 10.1159/000481227] [Medline: 29258081]

8.   Di Segni M, de Soccio V, Cantisani V, Bonito G, Rubini A, Di Segni G, et al. Automated classification of focal breast
     lesions according to S-detect: validation and role as a clinical and teaching tool. J Ultrasound 2018 Jun;21(2):105-118.
     [doi: 10.1007/s40477-018-0297-2] [Medline: 29681007]

9.   Takahashi R, Kajikawa Y. Computer-aided diagnosis: A survey with bibliometric analysis. Int J Med Inform 2017
     May;101:58-67. [doi: 10.1016/j.ijmedinf.2017.02.004] [Medline: 28347448]

10. Choi J, Kang BJ, Baek JE, Lee HS, Kim SH. Application of computer-aided diagnosis in breast ultrasound interpretation:
     improvements in diagnostic performance according to reader experience. Ultrasonography 2018 Jul;37(3):217-225. [doi:
     10.14366/usg.17046] [Medline: 28992680]

11. Liu D, Ba Z, Ni X, Wang L, Yu D, Ma X. Apparent Diffusion Coefficient to Subdivide Breast Imaging Reporting and Data
     System Magnetic Resonance Imaging (BI-RADS-MRI) Category 4 Lesions. Med Sci Monit 2018 Apr 12;24:2180-2188.
     [doi: 10.12659/msm.907000] [Medline: 29644993]

12. Spinelli Varella MA, Teixeira da Cruz J, Rauber A, Varella IS, Fleck JF, Moreira LF. Role of BI-RADS Ultrasound
     Subcategories 4A to 4C in Predicting Breast Cancer. Clin Breast Cancer 2018 Aug;18(4):e507-e511. [doi:
     10.1016/j.clbc.2017.09.002] [Medline: 29066139]

13. Wang M, Yang Z, Liu C, Yan J, Zhang W, Sun J, et al. Differential Diagnosis of Breast Category 3 and 4 Nodules Through
     BI-RADS Classification in Conjunction with Shear Wave Elastography. Ultrasound Med Biol 2017 Mar;43(3):601-606.
     [doi: 10.1016/j.ultrasmedbio.2016.10.004] [Medline: 27988221]

14. Ma H, Bandos AI, Gur D. On the use of partial area under the ROC curve for comparison of two diagnostic tests. Biom J
     2015 Mar;57(2):304-320. [doi: 10.1002/bimj.201400023] [Medline: 25537143]

15. Payne S, Large S, Jarrett N, Turner P. Written information given to patients and families by palliative care units: a national
     survey. Lancet 2000 May 20;355(9217):1792. [doi: 10.1016/S0140-6736(00)02272-8] [Medline: 10832835]

16. Migowski A. [Early detection of breast cancer and the interpretation of results of survival studies]. Cien Saude Colet 2015
     Apr;20(4):1309 [FREE Full text] [doi: 10.1590/1413-81232015204.17772014] [Medline: 25923642]

17. Guo R, Lu G, Qin B, Fei B. Ultrasound Imaging Technologies for Breast Cancer Detection and Management: A Review.
     Ultrasound Med Biol 2018 Jan;44(1):37-70 [FREE Full text] [doi: 10.1016/j.ultrasmedbio.2017.09.012] [Medline: 29107353]

18. Shen W, Chang R, Moon WK. Computer aided classification system for breast ultrasound based on Breast Imaging Reporting
     and Data System (BI-RADS). Ultrasound Med Biol 2007 Nov;33(11):1688-1698. [doi: 10.1016/j.ultrasmedbio.2007.05.016]
     [Medline: 17681678]

19. Wang Y, Jiang S, Wang H, Guo YH, Liu B, Hou Y, et al. CAD algorithms for solid breast masses discrimination: evaluation
     of the accuracy and interobserver variability. Ultrasound Med Biol 2010 Aug;36(8):1273-1281. [doi:
     10.1016/j.ultrasmedbio.2010.05.010] [Medline: 20691917]

20. Morra L, Sacchetto D, Durando M, Agliozzo S, Carbonaro LA, Delsanto S, et al. Breast Cancer: Computer-aided Detection
     with Digital Breast Tomosynthesis. Radiology 2015 Oct;277(1):56-63. [doi: 10.1148/radiol.2015141959] [Medline:
     25961633]

21. Chabi M, Borget I, Ardiles R, Aboud G, Boussouar S, Vilar V, et al. Evaluation of the accuracy of a computer-aided
     diagnosis (CAD) system in breast ultrasound according to the radiologist's experience. Acad Radiol 2012 Mar;19(3):311-319.
     [doi: 10.1016/j.acra.2011.10.023] [Medline: 22310523]

22. Harbeck N, Gnant M. Breast cancer. Lancet 2017 Mar 18;389(10074):1134-1150. [doi: 10.1016/S0140-6736(16)31891-8]
     [Medline: 27865536]

23. Wang Y, Fan W, Zhao S, Zhang K, Zhang L, Zhang P, et al. Qualitative, quantitative and combination score systems in
     differential diagnosis of breast lesions by contrast-enhanced ultrasound. Eur J Radiol 2016 Jan;85(1):48-54. [doi:
     10.1016/j.ejrad.2015.10.017] [Medline: 26724648]

24. Choi EJ, Choi H, Park EH, Song JS, Youk JH. Evaluation of an automated breast volume scanner according to the fifth
     edition of BI-RADS for breast ultrasound compared with hand-held ultrasound. Eur J Radiol 2018 Feb;99:138-145. [doi:
     10.1016/j.ejrad.2018.01.002] [Medline: 29362145]

Abbreviations

AUC: area under the receiving operating characteristic curve
BI-RADS: Breast Imaging Reporting and Data System
CAD: computer-assisted diagnosis
NPV: negative predictive value
PPV: positive predictive value

Original Paper

Use of an Electronic Clinical Decision Support System in Primary Care to Assess Inappropriate Polypharmacy in Young Seniors With Multimorbidity: Observational, Descriptive, Cross-Sectional Study

Eloisa Rogero-Blanco[1,2,3*], MD; Juan A Lopez-Rodriguez[1,2,3,4*], MD, MSc, PhD; Teresa Sanz-Cuesta[3,4*], MD, PhD; Mercedes Aza-Pascual-Salcedo[5*], PharmD; M Jose Bujalance-Zafra[6*], MD, PhD; Isabel Cura-Gonzalez[2,3,4*], MD, MPH, PhD; MultiPAP Group[7*]

[1]General Ricardos Primary Health Care Centre, Madrid, Spain

[2]Medical Specialties and Public Health, School of Health Sciences, University Rey Juan Carlos, Alcorcón, Madrid, Spain

[3]Health Services Research on Chronic Patients Network (REDISSEC), Madrid, Spain

[4]Research Support Unit, Primary Care Management, Madrid, Spain

[5]Dirección Farmacia Atención Primaria Sector Zaragoza III, Zaragoza, Spain

[6]Dirección Unidad Gestión Clínica Victoria en Málaga, Servicio Andaluz de Salud, Málaga, Spain

[7]See Acknowledgments

[*]all authors contributed equally

Corresponding Author:
Juan A Lopez-Rodriguez, MD, MSc, PhD
General Ricardos Primary Health Care Centre
Calle General Ricardos
Madrid, 28019
Spain
Phone: 34 685197913
Email: juanantonio.lopez@salud.madrid.org

Related Article:

This is a corrected version. See correction statement: http://medinform.jmir.org/2020/11/e25678/

Abstract

Background: Multimorbidity is a global health problem that is usually associated with polypharmacy, which increases the risk of potentially inappropriate prescribing (PIP). PIP entails higher hospitalization rates and mortality and increased usage of services provided by the health system. Tools exist to improve prescription practices and decrease PIP, including screening tools and explicit criteria that can be applied in an automated manner.

Objective: This study aimed to describe the prevalence of PIP in primary care consultations among patients aged 65-75 years with multimorbidity and polypharmacy, detected by an electronic clinical decision support system (ECDSS) following the 2015 American Geriatrics Society Beers Criteria, the European Screening Tool of Older Person's Prescription (STOPP), and the Screening Tool to Alert doctors to Right Treatment (START).

Methods: This was an observational, descriptive, cross-sectional study. The sample included 593 community-dwelling adults aged 65-75 years (henceforth called young seniors), with multimorbidity (≥3 diseases) and polypharmacy (≥5 medications), who had visited their primary care doctor at least once over the last year at 1 of the 38 health care centers participating in the Multimorbidity and Polypharmacy in Primary Care (Multi-PAP) trial. Sociodemographic data, clinical and pharmacological treatment variables, and PIP, as detected by 1 ECDSS, were recorded. A multivariate logistic regression model with robust estimators was built to assess the factors affecting PIP according to the STOPP criteria.

Results: PIP was detected in 57.0% (338/593; 95% CI 53-61) and 72.8% (432/593; 95% CI 69.3-76.4) of the patients according to the STOPP criteria and the Beers Criteria, respectively, whereas 42.8% (254/593; 95% CI 38.9-46.8) of the patients partially

met the START criteria. The most frequently detected PIPs were benzodiazepines (BZD) intake for more than 4 weeks (217/593, 36.6%) using the STOPP version 2 and the prolonged use of proton pump inhibitors (269/593, 45.4%) using the 2015 Beers Criteria. Being a woman (odds ratio [OR] 1.43, 95% CI 1.01-2.01; *P*=.04), taking a greater number of medicines (OR 1.25, 95% CI 1.14-1.37; *P*<.04), working in the primary sector (OR 1.91, 95% CI 1.25-2.93; *P*=.003), and being prescribed drugs for the central nervous system (OR 3.75, 95% CI 2.45-5.76; *P*<.001) were related to a higher frequency of PIP.

Conclusions: There is a high prevalence of PIP in primary care as detected by an ECDSS in community-dwelling young seniors with comorbidity and polypharmacy. The specific PIP criteria defined by this study are consistent with the current literature. This ECDSS can be useful for supervising prescriptions in primary health care consultations.

Introduction

Background

Multimorbidity is defined by the World Health Organization as the coexistence of two or more chronic illnesses in a person [1]. It is a growing phenomenon that has become a health problem and an international health challenge [2] resulting from increased life expectancy and nontransmittable disease rates, among other factors. Patients with multimorbidity usually present with polypharmacy, defined as the simultaneous usage of several medicines. This definition is controverted, and there is no standard agreement on the number of drugs that indicates that the patient is polymedicated [3].

Polypharmacy entails a greater risk for potentially inappropriate prescribing (PIP), which is defined as "the prescribing of medications or medication classes that should generally be avoided in persons 65 years or older because they are either ineffective or they pose unnecessarily high risk for older persons and a safer alternative is available (potentially inappropriate medication, PIM)" [4]. PIP also includes prescription omissions as the lack of use of potential right treatments. Data from the US National Health Survey yielded a prevalence of polypharmacy of 39% in the population aged older than 65 years [5]. A European study on medication dispensing for 310,000 adults within the UK National Health Service observed that the proportion of adult patients prescribed with 5 or more medicines doubled between 1995 and 2014, reaching 20.8% of the population [6]. For the same population age range in Spain, the percentage of polymedicated patients was approximately 50% [7]. The prevalence of patients with PIP is high, although it differs among studies, many of which are aimed at hospitalized, institutionalized, or immobilized elderly patients [8]. In the primary care setting in Europe, the PIP prevalence varies between 36% in Ireland [9] and 39% or 40.4% reported in 2 studies in Spain [10,11]. The existence of PIP has been related to a higher risk of being hospitalized [12], greater use of health emergency services and family doctor visits, reduced functionality [13], and increased mortality rates [14]. In terms of economic impact, some European studies estimate that the direct cost associated with PIP ranges between €188 and €318 (US $213-US $370) per patient per year [9,15].

A number of tools have been developed for studying appropriate prescribing and facilitating the detection of PIP. There are implicit methods based on clinical judgment, such as the Medication Appropriateness Index [16], and explicit methods based on predefined criteria, such as the American Geriatrics Society Beers Criteria, the European Screening Tool of Older Person's Prescribing (STOPP) Criteria, and the Screening Tool to Alert Doctors to the Right Treatment (START) Criteria. Although the Beers Criteria were the first to be published in 1991 [4] and have been the most widely employed to date, their application in the European setting is limited because they include a high percentage of pharmacological drugs not recognized in the majority of catalogs for European countries. To address this limitation, the criteria for the STOPP and START were set in 2008. Studies comparing the STOPP Criteria with other explicit criteria show that the STOPP Criteria are more exhaustive [9,15,17]. The STOPP, START, and Beers Criteria were updated and validated in Spanish in 2015 [18,19], but there are no studies comparing them.

Some studies indicate that using the abovementioned tools to check prescribed medication translates into improved health outcomes for patients (falls, functionality, hospitalizations, number of consultations, and use of emergency services) [20]. However, applying these criteria requires extensive and updated clinical knowledge, time, and access to various simultaneous sources of information. For their use to be viable at primary care consultations, health professionals need automated, quick-to-consult tools that facilitate and supervise the process of medication prescribing. Electronic clinical decision support systems (ECDSSs) can improve prescription quality and reduce medication errors [21,22]. Although electronic clinical records are largely implemented in the Spanish primary care system, no adequate ECDSSs are currently available. Several clinical trials are being conducted to evaluate the use of ECDSSs in clinical practice [23-27].

In Spain, an available ECDSS (CheckTheMeds Technology SL) has been used in the hospital setting and in pharmacological services offering a personalized system for medication dispensing [28,29]. This Web-based tool globally processes each patient's information, combining clinical and pharmacological data, thus offering the health professional an analysis of criteria (Beers, STOPP, and START) for detecting PIP in real time.

Objective

The main aim of this study was to estimate the prevalence of PIP in young seniors (aged 65 to 75 years) with multimorbidity and polypharmacy, according to the updated criteria by the STOPP, START, and Beers, as detected by the ECDSS, at primary care consultations. The secondary objectives were to assess the relationships between clinical and sociodemographic variables and the presence of PIP.

Methods

Design

This was an observational, cross-sectional, descriptive, multicentric study conducted in the Spanish primary care setting.

Population

This study included patients aged 65-75 years with multimorbidity (≥3 diseases) and polypharmacy (≥5 consumed medicines for at least 3 months), who had attended their doctor consultation at least once over the last year and provided written consent to participate in the Multimorbidity and Polypharmacy in Primary Care (Multi-PAP) trial [30]. Institutionalized patients—those whose life expectancy was less than 12 months, as estimated by their doctor—and patients with any severe mental disorder were excluded. Recruitment physicians did not receive any economic incentives. Each doctor offered participation to patients from a random list of patients potentially meeting the inclusion criteria. A total of 117 family doctors from 38 health care centers from 3 Spanish regions (Andalucia, Aragon, and Madrid) included a total of 593 subjects who agreed to participate. On the basis of the previous studies reporting a PIP percentage of 39% [10], with this sample size, a maximum type 1 error of 3.9% with a 95% CI was reported.

Variables

The following sociodemographic variables were recorded for patients: age, gender, marital status, educational level, social class according to the Spanish classification [31], and family income in thousands of euros adjusted by the number of people in the household using the method proposed by the Organisation for Economic Cooperation and Development. In addition, the following clinical variables were collected: number of active pharmaceutical ingredients per patient according to the Anatomical Therapeutic Chemical (ATC) classification system and chronic conditions in accordance with the International Classification of Primary Care, with the most relevant ones selected according to the criterion by O'Halloran et al [32]. The assessed PIP variables were existence of STOPP and START criteria following its Spanish-adapted 2014 version [18], existence of Beers 2015 version criteria [19], number of STOPP criteria, number of START criteria, number of Beers Criteria, and classification of the type of detected criteria.

Sociodemographic and clinical data were obtained between December 2016 and January 2017 through an interview by each patient's doctor and recorded in a data collection notebook (DCN). Subsequently, in April 2018, the data were uploaded from the DCN into the ECDSS. One researcher with vast clinical and therapeutic drug monitoring experience supervised the information transfer to the CheckTheMeds Technology SL and used this tool to globally review the treatment of all patients. The latest versions of the STOPP, START, and Beers Criteria were employed.

All the STOPP and Beers Criteria were analyzed. In agreement with previous studies [33] and to avoid potential information bias, this research team agreed on omitting the A1 STOPP criterion from the analysis (any drug prescribed but not indicated by clinical evidence) to prevent its overestimation [33].

Statistical Analysis

Categorical variables were described by their frequencies and percentages. Quantitative variables were described by their mean and SD, with their corresponding 95% CI when they fit a normal distribution, or by their median and IQR in the case of asymmetric distributions. The presence of one or more PIP was identified, and the association between groups was assessed for the main variables using a chi-square test in the case of categorical variables or the Student t test (Mann-Whitney U test when the variable did not fit a normal distribution) in the case of quantitative ones. A multivariate logistic regression model was built to study factors related to PIP, with robust estimators that controlled for the effect of cluster sampling. The dependent variable was the presence of one or more PIP according to the STOPP version 2 (V2) criteria, and independent variables were those reaching statistical significance in the bivariate analysis or considered of clinical relevance. Stata version 14.0 (StataCorp LLC) and IBM SPSS 21 software (IBM Corp) were employed in the statistical analyses.

Results

A total of 4386 prescriptions were recorded for the 593 included patients (593/635, 93.4% of the total offered for participation). The average age of patients was 69.7 (SD 2.7) years; 56.3% (334/593) of the patients were women, 75.4% (447/593) were married, and 17.9% (106/593) lived by themselves. Multimedia Appendix 1 describes the main sociodemographic and clinical characteristics for each group according to the presence or absence of PIP or medication omissions according to the START Criteria. The most frequent pathologies were high blood pressure that amounted to 78.9% cases (468/593) and hypercholesterolemia with 50.2% cases (301/593). Of the 593 patients, 250 (42.2%) were diabetic, 215 (36.3%) had arthritis (knee, hip, or other joints), and 220 (37.1%) had mental disorders (for a complete list of disease prevalence, see Multimedia Appendix 2). The median number of chronic illnesses per patient was 5 (IQR 4-7). The median number of medicines per patient was 7 (IQR 6-9), and 17.9% (106/593) of the patients were prescribed ≥10 drugs. In terms of ATC groups, the cardiovascular one was the most frequent, with 95.3% (565/593) subjects taking at least one drug, followed by the digestive system/metabolism and the nervous system groups. The most prescribed drug was omeprazole, which was taken by 49.2% (292/593) of patients, followed by acetylsalicylic acid by 36.8% (218/593) of patients, metformin by 34.4% (204/593) of patients, simvastatin by 32.9% (195/593) of patients, and enalapril by 27.2% (161/593) of patients.

XSL•FO

RenderX

The frequency for the ECDSS to detect at least one explicit criterion was 57.0% (95% CI 53-61), 42.8% (95% CI 38.9-46.8), and 72.8% (95% CI 69.3-76.4) applying the STOPP, START, and Beers Criteria, respectively, all in their latest versions at the point when this study was conducted (Table 1). The percentage of patients that met three or more Beers Criteria was 16.5% (98/593). Of the overall criteria evaluated for the studied sample, 30 different criteria were detected in the STOPP, 21 in the START, and 34 in the Beers Criteria (see Multimedia Appendix 3). The most frequently found PIP according to the STOPP criteria were the prolonged use of benzodiazepines (BZD) in 36.6% of patients (217/593), followed by beta blockers in 12.5% (74/593) of patients with diabetes mellitus with frequent episodes of hypoglycemia and the prescribed use of opioids without associated laxatives found in 5.4% (32/593) of patients. Following the Beers Criteria, the most frequent PIP were the prolonged use of proton pump inhibitors (PPIs) by 45.4% (269/593) of patients, prolonged use of BZD and hypnotic drugs by 27% (160/593) of patients, and prolonged use of antidepressants or the combined use of 2 or more central nervous system depressants by 17.4% (130/593) of patients (Multimedia Appendix 2).

At least one PIP was detected using both methods in 40.8% (242/593) of patients, and disagreement was found in 32.0% (190/593) of patients for whom 1 PIP was detected according to the Beers Criteria but did not meet any criterion in the STOPP. On the contrary, 16.2% (96/593) of patients presented with some PIP using the STOPP criteria, but none using the Beers Criteria.

Factors that were found to relate to the presence of PIP were being a woman, greater medicine intake, belonging to social class 2 (qualified primary sector), and using a pharmacological treatment for the central nervous system (Anatomical Therapeutic Chemical classification system-nervous system [ATC-N]; see Table 2).

The most frequent omission of medicines detected by the START criteria in 593 patients were inhaled corticosteroids in 55 (9.3%) patients with asthma or severe chronic obstructive pulmonary disease (COPD); prostaglandins, prostamide, or topical beta blockers in 40 (6.7%) patients with primary open-angle glaucoma; calcium and vitamin D supplements in 37 (6.2%) patients with osteoporosis; and inhibitors of 5α-reductase also in 37 (6.2%) patients for the treatment of symptomatic prostatism when prostatectomy is not necessary (see Multimedia Appendix 2).

Table 1. Percentage of detected potentially inappropriate prescribing (potentially inappropriate medication or prescription omission) according to the different criteria sets.

Potentially inappropriate prescribing	Screening Tool of Older Person's potentially inappropriate prescriptions version 2		Beers 2015 version		Screening Tool to Alert doctors to Right Treatment, version 2	
	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI
Patients with at least one PIM[a] or one MO[b]	338 (57.0)	53-61	432 (72.8)	69.3-76.4	254 (42.0)	38.8-46.8
Number of PIM/MO per patient						
1	257 (43.3)	N/A[c]	214 (36.1)	N/A	160 (27.0)	N/A
2	65 (11.0)	N/A	120 (20.2)	N/A	64 (10.8)	N/A
≥3	16 (2.7)	N/A	98 (16.5)	N/A	28 (4.7)	N/A

[a]PIM: potentially inappropriate medication.

[b]MO: medication omission.

[c]N/A: not applicable.

Table 2. Factors associated with a potentially inappropriate medication according to the Screening Tool of Older Person's Prescription criteria.

Associated factors	Odds ratio (95% CI)	P value
Woman	1.43 (1.01-2.01)	.04
Number of drugs	1.25 (1.14-1.37)	<.001
Social class		
Unskilled	Reference	N/A[a]
Skilled worker in the primary sector	1.45 (0.98-2.16)	.07
Supervisors, managers, and directors	1.91 (1.25-2.93)	.003
Usage of drugs in Anatomical Therapeutic Chemical classification system-nervous system group	3.75 (2.43-5.76)	<.001

[a]N/A: not applicable.

Discussion

Principal Findings

The percentage of patients in the sample with one or more PIP was 57.0% (338/593) using STOPP (95% CI 53-61), 72.8% (432/593) with Beers (95% CI 69.3-76.4), and 42.8% (254/593) with START (95% CI 38.9-46.8). The most frequent PIPs were the use of BZD for more than 4 weeks according to the STOPP V2 criteria (217/593, 36.6%) and the prolonged use of PPI using the Beers 2015 version (269/593, 45.4%). Factors associated with a higher PIP frequency were being a woman (OR 1.43, 95% CI 1.01-2.01; P=.04), greater medicine intake (OR 1.25, 95% CI 1.14-1.37; P<.001), being a primary sector worker (OR 1.91, 95% CI 1.25-2.93; P=.003), and using pharmacological treatments for the central nervous system (OR 3.75, 95% CI 2.45-5.76; P<.001). Both the STOPP and Beers tools simultaneously detected one or more PIP in 40.8% (242/593) of the patients. Differences were found using the STOPP and Beers sets of criteria, and the difficulties for using some of the explicit criteria were found by the ECDSS.

Prevalence of Potentially Inappropriate Prescribing

The prevalence of PIP among community-dwelling young seniors varies depending on the explicit criteria employed as well as the different study designs and included populations [15,18,34]. This study obtained a prevalence of 57.0% in accordance with the STOPP criteria, which is higher than that of previous studies [9,10] using previous STOPP versions in the primary care setting and the European population (36% and 39%, respectively). Studies using the STOPP 2014 version found prevalence values of 8.7% and 40.4% [11,33]. The study by Blanco-Reina et al [11], with similar data to ours, obtained a prevalence of 40.4% in contrast to the 18.7% obtained using the original STOPP version and concluded that the latest version is more sensitive for PIP detection. The greater prevalence observed in this study can result from the included population, which had to meet the polypharmacy criterion to participate in the Multi-PAP trial group, whereas only 72.9% of the patients in the abovementioned study were polymedicated.

Following the Beers 2015 version criteria, the PIP prevalence (77.2%) was higher than that described in some previous studies. Owing to the repeated updates of these criteria (5 up to 2015) and different study designs and studied populations, there is great variability in the obtained prevalence values [34]. Studies using the Beers 2015 version reported a prevalence ranging from 53.5% in the hospitalized population [35] to 72.8% in studies based on big data [36]. In Spain, using the 2008 and 2012 versions, the prevalence of PIP varied from 24.3% to 44.0%, respectively [37].

Studies comparing the 2015 version with the 2012 versions of Beers Criteria [35] confirmed that the latest version has greater sensitivity for detecting PIP. The increased number of criteria in the Beers 2015 version and the adaptation of the medications catalog [17], together with the application of these criteria to a polymedicated population, can explain the higher prevalence found in our sample.

The percentage of patients with one or more medication omission was 42.8%, a figure superior to those in previous studies, where it ranged between 22.0% and 39.9% using the START criteria [10,11,33,37].

Most Frequent Screening Tool of Older Person's Prescription and Screening Tool to Alert Doctors to Right Treatment, the Beers Criteria

In agreement with most previous studies, BZD and PPI were the most frequent PIP detected [10,11,15,33,37]. The proportion of patients using BZD for more than 4 weeks in the studied sample using STOPP V2 is similar to that found by Blanco-Reina et al [11] (36.6% vs 38.6%, respectively), whereas the use of BZD and hypnotic drugs using the Beers Criteria was lower than that observed in the mentioned study (26.9% vs 52.4%, respectively) but similar to that in a study by Zhang et al (29.8%) [35]. The reason for this difference can be that the most used BZD by patients in this study were lorazepam (13.5%), bromazepam (8.6%), and lormetazepam (6.9%), with the 2 latest being included in the catalog of medicines detected by the STOPP criteria but not in the catalog for the Beers. The work by Blanco-Reina et al [11], which did not specify what the most used BZD were, reported that 46.7% of patients had insomnia, a figure much higher than the one observed in our sample, which could explain the differences between them.

This study found that the prolonged use of PPI was the most frequent PIP detected by the Beers latest version (45.5%), although it was not detected using STOPP. Several systematic reviews report it as one of the most frequent PIP observed in the polymedicated population in accordance with different explicit criteria [15,34,38]. The Beers 2015 updated version included this criterion that is already described as frequent (41.9%) by some authors [35], although their samples are not comparable with ours. Some studies that identified this PIP as frequent [9,10] employed a modified version of STOPP that facilitates its detection, instead of the original criterion that is associated with a specific clinical condition that limits its capacity for such purpose (PPI for treating peptic ulcer disease without complications or peptic erosive esophagitis with full therapeutic doses for more than 8 weeks). For example, 40 mg of omeprazole was considered the full therapeutic dose by the ECDSS.

The use of nonsteroidal anti-inflammatory drugs (NSAIDs) also appears in the current literature as one the most frequently found PIP in the polymedicated population [15,34,38]. This criterion only was met by 1.5% of patients in the study by Blanco-Reina et al [11], which used a similar methodology to ours. As their own family doctors were the ones entering the pharmacological information for their patients in the DCN, a bias can be present in the outcome if the patient was taking the NSAIDs, but the doctor did not consider it as their regular chronic medication and hence did not include it in the DCN. This record omission, together with the ample information on increased cardiovascular risk associated with these drugs, can explain the observed differences.

In studies using the START 2014, the most relevant omission criterion is the use of inhaled corticosteroids in patients with

asthma or severe COPD as well as the use of calcium and vitamin D supplements in patients with osteoporosis, as was the case in our sample. The criterion for prostaglandin, prostamides, or topical beta blockers in patients with primary open-angle glaucoma can be overestimated because topical medication was not recorded, which can partially explain the high prevalence of omissions in this trial.

Factors relating to the presence of PIP as detected using the STOPP are similar to those in previous studies [15,34] that have identified a relationship between being a woman and taking a greater number of medicines with the presence of PIP. These studies also found a correlation with poor economic status, which was not the case in our study, although the classification systems for the economic level and social class are not comparable. The ATC-N includes all the drugs in the BZD group as well as antidepressants, thus its association with the presence of PIP.

Agreement Between Methods

Only 40.0% (237/593) of patients were simultaneously detected with one or more PIP according to the criteria of the Beers and STOPP methods. The use of different medication catalogs by these tools does not entirely explain the observed discrepancies. Although previous studies have analyzed the agreement between both the tools in terms of the overall percentage of detected PIP [11,37,39], they did not take into consideration that the drugs and clinical scenarios assessed by each method are different and changed throughout their updated versions. In a review by Motter et al [17], a detailed analysis of differences among the existing explicit criteria since 1991 (STOPP, Beers, and others) was conducted; they concluded that these tools assess different parameters, medicines, and clinical scenarios, which are not readily comparable. Given the differences between these tools, their combined use in a complementary manner is probably the most adequate approach.

Algorithms Used by the Electronic Clinical Decision Support System

ECDSSs are very helpful tools for applying different explicit criteria simultaneously [40,41]. Translating the STOPP criteria into computerized algorithms to be used by an ECDSS can be more complicated than doing so for the Beers Criteria. This was evident in our study in terms of the difficulty found in detecting the usage of PPIs for more than 8 weeks using the STOPP criteria, as previously mentioned. The majority of the STOPP criteria are linked to a clinical condition that is often difficult

to codify or extract from the patient's electronic clinical records [42]. Studies evaluating an ECDSS [43] that applies the explicit criteria in the latest versions of these tools estimate that 67% of the STOPP criteria require additional clinical information vs 31% for Beers. Many studies employ several criteria instead of the complete list; hence, the analysis of prescriptions remains incomplete [44], whereas others [10] use free adaptations of the criteria to facilitate the development of computerized algorithms.

Strengths and Limitations

Not exporting the data directly from clinical records but from a DCN designed ad hoc could have potentially introduced an information bias regarding illnesses and pharmacological treatment. Given the design of the Multi-PAP trial, health professionals may have prioritized pathologies they considered chronic and were included in the classification employed in the trial instead of registering the total number of patients' diseases. Something similar may have occurred with topical medication as the research team decided not to include it in the study because it is not associated with significant adverse effects. This bias could have resulted in an overestimation of the prevalence of certain STOPP or START criteria that the research team tried to correct by adjusting the A1P criterion.

The pragmatic design of the study is among our strengths, with patients comprising a representative sample of the community-dwelling young seniors with multimorbidity and polypharmacy, who were interviewed and evaluated by their usual doctors during clinical practice. Overall, this age group has a good quality of life and substantial potential for early interventions.

Applicability of the Outcome

Using the ECDSS allows for a rapid, complete, and simultaneous evaluation of PIP based on different explicit criteria. This tool can support family doctors when deciding what the best therapeutic choice is for each patient, not only to review the therapeutic plan but also to prescribe new treatments. In any event, even after the information is provided via the ECDSS, the professional must always evaluate it and make the final decision. The decision to continue or discontinue certain treatments must be made based on clinical conditions and within a framework of a doctor-patient shared decision making. The evaluation of this tool in the future in clinical effectiveness and implementation of hybrid design studies will help in integrating them into the clinical electronic records of patients.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sociodemographic, clinical, and pharmacological characteristics of patients.
[XLSX File (Microsoft Excel File), 16 KB - medinform_v8i3e14130_app1.xlsx]

Multimedia Appendix 2
Disease List Prevalence according to International Classification in Primary Care.
[XLSX File (Microsoft Excel File), 14 KB - medinform_v8i3e14130_app2.xlsx]

Multimedia Appendix 3
List of all screening tool of older people's prescriptions, screening tool to alert to right treatment, and Beers criteria.
[XLSX File (Microsoft Excel File), 16 KB - medinform_v8i3e14130_app3.xlsx]

XSL•FO
RenderX

References

1. World Health Organization. 2008. The World Health Report 2008 - primary Health Care (Now More Than Ever) URL: https://www.who.int/whr/2008/en/

2. The Academy of Medical Sciences. 2018 Apr. Multimorbidity: a priority for global health research URL: https://acmedsci.ac.uk/file-download/82222577

3. Rankin A, Cadogan CA, Patterson SM, Kerse N, Cardwell CR, Bradley MC, et al. Interventions to improve the appropriate use of polypharmacy for older people. Cochrane Database Syst Rev 2018 Dec 03;9:CD008165. [doi: 10.1002/14651858.CD008165.pub4] [Medline: 30175841]

4. Beers MH, Ouslander JG, Rollingher I, Reuben DB, Brooks J, Beck JC. Explicit criteria for determining inappropriate medication use in nursing home residents. UCLA Division of Geriatric Medicine. Arch Intern Med 1991 Sep;151(9):1825-1832. [Medline: 1888249]

5. Kantor ED, Rehm CD, Haas JS, Chan AT, Giovannucci EL. Trends in prescription drug use among adults in the United States from 1999-2012. J Am Med Assoc 2015 Nov 03;314(17):1818-1831 [FREE Full text] [doi: 10.1001/jama.2015.13766] [Medline: 26529160]

6. Guthrie B, Makubate B, Hernandez-Santiago V, Dreischulte T. The rising tide of polypharmacy and drug-drug interactions: population database analysis 1995-2010. BMC Med 2015 Apr 07;13:74 [FREE Full text] [doi: 10.1186/s12916-015-0322-7] [Medline: 25889849]

7. Molina López T, Caraballo Camacho MD, Palma Morgado D, López Rubio S, Domínguez Camacho JC, Morales Serna JC. [Prevalence of multiple medication and cardiovascular risk in patients over 65 years]. Aten Primaria 2012 Apr;44(4):216-222 [FREE Full text] [doi: 10.1016/j.aprim.2011.03.010] [Medline: 21924797]

8. Gavilán Moral E, Morales Suárez-Varela MT, Hoyos Esteban JA, Pérez Suanes AM. Polimedicación y prescripción de fármacos inadecuados en pacientes ancianos inmovilizados que viven en la comunidad. Atención Primaria 2006 Nov;38(9):476-480. [doi: 10.1157/13095047]

9. Cahir C, Fahey T, Teeling M, Teljeur C, Feely J, Bennett K. Potentially inappropriate prescribing and cost outcomes for older people: a national population study. Br J Clin Pharmacol 2010 May;69(5):543-552 [FREE Full text] [doi: 10.1111/j.1365-2125.2010.03628.x] [Medline: 20573091]

10. Cruz-Esteve I, Marsal-Mora JR, Galindo-Ortego G, Galván-Santiago L, Serrano-Godoy M, Ribes-Murillo E, en representación del grupo de investigación GRETAP. [Potentially inappropriate prescribing in older Spanish population according to STOPP/START criteria (STARTREC study)]. Aten Primaria 2017 Mar;49(3):166-176 [FREE Full text] [doi: 10.1016/j.aprim.2016.02.013] [Medline: 27693033]

11. Blanco-Reina E, García-Merino MR, Ocaña-Riola R, Aguilar-Cano L, Valdellós J, Bellido-Estévez I, et al. Assessing potentially inappropriate prescribing in community-dwelling older patients using the updated version of STOPP-START criteria: a comparison of profiles and prevalences with respect to the original version. PLoS One 2016;11(12):e0167586 [FREE Full text] [doi: 10.1371/journal.pone.0167586] [Medline: 27907210]

12. Pérez T, Moriarty F, Wallace E, McDowell R, Redmond P, Fahey T. Prevalence of potentially inappropriate prescribing in older people in primary care and its association with hospital admission: longitudinal study. Br Med J 2018 Dec 14;363:k4524 [FREE Full text] [doi: 10.1136/bmj.k4524] [Medline: 30429122]

13. Moriarty F, Bennett K, Cahir C, Kenny RA, Fahey T. Potentially inappropriate prescribing according to STOPP and START and adverse outcomes in community-dwelling older people: a prospective cohort study. Br J Clin Pharmacol 2016 Dec;82(3):849-857 [FREE Full text] [doi: 10.1111/bcp.12995] [Medline: 27136457]

14. Leelakanok N, Holcombe AL, Lund BC, Gu X, Schweizer ML. Association between polypharmacy and death: a systematic review and meta-analysis. J Am Pharm Assoc (2003) 2017;57(6):729-738.e10. [doi: 10.1016/j.japh.2017.06.002] [Medline: 28784299]

15. Hill-Taylor B, Sketris I, Hayden J, Byrne S, O'Sullivan D, Christie R. Application of the STOPP/START criteria: a systematic review of the prevalence of potentially inappropriate prescribing in older adults, and evidence of clinical, humanistic and economic impact. J Clin Pharm Ther 2013 Oct;38(5):360-372. [doi: 10.1111/jcpt.12059] [Medline: 23550814]

16. Hanlon JT, Schmader KE, Samsa GP, Weinberger M, Uttech KM, Lewis IK, et al. A method for assessing drug therapy appropriateness. J Clin Epidemiol 1992 Oct;45(10):1045-1051. [Medline: 1474400]

17. Motter FR, Fritzen JS, Hilmer SN, Paniz É, Paniz VM. Potentially inappropriate medication in the elderly: a systematic review of validated explicit criteria. Eur J Clin Pharmacol 2018 Jun;74(6):679-700. [doi: 10.1007/s00228-018-2446-0] [Medline: 29589066]

18. Delgado Silveira E, Montero Errasquín B, Muñoz García M, Vélez-Díaz-Pallarés M, Lozano Montoya I, Sánchez-Castellano C, et al. [Improving drug prescribing in the elderly: a new edition of STOPP/START criteria]. Rev Esp Geriatr Gerontol 2015;50(2):89-96. [doi: 10.1016/j.regg.2014.10.005] [Medline: 25466971]

19. American Geriatrics Society 2015 Beers Criteria Update Expert Panel. American Geriatrics Society 2015 updated Beers criteria for potentially inappropriate medication use in older adults. J Am Geriatr Soc 2015 Nov;63(11):2227-2246. [doi: 10.1111/jgs.13702] [Medline: 26446832]

20. Hill-Taylor B, Walsh KA, Stewart S, Hayden J, Byrne S, Sketris IS. Effectiveness of the STOPP/START (Screening Tool of Older Persons' potentially inappropriate Prescriptions/Screening Tool to Alert doctors to the Right Treatment) criteria:

systematic review and meta-analysis of randomized controlled studies. J Clin Pharm Ther 2016 Apr;41(2):158-169. [doi: 10.1111/jcpt.12372] [Medline: 26990017]

21. Reis W, Bonetti A, Bottacin W, Reis A, Souza T, Pontarolo R, et al. Impact on process results of clinical decision support systems (CDSSs) applied to medication use: overview of systematic reviews. Pharm Pract (Granada) 2017;15(4):1036 [FREE Full text] [doi: 10.18549/PharmPract.2017.04.1036] [Medline: 29317919]

22. Clyne B, Bradley MC, Hughes C, Fahey T, Lapane KL. Electronic prescribing and other forms of technology to reduce inappropriate medication use and polypharmacy in older people: a review of current evidence. Clin Geriatr Med 2012 May;28(2):301-322. [doi: 10.1016/j.cger.2012.01.009] [Medline: 22500545]

23. Soiza RL, Subbarayan S, Antonio C, Cruz-Jentoft AJ, Petrovic M, Gudmundsson A, et al. The SENATOR project: developing and trialling a novel software engine to optimize medications and nonpharmacological therapy in older people with multimorbidity and polypharmacy. Ther Adv Drug Saf 2017 Mar;8(3):81-85 [FREE Full text] [doi: 10.1177/2042098616675851] [Medline: 28382196]

24. Sönnichsen A, Trampisch US, Rieckert A, Piccoliori G, Vögele A, Flamm M, et al. Polypharmacy in chronic diseases-Reduction of Inappropriate Medication and Adverse drug events in older populations by electronic Decision Support (PRIMA-eDS): study protocol for a randomized controlled trial. Trials 2016 Jan 29;17:57 [FREE Full text] [doi: 10.1186/s13063-016-1177-8] [Medline: 26822311]

25. Christiane M. U.S National Library of Medicine. 2018. Application for an Electronic Medication Management Support System (AdAM) URL: https://clinicaltrials.gov/ct2/show/NCT03430336 [accessed 2020-01-13]

26. Stefanie HM. U.S National Library of Medicine. OPtimising thERapy to Prevent Avoidable Hospital Admissions in the Multimorbid Older People (OPERAM) Internet URL: https://clinicaltrials.gov/ct2/show/NCT02986425

27. Young A, Tordoff J, Dovey S, Reith D, Lloyd H, Tilyard M, et al. Using an electronic decision support tool to reduce inappropriate polypharmacy and optimize medicines: rationale and methods. JMIR Res Protoc 2016 Jun 10;5(2):e105 [FREE Full text] [doi: 10.2196/resprot.5543] [Medline: 27288200]

28. Barris Blundell D. Revisión de medicación según criterios STOPP/START en pacientes mayores del servicio de sistema personalizado de dosificación de medicamentos de una farmacia comunitaria. Farm Comunitarios 2015 Jun 01;7(2):31-36. [doi: 10.5672/FC.2173-9218.(2015/Vol7).002.05]

29. González-Rojano E, Acosta SV, Palacián DR, Hernández M, Abad-Santos F. Utility of Checkthemeds® for detection of potentially inappropriate medication in hospitalized patients. Clin Ther 2017 Aug;39(8):e37. [doi: 10.1016/j.clinthera.2017.05.116]

30. Prados-Torres A, Del Cura-González I, Prados-Torres D, López-Rodríguez JA, Leiva-Fernández F, Calderón-Larrañaga A, Multi-PAP Group. Effectiveness of an intervention for improving drug prescription in primary care patients with multimorbidity and polypharmacy: study protocol of a cluster randomized clinical trial (Multi-PAP project). Implement Sci 2017 Apr 27;12(1):54 [FREE Full text] [doi: 10.1186/s13012-017-0584-x] [Medline: 28449721]

31. Domingo-Salvany A, Bacigalupe A, Carrasco JM, Espelt A, Ferrando J, Borrell C, del Grupo de Determinantes Sociales de Sociedad Española de Epidemiología. [Proposals for social class classification based on the Spanish National Classification of Occupations 2011 using neo-Weberian and neo-Marxist approaches]. Gac Sanit 2013;27(3):263-272 [FREE Full text] [doi: 10.1016/j.gaceta.2012.12.009] [Medline: 23394892]

32. O'Halloran J, Miller G, Britt H. Defining chronic conditions for primary care with ICPC-2. Fam Pract 2004 Aug;21(4):381-386. [doi: 10.1093/fampra/cmh407] [Medline: 15249526]

33. O Riordan D, Aubert CE, Walsh KA, Van Dorland A, Rodondi N, Du Puy RS, et al. Prevalence of potentially inappropriate prescribing in a subpopulation of older European clinical trial participants: a cross-sectional study. BMJ Open 2018 Mar 22;8(3):e019003 [FREE Full text] [doi: 10.1136/bmjopen-2017-019003] [Medline: 29567842]

34. Tommelein E, Mehuys E, Petrovic M, Somers A, Colin P, Boussery K. Potentially inappropriate prescribing in community-dwelling older people across Europe: a systematic literature review. Eur J Clin Pharmacol 2015 Dec;71(12):1415-1427. [doi: 10.1007/s00228-015-1954-4] [Medline: 26407687]

35. Zhang X, Zhou S, Pan K, Li X, Zhao X, Zhou Y, et al. Potentially inappropriate medications in hospitalized older patients: a cross-sectional study using the Beers 2015 criteria versus the 2012 criteria. Clin Interv Aging 2017;12:1697-1703 [FREE Full text] [doi: 10.2147/CIA.S146009] [Medline: 29066875]

36. Kim GJ, Lee KH, Kim JH. South Korean geriatrics on Beers Criteria medications at risk of adverse drug events. PLoS One 2018;13(3):e0191376 [FREE Full text] [doi: 10.1371/journal.pone.0191376] [Medline: 29543860]

37. Blanco-Reina E, Ariza-Zafra G, Ocaña-Riola R, León-Ortiz M. 2012 American Geriatrics Society Beers criteria: enhanced applicability for detecting potentially inappropriate medications in European older adults? A comparison with the Screening Tool of Older Person's Potentially Inappropriate Prescriptions. J Am Geriatr Soc 2014 Jul;62(7):1217-1223. [doi: 10.1111/jgs.12891] [Medline: 24917083]

38. Lucchetti G, Lucchetti AL. Inappropriate prescribing in older persons: a systematic review of medications available in different criteria. Arch Gerontol Geriatr 2017;68:55-61. [doi: 10.1016/j.archger.2016.09.003] [Medline: 27649514]

39. Pastor Cano J, Aranda García A, Gascón Cánovas JJ, Rausell Rausell VJ, Tobaruela Soto M. [Beers versus STOPP criteria and the possible implications of the Beers criteria's Spanish adaptation]. Farm Hosp 2017 Jan 01;41(n01):130-131 [FREE Full text] [doi: 10.7399/fh.2017.41.1.10568] [Medline: 28045656]

40. Meulendijk MC, Spruit MR, Drenth-van MA, Numans ME, Brinkkemper S, Jansen PA, et al. Computerized decision support improves medication review effectiveness: an experiment evaluating the STRIP assistant's usability. Drugs Aging 2015 Jun;32(6):495-503 [FREE Full text] [doi: 10.1007/s40266-015-0270-0] [Medline: 26025118]

41. Niehoff KM, Rajeevan N, Charpentier PA, Miller PL, Goldstein MK, Fried TR. Development of the tool to reduce inappropriate medications (TRIM): a clinical decision support system to improve medication prescribing for older adults. Pharmacotherapy 2016 May 28;36(6):694-701. [doi: 10.1002/phar.1751]

42. Nauta KJ, Groenhof F, Schuling J, Hugtenburg JG, van Hout HP, Haaijer-Ruskamp FM, et al. Application of the STOPP/START criteria to a medical record database. Pharmacoepidemiol Drug Saf 2017 Oct;26(10):1242-1247. [doi: 10.1002/pds.4283] [Medline: 28799295]

43. Anrys P, Boland B, Degryse J, De Lepeleire J, Petrovic M, Marien S, et al. STOPP/START version 2-development of software applications: easier said than done? Age Ageing 2016 Sep;45(5):589-592. [doi: 10.1093/ageing/afw114] [Medline: 27496930]

44. Price M, Davies I, Rusk R, Lesperance M, Weber J. Applying STOPP guidelines in primary care through electronic medical record decision support: randomized control trial highlighting the importance of data quality. JMIR Med Inform 2017 Jun 15;5(2):e15 [FREE Full text] [doi: 10.2196/medinform.6226] [Medline: 28619704]

Abbreviations

ATC: Anatomical Therapeutic Chemical
ATC-N: Anatomical Therapeutic Chemical classification system-nervous system
BZD: benzodiazepines
COPD: chronic obstructive pulmonary disease
DCN: data collection notebook
ECDSS: electronic clinical decision support system
Multi-PAP: Multimorbidity and Polypharmacy in Primary Care
NSAID: nonsteroidal anti-inflammatory drug
OR: odds ratio
PIP: potentially inappropriate prescribing
PPI: proton pump inhibitor
START: Screening Tool to Alert doctors to Right Treatment
STOPP: Screening Tool of Older Person's Prescription
V2: version 2

Original Paper

A Deep-Learning Algorithm (ECG12Net) for Detecting Hypokalemia and Hyperkalemia by Electrocardiography: Algorithm Development

Chin-Sheng Lin[1*], MD, PhD; Chin Lin[2,3,4*], PhD; Wen-Hui Fang[5], MD; Chia-Jung Hsu[6], MSc; Sy-Jou Chen[7,8], MD; Kuo-Hua Huang[6], BSc; Wei-Shiang Lin[1], MD; Chien-Sung Tsai[9], MD; Chih-Chun Kuo[10], BSc; Tom Chau[11], MD; Stephen JH Yang[12], PhD; Shih-Hua Lin[2,13], MD

[1]Division of Cardiology, Department of Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[2]Graduate Institute of Life Sciences, National Defense Medical Center, Taipei, Taiwan

[3]School of Public Health, National Defense Medical Center, Taipei, Taiwan

[4]Department of Research and Development, National Defense Medical Center, Taipei, Taiwan

[5]Department of Family and Community Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[6]Planning and Management Office, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[7]Department of Emergency Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[8]Graduate Institute of Injury Prevention and Control, College of Public Health and Nutrition, Taipei Medical University, Taipei, Taiwan

[9]Division of Cardiovascular Surgery, Department of Surgery, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[10]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

[11]Department of Medicine, Providence St Vincent Medical Center, Portland, OR, United States

[12]Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan

[13]Division of Nephrology, Department of Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[*]these authors contributed equally

Corresponding Author:
Shih-Hua Lin, MD
Division of Nephrology, Department of Medicine
Tri-Service General Hospital
National Defense Medical Center
No 325, Section 2, Cheng-Kung Road, Neihu 114
Taipei, 114
Taiwan
Phone: 886 +886287927134
Email: l521116@gmail.com

Abstract

Background: The detection of dyskalemias—hypokalemia and hyperkalemia—currently depends on laboratory tests. Since cardiac tissue is very sensitive to dyskalemia, electrocardiography (ECG) may be able to uncover clinically important dyskalemias before laboratory results.

Objective: Our study aimed to develop a deep-learning model, ECG12Net, to detect dyskalemias based on ECG presentations and to evaluate the logic and performance of this model.

Methods: Spanning from May 2011 to December 2016, 66,321 ECG records with corresponding serum potassium (K^+) concentrations were obtained from 40,180 patients admitted to the emergency department. ECG12Net is an 82-layer convolutional neural network that estimates serum K^+ concentration. Six clinicians—three emergency physicians and three cardiologists—participated in human-machine competition. Sensitivity, specificity, and balance accuracy were used to evaluate the performance of ECG12Net with that of these physicians.

Results: In a human-machine competition including 300 ECGs of different serum K+ concentrations, the area under the curve for detecting hypokalemia and hyperkalemia with ECG12Net was 0.926 and 0.958, respectively, which was significantly better than that of our best clinicians. Moreover, in detecting hypokalemia and hyperkalemia, the sensitivities were 96.7% and 83.3%,

respectively, and the specificities were 93.3% and 97.8%, respectively. In a test set including 13,222 ECGs, ECG12Net had a similar performance in terms of sensitivity for severe hypokalemia (95.6%) and severe hyperkalemia (84.5%), with a mean absolute error of 0.531. The specificities for detecting hypokalemia and hyperkalemia were 81.6% and 96.0%, respectively.

Conclusions: A deep-learning model based on a 12-lead ECG may help physicians promptly recognize severe dyskalemias and thereby potentially reduce cardiac events.

Introduction

Dyskalemias—hyperkalemia and hypokalemia—are common causes of sudden cardiac death in clinical practice [1]. Prompt recognition and rapid correction of these potassium (K^+) derangements are needed to prevent catastrophic outcomes [2]. Currently, the detection of dyskalemia relies on laboratory tests. Point-of-care blood testing provides rapid analysis of electrolyte levels, however, its accuracy and precision may not be as reliable as that from a clinical central laboratory; this is mainly due to dilution, which would underestimate plasma K^+ concentration, and the inability to discern hemolysis from pseudohyperkalemia [3,4]. Electrocardiography (ECG) is universally needed in patients with emergent cardiac or noncardiac conditions, which may exhibit the typical changes seen in dyskalemia since cardiac tissue is very sensitive to this disease. The main ECG changes associated with hypokalemia include a decreased T wave amplitude, ST-segment depression, T wave inversion, a prolonged PR interval, and an increased corrected QT interval (QTc) [5]. The typical ECG findings for hyperkalemia progress from tall *peaked* T waves and a shortened QT interval to a lengthened PR interval and a loss of the P wave, followed by a widening QRS complex and ultimately a *sine wave* morphology [5,6]. Although these morphologic changes are well known in dyskalemias, even experienced clinicians frequently do not notice all of these subtle details [7].

Previous researchers have developed ECG quantification algorithms to predict serum K^+ concentration based on T wave morphology, mainly using the slope and width of T waves. Hyperkalemia is associated with tall, narrow, and symmetrical T waves, whereas hypokalemia is associated with flat T waves [8-12]. The algorithms were mostly derived from continuous patient monitoring, such as during hemodialysis, with homogeneous ECG morphologies from a limited set of patients [8-12]. Recently, applying the processing of T wave morphologies manually has been used to improve the diagnosis of hyperkalemia [13]. Nevertheless, using T wave changes alone to detect dyskalemias is less sensitive and specific than a comprehensive ECG interpretation [14].

With the revolution in artificial intelligence (AI), several advanced deep-learning models, such as Oxford's VGGNet [15], Inception Net [16], ResNet [17], and DenseNet [18], have been developed, providing an unprecedented opportunity to improve health care; this was initiated by AlexNet's victory in the ImageNet Large Scale Visual Recognition Challenge in 2012 [19]. Existing deep-learning models have been shown to achieve human-level performance and be effective in medical applications when large annotated datasets are available [17,20-22]. This potential to improve diagnosis and patient care prompted us to develop a deep-learning model to assist emergency physicians in recognizing ECG changes associated with dyskalemias.

Our study aimed to train a deep-learning model, ECG12Net, to predict serum K^+ concentration by ECG. The deep-learning model was an 82-layer convolutional neural network that underwent a series of training processes to optimize model performance. The AI system, which will learn from more than 50,000 electrocardiograms to identify critical morphologic changes, will help to reduce medical errors in emergency departments (EDs) resulting from intense time pressure and harried ED staff during busy periods in ED environments [23]. Facilitated by the system's powerful computing ability, the performance of the trained model was compared with that of emergency physicians and cardiologists. Finally, we visualized ECG12Net's calculation process to understand why and how it works.

Methods

Data Source

The data were obtained from Tri-Service General Hospital, Taiwan, and research approval was given by the Institutional Review Board (IRB) (IRB No. 1-107-05-047). From May 11, 2011, to December 31, 2016, 40,180 emergency patients were enrolled who had 66,321 ECG records within 1 hour before or after serum K^+ concentration for reference. Serum K^+ concentrations were measured in the laboratory using indirect ion-selective electrode methods that had been accredited by the International Organization for Standardization (ISO) standard ISO-15189 and the College of American Pathologists' Laboratory Accreditation Program. All hemolyzed samples were excluded. Potential confounders, such as patients with chest pain or thyroid disorders, were not excluded from the study. We divided the dataset into training (~70%), validation (~10%), and test (~20%) sets by date. Emergency patients presenting before April 30, 2016, were included in the training set; those presenting between May 1 and July 20, 2016, were in the validation set; and those presenting after July 21, 2016, were in the test set to assess model performance. All records included in the training set were excluded from the validation and test sets; thus, there was no overlap among the three datasets. The ECG recordings were collected using a Philips 12-Lead ECG machine (PH080A). The ECG signal was recorded

in a digital format. The sampling frequency was 500 Hz with 2.5 seconds recorded in each lead. The estimated K^+ concentrations ranged from 1.5 mEq/L to 7.5 mEq/L. Predicted K^+ concentrations less than 1.5 mEq/L or greater than 7.5 mEq/L were indicated accordingly without further detail (ie, as either <1.5 mEq/L or >7.5 mEq/L). Patient characteristics and laboratory results were collected using an electronic health record system. The estimated glomerular filtration rate was calculated using the Chronic Kidney Disease Epidemiology Collaboration formula [24]. Eight basic ECG morphology parameters (EMPs) were calculated by the Philips 12-Lead ECG machine: heart rate, PR interval, QRS duration, QT interval, QTc, P wave axis, RS wave axis, and T wave axis.

The Implementation of ECG12Net

We developed a 12-channel sequence-to-sequence model, which is modified from DenseNet [18]. The details are shown in Multimedia Appendix 1. The architecture of ECG12Net is shown in Figure 1. We designed an ECG lead block with 80 trainable layers whose architecture is shown in Figure 1 A. This ECG lead block was used to extract 864 features from each ECG lead, making a basic output prediction based on each lead. Figure 1 B shows how ECG12Net integrates all the information from the ECG leads to make an overall prediction. ECG12Net is composed of 12 of these ECG lead blocks corresponding to each lead sequence. We designed an attention mechanism based on a hierarchical attention network to concatenate these blocks, increasing the interpretive power of ECG12Net [25]. ECG12Net-1, which uses only ECG wave information, contains 82 trainable layers. To improve prediction performance, we added an EMPNet, which is a multilayer perceptron with two hidden layers containing eight EMPs, to ECG12Net-1 to create ECG12Net-2.

Figure 1. Architecture of ECG12Net. A. Electrocardiography (ECG) lead block with 80 trainable layers. B. ECG12Net integrates all the information from the ECG leads to make an overall prediction. The bolded and colored words indicate the output dimensions of the layers, and the words in black are the required parameters for the layers. conv: convolution; BN: batch normalization; ReLU: rectified linear unit; FC: fully connected; MLP: multilayer perceptron.
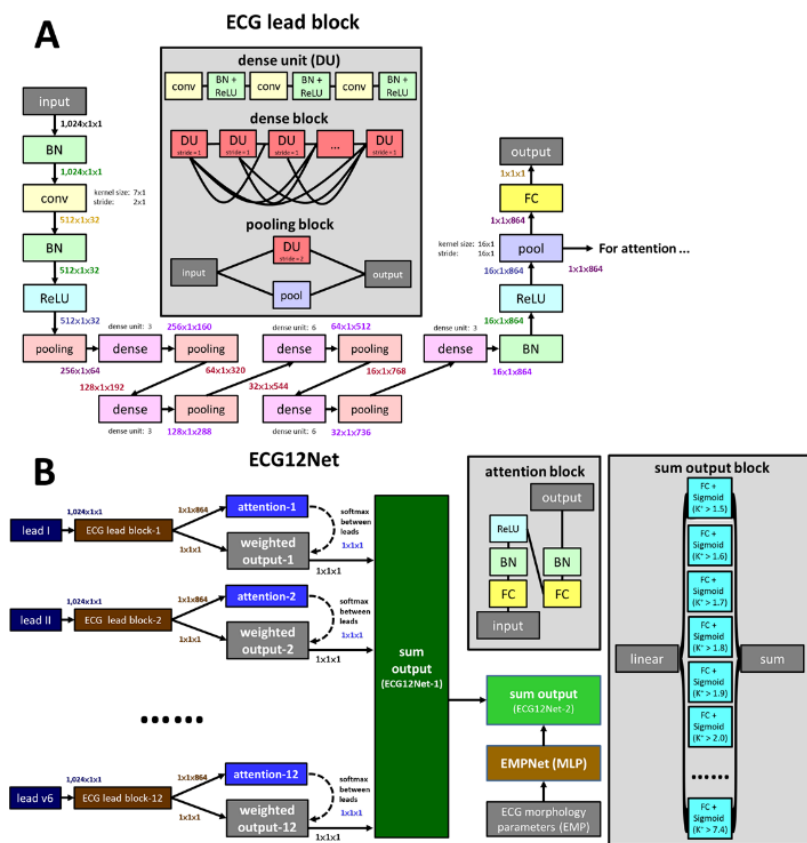
Human-Machine Competition

We evaluated the performance of practicing physicians using a subtest set. We divided the data into five categories based on the serum K^+ concentration: (1) K^+ ≤2.5 mEq/L, (2) 2.5< K^+ ≤3.5 mEq/L, (3) 3.5< K^+ <5.5 mEq/L, (4) 5.5≤ K^+ <6.5 mEq/L, and (5) K^+ ≥6.5 mEq/L. Stratified sampling was used to create the subtest set due to the rarity of cases in the first and fifth categories. Each category of K^+ concentration comprised 60 cases, and a total of 300 cases were used in the test. The participating physicians included an emergency physician under training (second-year resident); two emergency physicians, one with 4 and the other with 13 years of experience; a chief resident in cardiology; and two cardiologists, one with 2 and the other with 9 years of experience. The physicians had no access to patient information and no knowledge of the data. The responses they provided were entered into an online standardized data entry program. We calculated their sensitivity and specificity and compared their results with those of ECG12Net.

Statistical Analysis and Model Performance Assessment

The study cohort was divided into training, validation, and test sets. We presented their characteristics as the means and standard deviations, the numbers of patients, or the percentages, where appropriate. This information was compared using either analysis of variance or the chi-square test as appropriate. We then analyzed the EMP differences between the five serum K^+ groups, and the EMPs were subjected to post hoc analysis. All the dyskalemia groups were compared to the normal group.

The primary analysis was done to evaluate the performance in dyskalemia prediction between ECG12Net and the clinicians in a machine-human competition. Receiver operating characteristic curves and the areas under the curve (AUCs) were applied to evaluate the competition results. Additionally, the sensitivity, specificity, and balance accuracy of dyskalemia prediction by ECG12Net and the clinical physicians were calculated. The balance accuracy is defined as the mean of the sensitivity and specificity obtained in the study. Due to the stratified sampling process destroying the original prevalence, the positive predictive value and negative predictive value for the competition results are not presented.

The secondary analyses were performed on our test set with the data obtained after July 21, 2016, which had not been used in the training process. This was a simulated prospective study to evaluate the performance of the AI models with the mean absolute error (MAE) as the major measurement index due to the continuous predictions. Moreover, categorized analyses are also presented. Sensitivity, specificity, positive predictive value, negative predictive value, and the squared weighted kappa were used to evaluate the performance of the models. Finally, we conducted a series of logistic models to identify the effects of patient demographic characteristics on the performance of our deep-learning model.

We used a significance level of $P<$ throughout the analysis. Bootstrap 95% CIs were calculated and presented for all measure indexes based on 10,000 permutations. No additional adjustments for multiple comparisons were used because of the small number of planned comparisons. The statistical analysis was carried out using the software environment R, version 3.4.3 (The R Foundation).

Results

Cohort Description

The training, validation, and test sets comprised records from 28,183; 3993; and 8004 patients, respectively. Table 1 shows the patient characteristics, which reveal similar distributions among the sets of gender, age, body mass index, marital status, education, and underlying comorbidities, including diabetes mellitus, coronary artery disease, hypertension, heart failure, hyperlipidemia, chronic kidney disease, chronic obstructive pulmonary disease, and pneumothorax. The training, validation, and test sets consisted of 46,692; 6407; and 13,222 pairs, respectively, of ECGs and K^+ concentrations. The details of the laboratory and EMP analyses are presented in Multimedia Appendix 1. The detailed dyskalemia distribution (see Multimedia Appendix 1) shows a hypokalemia/hyperkalemia prevalence of 22.7%/2.6%, 22.9%/2.3%, and 22.7%/2.8% in the training, validation, and test sets, respectively.

Table 1. Patients' characteristics in the training, validation, and test sets.

Characteristic	Training set (N=28,183)	Validation set (N=3993)	Test set (N=8004)	P value
Gender, n (%)				.08
Female	13,828 (49.07)	1942 (48.64)	3814 (47.65)	
Male	14,350 (50.92)	2049 (51.31)	4190 (52.35)	
Age (years), mean (SD)	62.57 (19.45)	62.47 (19.33)	62.61 (19.25)	.93
Height (cm), mean (SD)	162.24 (9.37)	162.19 (9.58)	163.29 (36.90)	.09
Weight (cm), mean (SD)	63.98 (14.12)	64.11 (14.16)	63.75 (13.79)	.78
BMI (kg/m^2), mean (SD)	24.32 (6.38)	24.39 (6.71)	24.07 (4.49)	.24
Underlying comorbidities, n (%)				
Diabetes mellitus	3553 (12.61)	476 (11.92)	1009 (12.61)	.47
Coronary artery disease	1694 (6.01)	257 (6.44)	485 (6.06)	.57
Hypertension	5219 (18.52)	741 (18.56)	1496 (18.69)	.94
Heart failure	825 (2.93)	124 (3.11)	239 (2.99)	.81
Hyperlipidemia	3868 (13.72)	520 (13.02)	1078 (13.47)	.45
Chronic kidney disease	6294 (22.33)	859 (21.51)	1786 (22.31)	.50
Chronic obstructive pulmonary disease	1351 (4.79)	193 (4.83)	408 (5.10)	.54
Pneumothorax	88 (0.31)	11 (0.28)	24 (0.30)	.92

Primary Analysis

The results of the human-machine competition are summarized in Figure 2. The AUCs of our ECG12Net-1 were 0.993, 0.926, 0.958, and 0.976 in the detection of severe hypokalemia, hypokalemia, hyperkalemia, and severe hyperkalemia, respectively. Due to the continuous nature of the K^+ concentration predictions from ECG12Net, we used clinical cut points as described in the Methods section for further analysis. Our clinicians detected severe hypokalemia with sensitivities and specificities of 45%-78.3% and 74.4%-83.9%, respectively, whereas ECG12Net-1 achieved a sensitivity of 96.7% (95% CI 91.7-100.0) and a specificity of 93.3% (95% CI 89.4-96.7). In detecting severe hyperkalemia, the clinicians had nearly perfect specificity (92.8%-100.0%) but low sensitivity (16.7%-43.3%), while ECG12Net-1 exhibited a sensitivity of 83.3% (95% CI 73.3-91.7) and a specificity of 97.8% (95% CI 95.6-99.4). Including mild-to-moderate dyskalemias, ECG12Net-1 had the highest sensitivity in detecting hypokalemia (67.5%, 95% CI 59.2-75.8) and hyperkalemia (67.5%, 95% CI 59.2-75.8) in the human-machine competition. The details of the human-machine competition are shown in Table 2. In terms of balance accuracy, ECG12Net-1's performance was significantly better than that of the best clinician (cardiologist 2) participating in the hypokalemia detection (80.4%, 95% CI 75.7-84.9, vs 66.7%, 95% CI 61.4-72.1). In detecting severe hyperkalemia, the balance accuracy of ECG12Net-1 was also significantly better than that of the best clinician (cardiologist 3) (82.7%, 95% CI 78.2-86.8, vs 70.6%, 95% CI 65.6-75.4). Although ECG12Net-2 exhibited lower performance compared with ECG12Net-1, it performed much better than all of the clinicians. The results of the consistency analysis are shown in Multimedia Appendix 1. When inconsistency arose between the predictions made by ECG12Net and the experts, ECG12Net was approximately 3.85 times more likely to be correct (P<.001 based on the McNemar test).

Figure 2. Performance comparison in detecting dyskalemias from the human-machine competition (n=300). The receiver operating characteristic curves are made by the predictions of ECG12Net-1. The red triangles and blue circles indicate emergency physicians and cardiologists, respectively, in the human-machine competition. K^+ ≤2.5 mEq/L, 2.5< K^+ ≤3.5 mEq/L, 3.5< K^+ <5.5 mEq/L, 5.5≤ K^+ <6.5 mEq/L, and K^+ ≥6.5 mEq/L were defined as severe hypokalemia (n=60), hypokalemia (n=120), normal (n=60), hyperkalemia (n=120), and severe hyperkalemia (n=60), respectively. AUC: area under the curve.
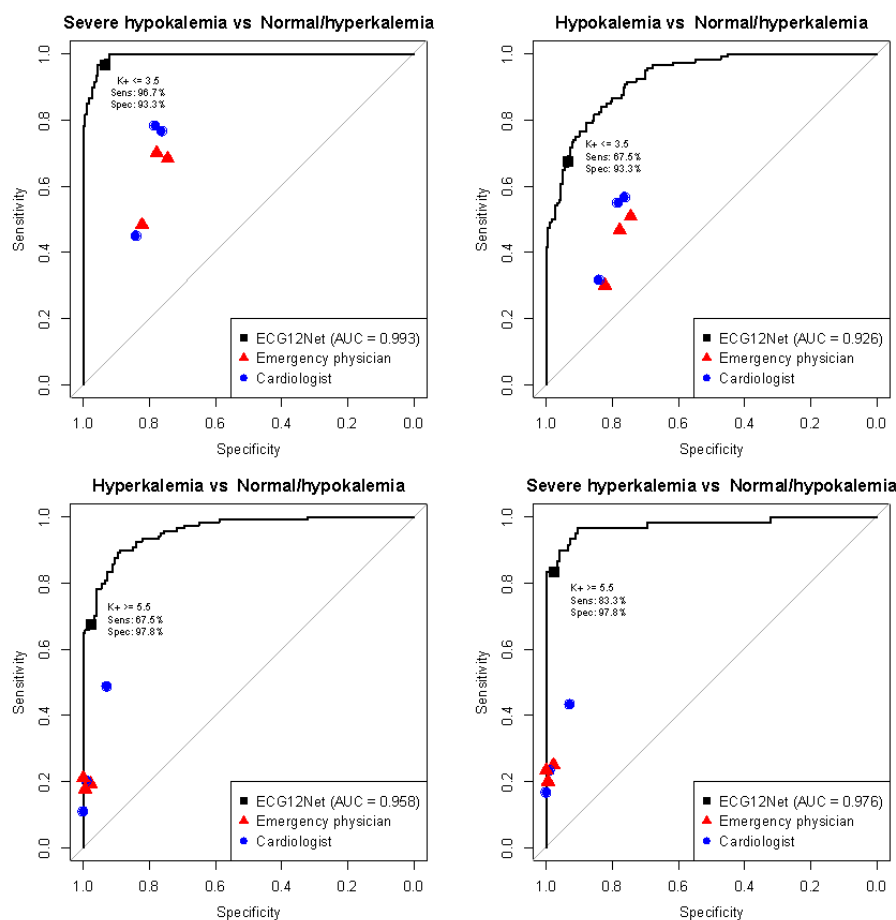
Table 2. Comparison between human experts and ECG12Net on the sensitivity and specificity in the subtest set (n=300).

Type of dyskalemia	Sensitivity[a], 95% CI			Specificity[a] (n=180), 95% CI	Balance accuracy[b], 95% CI
	Overall (n=120)	Severe (n=60)	Mild to moderate (n=60)		
Hypokalemia (K$^+$≤3.5 mEq/L)					
Emergency physician 1[c]	0.300 (0.219-0.385)	0.483 (0.356-0.613)	0.117 (0.040-0.206)	0.822 (0.765-0.875)	0.561 (0.512-0.611)
Emergency physician 2[d]	0.508 (0.420-0.598)	0.683 (0.562-0.797)	0.333 (0.217-0.455)	0.744 (0.680-0.807)	0.626 (0.572-0.682)
Emergency physician 3[e]	0.467 (0.378-0.554)	0.700 (0.581-0.812)	0.233 (0.131-0.345)	0.778 (0.717-0.835)	0.622 (0.569-0.676)
Cardiologist 1[f]	0.317 (0.236-0.403)	0.450 (0.323-0.579)	0.183 (0.091-0.288)	0.839 (0.782-0.892)	0.578 (0.528-0.628)
Cardiologist 2[g]	0.550 (0.462-0.637)	0.783 (0.673-0.885)	0.317 (0.204-0.439)	0.783 (0.722-0.842)	0.667 (0.614-0.721)
Cardiologist 3[h]	0.567 (0.477-0.654)	0.767 (0.654-0.870)	0.367 (0.246-0.492)	0.761 (0.697-0.820)	0.664 (0.608-0.718)
ECG12Net-1	0.675 (0.592-0.758)	0.967 (0.917-1.000)	0.383 (0.267-0.500)	0.933 (0.894-0.967)	0.804 (0.757-0.849)
ECG12Net-2	0.675 (0.592-0.758)	0.967 (0.917-1.000)	0.383 (0.267-0.500)	0.922 (0.883-0.961)	0.799 (0.751-0.843)
Hyperkalemia (K$^+$≥5.5 mEq/L)					
Emergency physician 1	0.192 (0.124-0.266)	0.250 (0.145-0.365)	0.133 (0.053-0.224)	0.978 (0.954-0.995)	0.585 (0.549-0.623)
Emergency physician 2	0.175 (0.110-0.244)	0.200 (0.103-0.304)	0.150 (0.065-0.250)	0.994 (0.982-1.000)	0.585 (0.552-0.620)
Emergency physician 3	0.208 (0.137-0.282)	0.233 (0.130-0.344)	0.183 (0.089-0.288)	1.000 (1.000-1.000)	0.604 (0.569-0.641)
Cardiologist 1	0.108 (0.056-0.167)	0.167 (0.077-0.266)	0.050 (0.000-0.113)	1.000 (1.000-1.000)	0.554 (0.528-0.583)
Cardiologist 2	0.200 (0.131-0.274)	0.233 (0.132-0.345)	0.167 (0.078-0.265)	0.989 (0.971-1.000)	0.594 (0.560-0.632)
Cardiologist 3	0.483 (0.393-0.571)	0.433 (0.305-0.558)	0.533 (0.403-0.661)	0.928 (0.888-0.963)	0.706 (0.656-0.754)
ECG12Net-1	0.675 (0.592-0.758)	0.833 (0.733-0.917)	0.517 (0.383-0.633)	0.978 (0.956-0.994)	0.827 (0.782-0.868)
ECG12Net-2	0.683 (0.600-0.767)	0.833 (0.733-0.917)	0.533 (0.400-0.650)	0.972 (0.944-0.994)	0.828 (0.783-0.869)

[a]The test provides three selections for prediction: hypokalemia (K$^+$ ≤3.5 mEq/L), normokalemia (3.5 mEq/L< K$^+$ <5.5 mEq/L), and hyperkalemia (K$^+$ ≥5.5 mEq/L).

[b]The balance accuracy value represents the average of the overall sensitivity and specificity.

[c]Emergency physician 1: second-year resident.

[d]Emergency physician 2: 4 years of experience.

[e]Emergency physician 3: 13 years of experience.

[f]Cardiologist 1: chief resident of cardiology.

[g]Cardiologist 2: 2 years of experience.

[h]Cardiologist 3: 9 years of experience.

Performance of ECG12Net on the Test Set

The model performance on the test set is shown in Multimedia Appendix 1. The performance of ECG12Net was better than that of each lead. ECG12Net-1 had the lowest MAE (0.531). Including EMP information did not improve the prediction of K$^+$ concentration (MAE ECG12Net-1: 0.531; MAE ECG12Net-2: 0.538). When categorizing among three

classes—hypokalemia, normokalemia, and hyperkalemia—and five classes, with the addition of severe hypokalemia and severe hyperkalemia, as described in Multimedia Appendix 1, a similar performance was observed by ECG12Net-1; this demonstrated the highest squared weighted kappa of 0.354 in the three-class categorization and 0.396 in the five-class categorization. For the detection of hypokalemia, the sensitivity, specificity, positive predictive value, and negative predictive value of ECG12Net-1

were 50.7%, 81.6%, 44.7%, and 85.0%, respectively; for hyperkalemia, they were 50.8%, 96.0%, 26.9%, and 98.5%, respectively. The confusion scatter plots for the predictions by the two ECG12Nets are shown in Figure 3. Importantly, in detecting severe hypokalemia and hyperkalemia, ECG12Net-1 demonstrated a sensitivity of 95.6% and 84.5%, respectively. ECG12Net-2 exhibited similar prediction capabilities for severe hypokalemia and hyperkalemia as ECG12Net-1.

Figure 3. Confusion scatter plots of ECG12Net-1 and ECG12Net-2 predictions on the test set (n=13,222). The x-axis indicates the true K+ concentration from laboratory testing. The y-axis presents the predicted K+ concentration by ECG12Net-1 (A) and ECG12Net-2 (B). Red points represent the highest density, followed by yellow, green, light blue, and dark blue. Perfect model performance would fall only along the red diagonal line. We categorized the K+ concentration into five groups (K$^+$ ≤2.5 mEq/L, 2.5< K$^+$ ≤ 3.5 mEq/L, 3.5< K$^+$ <5.5 mEq/L, 5.5≤ K$^+$ <6.5 mEq/L, and K$^+$ ≥6.5 mEq/L) and calculated the case counts in each grid.
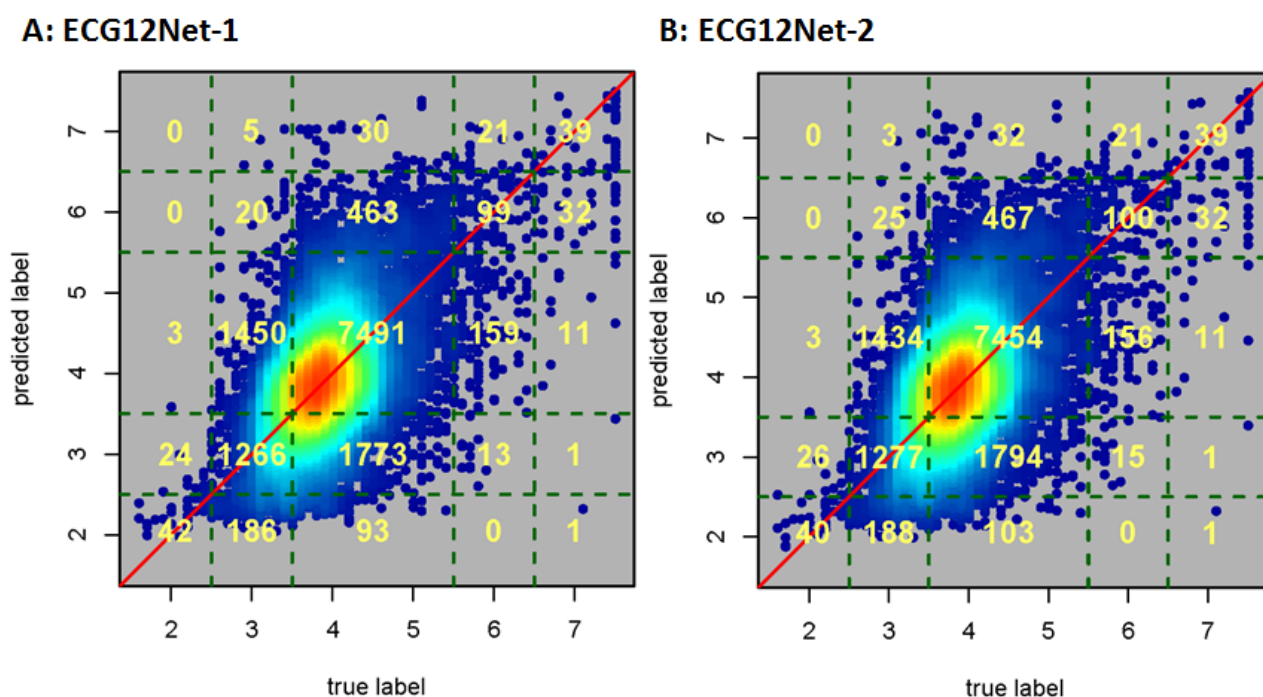
Model Interpretation

A total of 58 severe hypokalemia cases were correctly detected by ECG12Net-1, of which 15 (26%) were overlooked by clinician consensus. The classical ECG findings of U wave and ST segment depression, especially in leads V2 and V3, were consistently recognized as severe hypokalemia by both the clinicians and ECG12Net-1 (see Figure 4 A). As shown in Figure 4 B, ECG12Net-1 predicted a case of severe hypokalemia from ST segment depression in the V3 lead; this case was misdiagnosed by all the clinicians. Two cases of severe hypokalemia were misclassified by ECG12Net-1 but diagnosed correctly by the clinicians (data not shown). These cases had

severe noise in the presented ECG; however, the clinicians made the correct diagnosis based on the presence of a prolonged QTc.

A total of 50 severe hyperkalemia cases were correctly detected by ECG12Net-1, with 36 (72%) of these cases overlooked by clinician consensus. Figure 4 C shows a typical ECG presentation of severe hyperkalemia with tented T waves accompanied by a long QRS complex duration, which was correctly diagnosed by all clinicians and ECG12Net-1. Figure 4 D shows a case of severe hyperkalemia correctly recognized by ECG12Net-1, with ST depression followed by a peaked T wave in lead V6, which was misdiagnosed as hypokalemia by all the clinicians. There were also 10 cases of severe hyperkalemia overlooked by ECG12Net-1 and all clinicians.

XSL•FO
RenderX

Figure 4. Visualization analysis for ECG12Net-1 in selected severe hypokalemia and hyperkalemia cases. The lighter areas (green or yellow) indicate areas of focus by ECG12Net-1. Clinicians consistently recognized panel A as a typical case of severe hypokalemia but overlooked panel B. Similarly, clinicians consistently recognized panel C as severe hyperkalemia but overlooked panel D. From A to D, the real K^+ concentrations were 2.3 mEq/L, 2.5 mEq/L, 9.1 mEq/L, and 7.1 mEq/L, respectively. AI: artificial intelligence.
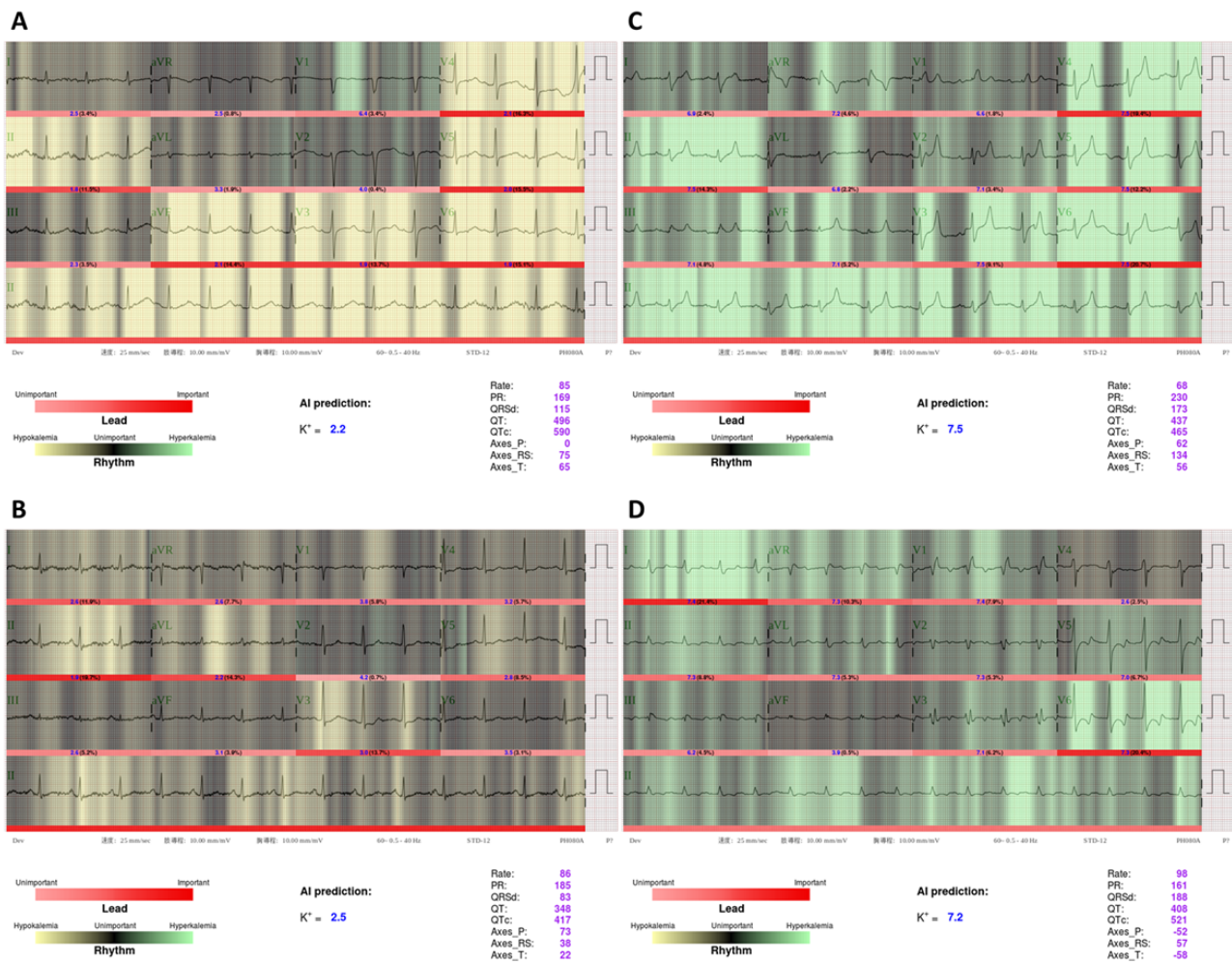
Discussion

In this study, we developed a deep-learning model, ECG12Net, to detect dyskalemias through ECG analysis. Using a deep convolutional network extracting many useful ECG features with a training set of more than 50,000 ECGs, ECG12Net performed better than clinicians in detecting dyskalemias. Notably, ECG12Net performed well with sensitivities of 95.6% and 84.5% in detecting severe hypokalemia and severe hyperkalemia, respectively.

ECG interpretation is one of the most important skills in medical practice. Previous studies have analyzed morphological features, for instance, the R wave peak [26] and the QRS complex [27], combined with machine learning approaches for disease detection, such as atrial fibrillation [28]. These systems were relatively imprecise, making it troublesome to quantify specific rhythm morphologies [29]. Although some recent studies have used deep convolutional neural networks and recurrent neural networks mainly for arrhythmia detection [30-35], most of the data were collected from wearable devices without offering all the important information provided by a 12-lead ECG [11]. The clinical value of these findings is also dampened by the lack of

laboratory-based diagnosis and annotation and the relatively small volumes of data. In contrast, our database was unprecedented, comprising 40,180 patients and 66,321 laboratory-annotated ECG records collected by standard 12-lead ECG machines.

Galloway et al recently developed a deep-learning model to screen for hyperkalemia in patients with chronic kidney disease, stage III or higher, using ECG [36]. We applied ECG12Net to a broad set of patients in the ED and developed a continuous prediction of both hypokalemia and hyperkalemia. Moreover, although the three-category classification task in our study is more difficult than the two-category classification task in theirs, our ECG12Net achieved an AUC greater than 0.9 in detecting hyperkalemia, which is similar to that of their model with an AUC of 0.85-0.88. This highlights the strength of ECG12Net.

The EMPs of different K^+ concentration groups yielded several interesting findings. The EMPs, such as the PR and QTc intervals, and the data used for analysis were all collected from the original ECGs (see Multimedia Appendix 1). The impact of hyperkalemia on the T wave axis was more profound and substantial than the axes of the P and RS waves. Hypokalemia

was actually associated with a widening of the QRS complex, which may be explained by the decrease in conduction velocity caused by reduced K^+ concentrations after hemodialysis [37]. Although the longest QTc occurred in the severe hypokalemia group, a well-documented finding, the QTc was longer in patients with hyperkalemia as well. In fact, for most of the intervals and durations, the nadir was in normokalemia, with increases on both forms of dyskalemia. Although the underlying mechanisms are unclear, these findings uncovered by big data may guide directions for further research.

Interestingly, the algorithm focusing only on morphologic changes (ie, ECG12Net-1) performed slightly better than that with additional EMP information (ie, ECG12Net-2). That the addition of EMP information did not improve the model's predictive ability corroborates prior research that found that deep-learning models can automatically extract useful features for prediction without preprocessing [17,20,21]. This also highlights the importance of morphologic changes in ECG over EMPs in the detection of dyskalemias.

There are several clinical applications of ECG12Net shown in Multimedia Appendix 1. First, severe dyskalemia could be identified by ECG12Net within 5 minutes, much faster than laboratory testing, leading to more prompt management. Second, pseudodyskalemia, defined as an abnormal reported serum or plasma K^+ concentration despite a normal in vivo K^+ concentration, can be excluded early by ECG12Net to avoid inappropriate treatment. Third, the performance of ECG12Net is more than 10% better than that of the best cardiologist in our study, whose performance was similar to other experts in prior studies [38,39]. This means that emergency physicians could have access to a consistent, beyond cardiologist-level decision aid available 24 hours a day to help diagnose and manage dyskalemic patients. Fourth, the developed ECG12Net model can be included in a wearable device for dyskalemia detection, especially for patients with advanced chronic kidney disease or uremia on dialysis. Finally, the ECG12Net model could be incorporated into ECG machines in ambulances or remote areas to facilitate telemedicine.

Explainable AI plays a critical role in clinical practice [40,41]. The so-called "black box" approach in the deep-learning models often precludes the understanding of the decision-making process [42]. To increase the interpretability of our model, we established heatmaps to visualize the focus in the ECG by ECG12Net using class activation mappings [25,43], which can help physicians understand the logic of the AI decisions. Although our ECG12Net was approximately 3.85 times more likely to be correct when inconsistencies occurred between the AI and human predictions (see Multimedia Appendix 1), physicians who can integrate the AI suggestions with the symptoms and signs of patients should make the final decision to take appropriate action.

Some limitations of this study should be mentioned. First, the studied patients were only enrolled from one academic medical center, despite the similar distribution of blood K^+ concentration in other large studies [44,45]. Multicenter validation is needed to confirm the value and application of this study. Second, only six clinicians participated in the competition with ECG12Net's performance. Although their performance in severe hyperkalemia detection was consistent with that of the previous studies [38,39], comparisons should be made with more experts to confirm the superiority of ECG12Net. Third, only the patients in the ED with both an ECG and a serum K^+ test were enrolled in this study, which may have caused selection bias and constrained the generalizability of the results. Fourth, although the sensitivity heatmap provides a glimpse into the basis for ECG12Net's prediction, the reason why the particular ECG segment was highlighted remains unclear. Finally, ECG12Net showed decreased sensitivity in detecting mild-to-moderate hypokalemia, which accounts for the majority of dyskalemias, leading to low *weighted averages* of the sensitivities. Hypokalemia-associated ECG changes usually occur when the serum K^+ level falls below 3 mEq/L [46], which may explain why our algorithm failed to accurately distinguish the ECG morphologies of mild-to-moderate hypokalemia from normokalemia.

In conclusion, we established a deep-learning model called ECG12Net to detect dyskalemias in the ED. The collaboration between physicians and AI can lead to better health care for our patients. This model will help emergency physicians promptly recognize severe dyskalemias and potentially reduce sudden cardiac death.

Acknowledgments

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 2261 KB - medinform_v8i3e15931_app1.docx]

References

1. No authors listed. Editorial: Slow-K, quick quick, slow. Lancet 1974 Nov 09;2(7889):1123-1124. [Medline: 4139417]

2. Priori SG, Blomström-Lundqvist C, Mazzanti A, Blom N, Borggrefe M, Camm J, ESC Scientific Document Group. 2015 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: The Task Force for the Management of Patients with Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death of the European Society of Cardiology (ESC). Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC). Eur Heart J 2015 Nov 01;36(41):2793-2867. [doi: 10.1093/eurheartj/ehv316] [Medline: 26320108]

3. Gavala A, Myrianthefs P. Comparison of point-of-care versus central laboratory measurement of hematocrit, hemoglobin, and electrolyte concentrations. Heart Lung 2017;46(4):246-250. [doi: 10.1016/j.hrtlng.2017.04.003] [Medline: 28477952]

4. Dylewski JF, Linas S. Variability of potassium blood testing: Imprecise nature of blood testing or normal physiologic changes? Mayo Clin Proc 2018 May;93(5):551-554. [doi: 10.1016/j.mayocp.2018.03.019] [Medline: 29728197]

5. Diercks DB, Shumaik GM, Harrigan RA, Brady WJ, Chan TC. Electrocardiographic manifestations: Electrolyte abnormalities. J Emerg Med 2004 Aug;27(2):153-160. [doi: 10.1016/j.jemermed.2004.04.006] [Medline: 15261358]

6. Slovis C, Jenkins R. ABC of clinical electrocardiography: Conditions not primarily affecting the heart. BMJ 2002 Jun 01;324(7349):1320-1323 [FREE Full text] [doi: 10.1136/bmj.324.7349.1320] [Medline: 12039829]

7. Van Mieghem C, Sabbe M, Knockaert D. The clinical value of the ECG in noncardiac conditions. Chest 2004 Apr;125(4):1561-1576. [doi: 10.1378/chest.125.4.1561] [Medline: 15078775]

8. Dillon JJ, DeSimone CV, Sapir Y, Somers VK, Dugan JL, Bruce CJ, et al. Noninvasive potassium determination using a mathematically processed ECG: Proof of concept for a novel "blood-less, blood test". J Electrocardiol 2015;48(1):12-18 [FREE Full text] [doi: 10.1016/j.jelectrocard.2014.10.002] [Medline: 25453193]

9. Attia ZI, DeSimone CV, Dillon JJ, Sapir Y, Somers VK, Dugan JL, et al. Novel bloodless potassium determination using a signal-processed single-lead ECG. J Am Heart Assoc 2016 Jan 25;5(1):e002746 [FREE Full text] [doi: 10.1161/JAHA.115.002746] [Medline: 26811164]

10. Greenlee M, Wingo CS, McDonough AA, Youn J, Kone BC. Narrative review: Evolving concepts in potassium homeostasis and hypokalemia. Ann Intern Med 2009 May 05;150(9):619-625 [FREE Full text] [doi: 10.7326/0003-4819-150-9-200905050-00008] [Medline: 19414841]

11. Corsi C, Cortesi M, Callisesi G, De Bie J, Napolitano C, Santoro A, et al. Noninvasive quantification of blood potassium concentration from ECG in hemodialysis patients. Sci Rep 2017 Feb 15;7:42492 [FREE Full text] [doi: 10.1038/srep42492] [Medline: 28198403]

12. Corsi C, DeBie J, Napolitano C, Priori S, Mortara D, Severi S. Validation of a novel method for non-invasive blood potassium quantification from the ECG. Comput Cardiol 2012;39:105-108 [FREE Full text]

13. Velagapudi V, O'Horo JC, Vellanki A, Baker SP, Pidikiti R, Stoff JS, et al. Computer-assisted image processing 12 lead ECG model to diagnose hyperkalemia. J Electrocardiol 2017;50(1):131-138. [doi: 10.1016/j.jelectrocard.2016.09.001] [Medline: 27662777]

14. Montague BT, Ouellette JR, Buller GK. Retrospective review of the frequency of ECG changes in hyperkalemia. Clin J Am Soc Nephrol 2008 Mar;3(2):324-330 [FREE Full text] [doi: 10.2215/CJN.04611007] [Medline: 18235147]

15. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). 2015 Presented at: 3rd International Conference on Learning Representations (ICLR 2015); May 7-9, 2015; San Diego, CA p. 1-14 URL: https://arxiv.org/pdf/1409.1556.pdf

16. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). 2015 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015); June 7-12, 2015; Boston, MA p. 1-9. [doi: 10.1109/cvpr.2015.7298594]

17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016); June 26-July 1, 2016; Las Vegas, NV p. 770-778. [doi: 10.1109/cvpr.2016.90]

18. Huang G, Liu Z, Weinberger KQ, van der Maaten L. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017); July 21-26, 2017; Hawaii, HI p. 2261-2269. [doi: 10.1109/cvpr.2017.243]

19. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90 [FREE Full text] [doi: 10.1145/3065386]

20. Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A. Microsoft. 2017 Feb. Achieving human parity in conversational speech recognition URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/ms_parity.pdf [accessed 2020-02-21]

21. Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin. In: Proceedings of the International Conference on Machine Learning (ICML 2016). 2016 Presented at: International Conference on Machine Learning (ICML 2016); June 19-24, 2016; New York, NY p. 1-10 URL: http://proceedings.mlr.press/v48/amodei16.pdf [doi: 10.1007/978-3-030-14596-5_12]

22. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 02;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]

23. Schenkel S. Promoting patient safety and preventing medical error in emergency departments. Acad Emerg Med 2000 Nov;7(11):1204-1222 [FREE Full text] [doi: 10.1111/j.1553-2712.2000.tb00466.x] [Medline: 11073469]

24. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, Feldman HI, CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A new equation to estimate glomerular filtration rate. Ann Intern Med 2009 May 05;150(9):604-612 [FREE Full text] [doi: 10.7326/0003-4819-150-9-200905050-00006] [Medline: 19414839]

25. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL). 2016 Presented at: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL); June 12-17, 2016; San Diego, CA p. 1480-1489 URL: https://www.aclweb.org/anthology/N16-1174.pdf [doi: 10.18653/v1/n16-1174]

26. Li C, Zheng C, Tai C. Detection of ECG characteristic points using wavelet transforms. IEEE Trans Biomed Eng 1995 Jan;42(1):21-28. [doi: 10.1109/10.362922] [Medline: 7851927]

27. Mukhopadhyay S, Biswas S, Roy A, Dey N. Wavelet based QRS complex detection of ECG signal. Int J Eng Res Appl 2012 May;2(3):2361-2365 [FREE Full text]

28. Zabihi M, Rad A, Katsaggelos A, Kiranyaz S, Narkilahti S, Gabbouj M. Detection of atrial fibrillation in ECG hand-held devices using a random forest classifier. Comput Cardiol 2017 Sep 24;44:1-4 [FREE Full text] [doi: 10.22489/cinc.2017.069-336]

29. Guglin ME, Thatai D. Common errors in computer electrocardiogram interpretation. Int J Cardiol 2006 Jan 13;106(2):232-237. [doi: 10.1016/j.ijcard.2005.02.007] [Medline: 16321696]

30. Rajpurkar P, Hannun A, Haghpanahi M, Bourn C, Ng A. Stanford ML Group. 2017. Cardiologist-level arrhythmia detection with convolutional neural networks URL: https://arxiv.org/pdf/1707.01836.pdf [accessed 2020-02-21]

31. Zihlmann M, Perekrestenko D, Tschannen M. Convolutional recurrent neural networks for electrocardiogram classification. Comput Cardiol 2017;44:1-4 [FREE Full text] [doi: 10.22489/cinc.2017.070-060]

32. Rubin J, Parvaneh S, Rahman A, Conroy B, Babaeizadeh S. Densely connected convolutional networks and signal quality analysis to detect atrial fibrillation using short single-lead ECG recordings. In: Proceedings of 2017 Computing in Cardiology Conference (CinC). 2017 Presented at: 2017 Computing in Cardiology Conference (CinC); September 24-27, 2017; Rennes, France p. 1-4 URL: https://arxiv.org/ftp/arxiv/papers/1710/1710.05817.pdf [doi: 10.22489/cinc.2017.160-246]

33. Acharya UR, Fujita H, Lih OS, Hagiwara Y, Tan JH, Adam M. Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. Inf Sci 2017 Sep;405:81-90. [doi: 10.1016/j.ins.2017.04.012]

34. Tan JH, Hagiwara Y, Pang W, Lim I, Oh SL, Adam M, et al. Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. Comput Biol Med 2018 Mar 01;94:19-26. [doi: 10.1016/j.compbiomed.2017.12.023] [Medline: 29358103]

35. Acharya UR, Fujita H, Oh SL, Hagiwara Y, Tan JH, Adam M. Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. Inf Sci 2017 Nov;415-416:190-198. [doi: 10.1016/j.ins.2017.06.027]

36. Galloway CD, Valys AV, Shreibati JB, Treiman DL, Petterson FL, Gundotra VP, et al. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. JAMA Cardiol 2019 May 01;4(5):428-436. [doi: 10.1001/jamacardio.2019.0640] [Medline: 30942845]

37. Severi S, Pogliani D, Fantini G, Fabbrini P, Viganò MR, Galbiati E, et al. Alterations of atrial electrophysiology induced by electrolyte variations: Combined computational and P-wave analysis. Europace 2010 Jun;12(6):842-849. [doi: 10.1093/europace/euq042] [Medline: 20185484]

38. Wrenn KD, Slovis CM, Slovis BS. The ability of physicians to predict hyperkalemia from the ECG. Ann Emerg Med 1991 Nov;20(11):1229-1232. [doi: 10.1016/s0196-0644(05)81476-3] [Medline: 1952310]

39. Acker CG, Johnson JP, Palevsky PM, Greenberg A. Hyperkalemia in hospitalized patients: Causes, adequacy of treatment, and results of an attempt to improve physician compliance with published therapy guidelines. Arch Intern Med 1998 Apr 27;158(8):917-924. [doi: 10.1001/archinte.158.8.917] [Medline: 9570179]

40. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Min Knowl Discov 2019 Apr 02;9(4):1-13 [FREE Full text] [doi: 10.1002/widm.1312]

41. Holzinger A, Kieseberg P, Weippl E, Tjoa A. Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable AI. In: Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction. 2018 Presented at: International Cross-Domain Conference for Machine Learning and Knowledge Extraction; August 27-30, 2018; Hamburg, Germany p. 1-8. [doi: 10.1007/978-3-319-99740-7_1]

42. Castelvecchi D. Can we open the black box of AI? Nature 2016 Oct 06;538(7623):20-23. [doi: 10.1038/538020a] [Medline: 27708329]

43. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). 2016 Presented at: IEEE Conference

on Computer Vision and Pattern Recognition (CVPR 2016); June 26-July 1, 2016; Las Vegas, NV URL: https://arxiv.org/pdf/1512.04150.pdf [doi: 10.1109/cvpr.2016.319]

44. Marti G, Schwarz C, Leichtle AB, Fiedler G, Arampatzis S, Exadaktylos AK, et al. Etiology and symptoms of severe hypokalemia in emergency department patients. Eur J Emerg Med 2014 Feb;21(1):46-51. [doi: 10.1097/MEJ.0b013e3283643801] [Medline: 23839104]

45. Nilsson E, Gasparini A, Ärnlöv J, Xu H, Henriksson KM, Coresh J, et al. Incidence and determinants of hyperkalemia and hypokalemia in a large healthcare system. Int J Cardiol 2017 Oct 15;245:277-284 [FREE Full text] [doi: 10.1016/j.ijcard.2017.07.035] [Medline: 28735756]

46. El-Sherif N, Turitto G. Electrolyte disorders and arrhythmogenesis. Cardiol J 2011;18(3):233-245 [FREE Full text] [Medline: 21660912]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
ECG: electrocardiography
ED: emergency department
EMP: electrocardiography morphology parameter
IRB: Institutional Review Board
ISO: International Organization for Standardization
MAE: mean absolute error
QTc: corrected QT interval

XSL•FO
RenderX

Original Paper

Temporal Pattern Detection to Predict Adverse Events in Critical Care: Case Study With Acute Kidney Injury

Mohammad Amin Morid[1], PhD; Olivia R Liu Sheng[2], PhD; Guilherme Del Fiol[3], MD, PhD; Julio C Facelli[3,4], PhD; Bruce E Bray[3,5], MD; Samir Abdelrahman[3,6], MSc, PhD

[1]Department of Information Systems and Analytics, Leavey School of Business, Santa Clara University, Santa Clara, CA, United States

[2]Department of Operations and Information Systems, David Eccles School of Business, University of Utah, Salt Lake City, UT, United States

[3]Department of Biomedical Informatics, School of Medicine, University of Utah, Salt Lake City, UT, United States

[4]Center for Clinical and Translational Science, University of Utah, Salt Lake City, UT, United States

[5]Division of Cardiovascular Medicine, School of Medicine, University of Utah, Salt Lake City, UT, United States

[6]Computer Science Department, Faculty of Computers and Information, Cairo University, Cairo, Egypt

Corresponding Author:
Samir Abdelrahman, MSc, PhD
Department of Biomedical Informatics
School of Medicine
University of Utah
Research Park
Salt Lake City, UT, 84108
United States
Phone: 1 8015875237
Email: samir.abdelrahman@utah.edu

Abstract

Background: More than 20% of patients admitted to the intensive care unit (ICU) develop an adverse event (AE). No previous study has leveraged patients' data to extract the temporal features using their structural temporal patterns, that is, trends.

Objective: This study aimed to improve AE prediction methods by using structural temporal pattern detection that captures global and local temporal trends and to demonstrate these improvements in the detection of acute kidney injury (AKI).

Methods: Using the Medical Information Mart for Intensive Care dataset, containing 22,542 patients, we extracted both global and local trends using structural pattern detection methods to predict AKI (ie, binary prediction). Classifiers were built on 17 input features consisting of vital signs and laboratory test results using state-of-the-art models; the optimal classifier was selected for comparisons with previous approaches. The classifier with structural pattern detection features was compared with two baseline classifiers that used different temporal feature extraction approaches commonly used in the literature: (1) symbolic temporal pattern detection, which is the most common approach for multivariate time series classification; and (2) the last recorded value before the prediction point, which is the most common approach to extract temporal data in the AKI prediction literature. Moreover, we assessed the individual contribution of global and local trends. Classifier performance was measured in terms of accuracy (primary outcome), area under the curve, and F-measure. For all experiments, we employed 20-fold cross-validation.

Results: Random forest was the best classifier using structural temporal pattern detection. The accuracy of the classifier with local and global trend features was significantly higher than that while using symbolic temporal pattern detection and the last recorded value (81.3% vs 70.6% vs 58.1%; $P<.001$). Excluding local or global features reduced the accuracy to 74.4% or 78.1%, respectively ($P<.001$).

Conclusions: Classifiers using features obtained from structural temporal pattern detection significantly improved the prediction of AKI onset in ICU patients over two baselines based on common previous approaches. The proposed method is a generalizable approach to predict AEs in critical care that may be used to help clinicians intervene in a timely manner to prevent or mitigate AEs.

(JMIR Med Inform 2020;8(3):e14272) doi:10.2196/14272

XSL·FO
RenderX

Introduction

Adverse Events Prediction

An adverse event (AE) refers to a patient's injury or complication caused by medical care [1]. Previous studies have shown that AEs are responsible for 44,000 to 98,000 deaths per year, an average of 31 days increase in hospital length and about US $3900 increase in the patient's hospital cost [2,3]. In intensive care unit (ICU) settings, the risk for AEs is even higher because of the complexity of care, the large number of interventions, and the patients' fragile medical status [4]. However, more than 50% of AEs in the ICU are preventable through timely medical interventions [2]. Therefore, it is important to predict the onset of AEs in ICU patients as early as possible [5].

Patient data are collected over time at varying time intervals to monitor the patient's status, provide situation awareness, and support medical decisions, leading to a wide variety of time series data (eg, vital signs, lab results) stored in electronic health record (EHR) systems. The most common approach to use time series data for AE prediction is to use static transformations (STs) to produce a representative value for each time series (eg, mean, first value, and last value in the series) [6]. Although the ST approach facilitates the prediction process by reducing dimensionality, it also results in information loss by ignoring the temporal trends in the time series, which could affect the accuracy of AE prediction [7]. An alternative approach, dynamic transformations (DTs), is to segment a time series into a sequence of fixed-sized, nonoverlapping, consecutive windows (or intervals) [8] and then identify the temporal pattern(s) of data values within and across windows. As a result, temporal pattern detection approaches reduce information loss by benefiting from hidden information embedded over different periods of the time series. The most common method to implement this is symbolic (categorical) temporal pattern detection, where each time interval is represented by the state of its values (eg, high, moderate, low blood pressure) and eventually patterns are extracted from the symbolic time intervals. Although this method can be effective when expert domain knowledge to discretize the values is available, it may lose accuracy from temporal discretization. An alternative method is using structural (numerical) temporal pattern detection where each time interval is represented by a set of numerical values capturing its pattern. This method overcomes the limitations of previous methods by benefiting from original data without any arbitrary discretization. To the best of our knowledge, there is no study in the literature investigating structural temporal pattern detection for the prediction of AEs in critical care.

The goal of this study was to leverage temporal data to predict AEs for ICU patients by using temporal pattern detection. As a case study, we focused on the prediction of acute kidney injury (AKI), one of the most common AEs in ICU settings [9]. More than 50% of all ICU patients develop acute kidney injury, which increases the risk of death in the hospital or shortly after their discharge [10]. Delays in the detection of AKI impair physicians' ability to intervene in a timely manner to prevent AKI and its complications. A study on patients who died in the hospital with a primary diagnosis of AKI showed an unacceptable delay in the detection of AKI in 43% and preventable death in at least 20% of the patients [11]. Over the past decade, AKI prediction methods have been proposed to detect high-risk patients that are candidates for early management [12]. However, the performance of these methods is suboptimal, partially because they use the last value in the input time series for the prediction task, thus missing the rich information contained within the time series.

In this study, we investigated approaches to predict AEs in ICU settings using structural temporal pattern detection methods for both local (ie, within each time window) and global (ie, across time windows) trends. Specifically, using a factorial design, we compared the accuracy of the last recorded value method (mentioned earlier) versus local, global, and both temporal pattern detection methods in the prediction of AKI.

Multivariate Time Series Representation

EHRs contain a rich resource of multivariate time series data providing an important opportunity to discover new knowledge using various data mining methods. However, the classifications of these multivariate time series, especially discrete time series (eg, blood pressure, calcium, magnesium), are challenging [13-15] because data points in EHR time series are often sampled at different and sometimes irregular time intervals. Also, it is very common to have large amounts of missing data points due to intentional (ie, due to medical reason) or unintentional (ie, human mistake or operational constraints) reasons [16].

The most common approach to overcome the aforementioned issues is to transform raw multivariate time series data into a different form where the time series values are uniformly represented [17]. This can be performed by two types of transformations: *static* and *dynamic* [16]. In STs, each time series is represented by a predefined set of features and their values (eg, most recent platelet measurement, maximum hemoglobin measurement). In DTs, each time series is transformed to a high-level qualitative (categorical) or quantitative (numerical) form [17]. The most common method for qualitative transformation is temporal abstraction. Using this method, each time series (eg, series of white blood cell counts) is transformed into a set of intervals using temporal discretization where an alphabet represents the qualitative measure of the values in that interval [18]. Temporal discretization can be done using domain knowledge or an automated method, such as aggregate approximation (SAX) [8] or equal-width discretization (EWD) [19]. Previous research has found that while SAX is the most effective automated method, it is not as effective as knowledge-based methods [18,20]. In quantitative transformations, each time series is segmented into fixed-size, nonoverlapping windows, and each window is summarized by one or more numeric aggregation measures (eg, average).

After transforming the original time series (unevenly sampled) to a time series of high-level qualitative or quantitative measures, various standard classification methods can be applied to classify or predict the expected outcome. This is usually done

by finding patterns that distinguish different classes of the outcome [17]. Depending on the qualitative or quantitative representation, the patterns can be qualitative or quantitative, as described in the next section.

Pattern Detection Methods

Pattern detection methods have been widely used for tasks such as image recognition, speech analysis, traffic analysis, smog detection, and health care predictive analytics [21]. The aim of pattern detection is to identify an object (eg, patient) as belonging to a particular class (eg, patients who develop AKI) by extracting patterns and regularities that are specific to the instances of that class [22]. The underlying idea is that the objects associated with a particular group share more common attributes (ie, patterns) than the objects in other groups [23]. A pattern detection procedure can be divided into 2 basic tasks: *description* and *classification* [24]. The description task extracts the features from each object using feature extraction techniques. The classification task assigns a group label to the object based on the extracted attributes using a classification method [25]. Two main types of feature extraction for the description task of pattern detection are described in the following sections: *symbolic* and *structural* pattern detection.

Symbolic Pattern Detection

Symbolic (also known as categorical or qualitative) patterns are extracted from multivariate time series represented by interval alphabets extracted through temporal abstraction. These patterns are mostly referred to as time interval–related patterns (TIRPs) [26-28]. The most common approach to extract TIRPs is using Allen's temporal relations [29]. These are seven different relations capturing the state of two alphabetic time intervals against each other (eg, *overlap*, *equals*, and *meets*). Several studies attempted to use these all or part of these relations for pattern extraction [16,30-32]. Moskovitch and Shahar [15]

proposed KarmaLego, a fast time interval–mining method, to exploit temporal relations [15]. KarmaLego includes two main steps: Karma and Lego. In the Karma step, all frequent two-sized TIRPs are discovered using a breadth-first-search approach. In the Lego step, the frequent two-sized TIRPs are extended into a tree of longer frequent TIRPs. Recently, the same authors proposed a set of three abstract temporal relations as disjunctions of Allen's relations (ie, before, overlap, and contain) and showed that it is more effective than using the full set of Allen's relations [20]. They called their general framework for classification of multivariate time series analysis as KarmaLegoSification (KLS). In this study, we used KLS with the three temporal relations to implement symbolic pattern detection as a baseline for comparison with our proposed structural pattern detection method.

Although symbolic pattern detection methods are promising, their performance is highly dependent on temporal discretization [33]. Domain knowledge is not always available for knowledge-based methods and automated methods are often not so effective, as automation may result in information loss [15]. For example, SAX labels time intervals by producing equal-sized areas under a Gaussian curve of normalized time series. Once time series are transformed to alphabetic time intervals (eg, low (L), medium (M), and high (H)) the original data are lost, which may mislead the classification process. For instance, Table 1 shows the breakpoint cutoffs of SAX applied to our dataset compared with physicians' domain knowledge according to [20] for hemoglobin A_{1c}. The SAX values might be different in other datasets.

Therefore, numerical patterns can be more effective than categorical patterns, at least in the lack of knowledge-based methods, as they benefit from the original data without any data manipulation or arbitrary discretization. More details on different discretization methods can be found elsewhere [34].

Table 1. Hemoglobin A_{1c} breakpoint cutoffs of aggregate approximation (SAX) in our dataset compared with the physician's domain knowledge. Automatic discretization cannot make proper categorical distinctions between very close values of a continuous variable.

State	Expert value range	SAX value range
1	<7	<5
2	7-9	5-8
3	9-10.5	8-9
4	>10.5	>9

Structural Pattern Detection

Structural pattern detection was intuited by human perception for object recognition [35]. Humans involve mental representations of structure-oriented characteristics of objects to detect them [36]. In a study by Biederman et al [37], the human object recognition process was explained by the following steps: (1) the object (eg, patients' time series) is segmented into separate regions (eg, time windows); (2) each segmented region is approximated by a simple geometric shape; (3) these shapes are combined to build a geometric composition; and (4) the similarity between the geometric composition and a set of predefined object groups in the human mind recognizes the object.

Following a process similar to the human mind, structural pattern detection methods split the data into smaller partitions, each with different subpatterns. Then, each subpattern is represented by one or more features to generate a feature vector. For temporal data, structural patterns can detect the local trends at different parts of a time series (eg, heart rate, temperature, and serum glucose) and represent each part with a different structural pattern (model). More specifically, the time series is segmented into a sequence of fixed-sized, nonoverlapping, consecutive windows (or intervals) [8]. Then, each window is represented with a specific set of features extracted to show the structural patterns of all data values within the window.

Although structural pattern detection has been widely used to capture the local trends in the data, they can be also applied to the windowed data to capture global trends [38]. To do so, each window is represented by a single aggregation measure. The most common method to implement this approach is *piecewise aggregation approximation*, which extracts the average of the data values in each time window [8]. Then, structural pattern detection is applied on the aggregated time series. The output is a set of quantitative features organized into a feature vector where each feature has its own position (eg, mean at the first position, SD at the second position).

The most frequently used structural patterns include (a) constant, (b) linear, (c) exponential, (d) sinusoidal, (e) triangular, and (f) rectangular [24]. These patterns can model most types of trends within time series data. In this study, we employed structural pattern detection for both local and global trend detection (see the Methods section).

Acute Kidney Injury Prediction

Although there is a rich literature on different prediction tasks in the context of AKI [39], this section is focused on those that primarily attempted to predict the occurrence of AKI. As a result, studies such as those predicting the progression of various stages of AKI [40] or prediction of AKI mortality [10] were excluded. We found three different categories of studies. The first category included approaches to predict AKI after surgical procedures using patients' data before the procedure. Wong et al [41] predicted AKI after cardiac surgery. To achieve this, different predictors were collected until the morning of the procedure, such as preoperative intra-aortic balloon pump, ejection fraction, the type of surgery, previous cardiac surgery, cardiopulmonary bypass time, clamp time, pump time, and the number of bypass grafts. A multivariate logistic regression [42] combined with a stepwise selection method achieved an area under the curve (AUC) of 0.78. Park et al [13] predicted AKI after living-donor liver transplantation (LDLT) surgery using predictors such as alcoholic liver disease, liver disease score, and Child-Turcotte-Pugh estimated graft to recipient body weight ratio. Similar to the previous study, they gathered the information before the procedure to predict AKI. A multivariate logistic regression analysis resulted in AUC=0.85. In both previous studies, most of the predictors were specific to the procedure and therefore not generalizable to other procedures.

The second category is AKI prediction in critical care settings. Kane-Gill et al [43] attempted to predict AKI for older adults with critical illness. The input to the model included susceptibilities and exposures consisting of age, sex, race, body mass, comorbid conditions, severity of illness, baseline kidney function, sepsis, and shock collected from the first 24 hours of patients' ICU admission. AKI was defined according to the Kidney Disease: Improving Global Outcomes [44] criteria and predicted by multivariable logistic regression. The approach obtained good performance with AUC=0.744. Schneider et al [45] predicted AKI in critically ill-burn patients in ICU settings. The authors defined AKI according to the risk, injury, failure, loss, and end-stage kidney criteria [46] to predict AKI using a classification and regression tree (CART) model [47]. The decision tree used the first 48 hours of admission data to predict

which subset of patients would develop AKI. The proposed method reached an overall accuracy of 73%. This was one of the first studies in AKI prediction to use a machine learning method rather than regression models. Both studies focused on specific types of patients and also used specific, nongeneralizable features. To our knowledge, there is no study in this category that attempted to predict AKI in all ICU patients.

The third AKI prediction category includes AKI prediction in hospitalized patients, regardless of unit. Kate et al [9] applied a variety of machine learning models to predict AKI in hospitalized older adults including logistic regression, Naïve Bayes [48], C4.5 decision tree [49], support vector machine [50], and an ensemble of all these methods. Laboratory results, demographics, medications, and comorbidities recorded in the first 24 hours were used as input. The logistic regression model outperformed other models with AUC=0.743. In a more recent study by Cheng et al [6], the authors attempted to early predict AKI 1, 2, and 3 days before its occurrence. They applied a variety of machine learning methods on all hospitalized patients using laboratory results, vital signs, demographics, medications, and comorbidities. The Random Forest classifier had the highest AUC for 1, 2, and 3 days (0.765, 0.733, and 0.709, respectively) before the AKI occurrence. Compared with studies in the categories mentioned earlier, the datasets used in this category had very imbalanced datasets (ie, <15% positive cases), as the incidence of AKI in the general hospital population is lower than the incidence in ICU settings.

In summary, there are three main limitations in prior AKI prediction methods. First, previous studies used only the last recorded value before the prediction point to represent temporal data. This approach can compromise the prediction performance by missing potentially useful data trends in the time series. Second, most of the studies have used predictors that are specific to certain types of patients (eg, burn) or procedures (eg, cardiac surgery, LDLT) and do not generalize to other prediction problems. Third, previous studies aimed to predict AKI in specific types of patients. To our knowledge, no previous attempt has been made to predict AKI in all patients admitted to the ICU using the entire time series data available in this setting, which is important given the high incidence of this AE in critical care settings.

Leveraging Time Series Data for Patient Status Predictions

As discussed earlier, there is a great need to develop and demonstrate methods that can take advantage of all temporal data existing in the EHR to predict as early as possible the onset of critical adverse events. To this end, in this paper, we report our work in using both local and global temporal pattern detection and classification techniques to better the prediction of AE using available time series data. We demonstrate our methods by leveraging patients' ICU temporal data for AKI prediction by extracting structural temporal pattern features. We used general predictors such as a set of laboratory results and vital signs, which are widely available as time-series values for any patient in ICU settings. We considered a cohort of all ICU patients without any exclusion to ensure generalizability

of the approach. To the best of our knowledge, there is no paper on AKI prediction choosing this cohort of patients.

Methods

Study Design

The study method consisted of the following steps: (1) dataset and data preparation; (2) implementation of local and global structural pattern detection; (3) AKI prediction; and (4) evaluation. Each of these steps is explained in detail in the following sections (see Figure 1).

Figure 1. An overview of the proposed method and evaluation. AKI: acute kidney injury; ICU: intensive care unit.
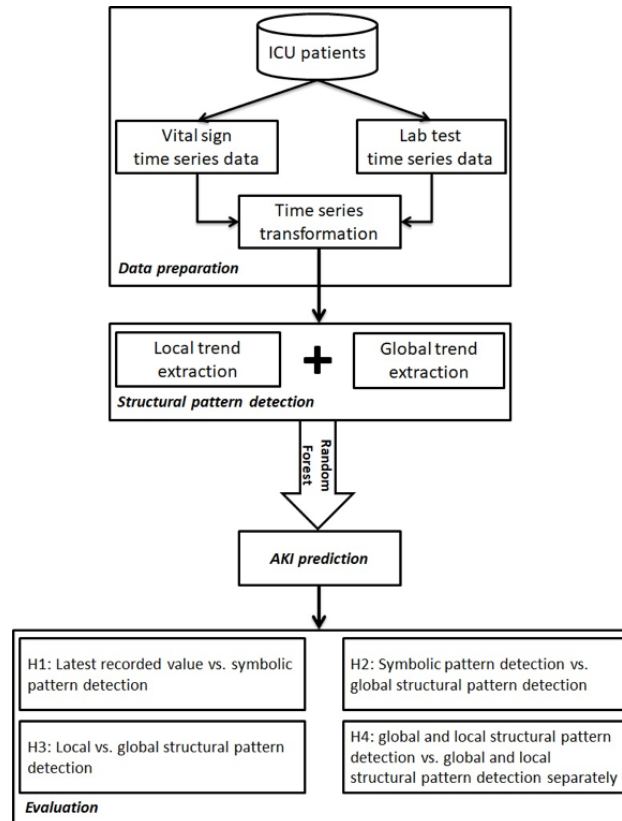
Dataset and Data Preparation

The Medical Information Mart for Intensive Care (MIMIC) III [51] dataset was used for this study. MIMIC III contains comprehensive clinical deidentified data of 38,597 patients admitted to the ICU. As in previous studies [6,43], we used data from the first 48 hours of ICU admission to predict if patients developed AKI before hospital discharge as the main study analysis (binary prediction). As secondary analyses, we also assessed the performance of our proposed model for different data collection periods (see Multimedia Appendix 1, Table S1).

Patients who died within the first 48 hours or developed AKI within the first 48 hours were excluded. Moreover, as in previous studies [6,10], patients with end-stage renal disease on admission (identified based on diagnosis codes and admission serum creatinine>4 mg/dL) were excluded. The resulting dataset contained 22,542 patients. On the basis of findings from previous studies [52,53] on AKI prediction, the 17 time series features listed in Table 2 were chosen as input. We focused on features that are not specific to any condition or procedure to maximize generalizability to other AEs.

Table 2. Input features.

Category and subcategory	Feature
Vital signs	Heart rate, temperature, systolic blood pressure, and diastolic blood pressure
Lab test	
Hematology	White blood cells, hemoglobin, and platelets
Biochemistry	Sodium, anion gap, blood urea nitrogen , potassium, prothrombin, calcium, magnesium, chloride, bicarbonate, and phosphate

Finally, following Mandelbaum et al [10], the onset of AKI was defined according to the AKIN [53] criteria as follows:

- Increase in serum creatinine by ≥0.3 mg/dL within 48 hours OR

- Increase in serum creatinine by ≥1.5 times the baseline within the previous 7 days,

where the lowest serum creatinine measurement during the ICU stay was used as the baseline level. All serum creatinine measurements, from patient admission to discharge, were used only as output parameters to determine the class label, that is, occurrence of AKI.

To prepare the input data, time series features (see Table 2) were transformed to uniform time intervals using a DT, where each time series was segmented into a sequence of fixed-sized nonoverlapping consecutive windows (or intervals) [8]. As suggested in previous research [20,33], we tried different window sizes with lengths of 1, 2, 4, 6, and 8 hours, which led to 48, 24, 12, 8, and 6 windows, respectively. The length of 2 hours (ie, 24 windows) had the best performance compared with others ($P<.05$ for all comparisons). Therefore, all subsequent experiments used this window size. Table S2 in the Multimedia Appendix 1 contains the experimental results of all window sizes.

We used 30% of the data as a development dataset for selecting the best classifiers after tuning their parameters. The remaining 70% was used to build and evaluate the models using 20-fold cross validation. The splitting process was random and stratified to keep the same ratio of the positive to the negative AKI classes (Table 3).

Table 3. Dataset description.

Data	Patients, N	With AKI[a], n (%)	No AKI, n (%)
Full dataset	22,542	12,848 (57.00)	9694 (43.00)
Model building and evaluation	15,779	8994 (57.00)	6785 (43.00)
Development	6763	3855 (57.00)	2908 (43.00)

[a]AKI: acute kidney injury.

Implementation of Local and Global Trend Detection Approaches

To implement structural pattern detection to detect local and global trends, we followed four steps. First, we divided the time series of each input measurement (eg, heart rate, temperature) into fixed sized windows. Second, for local trend detection, structural pattern detection methods were applied on each window to find the structure that best fits that window, including constant, linear, exponential, sinusoidal, triangular (see Figure 2). Third, for global trend detection, the average value of each window was extracted building a time series of the average values. Then, the same structural pattern detection methods were applied to the time series of averages to find the best fitting structure. Finally, the local and global trend detection outputs were used as features to build a classification model for prediction.

Figure 2. Structure detectors' shapes and formulas.

Structure detector	Formula	Shape
Constant	$f(Y_t) = c, \qquad c = \dfrac{1}{n}\sum_{t=1}^{n} Y_t$	
Linear	$f(Y_t) = a*t + b$	
Exponential	$f(Y_t) = a*|b|^t + c$	
Sinusoidal	$f(Y_t) = a*\sin(t+b) + c$	
Triangular	$f(Y_t) = \begin{cases} a + b*t & t < c \\ (a + 2*b*c) - (b*t) & t \ge c \end{cases}$	

The process of structural pattern detection was the same for both *local* and *global* trend detection. That is, a time series was provided as the input and a new time series was generated as the output containing a set of values that describe the identified structure. This process is explained below.

The input of structural pattern detection was a time series of ordered data points, $Y(t)$. The structural pattern detection task was to apply different structures and find the one that best fits $Y(t)$. Each individual structure was a function that approximates $Y(t)$ with a specific pattern [54]. This approximation function is defined as: $f(Y(t)) = \hat{Y}(t)$. Then, to find the structure with the best fit, an error function, E, evaluates how closely $\hat{Y}(t)$ approximates Y_t for each structure. The following error function was used in this study:

The formulas of the structures (ie, approximation functions) used in this study are shown in Figure 2 [24].

These functions are the most commonly used in the literature [55]. Using more sophisticated functions would require a much higher number of data points in each window than what was available in the study dataset. Figure 3 shows an example of data points for a 24 data series of Hemoglobin. The best fit pattern for these values is the *linear* model with *a*=0.128 and *b*=7.133.

To find the optimal parameters (ie, *a*, *b*, and *c*) of a time series' structural pattern, standard linear regression equations were used for Constant and Linear structure detectors. For the remaining structure detectors, we needed to search for the best parameter values that minimized the error function. To achieve this, we used Simplex search [56], which is a direct search method guided by evaluating the error function with various combinations of values for the three parameters (ie, *a*, *b*, and *c*).

Finally, after finding the best approximation function, the structure pattern detection generates 4 values as the final output including the three parameters *a*, *b*, and *c* and the *index* of the best fitted structure detector ranging from 1 to 5 (eg, *Triangular* structure is 5). Therefore, the output of structure pattern detection for *global* trend detection is in the form of G_a, G_b, G_c, and G_{index}, where G_a, G_b, and G_c are *a*, *b*, and *c* parameters of the best fitted structure and G_{index} is the index of that structure. If there is a time series with just 1 data point the *Constant* detector was used to represent the time series. Similarly, the output of structure pattern detection for local trend detection over *m* windows is in the form of

Here, are the *a*, *b*, and *c* parameters of the best fitted structure on window *i* and is the index of that structure. Combining both *local* and *global* trend detection outputs, the final time series looked like the following:

Figure 3. An example of extracted linear structure from a Hemoglobin time series.
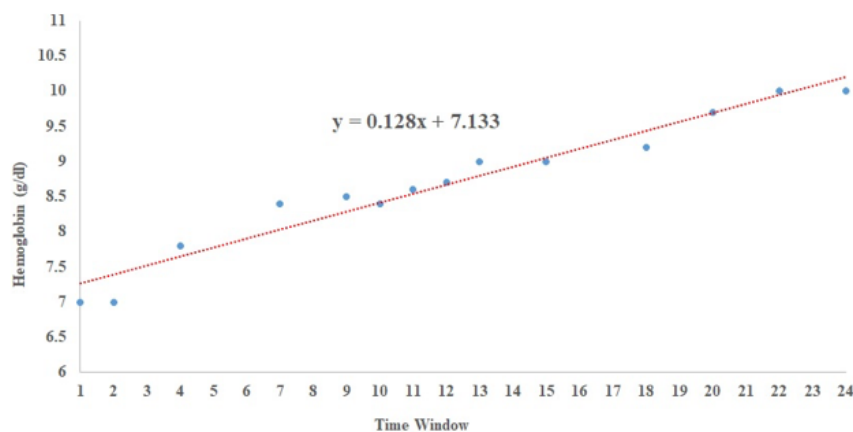
Acute Kidney Injury Prediction

For the classification task, several state-of-art machine learning algorithms were applied to predict AKI. To achieve this, each algorithm was tuned to find its best performance [57]. These algorithms include Random Forest [58], Extreme Gradient Boosting Tree [59], Kernel-based Bayesian Network [60], Support Vector Machine (SVM) [61], Logistic Regression [42], Naïve Bayes [62], K-Nearest Neighbor [63], and Artificial Neural Network (ANN) [64]. Algorithms were evaluated with the following parameter tuning settings: *maximum depth*, *number of bins*, and *learning rate* were varied for the extreme gradient boosting tree, *kernel type* and *number of kernels* were varied for the Kernel-based Bayesian Network; *number of hidden layers*, *number of nodes* in each layer, and *learning rate* were varied for the Neural Network; *Kernel type* along with the corresponding parameters of each kernel type were varied for the SVM; the *value of k* and the *weighted voting* method were changed for the K-Nearest Neighbor algorithm; and *the number of trees* was varied for Random Forest. Similar to previous research on AKI prediction [6], a Random Forest classifier

achieved the best performance. This is an ensemble learning algorithm that fits several decision trees on different subsamples of the data. The mode value of the decision tree outcomes determines the final predicted label of the algorithm [58]. Therefore, this classifier was used in all experiments described below. The performance results comparison of all classifiers can be found in Table S3 of the Multimedia Appendix 1.

Evaluation

In the evaluation step, we tested four hypotheses that were defined a priori. The hypotheses were tested according to a 2×2 factorial study design [65] with local structural pattern and global structural pattern as dimensions. The factorial design allowed us to compare the performance of all possible combinations from a baseline to an approach with both *local* and *global* structural patterns.

- Hypothesis 1: A baseline symbolic temporal pattern detection method has higher accuracy than a nontemporal pattern detection method (ie, last recorded value before the prediction point) in the prediction of AKI in ICU patients.

- Hypothesis 2: Global structural pattern detection has higher accuracy than symbolic pattern detection in the prediction of AKI in ICU patients.
- Hypothesis 3: Local structural pattern detection has higher accuracy than global structural pattern detection in the prediction of AKI in ICU patients.
- Hypothesis 4: Global and local structural pattern combined has higher accuracy than global and local structural pattern detection separately in the prediction of AKI in ICU patients.

As the baseline, we implemented the symbolic pattern detection according to the KLS framework by Moskovitch and Shahar (the most common approach for multivariate time series classification) [20]. This framework includes four main components: temporal abstraction, time-interval mining, TIRP-based feature representation, and classifier selection, where each component has its own settings. Aligned with the authors suggestion after trying different settings in several evaluations [20], we used the following parameter settings: *SAX* was used for temporal discretization with *four* bins; KarmaLego with epsilon value of *0* and minimal vertical threshold of *60%* was used for three time-intervals mining; the *three* abstract relations (ie, before, overlaps, and contains) proposed by the authors were used for temporal relations; *mean duration* was used to represent TIRPs (without any feature selection); and *Random Forest* was used as the classifier. We also tried EWD

as the second-best method for temporal discretization suggested by Moskovitch and Shahar [20], but it is was outperformed by SAX (accuracy of 0.706 vs 0.667; *P*<.001).

To test the significance of the differences between the classifiers, we used ANOVA (analysis of variance) repeated measures test [66], with classification *accuracy* as the primary outcome. This approach allowed us to test for a potential interaction (ie, dependency) between parameters of structure detectors for local and global trends (see Section 3.1). We used the baseline (ie, symbolic) as the control group and the local and global trends as the two treatment factors, with the 20 folds as the observations. In other words, for each fold, we have results for the baseline (ie, control group), local structural pattern (ie, factor), global local structural pattern (ie, factor) and the combination of local and global local structural patterns (ie, interaction). This experimental design is similar to previous studies on AKI prediction [13,41].

Results

Hypothesis 1: Symbolic Pattern Detection Versus Last Recorded Value

The accuracy of symbolic pattern detection in predicting AKI was significantly higher than the last recorded value method (0.706 vs 0.581; *P*<.001). Similar significant differences were found in terms of F-measure and AUC (Figure 4).

Figure 4. Accuracy, F-measure, and AUC of the latest recorded value method versus symbolic pattern detection for AKI prediction. AUC: area under the curve; AKI: acute kidney injury.
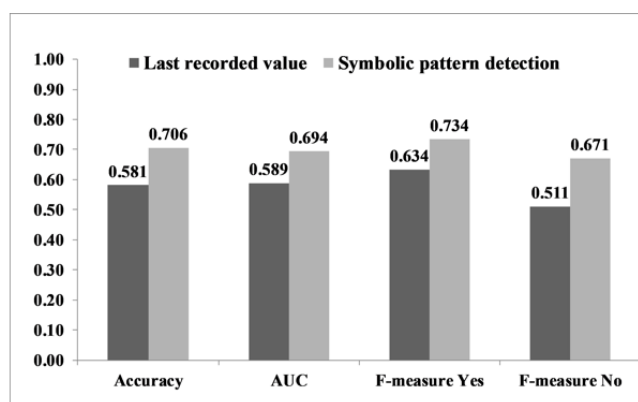
Hypothesis 2: Global Structural Pattern Detection Versus Symbolic Pattern Detection

The accuracy of global structural pattern detection in predicting AKI was significantly higher than symbolic pattern detection

(0.744 vs 0.706; *P*<.001). Similar significant differences were found in terms of F-measure and AUC (Figure 5).

Figure 5. Accuracy, F-measure, and AUC of the global structural pattern detection versus symbolic pattern detection for AKI prediction. AUC: area under the curve; AKI: acute kidney injury.
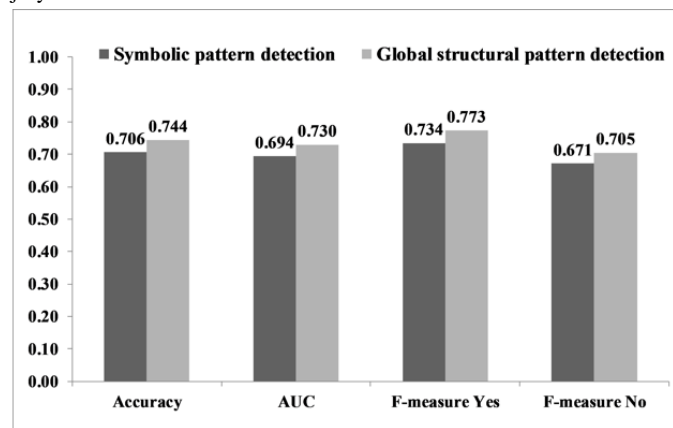
Hypothesis 3: Local Versus Global Structural Pattern Detection

The accuracy of local structural pattern detection in predicting AKI was significantly higher than global structural pattern detection (0.781 vs 0.744; $P<.001$). Similar significant differences were found in terms of F-measure and AUC (Figure 6).

Figure 6. Accuracy, F-measure, and AUC of the local versus global structural pattern detection method for AKI prediction. AUC: area under the curve; AKI: acute kidney injury.
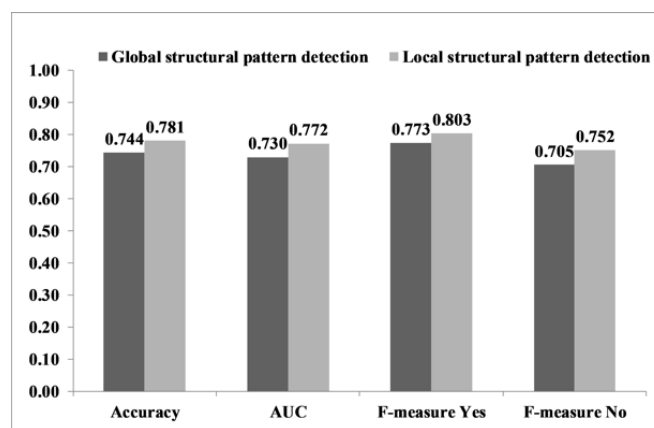
Hypothesis 4: Global and Local Structural Pattern Detection Combined Versus Global and Local Structural Pattern Detection Separately

The accuracy of combined global and local structural pattern detection in predicting AKI was significantly higher than global and local structural pattern detection separately (0.813 vs 0.744 and 0.781; $P<.001$). Similar significant differences were found in terms of F-measure and AUC (Figure 7). Also, Table 4 shows the distribution of extracted patterns for local and global structure detectors.

Figure 7. Accuracy, F-measure, and AUC of the global and local structural pattern detection combined versus global and local structural pattern detection separately for AKI prediction. AUC: area under the curve; AKI: acute kidney injury.
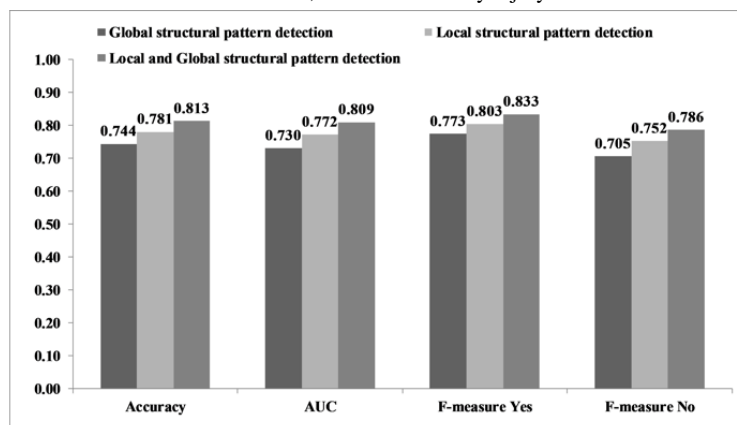
Table 4. Distribution (%) of different structure detectors for local and global patterns.

Features size	Constant	Straight	Exponential	Triangular	Sinusoidal
Local	55.82	34.73	5.26	3.03	1.16
Global	24.18	37.67	7.53	27.18	3.44

Discussion

Principal Findings

We investigated methods for extracting temporal patterns from patients' data to predict AEs in critical care settings. Overall, we found that local and global structural pattern detection methods outperformed the accuracy of symbolic pattern detection in AKI prediction (78.1% vs 74.4% vs 70.6%), with local and global structural patterns combined yielding the highest accuracy of all methods investigated (81.3%). Such finds are clinically important, as early prediction of AEs may warn clinicians to implement interventions or closer monitoring strategies to help prevent AEs in a timely manner. In fact, compared with symbolic pattern detection, the combined local and global approach resulted in 1076 out of 9392 additional AKI patients correctly identified. This is a remarkable improvement that, if integrated with routine clinical care, has the potential to reduce hospital morbidity and mortality.

We conducted four experiments to test four hypotheses. The first experiment demonstrated the value of temporal data using symbolic pattern detection, which significantly outperformed the last recorded value (70.6% vs 58.1%), which is the most common approach in the literature. As the incidence of AKI in the dataset was very high (57%), the accuracy of the classifier based on last recorded values was similar to always predicting cases as positive. Thus, this finding highlights the importance of using all information in time series data rather than using a single value.

The second and third experiments showed that detecting local and global trends using structural pattern detection improves the accuracy of the baseline symbolic pattern detection method (78.1% vs 74.4% vs 70.6%). This suggests that information loss caused by temporal discretization has significant negative effect on the performance of symbolic pattern detection. Also, local trends provided a significantly contribution to the increase in accuracy compared to global trends. Most important, the fourth experiment found that combining local and global trends achieved the best accuracy of all methods in this study. This finding highlights the importance of detecting trends at different data segments rather than one trend over the whole time series.

Shedding more light on the local and global trends detected by the structural pattern detection methods, Table 4 shows the distribution of different types of structure detectors. As seen, more than 90 percent of the fitted structure detectors for local trends are either constant or straight line. One reason for the small percentage of other structure detectors could be because there is a small number of data values on each window, which is not suitable for sophisticated models (eg, sinusoidal). Similarly, Constant, Straight and Triangular are the most common patterns in global trends. The implication is that patients' health status at local and global time windows may not need complicated structure detectors.

Our study had some limitations. First, we focused on the prediction of AKI as a case study and did not test the generalizability to other AEs. Nevertheless, to maximize generalizability, we used a set of input features that are widely used in critical care settings and are not specific to any condition or procedure. Future studies are needed to assess generalizability to other AEs and datasets. Second, as we did not have access to serum creatinine data before ICU admission, as in previous studies, the lowest serum creatinine level after ICU admission was used as the baseline. Third, as the focus of our study was on testing different temporal pattern detection methods, we limited our dataset to numeric variables that change frequently overtime, which is not the case of variables such as age, gender, and comorbidities. As AKI is a frequent comorbidity, expanding the model input to include medical conditions—for example, sepsis, heart failure, and age would likely improve model accuracy but might not significantly change the relative performance levels of different structural patterns. Fourth, as we applied structural pattern detection on granular time series data, clinical interpretation of the patterns associated with AKI prediction was very difficult. There is always a trade-off between accuracy and explainability, and in this study, we focused on accuracy. Tackling the explainability limitation is a subject for future studies. Currently, we are investigating the use of deep learning approaches, especially recurrent neural networks, with a larger number of predictors, along with the proposed local and global pattern detection method.

Conclusions

We investigated the effect of temporal pattern detection methods on AE prediction, using AKI as a case study. Capturing patterns in local and global trends with structural pattern detection significantly improved the accuracy of AKI prediction in ICU settings. Besides the technical contributions, accurate prediction of patients with a high risk for AEs has the potential to decrease hospital morbidity and mortality.

XSL•FO

RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Experimental results of all window sizes.
[DOCX File , 16 KB - medinform_v8i3e14272_app1.docx]

References

1. Gurjar SV, Roshanzamir S, Patel S, Harinath G. General surgical adverse events in a UK district general hospital-lessons to learn. Int J Surg 2011;9(1):55-58 [FREE Full text] [doi: 10.1016/j.ijsu.2010.08.008] [Medline: 20817049]

2. Forster AJ, Kyeremanteng K, Hooper J, Shojania KG, van Walraven C. The impact of adverse events in the intensive care unit on hospital mortality and length of stay. BMC Health Serv Res 2008 Dec 17;8:259 [FREE Full text] [doi: 10.1186/1472-6963-8-259] [Medline: 19091089]

3. Kaushal R, Bates DW, Franz C, Soukup JR, Rothschild JM. Costs of adverse events in intensive care units. Crit Care Med 2007 Nov;35(11):2479-2483. [doi: 10.1097/01.CCM.0000284510.04248.66] [Medline: 17828035]

4. Pagnamenta A, Rabito G, Arosio A, Perren A, Malacrida R, Barazzoni F, et al. Adverse event reporting in adult intensive care units and the impact of a multifaceted intervention on drug-related adverse events. Ann Intensive Care 2012 Nov 22;2(1):47 [FREE Full text] [doi: 10.1186/2110-5820-2-47] [Medline: 23174137]

5. King A, Bottle A, Faiz O, Aylin P. Investigating adverse event free admissions in medicare inpatients as a patient safety indicator. Ann Surg 2017 May;265(5):910-915. [doi: 10.1097/SLA.0000000000001792] [Medline: 27192350]

6. Cheng P, Waitman LR, Hu Y, Liu M. Predicting inpatient acute kidney injury over different time horizons: how early and accurate? AMIA Annu Symp Proc 2017;2017:565-574 [FREE Full text] [Medline: 29854121]

7. Lee Y, Wei C, Cheng T, Yang C. Nearest-neighbor-based approach to time-series classification. Decis Support Syst 2012;53(1):207-217 [FREE Full text] [doi: 10.1016/j.dss.2011.12.014]

8. Lin J, Keogh E, Lonardi S, Chiu B. A Symbolic Representation of Time Series, With Implications for Streaming Algorithms. In: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. 2003 Presented at: DMKD'03; June 13, 2003; California, San Diego p. 2-11. [doi: 10.1145/882082.882086]

9. Kate RJ, Perez RM, Mazumdar D, Pasupathy KS, Nilakantan V. Prediction and detection models for acute kidney injury in hospitalized older adults. BMC Med Inform Decis Mak 2016 Mar 29;16:39 [FREE Full text] [doi: 10.1186/s12911-016-0277-4] [Medline: 27025458]

10. Mandelbaum T, Lee J, Scott DJ, Mark RG, Malhotra A, Howell MD, et al. Empirical relationships among oliguria, creatinine, mortality, and renal replacement therapy in the critically ill. Intensive Care Med 2013 Mar;39(3):414-419 [FREE Full text] [doi: 10.1007/s00134-012-2767-x] [Medline: 23223822]

11. Bhagwanani A, Carpenter R, Yusuf A. Improving the management of acute kidney injury in a district general hospital: introduction of the DONUT bundle. BMJ Qual Improv Rep 2014;2(2):- [FREE Full text] [doi: 10.1136/bmjquality.u202650.w1235] [Medline: 26734238]

12. Sutherland SM, Chawla LS, Kane-Gill SL, Hsu RK, Kramer AA, Goldstein SL, 15 ADQI Consensus Group. Utilizing electronic health records to predict acute kidney injury risk and outcomes: workgroup statements from the 15(th) ADQI Consensus Conference. Can J Kidney Health Dis 2016;3:11 [FREE Full text] [doi: 10.1186/s40697-016-0099-4] [Medline: 26925247]

13. Park MH, Shim HS, Kim WH, Kim H, Kim DJ, Lee S, et al. Clinical risk scoring models for prediction of acute kidney injury after living donor liver transplantation: a retrospective observational study. PLoS One 2015;10(8):e0136230 [FREE Full text] [doi: 10.1371/journal.pone.0136230] [Medline: 26302370]

14. Strobl AN, Vickers AJ, van Calster B, Steyerberg E, Leach RJ, Thompson IM, et al. Improving patient prostate cancer risk assessment: moving from static, globally-applied to dynamic, practice-specific risk calculators. J Biomed Inform 2015 Aug;56:87-93 [FREE Full text] [doi: 10.1016/j.jbi.2015.05.001] [Medline: 25989018]

15. Moskovitch R, Shahar Y. Medical temporal-knowledge discovery via temporal abstraction. AMIA Annu Symp Proc 2009 Nov 14;2009:452-456 [FREE Full text] [Medline: 20351898]

16. Batal I, Fradkin D, Harrison J, Moerchen F, Hauskrecht M. Mining recent temporal patterns for event detection in multivariate time series data. Knowl Discov Data Min 2012;2012:280-288 [FREE Full text] [doi: 10.1145/2339530.2339578] [Medline: 25937993]

17. Shknevsky A, Shahar Y, Moskovitch R. Consistent discovery of frequent interval-based temporal patterns in chronic patients' data. J Biomed Inform 2017 Nov;75:83-95 [FREE Full text] [doi: 10.1016/j.jbi.2017.10.002] [Medline: 28987378]

18. Moskovitch R, Shahar Y. Vaidurya: a multiple-ontology, concept-based, context-sensitive clinical-guideline search engine. J Biomed Inform 2009 Feb;42(1):11-21 [FREE Full text] [doi: 10.1016/j.jbi.2008.07.003] [Medline: 18721900]

19. Dougherty J, Kohavi R, Sahami M. Supervised and Unsupervised Discretization of Continuous Features. In: Proceedings of the Twelfth International Conference on Machine Learning. 1995 Presented at: ICML'95; July 9–12, 1995; Tahoe City,

California p. 194-202 URL: https://doi.org/10.1016/B978-1-55860-377-6.50032-3 [doi: 10.1016/b978-1-55860-377-6.50032-3]

20. Moskovitch R, Shahar Y. Classification of multivariate time series via temporal abstraction and time intervals mining. Knowl Inf Syst 2015;45:35-74. [doi: 10.1007/s10115-014-0784-5]

21. Lin J, Williamson S, Borne K, DeBarr D. Pattern recognition in time series. In: Way MJ, Scargle JD, Ali KM, Srivastava AN, editors. Advances in Machine Learning and Data Mining for Astronomy. London: Chapman and Hall; 2012:617-645.

22. Rodríguez JJ, Alonso CJ, Maestro JA. Support vector machines of interval-based features for time series classification. Knowl Based Syst 2005 Aug;18(4-5):171-178 [FREE Full text] [doi: 10.1016/j.knosys.2004.10.007]

23. Weng X, Shen J. Classification of multivariate time series using two-dimensional singular value decomposition. Knowl Based Syst 2008 Oct;21(7):535-539 [FREE Full text] [doi: 10.1016/j.knosys.2008.03.014]

24. Olszewski RT. Carnegie Mellon School of Computer Science. 2001 Feb. Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data URL: https://www.cs.cmu.edu/~bobski/pubs/tr01108-twosided.pdf [accessed 2020-02-11]

25. Morid MA, Sheng OR, Kawamoto K, Ault T, Dorius J, Abdelrahman S. Healthcare cost prediction: leveraging fine-grain temporal patterns. J Biomed Inform 2019 Mar;91:103113. [doi: 10.1016/j.jbi.2019.103113] [Medline: 30738188]

26. Papapetrou P, Kollios G, Sclaroff S, Gunopulos D. Mining frequent arrangements of temporal intervals. Knowl Inf Syst 2009;21(2):133-171. [doi: 10.1007/s10115-009-0196-0]

27. Moskovitch R, Shahar Y. Fast time intervals mining using the transitivity of temporal relations. Knowl Inf Syst 2015;42(1):21-48 [FREE Full text] [doi: 10.1007/s10115-013-0707-x]

28. Moerchen F. Algorithms for Time Series Knowledge Mining. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006 Presented at: KDD'06; August 20 - 23, 2006; Philadelphia, USA p. 668-673. [doi: 10.1145/1150402.1150485]

29. Allen JF. Maintaining knowledge about temporal intervals. Commun ACM 1983;26(11):510-521. [doi: 10.1145/182.358434]

30. Moskovitch R, Choi H, Hripcsak G, Tatonetti N. Prognosis of clinical outcomes with temporal patterns and experiences with one class feature selection. IEEE/ACM Trans Comput Biol Bioinform 2017;14(3):555-563 [FREE Full text] [doi: 10.1109/TCBB.2016.2591539] [Medline: 27429447]

31. Batal I, Valizadegan H, Cooper GF, Hauskrecht M. A temporal pattern mining approach for classifying electronic health record data. ACM Trans Intell Syst Technol 2013 Sep;4(4) [FREE Full text] [doi: 10.1145/2508037.2508044] [Medline: 25309815]

32. Verduijn M, Sacchi L, Peek N, Bellazzi R, de Jonge E, de Mol BA. Temporal abstraction for feature extraction: a comparative case study in prediction from intensive care monitoring data. Artif Intell Med 2007 Sep;41(1):1-12. [doi: 10.1016/j.artmed.2007.06.003] [Medline: 17698331]

33. Singh A, Nadkarni G, Gottesman O, Ellis SB, Bottinger EP, Guttag JV. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. J Biomed Inform 2015 Feb;53:220-228 [FREE Full text] [doi: 10.1016/j.jbi.2014.11.005] [Medline: 25460205]

34. Garcia S, Luengo J, Sáez JA, López V, Herrera F. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. IEEE Trans Knowl Data Eng 2013 Apr;25(4):734-750. [doi: 10.1109/tkde.2012.35]

35. Yu G, Liu Z, Yuan J. Discriminative Orderlet Mining for Real-Time Recognition of Human-Object Interaction. In: Proceedings of the Asian Conference on Computer Vision. 2014 Presented at: ACCV'14; November 1-5, 2014; Singapore p. 50-65 URL: https://doi.org/10.1007/978-3-319-16814-2_4 [doi: 10.1007/978-3-319-16814-2_4]

36. Cichy RM, Pantazis D, Oliva A. Resolving human object recognition in space and time. Nat Neurosci 2014 Mar;17(3):455-462 [FREE Full text] [doi: 10.1038/nn.3635] [Medline: 24464044]

37. Biederman I. Recognition-by-components: a theory of human image understanding. Psychol Rev 1987 Apr;94(2):115-147. [doi: 10.1037/0033-295X.94.2.115] [Medline: 3575582]

38. Michaelsen E, Meidow J. Stochastic reasoning for structural pattern recognition: an example from image-based UAV navigation. Pattern Recogn 2014;47(8):2732-2744 [FREE Full text] [doi: 10.1016/j.patcog.2014.02.009]

39. Luong DT, Chandola V. A K-Means Approach to Clustering Disease Progressions. In: Proceedings of the 2017 IEEE International Conference on Healthcare Informatics. 2017 Presented at: ICHI'17; August 23-26, 2017; Park City, UT, USA. [doi: 10.1109/ICHI.2017.18]

40. Chiofolo C, Chbat N, Ghosh E, Eshelman L, Kashani K. Automated continuous acute kidney injury prediction and surveillance: a random forest model. Mayo Clin Proc 2019 May;94(5):783-792. [doi: 10.1016/j.mayocp.2019.02.009] [Medline: 31054606]

41. Wong B, St Onge J, Korkola S, Prasad B. Validating a scoring tool to predict acute kidney injury (AKI) following cardiac surgery. Can J Kidney Health Dis 2015;2:3 [FREE Full text] [doi: 10.1186/s40697-015-0037-x] [Medline: 25780626]

42. Hosmer DW, Lemeshow S. Applied Logistic Regression. Hoboken, New Jersey: Wiley-Interscience Publication; 2013.

43. Kane-Gill SL, Sileanu FE, Murugan R, Trietley GS, Handler SM, Kellum JA. Risk factors for acute kidney injury in older adults with critical illness: a retrospective cohort study. Am J Kidney Dis 2015 Jun;65(6):860-869 [FREE Full text] [doi: 10.1053/j.ajkd.2014.10.018] [Medline: 25488106]

44. Cattran DC, Feehally J, Cook HT, Liu ZH, Fervenza FC, Mezzano SA, et al. Kidney disease: Improving global outcomes (KDIGO) glomerulonephritis work group. KDIGO clinical practice guideline for glomerulonephritis. Kidney Int Suppl 2012 Jun;2(2):139-274 [FREE Full text] [doi: 10.1038/kisup.2012.9]

45. Schneider DF, Dobrowolsky A, Shakir IA, Sinacore JM, Mosier MJ, Gamelli RL. Predicting acute kidney injury among burn patients in the 21st century: a classification and regression tree analysis. J Burn Care Res 2012;33(2):242-251 [FREE Full text] [doi: 10.1097/BCR.0b013e318239cc24] [Medline: 22370901]

46. Ricci Z, Cruz D, Ronco C. The RIFLE criteria and mortality in acute kidney injury: a systematic review. Kidney Int 2008 Mar;73(5):538-546 [FREE Full text] [doi: 10.1038/sj.ki.5002743] [Medline: 18160961]

47. Loh W. Classification and regression trees. WIREs Data Mining Knowl Discov 2011;1(1):14-23. [doi: 10.1002/widm.8]

48. Lewis DD. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In: Proceedings of the European Conference on Machine Learning. 1998 Presented at: ECML'98; April 21-23, 1998; Chemnitz, Germany p. 4-15 URL: https://doi.org/10.1007/BFb0026666 [doi: 10.1007/BFb0026666]

49. Quinlan JR. C4.5: Programs for Machine Learning. San Mateo, California: Elsevier; 2014.

50. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge: Cambridge University Press; 2000.

51. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]

52. Demirjian S, Schold JD, Navia J, Mastracci TM, Paganini EP, Yared J, et al. Predictive models for acute kidney injury following cardiac surgery. Am J Kidney Dis 2012 Mar;59(3):382-389. [doi: 10.1053/j.ajkd.2011.10.046] [Medline: 22206745]

53. Mehta RL, Kellum JA, Shah SV, Molitoris BA, Ronco C, Warnock DG, Acute Kidney Injury Network. Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care 2007;11(2):R31 [FREE Full text] [doi: 10.1186/cc5713] [Medline: 17331245]

54. Berndt DJ, Clifford J, Briedé JJ. Using Dynamic Time Warping to Find Patterns in Time Series. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining. 1994 Presented at: AAAIWS'94; April 26, 1994; New York, New York p. 359-370 URL: https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf

55. Pavlidis T. Structural Pattern Recognition. Berlin: Springer; 2013.

56. Wright M. Direct search methods: Once scorned, now respectable. In: Griffiths DF, Watson GA, editors. Numerical Analysis. Harlow, United Kingdom: Addison-Wesley; 1996:191-208.

57. Morid MA, Fiszman M, Raja K, Jonnalagadda SR, Del Fiol G. Classification of clinically useful sentences in clinical evidence resources. J Biomed Inform 2016 Apr;60:14-22 [FREE Full text] [doi: 10.1016/j.jbi.2016.01.003] [Medline: 26774763]

58. Strobl C, Boulesteix A, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 2007 Jan 25;8:25 [FREE Full text] [doi: 10.1186/1471-2105-8-25] [Medline: 17254353]

59. Chen T, He T. CRAN. 2019 Aug 1. xgboost: eXtreme Gradient Boosting URL: http://cran.fhcrc.org/web/packages/xgboost/vignettes/xgboost.pdf [accessed 2019-08-01]

60. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn 1997 Nov;29(2-3):131-163 [FREE Full text] [doi: 10.1023/A:1007465528199]

61. Zhang W, Zhang X, Xiao W. Portfolio selection under possibilistic mean–variance utility and a SMO algorithm. Eur J Oper Res 2009 Sep;197(2):693-700 [FREE Full text] [doi: 10.1016/j.ejor.2008.07.011]

62. Zhang H. The Optimality of Naive Bayes. In: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference. 2004 Presented at: FLAIRS'04; May 17-19, 2004; Miami Beach, Florida, USA.

63. Peterson L. K-nearest neighbor. Scholarpedia 2009;4(2):1883. [doi: 10.4249/scholarpedia.1883]

64. Yegnanarayana B. Artificial Neural Networks. India: PHI Learning Pvt Ltd; 2009.

65. Collins LM, Dziak JJ, Kugler KC, Trail JB. Factorial experiments: efficient tools for evaluation of intervention components. Am J Prev Med 2014 Oct;47(4):498-504 [FREE Full text] [doi: 10.1016/j.amepre.2014.06.021] [Medline: 25092122]

66. Weinfurt KP. Repeated measures analysis: ANOVA, MANOVA, and HLM. In: Grimm LG, Yarnold PR, editors. Reading and Understanding More Multivariate Statistics. Washington, DC: American Psychological Association; 2000:317-361.

Abbreviations

AKI: acute kidney injury
ANN: artificial neural network
ANOVA: analysis of variance
AUC: area under the curve
CART: classification and regression tree
DT: dynamic transformation
EHR: electronic health record
EWD: equal-width discretization
ICU: intensive care unit

XSL•FO
RenderX

KLS: KarmaLegoSification
LDLT: living-donor liver transplantation
MIMIC: Medical Information Mart for Intensive Care
ST: static transformation
SVM: support vector machine
TIRP: time interval–related pattern

XSL•FO
RenderX

Original Paper

Toward Standardized Monitoring of Patients With Chronic Diseases in Primary Care Using Electronic Medical Records: Development of a Tool by Adapted Delphi Procedure

Leandra Falck[1], MD; Marco Zoller[1], MD; Thomas Rosemann[1], MD, PhD; Nahara Anani Martínez-González[1], MSc; Corinne Chmiel[1], MD, PD

Institute of Primary Care, University of Zurich and University Hospital of Zurich, Zurich, Switzerland

Corresponding Author:
Leandra Falck, MD
Institute of Primary Care
University of Zurich and University Hospital of Zurich
Pestalozzistrasse 24
Zurich, 8091
Switzerland
Phone: 41 44 255 98 55
Email: Leandra.Falck@usz.ch

Abstract

Background: Long-term care for patients with chronic diseases poses a huge challenge in primary care. There are deficits in care, especially regarding monitoring and creating structured follow-ups. Appropriate electronic medical records (EMR) could support this, but so far, no generic evidence-based template exists.

Objective: The aim of this study is to develop an evidence-based standardized, generic template that improves the monitoring of patients with chronic conditions in primary care by means of an EMR.

Methods: We used an adapted Delphi procedure to evaluate a structured set of evidence-based monitoring indicators for 5 highly prevalent chronic diseases (ie, diabetes mellitus type 2, asthma, arterial hypertension, chronic heart failure, and osteoarthritis). We assessed the indicators' utility in practice and summarized them into a user-friendly layout.

Results: This multistep procedure resulted in a monitoring tool consisting of condensed sets of indicators, which were divided into sublayers to maximize ergonomics. A cockpit serves as an overview of fixed goals and a set of procedures to facilitate disease management. An additional tab contains information on nondisease-specific indicators such as allergies and vital signs.

Conclusions: Our generic template systematically integrates the existing scientific evidence for the standardized long-term monitoring of chronic conditions. It contains a user-friendly and clinically sensible layout. This template can improve the care for patients with chronic diseases when using EMRs in primary care.

Introduction

Long-term care for patients with chronic diseases poses a huge challenge. There are deficits regarding monitoring and creating structured follow-ups. In Switzerland, unlike other countries, there are a plethora of different electronic medical record (EMR) providers. Although no official registry is maintained, the estimated number of current EMR providers is 60. Due to the lack of unified software standards, interaction between providers is impossible and migration of data is practically unfeasible. This fact illustrates the need for standardization across EMRs, and this may also be the reason why many practices in Switzerland still prefer paper-based records.

To introduce a tool for monitoring patients with chronic diseases within the EMR, it is essential to know the needs of potential users and to develop a customized tool. A survey of physicians not using EMRs showed that most concerns relate to the improvement of quality of care, the workflow process, and the

XSL•FO
RenderX

physician-patient relationship [1-3]. Thus, to increase EMR use, it is crucial to enhance the benefits. This can be achieved by customizing the EMR for increased productivity. For example, combining monitoring elements with time saving features, ergonomic navigation, and clear design could facilitate fast retrieval of all relevant information.

EMRs are not only a practical bookkeeping tool; they can also improve disease management. For chronic diseases, EMRs enable more thorough record keeping and surveillance of treatment intensification, thus improving monitoring [4-6]. In addition, EMRs can help with both documenting and reducing errors that are common in paper-based medical records, such as legibility, prescription, and transcription errors [7,8]. Further advantages of EMRs include the graphic representation of monitoring indicators and, more importantly, the migration of data for care coordination between different providers and between providers and patients. However, the poor dissemination and lack of standardization of EMRs poses a huge obstacle for research in primary care.

We have previously identified a structured set of evidence-based indicators for five common chronic conditions [9]. In this study we aimed to develop an evidence-based standardized, generic template that improves monitoring of patients with chronic conditions in primary care by means of an EMR.

Methods

Figure 1 shows an overview of the methodology.

Figure 1. Study flow: Summary of all steps of the monitoring tool project.
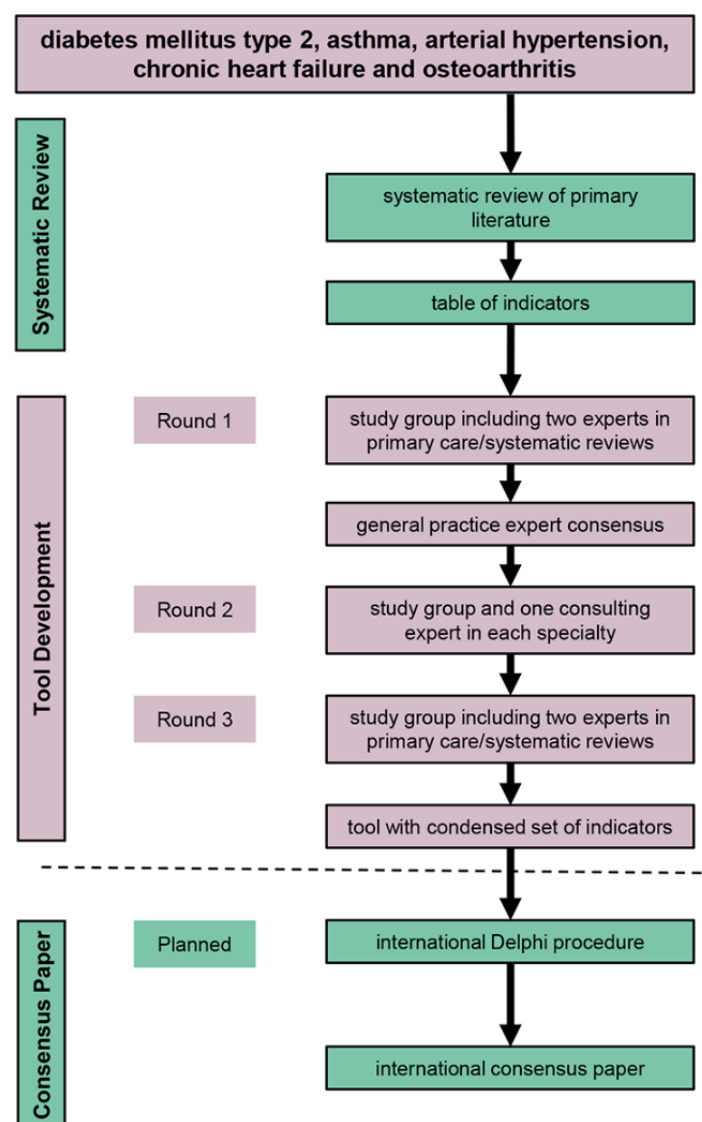
Prior Work: The Systematic Review

In the first phase of this study [9], we performed a systematic review to identify and assess a set of clinical indicators that can be used for monitoring chronic diseases in primary care. In this study, we selected clinical indicators for five diseases that have high prevalence globally and are associated with multimorbidity and polypharmacy including type 2 diabetes mellitus, arterial hypertension, chronic heart failure, asthma, and osteoarthritis [10,11]. The review consisted of the appraisal of clinical guidelines and primary peer-reviewed studies of any design that were carried out mainly in primary care.

Tool Development

The second phase in this study was an adapted Delphi procedure conducted to evaluate the usability of the indicators in practice and to develop a monitoring tool by integrating the indicators in a user-friendly EMR layout. Figure 1 presents the three phases of the study and the associated research steps. The steps for the development of the tool included consensus from general practice experts, an evaluation by specialists, and a draft of a tool that included the relevant indicators. In the following sections, we present the methods in more detail.

General Practice and Expert Consensus

Experts from the region of Zurich were contacted by email. We identified "well-renowned" expert physicians in the region of Zurich that had a special interest in chronic care and work experience in an outpatient setting. The experts were either well-known by some members of our team, who are clinically active physicians, or were identified by asking colleagues. All contacted experts agreed to participate. The study team included two experts in primary care and EMRs (CC and MZ) and an expert in systematic reviews (NM-G) as well as a medical student who was not considered as an expert (LF). For asthma and arterial hypertension, two experts in outpatient care with a special interest in chronic care, Dr. Claudia Steurer-Stey and Dr. Paolo Suter, participated in the study. For type 2 diabetes mellitus, chronic heart failure, and osteoarthritis, Dr. Henryk Zulewski, Dr. Tobias Höfflinghaus, and Dr. Lukas Wildi, PD participated, respectively. We contacted five additional experts

in total, one for each chronic disease. Multimedia Appendix 1 shows a list of all experts and affiliations.

Assessment of the Monitoring Indicators

Experts in any of the five chronic diseases evaluated the set of indicators by means of an adapted Delphi procedure. The Delphi procedure is a structured communication method consisting of a panel of participants and experts on a certain topic [12]. Multiple rounds are conducted on a specific topic, and in every new round the decision of one expert is influenced by the anonymous decisions of the rest of the expert panel in the previous round. The Delphi procedure has been proven to be a feasible method in evaluating indicators for chronic diseases and for generating a consensus [13,14]. We use the term *adapted Delphi procedure*, because the evaluation rounds were not anonymous and consisted of only a few participants. In total, we performed three rounds.

The first round consisted of face-to-face meeting sessions with our study group of four, including two experts in primary care and EMRs, one expert in systematic reviews, and a doctoral student (Multimedia Appendix 1). Each indicator was categorized into one of four types of data elements: 1) to be part of the monitoring at least annually, 2) data normally included in EMRs, 3) data not to be collected at a regular basis, and 4) data that should not be collected at all. Table 1 exemplifies this approach for diabetes mellitus. Before the second discussion round, we excluded data the experts indicated should not be collected.

Table 1. Extract from the set of indicators for diabetes mellitus type 2 identified by the systematic review and categorized based on the adapted Delphi procedure.

Indicators	Considered to be part of the monitoring at least annually	Data normally included in EMR	Data not to be collected at a regular basis	Data that should not be collected
Diabetes education history	X	—[a]	—	—
Current treatment	—	X	—	—
Weight history	—	—	—	X
Vaccination status	—	X	—	—
Physical activity patterns	—	—	X	—
Heart rate	—	—	—	X
ECG	—	—	—	X
Self-monitoring of urine glucose	—	—	—	X
Nutritional status	—	—	—	X
Teeth's condition	—	—	X	—
Eye examination	X	—	—	—
HbA$_{1c}$	X	—	—	—
Inspection of skin	X	—	—	—
Hyperkeratosis	X	—	—	—
Dryness	X	—	—	—
Dilated veins	X	—	—	—
Skin examination for insulin injection sites	X	—	—	—

[a]Not applicable.

In a second round, the study group and one consulting expert of each specialty discussed the condensed set of indicators resulting from the first round (Multimedia Appendix 1). The discussion led to an even more condensed set of indicators, which the study group further re-evaluated in a third round to focus on feasibility and exclusion of redundant indicators.

Design and Development of the Monitoring Tool

Based on a table of the condensed set of indicators, we developed a framework table for each condition, including 1 to 4 sublayers, to provide a structure for the indicators. We introduced the different layers to optimize usability and improve the overview. Layer one is the first visible layer when the tool opens. Each indicator represents itself or its own category of sub-indicators. When an indicator is selected from the first layer, the subsequent layer becomes visible. We designed the monitoring tool in a layout format that enables its integration into an EMR.

Results

Delphi Procedure

The adapted Delphi procedure resulted in a thorough set of indicators, since only relevant and practical (ie, useful and operable) indicators were selected. The systematic review provided 1162 indicators for the five chronic conditions; however, only 25.47% (296/1162) were considered by the experts as being of high enough relevance and feasibility to be

implemented in the monitoring tool, including 20.48% (51/249) of diabetes mellitus indicators (ie, 12 in additional tab, 4 in cockpit, 11 in first layer, 21 in second layer, and 3 in third layer), 26.78% (49/183) of asthma indicators (ie, 12 in additional tab, 3 in cockpit, 7 in first layer, 11 in second layer, and 16 in third layer), 14.63% (49/335) of arterial hypertension indicators (ie, 12 in additional tab, 4 in cockpit, 3 in first layer, 23 in second layer, and 7 in third layer), 33.33% (77/231) of chronic heart failure indicators (ie, 12 in additional tab, 6 in cockpit, 6 in first layer, 40 in second layer, and 13 in third layer), and 42.68% (70/164) of osteoarthritis indicators (ie, 12 in additional tab, 7 in cockpit, 7 in first layer, 17 in second layer, 20 in third layer, and 7 in fourth layer).

The Monitoring Tool

The face-to-face discussion sessions about the eligibility and relevance of the indicators resulted in a condensed set of relevant and practical indicators as part of an EMR monitoring tool. During the Delphi procedure, which was primarily meant to discuss the relevance and feasibility of the indicators, an additional subject that was determined to be important was the ergonomics of how the indicators should be displayed. Ergonomics within the tool were uniformly identified by all involved specialists as an essential element to achieve acceptance of a new monitoring tool. For each selected disease, the indicators were categorized and linked into sublayers, which can be accessed depending on the requirements of the user. The first layer gives an overview of the most important indicators

or categories. Clicking on each layer opens a set of further indicators. For a clear design and ergonomic use, we did not exceed four layers. The tool contains an additional tab with nondisease-specific information such as allergies, smoking, or drinking habits. This tab can be accessed at any point during tool use. The tool is completed by a "cockpit", which serves as a guidance in the process of disease management. The cockpit includes individually predefined treatment goals, and thus enables benchmarking in the monitoring process. In addition, the monitoring interval can also be documented. To guarantee individual adjustments, blank spaces described as "free text" are added in every layer. Figures 2-6 show the final design of the suggested monitoring tool that evolved during the Delphi procedure. Figure 7 shows an example of how some of the indicators of type 2 diabetes mellitus could be integrated into the EMR.

Figure 2. Monitoring of diabetes mellitus type 2. Light blue to dark blue represents layers one to three. HbA$_{1c}$: glycated haemoglobin; ACE: angiotensin-converting enzyme; CKD-EPI: Chronic Kidney Disease Epidemiology Collaboration.

Figure 3. Monitoring of asthma. Light blue to dark blue represents layers one to three.

Cockpit

- Aim: Asthma control
- Self-monitoring/self-mangegement indicated? How? How often?
- Interval of monitoring
- Free text

Additional tab

- Personal history
- Medication (including insulin therapy)
- Compliance (how confident are you that you take all medication as prescribed? [scale 1-10])
- BMI (date of last measurement)
- Waist/hip ratio (date of last measurement)
- Blood pressure (date of last measurement)
- Heart rate (date of last measurement)
- Breathing rate (date of last measurement)
- Allergies/side effects
- Smoking, alcohol, drugs
- Vaccination
- Special features of patient (e.g. difficult to treat disease?)

Asthma control

Diary —
- Peak expiratory flow self-monitoring
- Coughing
- Dyspnea
- Sputum
- Other symptoms (e.g. awakening at night, limitation of activity)
- On-demand medication

Problems with inhalation medications

- Concerns with regard to cortisone
- Experience with cortisone?
- Side effects
- Monitor the inhaler technique (let patient demonstrate technique) compliance (how confident are you that you take all medication as prescribed? [scale 1-10)])

  - Shaking
  - Nervousness
  - Thrush
  - Hoarseness
  - Haematoma
  - Glaucoma/ cataract

Action plan for self management

Auscultation

Competence of self-management given?

- Patient education in general
- Patient education regarding self-monitoring

Controlled, partially controlled, not controlled?

- Exacerbation — Depending on symptom control (investigation of allergy taken place?) Trigger
- Continuative diagnostics indicated? — Lung function Stress test

Smoking

Smoking stop (medicinal/professional consultancy taken place?)

XSL•FO

RenderX

Figure 4. Monitoring of arterial hypertension. Light blue to dark blue represents layers one to three. TIA: transient ischemic attack.

Cockpit

- Aim: Asthma control
- Self-monitoring/self-mangegement indicated? How? How often?
- Interval of monitoring
- Free text

Additional tab

- Personal history
- Medication (including insulin therapy)
- Compliance (how confident are you that you take all medication as prescribed? [scale 1-10])
- BMI (date of last measurement)
- Waist/hip ratio (date of last measurement)
- Blood pressure (date of last measurement)
- Heart rate (date of last measurement)
- Breathing rate (date of last measurement)
- Allergies/side effects
- Smoking, alcohol, drugs
- Vaccination
- Special features of patient (e.g. difficult to treat disease?)

Asthma control

Diary —
- Peak expiratory flow self-monitoring
- Coughing
- Dyspnea
- Sputum
- Other symptoms (e.g. awakening at night, limitation of activity)
- On-demand medication

Problems with inhalation medications

- Concerns with regard to cortisone
- Experience with cortisone?
- Side effects
- Monitor the inhaler technique (let patient demonstrate technique) compliance (how confident are you that you take all medication as prescribed? [scale 1-10)])

- Shaking
- Nervousness
- Thrush
- Hoarseness
- Haematoma
- Glaucoma/cataract

Action plan for self management

Auscultation

Competence of self-management given?

- Patient education in general
- Patient education regarding self-monitoring

Controlled, partially controlled, not controlled?

- Exacerbation — Depending on symptom control (investigation of allergy taken place?) Trigger

- Continuative diagnostics indicated? — Lung function Stress test

Smoking

Smoking stop (medicinal/professional consultancy taken place?)

XSL•FO
RenderX

Figure 5. Monitoring of chronic heart failure. Light blue to dark blue represents layers one to three. NTproBNP: N-terminal pro b-type natriuretic peptide; ICD: International Classification of Diseases; NYHA: New York Heart Association; COPD: chronic obstructive pulmonary disease; NSAR: nonsteroidal antirheumatics; CKD-EPI: Chronic Kidney Disease Epidemiology Collaboration; TSH: thyroid stimulating hormone.

Cockpit

- Target blood pressure
- Target weight
- Self-monitoring indicated? (how? how often?)
- Interval of monitoring
- Free text

Additional tab

- Personal history
- Medication (including insulin therapy)
- Compliance (how confident are you that you take all medication as prescribed? [scale 1-10])
- BMI (date of last measurement)
- Waist/hip ratio (date of last measurement)
- Blood pressure (date of last measurement)
- Heart rate (date of last measurement)
- Breathing rate (date of last measurement)
- Allergies/side effects
- Smoking, alcohol, drugs
- Vaccination
- Special features of patient (e.g. difficult to treat disease?)

Red flags

- Impaired vision (frequent change of glasses or contact lenses?)
- Dizziness when changing position? ── Orthostatic blood pressure measurement
- Loss of sensitivity or strength? (TIA)
- Loss of strength
- Chest pain
- Dyspnea (on exertion/at rest/ orthopnea)
- Ankle edema
- Erectile dysfunction
- Change in weight
- End organ damage ── Microalbuminuria / Left ventricular hypertrophy / Fundus hypertonicus
- Hospitalization
- Hyperkalemia

Measurements

- Self monitoring ── Blood pressure sitting and standing (patient calm and relaxed) / Monitor weight
- In office
- 24-hour- blood pressure measurement ── Dipper/non-dipper

If patient is not responding to therapy consider

- Problem of compliance? (e.g. does patient believe in the efficiency of the therapy?)
- Ask patient for any possible reason for high blood pressure
- Overconsumption of salt
- Family history of hypertension (e.g. M. Conn)
- Inexplicable hypokalemia
- Increase of creatinine
- Poorly adjusted risk factors
- Depression

XSL·FO
RenderX

Figure 6. Monitoring of osteoarthritis. Light blue to dark blue represents layers one to four. CPPD: calcium pyrophosphate dihydrate.

Cockpit

- Target dose of medication
- Target weight
- NT-pro BNP value
- Cardiological examination indicated? (if acute decompensation, any explanation?)
- ICD (yes/no, when? dicuss consequences)
- Discuss other options (transplantation, assisted device)
- Discuss palliation (switch off ICD)
- Free text

Additional tab

- Personal history
- Medication (including insulin therapy)
- Compliance (how confident are you that you take all medication as prescribed? [scale 1-10])
- BMI (date of last measurement)
- Waist/hip ratio (date of last measurement)
- Blood pressure (date of last measurement)
- Heart rate (date of last measurement)
- Breathing rate (date of last measurement)
- Allergies/side effects
- Smoking, alcohol, drugs
- Vaccination
- Special features of patient (e.g. difficult to treat disease?)

Red flags

- Dyspnea/orthopnea (NYHA)
- Dizziness/orthostasis
- Fatigue
- Rhythm disturbance ── Palpitation / ICD shock application / Bradycardia
- Edema/change in weigth
- Loss of appetite
- Nycturia
- Side effects of medication ── Confusion / Delirium / Impaired consciousness / Cognitive disorder
- Change in mental status

Electrocardiography

Laboratory

- Sodium
- Potassium
- Creatinine clearance (CKD-EPI)
- Liver enzymes
- Albumin
- Urin analysis
- Ferritin
- TSH
- NT-pro BNP
- If acute deterioration evaluate troponin and creatine kinase

Comorbidities

- Sleep apnea (especially central)
- COPD
- Depression
- Renal failure
- Stroke
- Anaemia
- Atrial fibrillation
- Diabetes mellitus
- Iron deficiency
- NSAR consumption (enhances heart failure)

Patient training

- Signs of decompensation
- Salt overconsumption
- Restrict water consumption ── CAVE: adjust to high temperature and fever
- Symptom diary/monitor weight and vital signs
- Travel advice ── High altitude / Airplane / Medical assistance abroad
- Rehabilitation (out- or inpatient)

Clinical examination

- Nutritional status
- Auscultation/palpation
- Temperature of lower extremity
- Jugular venous distension/ hepatojugular reflux
- Blood pressure (date of last measurement, standing vs. sitting or lying) ── Symptomatic orthostatic response of blood pressure

XSL•FO

RenderX

Figure 7. An example of how indicators (here type 2 diabetes mellitus) could be integrated into EMR. Numbers 1-3 represent layers one to three. The top left box is the cockpit. The image has been generated by Moqups (S.C Evercoder Software S.R.L.).
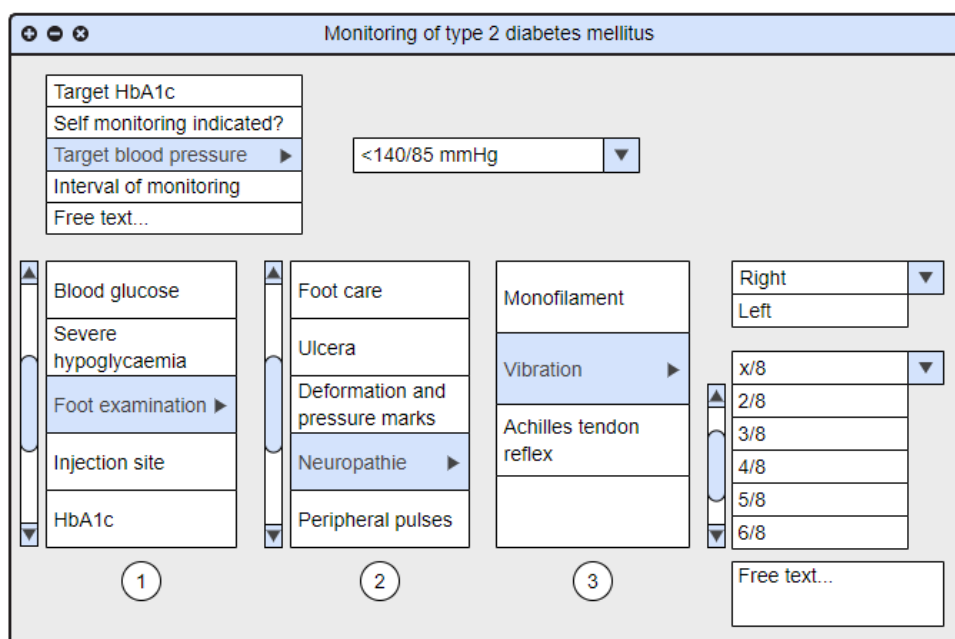
Discussion

To our knowledge, this study represents the first scientifically-based recommendation for the generic, standardized long-term monitoring of patients with chronic illnesses in primary care by means of a user-friendly and clinically sensible EMR layout. Previously, Steenkiste et al [13] used an adapted Delphi procedure to identify items for diagnosis and treatment of asthma for a Dutch electronic patient record. The adapted Delphi procedure proved to be a feasible method for selecting the indicators. However, a complete list of the items or a graphical presentation of the items is lacking in their study. Similar to our study, Lougheed et al [15] used the Delphi procedure and modified RAND/UCLA Appropriateness Method to identify core and optional data elements for asthma and chronic obstructive pulmonary disease to be integrated into electronic heal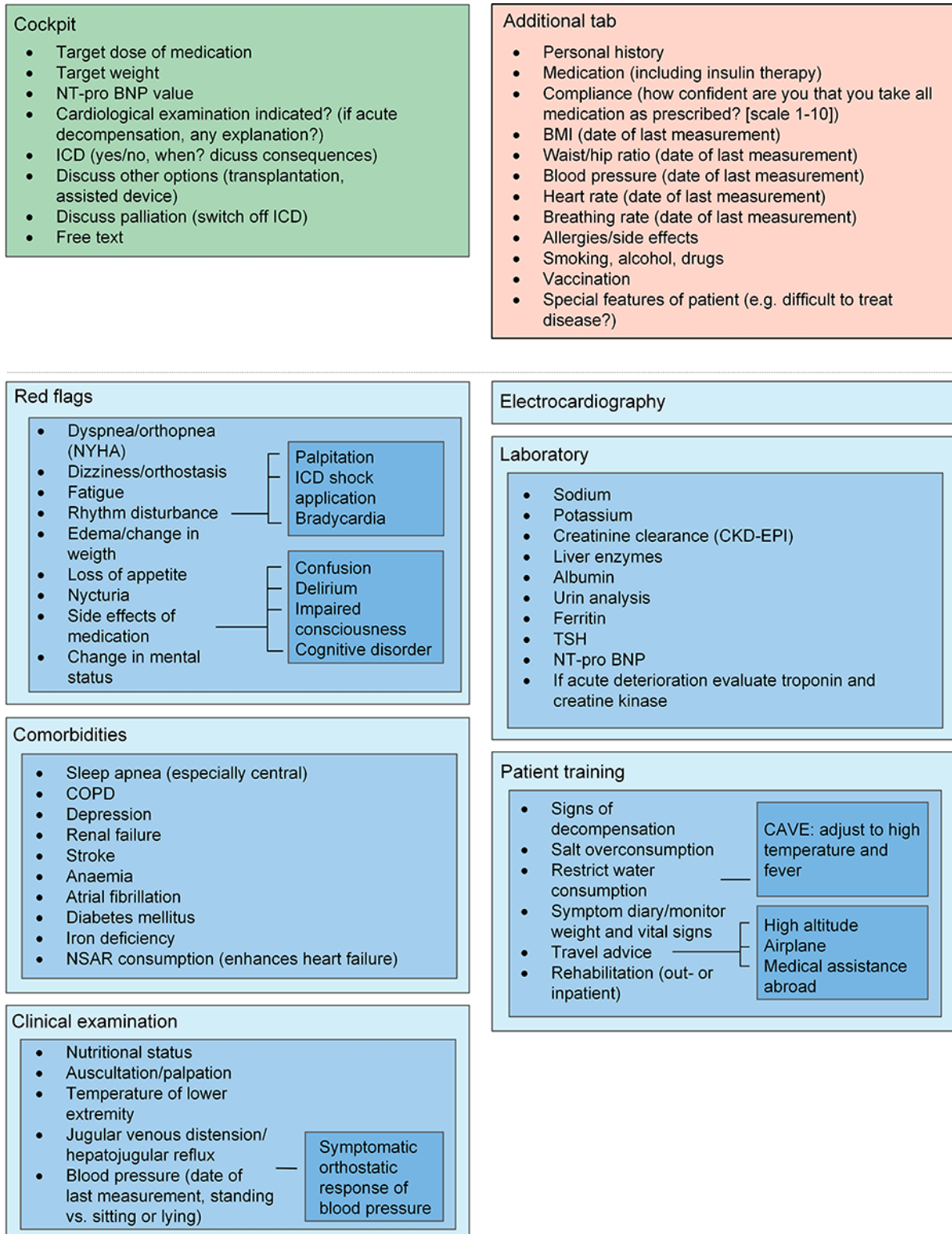th records (EHR) for primary and tertiary care. In contrast to our study, their selection of indicators was not based on a systematic review but on several data sources. Their method also lacked a presentation of potential implementation.

The tool that we developed facilitates the monitoring of patients with chronic diseases by providing all the essential monitoring indicators, which should be assessed at regular intervals according to the current scientific evidence. All indicators are arranged in up to four sublayers that contain only the most relevant indicators. This layout avoids a surplus of information and ensures that the patient, not the computer screen, remains the focus of the consultation. Therefore, all sublayers exceeding the first one are only visible if actively clicked on. The cockpit gives an overview of all preset goals as well as the current situation and, therefore, can serve as a benchmarking tool. By clearly displaying this critical information, the cockpit facilitates a patient handover or holiday replacement and, therefore, enables the continuity of care. Preset goals also help to overcome clinical inertia, a widespread problem in the care of patients

with chronic diseases [16]. Additional tabs that contain nondisease-specific information, such as allergies or body mass index, give an overview on basic but relevant patient information.

In Switzerland, where the study was conducted, EMRs are less developed than in other countries, and the number of physicians still using paper-based medical records is higher than elsewhere [1,3]. Therefore, it is necessary to first provide a functioning EMR basis within primary care. In a second step, it will be desirable to integrate modern applications into the EMR, such as mobile devices that allow patients to be in more control of their chronic conditions. The standardization of EHR tools such as the ones presented here could enable the provision of decision-support tools and add an extension to an EHR. This will link physicians and patients to provide a holistic approach to the process of monitoring.

In long-term care, the involvement of several professionals of different health care disciplines is common. Skill-mix models involving nonphysician disciplines, such as practice nurses, dieticians, or physiotherapists are on the rise; however, due to the specific regulations in different countries concerning allocation of responsibilities, it is not feasible to establish an international standard. This tool will thus have to be adapted according to different health care systems and their needs.

Strengths and Limitations

A strength of this project is the iterative Delphi process that identified the importance of the ergonomic layout of the monitoring tool. Consideration of ergonomics can enhance user-friendliness and facilitate chronic care within an EMR. This tool offers a practical approach for implementing scientific results into everyday practice. By involving generalists with extensive practice experience as well as specialists in different medical fields through an adapted Delphi procedure, a condensed set of indicators were identified as relevant for

everyday use in primary care. A potential limitation of our study is that the adapted Delphi procedure did not meet all criteria of a typical Delphi procedure. Since the discussions were not performed anonymously and the persons involved might have selected the indicators they were most familiar with, this may have increased the chances of bias. Furthermore, the adapted Delphi procedure included in total only ten experts and, in certain phases, only four experts (three experts from the study group including two in primary care and EMRs and one in systematic reviews, and an additional expert in the respective field), which might reduce external validity. Additionally, the complexity of the technical integration of indicators (beyond the visual layout on the screen) is not addressed in this manuscript. Details of EMR integration will vary between the different software companies and the needs of different health care systems.

International Consensus Paper

As mentioned in Figure 1, our final goal is an international consensus paper. This will be achieved by an international Delphi procedure with a larger number of experts in general practice and other specialties. To meet the demands of monitoring the rising number of multimorbid patients, the method we present here is meant to be extended to more diseases that will also be linked to each other. In the future, we hope the tool will become an integral part of the process of collecting patient data, as well as a clinical decision support system that will directly link current guidelines and algorithms with therapy suggestions.

Conclusion

Our generic template systematically integrates the existing scientific evidence for the standardized long-term monitoring of chronic conditions. It contains a user-friendly and clinically sensible layout and can improve the care of patients with chronic diseases by means of an EMR in primary care.

Authors' Contributions

LF was involved in designing the search strategy, performed the systematic screening and review of the literature, the selection of monitoring indicators through the Delphi procedure, and the writing of the manuscript. MZ was involved in study design, took part in the Delphi procedure and revised the manuscript. NAMG was involved in study design and search strategy, systematic review and revised and improved the manuscript. TR supervised the development and methodology of the study and helped improving the final version of the manuscript. CC was involved in study design, the study selection and prioritization, verified the extracted data, took part in the Delphi procedure, supervised and revised the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A list of all experts and their affiliations.
[DOCX File , 17 KB - medinform_v8i3e14483_app1.docx]

References

1. Djalali S, Ursprung N, Rosemann T, Senn O, Tandjung R. Undirected health IT implementation in ambulatory care favors paper-based workarounds and limits health data exchange. Int J Med Inform 2015 Nov;84(11):920-932. [doi: 10.1016/j.ijmedinf.2015.08.001] [Medline: 26296433]
2. Jamoom EW, Patel V, Furukawa MF, King J. EHR adopters vs. non-adopters: Impacts of, barriers to, and federal initiatives for EHR adoption. Healthc (Amst) 2014 Mar;2(1):33-39 [FREE Full text] [doi: 10.1016/j.hjdsi.2013.12.004] [Medline: 26250087]
3. De Rosis S, Seghieri C. Basic ICT adoption and use by general practitioners: an analysis of primary care systems in 31 European countries. BMC Med Inform Decis Mak 2015 Aug 22;15:70 [FREE Full text] [doi: 10.1186/s12911-015-0185-z] [Medline: 26296994]
4. Mitchell E, McConnahie A, Sullivan F. Consultation computer use to improve management of chronic disease in general practice: a before and after study. Inform Prim Care 2003;11(2):61-68 [FREE Full text] [doi: 10.14236/jhi.v11i2.557] [Medline: 14567872]

XSL•FO
RenderX

5.   Cebul RD, Love TE, Jain AK, Hebert CJ. Electronic health records and quality of diabetes care. N Engl J Med 2011 Sep 01;365(9):825-833. [doi: 10.1056/NEJMsa1102519] [Medline: 21879900]

6.   Reed M, Huang J, Graetz I, Brand R, Hsu J, Fireman B, et al. Outpatient electronic health records and the clinical care and outcomes of patients with diabetes mellitus. Ann Intern Med 2012 Oct 02;157(7):482-489 [FREE Full text] [doi: 10.7326/0003-4819-157-7-201210020-00004] [Medline: 23027319]

7.   Hartel MJ, Staub LP, Röder C, Eggli S. High incidence of medication documentation errors in a Swiss university hospital due to the handwritten prescription process. BMC Health Serv Res 2011 Aug 18;11:199 [FREE Full text] [doi: 10.1186/1472-6963-11-199] [Medline: 21851620]

8.   Albarrak AI, Al Rashidi EA, Fatani RK, Al Ageel SI, Mohammed R. Assessment of legibility and completeness of handwritten and electronic prescriptions. Saudi Pharm J 2014 Dec;22(6):522-527 [FREE Full text] [doi: 10.1016/j.jsps.2014.02.013] [Medline: 25561864]

9.   Falck L, Zoller M, Rosemann T, Martínez-González NA, Chmiel C. Toward standardized monitoring of patients with chronic diseases in primary care using electronic medical records: systematic review. JMIR Med Inform 2019 May 24;7(2):e10879 [FREE Full text] [doi: 10.2196/10879] [Medline: 31127717]

10. World Health Statistics 2018: Monitoring Health For The Sustainable Development Goals (sdgs). Geneva: World Health Organization; 2018.

11. Schweizerisches Gesundheitsobservatorium, editor. Gesundheit In Der Schweiz - Fokus Chronische Erkrankungen: Nationaler Gesundheitsbericht 2015. Bern: Hogrefe Verlag; 2015.

12. Fink A, Kosecoff J, Chassin M, Brook RH. Consensus methods: characteristics and guidelines for use. Am J Public Health 1984 Sep;74(9):979-983. [doi: 10.2105/ajph.74.9.979] [Medline: 6380323]

13. van Steenkiste BC, Jacobs JE, Verheijen NM, Levelink JH, Bottema BJ. A Delphi technique as a method for selecting the content of an electronic patient record for asthma. Int J Med Inform 2002 Apr;65(1):7-16. [doi: 10.1016/s1386-5056(01)00223-4] [Medline: 11904244]

14. Grypdonck L, Aertgeerts B, Luyten F, Wollersheim H, Bellemans J, Peers K, et al. Development of quality indicators for an integrated approach of knee osteoarthritis. J Rheumatol 2014 Jun;41(6):1155-1162. [doi: 10.3899/jrheum.130680] [Medline: 24737907]

15. Lougheed MD, Taite A, ten Hove J, Morra A, Van Dam A, Ducharme FM, et al. Pan-Canadian asthma and COPD standards for electronic health records: A Canadian Thoracic Society Expert Working Group Report. Canadian Journal of Respiratory, Critical Care, and Sleep Medicine 2018 Nov 05;2(4):244-250. [doi: 10.1080/24745332.2018.1517623]

16. Paul SK, Klein K, Thorsted BL, Wolden ML, Khunti K. Delay in treatment intensification increases the risks of cardiovascular events in patients with type 2 diabetes. Cardiovasc Diabetol 2015 Aug 07;14:100 [FREE Full text] [doi: 10.1186/s12933-015-0260-x] [Medline: 26249018]

Abbreviations

EHR: electronic health record
EMR: electronic medical record.

XSL•FO
RenderX

Review

The Use of Patient-Facing Teleconsultations in the National Health Service: Scoping Review

Micheal O'Cathail[1,2], MB BCh; M Ananth Sivanandan[2], MBBS; Claire Diver[3], PhD; Poulam Patel[1], PhD; Judith Christian[2,3], MD

[1]School of Medicine, University of Nottingham, Nottingham, United Kingdom

[2]Department of Oncology & Radiotherapy, Nottingham University Hospitals NHS Trust, Nottingham, United Kingdom

[3]School of Medicine & Health Sciences, University of Nottingham, Nottingham, United Kingdom

Corresponding Author:
Micheal O'Cathail, MB BCh
Department of Oncology & Radiotherapy
Nottingham University Hospitals NHS Trust
Hucknall Road
Nottingham
United Kingdom
Phone: 44 01159691169 ext 72149
Email: mocathail@gmail.com

Abstract

Background: The National Health Service (NHS) *Long-Term Plan* has set out a vision of enabling patients to access digital interactions with health care professionals within 5 years, including by video link.

Objective: This review aimed to examine the extent and nature of the use of patient-facing teleconsultations within a health care setting in the United Kingdom and what outcome measures have been assessed.

Methods: We conducted a systematic scoping review of teleconsultation studies following the Joanna Briggs Institute methodology. PubMed, Scopus, the Cochrane Library, and the Cumulative Index to Nursing and Allied Health Literature were searched up to the end of December 2018 for publications that reported on the use of patient-facing teleconsultations in a UK health care setting.

Results: The search retrieved 3132 publications, of which 101 were included for a full review. Overall, the studies were heterogeneous in design, in the specialty assessed, and reported outcome measures. The technology used for teleconsultations changed over time with earlier studies employing bespoke, often expensive, solutions. Two-thirds of the studies, conducted between 1995 and 2005, used this method. Later studies transitioned to Web-based commercial solutions such as Skype. There were five outcome measures that were assessed: (1) technical feasibility, (2) user satisfaction, (3) clinical effectiveness, (4) cost, (5) logistical and operational considerations. Due to the changing nature of technology over time, there were differing technical issues across the studies. Generally, teleconsultations were acceptable to patients, but this was less consistent among health care professionals. However, among both groups, face-to-face consultations were still seen as the gold standard. A wide range of clinical scenarios found teleconsultations to be clinically useful but potentially limited to more straightforward clinical interactions. Due to the wide array of study types and changes in technology over time, it is difficult to draw definitive conclusions on the cost involved. However, cost savings for health care providers have been demonstrated by the goal-directed implementation of teleconsultations. The integration of technology into routine practice represents a complex problem with barriers identified in funding and hospital reimbursement, information technologies infrastructure, and integration into clinicians' workflow.

Conclusions: Teleconsultations appear to be safe and effective in the correct clinical situations. Where offered, it is likely that patients will be keen to engage, although teleconsultations should only be offered as an option to support traditional care models rather than replace them outright. Health care staff should be encouraged and supported in using teleconsultations to diversify their practice. Health care organizations need to consider developing a digital technology strategy and implementation groups to assist health care staff to integrate digitally enabled care into routine practice. The introduction of new technologies should be assessed after a set period with service evaluations, including feedback from key stakeholders.

Introduction

Telemedicine is a branch of medicine, which concerns the use of information technologies (IT) in all aspects of medical care and education. A literary consensus defined telemedicine as: "…a subset of telehealth, uses communications networks for delivery of health care services and medical education from one geographical location to another, primarily to address challenges such as uneven distribution and shortage of infrastructural and human resources" [1]. Common examples of telemedicine include using telephones for patient interaction, videoconferencing with multidisciplinary team meetings, and the use of email in professional practice. Many of these technologies are considered integral to routine clinical practice. The National Health Service (NHS) refers to telemedicine as being synonymous with teleconsultations, involving a video link with patients [2]. To avoid confusion with other definitions, this review shall use the term teleconsultations rather than telemedicine.

Teleconsultations have the potential to improve access to medical care and reduce travel and costs for patients while maintaining the quality of care [3]. The NHS's recently published *Long-Term Plan* has set out a vision of how to transform outpatient care using technology. It states the desire to offer all patients the choice of digital interaction, including the use of teleconsultations, within 5 years, and to remove 30 million face-to-face appointments [4]. With such an ambitious plan, this review looks at the UK evidence of teleconsultation use for patient-facing interactions.

Although a systematic review may provide evidence for how effective an intervention is based on a predetermined study type, usually a randomized controlled trial (RCT), a scoping review can answer the broader question of what is already known; what the extent, nature, and range of intervention use is, and allows for greater inclusivity of different study types [5]. The objective of this review was to map the available evidence in relation to the use of patient-facing teleconsultations in the NHS. A review of the literature before commencing this review identified no existing systematic or scoping review that addressed this issue.

Methods

Methodological Framework

This review was guided by the methodological framework devised by Arksey and O'Malley [5], and further amendments that were contributed by Levac et al [6] and the Joanna Briggs Institute on conducting systematic scoping reviews [7]. This framework consists of a number of consecutive stages: (1) identifying the research question, (2) identifying relevant studies; (3) study selection; (4) charting the data; (5) collating, summarizing, and reporting results. This methodology summarizes the evidence available on a topic to convey the breadth and depth of that topic. We used the Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA) extension for Scoping Reviews checklist to report our results [8]. At present, the international Prospective Register of Systematic Reviews does not publish protocols for scoping reviews.

Identifying the Research Question

The purpose of this review was to find out what health care settings in the United Kingdom teleconsultations have been used in. The broad research questions of this review were as follows: What is the extent and nature of use of patient-facing teleconsultations within a health care setting in the United Kingdom and what outcome measures have been assessed?

Identifying the Relevant Studies

Information Sources and Search Strategies

As this review is interested only in the UK-based experience of teleconsultations, the study's search strategy was restricted to the United Kingdom or NHS affiliated authors. Databases searched were PubMed, Scopus, Cochrane library, and the Cumulative Index of Nursing and Allied Health Literature. Studies up to the end of 2018 were included with no predetermined lower range. The search strategy was developed in PubMed and translated into other databases. This is outlined in Table 1. Search results were exported to the Mendeley reference manager (Elsevier) and duplicated results were removed.

Table 1. PubMed search strategy.

Search field	Search term
Intervention	(Teleconsultation* OR telemedicine OR virtual clinic* OR video clinic* OR virtual consultation* OR video consultation*)
Restricted to UK-based authors	AND (UK[Affiliation] OR NHS[Affiliation] OR United Kingdom[Affiliation])
Date restrictions	Studies up to December 31, 2018
Total articles	2065

Eligibility Criteria

Types of Participants

This review included all participants that used teleconsultations in a health care setting. The only restriction was geography, as the area of the interest is specifically the NHS in the United Kingdom. All studies which used teleconsultations for direct patient-facing care were included. Any studies that used teleconsultations in a nonpatient facing capacity (eg, professional to professional teleconsultations for multidisciplinary meetings) were excluded.

This review was limited to studies that were conducted in the United Kingdom, and there was no restriction on the specialty or type of professionals involved in the consultations. Studies up to the end of 2018 were eligible for inclusion. All studies up to December 31, 2018, were included, and the date of the last search was on February 7, 2019. Study titles and abstracts were independently screened by two reviewers (MOC and MAS) based on predetermined inclusion and exclusion criteria, which are outlined in Table 2. Where abstracts were not available, these articles were excluded.

Table 2. Inclusion and exclusion criteria.

Search Parameter	Inclusion criteria	Exclusion criteria
Population	• Any health care setting	• Non–health care setting
Date	• Up to December 31, 2018	• N/A[a]
Study type	• No restrictions	• N/A
Intervention	• Teleconsultations involving real-time video link with patients	• No video link–based telemedical intervention • Not patient-facing (eg, teleconference multidisciplinary team meeting) • Not real time (eg, store and forward models in teledermatology)
Location	• United Kingdom/NHS[b]	• Non-UK-based studies

[a]N/A: not applicable.

[b]NHS: National Health Service.

Study Type

There was no restriction on the study type eligible for inclusion.

Study Selection

All studies up to December 31, 2018, were included, and the date of the last search was on February 7, 2019. Study titles and abstracts were independently screened by two reviewers (MC and MS) based on predetermined inclusion and exclusion criteria, which are outlined in Table 2. Where abstracts were not available, these articles were excluded. If the study suitability was not clear from the abstract, the full paper was requested for review. Disagreements between reviewers were resolved through consensus. The reasons for exclusion were only recorded at the full-text stage.

Charting the Data

The research team developed a data extraction tool that included the following items: (1) article identifiers (ie, year of publication, author, and title), (2) study identifiers (ie, study design and sample size), (3) setting/population (ie, area of medical specialty), (4) outcome measures assessed, and (5) brief article synopsis.

Data were extracted by one reviewer and verified by a second. The table charting of these articles in temporal order is shown in Multimedia Appendix 1 [9-109].

Collating, Summarizing, and Reporting the Results

A descriptive numerical summary of the characteristics of the included studies was performed. Tables and graphs were created to reflect the overall number of studies included, study designs and settings, publication years, and the outcomes reported. In line with the methodology of scoping reviews, an assessment of the quality of the included studies was not performed.

Statement of Patient and Public Involvement

This research was conducted without patient involvement. Patients were not invited to comment on the study design and were not consulted to develop patient-relevant outcomes or interpret the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy.

Results

Study Characteristics

A total of 3132 articles were retrieved. In total, 140 full texts were retrieved, with 101 meeting the inclusion criteria for review. The PRISMA flowchart is shown in Figure 1.

Figure 1. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow chart—teleconsultations in the National Health Service.
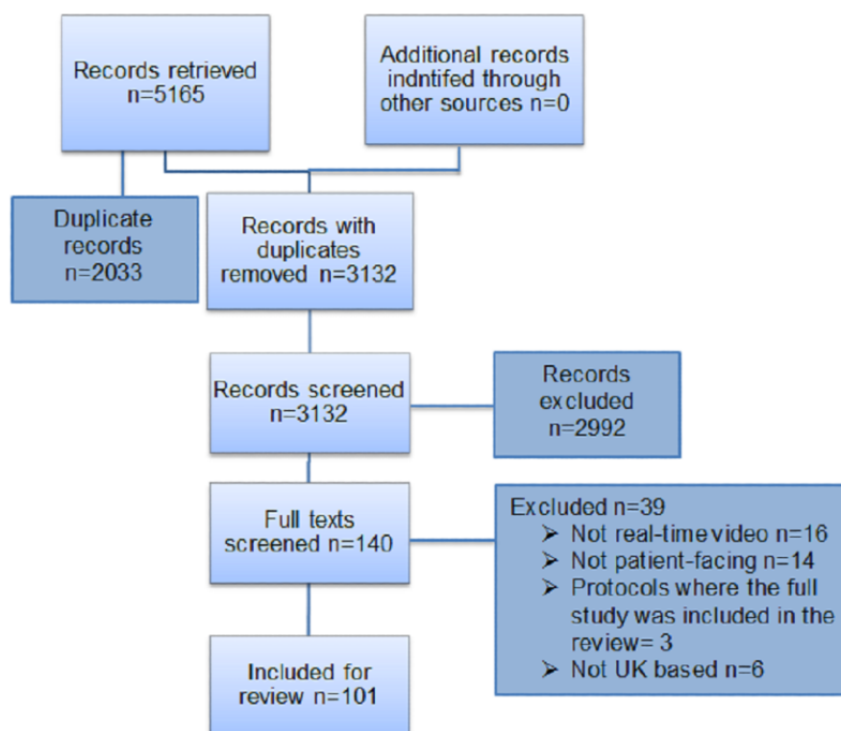
What is the Extent and Nature of the Use of Patient-Facing Teleconsultations Within a Health Care Setting in the United Kingdom?

There were a total of 101 studies across 24 different specialties included in the review starting in 1995 and ending in 2018. There was a large variation in study type, shown in Table 3. Pilots, audits, service reports, and case series/reports represented more than half of the articles included for review, whereas there were 13 RCTs [21,25,26,40,41,59-61,64,71,74,84,87]. Of these RCTs, there were 8 papers written about 4 RCTs [25,26,40,41,59,60,64,71]; therefore, only 9 could be considered unique study populations. Of these, 4 [61,74,84,87] had 30 or fewer participants, including one with 6 [61] and another with 11 participants [87].

Accident and emergency (A&E) was the most common single specialty studied with 19 articles, followed by psychiatry/psychology, neurology, and acute stroke. There were 15 studies that included more than one specialty, sometimes with a mix of primary and secondary care. Most specialties, however, were represented in 2 or fewer papers. These are shown in Table 4.

The timeline of published articles shows that there was significant interest in the potential of teleconsultations in the early 2000s, but this reduced significantly from 2003 until 2017 when publication numbers began to rise again. This is shown in Table 5.

Table 3. Summary of article types.

Article type	Count, n
Systematic review	4
Reviews	7
Randomized control trial	13
Cohort	2
Single-cluster, balanced crossover, blind	1
Mixed method	2
Qualitative study	12
Case-control	8
Retrospective study	5
Service report	2
Audit	3
Pilot	30
Case report/series	6
Descriptive study	3
Study protocol	3
Total	101

Table 4. Publications by specialty.

Specialty	Publications, n	Participants, n
Accident and emergency	19	7394
Multiple	15	2832
Psychiatry/psychology	15	264
Neurology	8	471
Stroke	7	356
General practice	6	884
Pediatrics	5	662
Dermatology	4	678
Orthopedics	2	71
Nephrology	2	16
Respiratory	2	71
Rehabilitation	2	1
Speech and language therapy	2	15
Ophthalmology	2	47
Rheumatology	2	120
Gastrointestinal/hepatology	1	80
Endocrinology	1	480
Care of the elderly	1	Unspecified
Genetics	1	37
Dietetics	1	30
Dentistry	1	25
Transplant medicine	1	180
Oncology	1	8

Table 5. Number of publications by year

Year	Publications, n
1995-2001	39
2002-2007	32
2008-2013	15
2014-2018	15
Total	101

What Outcome Measures Have Been Assessed in UK-Based Studies?

There are five main outcomes that were identified. These are (1) technical feasibility; (2) user acceptability; (3) clinical effectiveness; (4) economic assessment; and (5) logistical and operational considerations. Each section will provide an overview of the article types and a narrative summary of the findings of these studies.

Technical Feasibility

Technical feasibility relates to practical issues with using the technology used for teleconsultations, for example, where a study reports difficulties with the audiovisual link. We identified 18 articles, which reported aspects related to technical feasibility of which there were 8 pilots [14,32,37,39,43,56,82,86], 2 case reports/series [17,36], 1 descriptive study [46], 1 RCT [61], 2 systematic reviews [73,75], 2 nonsystematic reviews [54,67], 1 mixed method study [105], and 1 study protocol [38].

In total, 11 studies, including both systematic reviews, reported acceptable quality and reliability of the teleconsultation link. In total, 3 studies reported problems with audio or video quality, and this tended to follow the perceived importance of that deficiency. For instance, one study reported that the image quality of a video link was so poor such that 46% of dermatologists felt the diagnostic accuracy would be adversely affected [14]. In another, among peritoneal dialysis patients, poor image quality affected its utility in assessing Tenckhoff catheter sites [61]. A teletherapy study found that sound lag affected the flow of therapy at times, although it concluded that this was not prohibitive to continuing the session [43].

In a study designed to assess the feasibility of a dedicated teleconsultation link in a police college, the authors reported significant reliability issues, where only half of the intended patients were able to be seen by teleconsultation due to unspecified technological failures [86]. A more recent study, using commercial Web-based video calling technology, found that technical issues were minor but often prohibitive to proceeding with the consultation if not rectified. Workarounds by the clinician rectified these; loss of sound on two occasions was overcome by using a telephone for audio, and lack of video was found to be due to the patient forgetting to turn on the Web camera [105].

User Satisfaction and Experience

Satisfaction and user experience with teleconsultations was reported in 43 articles of which there were 17 pilots [9,14-16,23,39,43,45,50,52,56,70,77,81,85,97,108], 11

qualitative studies [33,51,57,76,89,93,98,100,101,107,109], 4 RCTs [21,40,59,61], 2 mixed method studies [102,105], 3 trial protocols [38,68,106], 2 case-control studies [42,69], 1 cohort [18], 1 systematic review [31], 1 nonsystematic review [54], and 1 descriptive study [47].

Satisfaction was assessed using feedback questionnaires in 23 articles, including 3 RCTs. Of these, 19 reported high levels of satisfaction with the medium. In 2 RCTs, satisfaction in the teleconsultation arm was actually greater than the face-to-face group [21,59], whereas another found no difference [40]. By contrast, only 1 small RCT found that patients were less satisfied with teleconsultations due to poor image and audio quality [61]. One pilot found that, while patients were satisfied, health care staff were uncomfortable with it; citing that they felt more *on show* to senior colleagues and families than would be normal in a face-to-face appointment [70].

A systematic review of patient satisfaction with teleconsultations concluded that although the published evidence suggests that teleconsultations appear to have high satisfaction rates in a variety of settings, we should be cautious about interpreting that as a true reflection of real life [31]. The authors suggest that most studies conducted tried to minimize the inconvenience for those taking part, and often, patients were seen both in person and by teleconsultation.

Patients and staff may be satisfied with teleconsultations, but that is not to say that they are preferable to face-to-face consultations. Several studies found that patients were satisfied with teleconsultations but also that they would still want the option to attend in person as they believe it to be the *gold standard* [77,107,109].

Qualitative studies exploring users' experiences of teleconsultations find that the main benefits commonly reported by patients are convenience, reduced travel, and greater accessibility to specialist care and improved flexibility of appointments, allowing minimal disruption to daily life [102,107]. Several studies found that the medium allowed patients to *open up* more than face-to-face consultations and that they felt empowered to ask more questions [57,107]. Among staff, a greater sense of job satisfaction and a reduced burden of travel have been reported [101]. Among a cancer population, participants reported a preference for receiving *bad news* in the comfort of their homes rather than in hospital [105].

By contrast, among a teenaged population being treated for chronic fatigue, participants raised concerns about privacy, fearing that they might be overheard by family. Their parents worried that the connection might not be secure enough to

ensure privacy, while some health professionals thought it was an invasion of patients' personal space [107]. There was the awareness that teleconsultations had certain *physical* limitations; the qualitative analysis from the large RCT by Wallace et al [59] found, either due to patient expectation or physician need, that the inability to perform physical examinations limited its usefulness [76]. A recent study found that physicians often restricted who was offered teleconsultations based on preconceived impracticalities, or they simply refused to participate in them [105].

One study, in which teleconsultations between patients and hospital specialists were facilitated by general practitioners (GPs), concluded that teleconsultations had a different dialogue flow than traditional face-to-face appointments. In particular, the opening phase of the consultation was found to be unfamiliar, leading to interruptions and apologies on both sides while a dialogue flow was established [89]. Morris et al [102] reported that patients and staff could find the medium awkward and uncomfortable when there was no previous relationship built up. The authors concluded that when there were staff changes in service or new-patient appointments, teleconsultations would not be appropriate. Haig-Ferguson et al [107] found that some participants felt teleconsultations were less personal and that the therapist was less *real* over a video link, with the screen acting as a physical and emotional barrier. Paradoxically, the same study found that being physically removed from the therapist allowed other participants to open up more easily. Due to these potential social difficulties, authors have suggested that teleconsultations are more appropriate for follow up appointments [102,107].

A telestroke study found that the utility of teleconsultations in facilitating timely care was acknowledged by families of acute stroke sufferers but that the clinical expertise of the on-site team was important for them to have confidence in the process [98].

Clinical Effectiveness

Where articles commented on the efficacy, safety or other clinical outcomes, these were categorized as assessing the clinical effectiveness of teleconsultations in delivering health care. There were 48 articles of which there were 12 pilots [14,15,30,34,43,48,62,63,65,86,88,108], 5 case reports/series [19,22,36,72,99], 7 case-controls [28,42,44,49,53,69,78], 1 qualitative study [10], 2 mixed method [102,105], 4 retrospective studies [35,58,80,92], 3 audits [83,94,95], 7 RCTs [26,40,59,64,74,84,87], 1 single-clustered, blinded crossover design study [79], 1 cohort [66], 2 systematic reviews [73,75], 1 nonsystematic review [54], 1 descriptive study [90], and 1 service report [96].

Within psychiatry, a single-cluster balanced crossover, blind study (where each patient had both a face-to-face and teleconsultation with a different researcher and each researcher was blind to the psychiatric assessment of the other) concluded that there was significant intermethod concordance, confirming its accuracy in psychiatric assessment [79]. This confirms the findings of an earlier systematic review [75].

Teleconsultations in acute stroke management networks are now widespread in the United Kingdom. In total, 3 retrospective studies of a combined 287 patients conducted in the United Kingdom confirm that its implementation has been safe; *door-to-needle time*, morbidity, mortality, and discharge rates were comparable to national standards for acute stroke management [92,94,95]. A novel study exploiting an inherent advantage of teleconsultations describes an international telestroke service between Scotland and New Zealand. In this small case series, there were no negative patient outcomes, and the authors suggest that utilization of the time difference would avoid doctor fatigue [99].

In total, 2 case-control studies in neurology assessed the concordance of diagnosis in both an inpatient and outpatient setting and found 96%-100% of cases were accurately diagnosed and managed via teleconsultation [28,29]. An RCT in a neurology outpatient setting compared face-to-face consultations with teleconsultations and found that the teleconsultation arm generated more investigations despite no difference in the diagnostic category of the cases seen. The authors conclude that this reflected a lack of confidence in their teleconsultation diagnosis [40]. A cohort study of 111 inpatients assessed by video link found no difference in 3-month mortality compared with all other hospital admissions during that time. On follow-up, no patient had their diagnosis or management changed when seen face-to-face, and no difference was seen in the use of hospital services in the following 3 months after discharge [66].

A large multispecialty RCT, by Wallace et al [59], enrolled over 2000 patients. They measured the number of investigations per patient and follow-up rates and, in contrast to the previously mentioned RCT, found that teleconsultations actually resulted in fewer investigations, at a rate of 0.79 per patient. However, this figure is offset by a higher rate of subsequent follow up seen in this group.

In an A&E setting, an RCT found no significant differences in diagnostic accuracy or management when teleconsultations were compared with the traditional model of care [64]. In minor injury units, the use of teleconsultations, connecting with a regional A&E center, allowed the majority of patients to be managed locally, with continued improvements seen with increasing technological familiarity [48,58,62,65].

In rheumatology, 2 studies reported conflicting findings. Graham et al [30] found rheumatologists—using a junior doctor as a proxy—were only 40% accurate in assessments via teleconsultation with physicians missing subtle but clinically important signs of inflammation. A year later, Leggett et al [42] concluded that teleconsultations—using a 3-way consultation between the patient, GP, and specialist—were 97% accurate in diagnosing fibromyalgia, degenerative arthritis, rheumatoid arthritis, and soft tissue disease.

In ophthalmology, some eye conditions such as simple ptosis and strabismus could be accurately assessed in up to 97% of cases. However, more complex eye conditions such as socket problems in patients who had a previous enucleation or those with nonspecific ocular pain were better assessed in a face-to-face consultation [44,49].

In Airedale NHS trust, providing a teleconsultation link between care homes and hospitals reduced nonelective admissions by 1731—a 37% reduction—compared with the same period before the intervention [96]. In a hospital diabetic clinic setting, over 4 years, appointment *did not attend* (DNA) rates were lower (13% vs 28%) in patients choosing to attend by teleconsultation with improved hemoglobin A_{1c} control [102]. In a prison inmate setting, the use of teleconsultations, coupled with other interventions, improved clinical outcomes for those being screened and treated for hepatitis C compared with controls [108].

Real-world data on teleconsultation appointments as a proportion of clinical activity has been rarely reported and ranges from 2% among a diabetic cohort to 22% among postoperative hepatobiliary cancer patients [105].

Cost

In total, 19 articles looked at health care provider cost, patient cost, or costs incurred by both. This comprised 5 RCTs [25,26,41,60,71], 3 systematic reviews [73,75,104], 2 nonsystematic reviews [54,91], 1 retrospective study [27], 1 case-control study [78], 1 service report [96], 3 pilots [12,43,55], and 3 protocols [38,68,106]. Nearly all of the studies reported higher costs for health care providers, including all RCTs. The issue of cost is closely related to the technology used, which has changed greatly over the period of this review—from expensive audiovisual systems to the use of smartphones and computers.

Early studies found that the initial cost of suitable videoconferencing equipment was prohibitively expensive. One early study quotes a figure of £48,000 (US $61,439) to establish a teleconsultation link, including videoconferencing unit and integrated services digital network connection charges [26]. Loane et al published the results of an RCT in 2 papers [25,26]. They found that real-time teleconsultations were 5 times more expensive to run for health providers than *store-and-forward* teleconsultation models; £132.10 (US $169) vs £29.60 (US $37.90). Although patients saved time and money due to reduced traveling, health care-associated costs were higher in the real-time arm as they took up more physician time than the store and forward model.

Direct comparison with face-to-face appointments in an outpatient setting has found teleconsultations to be more expensive for health care providers. In total, 3 RCTs and 1 case-control study set in pediatrics, neurology, secondary care outpatients, and A&E concluded that teleconsultations were between 15% and 100% more costly to run [41,60,71,78]. A systematic review from an A&E setting found only 23% of studies reported that teleconsultations were cost-effective [73].

Costs could be saved by improving access to specialist care in areas with limited local access to services. A systematic review of telepsychiatry concluded that cost savings would be made by doing just that, and they speculate that technology would become cheaper in the future [75]. A pilot study of teleconsultations in a rural dentistry setting found that up to £270 per patient appointment could be saved by the health service if it adopted teleconsultations to allow rural patients

access specialist services [55]. In Scotland, a report on telehealth services found teleconsultations for a 10-week rehabilitation course could be delivered for 3% to 10% of the cost associated with an outreach model (where the therapist travels) or a centralized model (where the patient travels), with the savings primarily being delivered through reduced travel costs [91].

A service report from a well-established teleconsultation service in Airedale, which links an acute hospital with several care homes, reported setup and maintenance costs of £175,000 (US $223,938). However, factoring in costs from avoided A&E attendances and reduced nonelective admissions, the project is estimated to have saved £1,194,083 (US $1,529,939)—a saving of £6.82 (US $8.73) for £1 (US $1.28) invested [110].

In all, 2 RCTs, in different settings, comparing patient costs reported conflicting results. Jacklin et al [60], assessing teleconsultations in multiple outpatient specialties found those who took part in teleconsultations, saved an average of £19 (US $24.34) compared with face-to-face appointments. However, Noble et al [71], assessing its use in a minor injuries unit setting, found patient costs were nearly £15 (US $19) more.

Logistical and Operational Considerations

There were 16 studies that either assessed or commented on aspects relating to logistics or operational challenges. These included 3 RCTs [26,59,64], 1 case-control study [53], 2 qualitative studies [57,101], 2 mixed method studies of real-world teleconsultation services [102,105], 4 pilots [34,62,70,108], 1 cohort [66], 1 audit [95], 1 nonsystematic review [54], and 1 report [103]. There is considerable variation in the extent to which this is assessed and exactly what logistical element of interest was described.

Several studies make reference to consultation length, in which teleconsultations are as much as 4 times as long as their face-to-face equivalent [53,57,64]. However, several others found them to be shorter in length [34,102,105]. Williams et al [103] found no difference in consultation length but reports that by avoiding travel to peripheral clinics, clinicians were able to provide more emergency care with the time saved, thus maximizing their clinical efficiency. In older studies, there was often an intermediary, either a GP or another health care professional who would sit with the patient and establish the teleconsultation link [26,59,66]. The extent to which this disruption affects the service provision of the health care professional that is acting in this capacity is not described. In addition to normal duties, clinical staff have reported that they are often needed to *triage* those who might be suitable for teleconsultations [105]. Furthermore, clinicians may not have dedicated time to do teleconsultations, having to fit them around their normal outpatient schedules instead [59]. Benger et al [64] found that waiting times to access A&E advice were shorter for patients seen by teleconsultation than face-to-face consultation as they bypassed the normal admission processes—in essence skipping the queue.

Altering the way patients are seen can lead to improved operational efficiencies; Ditchburn et al [101] describe a service established to support patients undergoing peritoneal dialysis at home. By avoiding the need to travel to individual patient

XSL·FO

RenderX

homes, staff reported that their time was used more efficiently as they were able to do other work on their computer while monitoring the patient via the video link. Nonattendance at hospital appointments is a source of lost revenue for health care providers and results in inefficient use of clinician time. By selectively choosing patient populations with high DNA rates, it is possible to achieve more operational efficiency. Morris et al [102], among a diabetic cohort, improved the DNA rate from 28% to 13%. Morey et al [108], among a prison population, describe a complete overhaul of a hepatitis C screening program pathway (along with other measures, teleconsultations were introduced), which led to a significant fall in DNA rates. These examples demonstrate clinical staff as drivers of change, but they can also be barriers to wider implementation [70].

Teleconsultations, as described by many of the studies included, can be seen as supporting a *hub and spoke* model of care, with district general hospital (DGH) *spokes* using teleconsultations to connect with more specialized *hub* hospitals. Agarwal et al [95] describe a telestroke *mesh* network of DGHs without a central *hub*, where out of hours stroke thrombolysis support was provided using telestroke rota shared across the region, thus reducing the frequency of a stroke physician's *on-call* nights. Furthermore, such a model meant that thrombolysis care could continue without significant investment in staff and reorganization of thrombolysis care into a *hub and spoke* model.

The extent to which these services have been integrated into routine practice has been largely superficial. This means that small scale services were often provided with *ad hoc* support from IT departments rather than formal arrangements [101]. To provide a wider rollout of teleconsultations would require dedicated support from IT. Greenhalgh et al [105], through key stakeholder interviews, reported that NHS IT processes would require major changes to speed up the introduction of new technologies into practice.

In the NHS, hospital trusts are reimbursed through tariffs, often based on a *per-patient seen* basis, with different tariffs in place for face-to-face consultations and phone consultations. No such tariff existed for teleconsultations, which means that managers are often unwilling or unable to justify diverting the cost of such services from increasingly stretched clinical budgets [70,105].

Discussion

This scoping review was aimed at assessing the extent of literature around UK-based teleconsultation patient interventions and the main outcome measures. The use of teleconsultations stretches back nearly 25 years, encompassing over 20 different specialties. Most of the specialties are represented in only a few articles and, though quality assessment of articles was not undertaken, pilots and case reports/series represent a significant proportion of that breadth. It is perhaps surprising to see that the decade between 1995 and 2005 accounts for two-thirds of the articles covered by this review, including all but 2 of the RCTs. The reasons behind this are not clear, but it may be a by-product of the challenging public finances since the 2008 economic recession. In that time, the NHS's budget has faced a sustained period of constrained annual growth of 1.1% to 2.3%, compared with an average annual rise of 6% in the

preceding years from 1996 to 2009 [111]. There was a narrowing of clinical focus during these years, perhaps to focus on where the need was most acute; the most enduring success of patient-facing teleconsultations in the United Kingdom is its use in acute stroke management, an intervention which was first reported in 2012 [92].

Nonetheless, the era in which most of these studies were carried out presents a number of problems for modern generalizability. Early studies used bespoke, expensive, complex, and cumbersome systems, which have now been largely superseded by the development of Web-based video calling technology such as Skype (Microsoft Corporation). This, coupled with the rise in smartphone use since 2007, means that videoconferencing technology is now in the pockets of millions of patients [112]. In a United Kingdom setting, however, relatively few studies have been done using this new technology.

Technical Feasibility

The clarity of many older audiovisual connections was criticized—particularly by professionals [14,43], though patients were not universally satisfied that they could see and hear everything that was needed either [39]. The technology employed in these studies has now been superseded by Web-based platforms. Using these modern solutions does not prevent technical issues, and contingency plans need to be considered to overcome common problems, such as poor internet speed and lack of an audiovisual stream with Web-based solutions [105]. Although potentially prohibitive, these technical issues were usually rectifiable to allow continued operation of these services, but it does raise concerns that clinician time is being used inefficiently in such cases.

In total, 9% of the population (disproportionately older people) have never used the internet [113]. This, among other reasons outlined below, makes teleconsultation services unlikely to be accepted as a replacement to traditional care models and more likely that it should be offered as a choice.

User Satisfaction and Experience

For the most part, patients seem to be satisfied with their experience of teleconsultations. Indeed, it seems that most teleconsultation interventions are aimed at improving aspects of the patient experience, such as convenience, rather than improving the experience of health care staff. They recognize its convenience and its utility when accessing specialist care in remote areas, or in time-sensitive matters. In some cases, patient satisfaction was higher than traditional clinic models, although as one author points out, this may be due to the increased accommodation provided to these patients for participation [31]. Avoidance of travel, although also convenient, may prove more pertinent; hospital-associated travel may cause stress in its own right, with 20% of older patients finding simply getting to and from appointments causes increased stress and anxiety [114].

It is interesting to note that among cancer patients, there was a preference for receiving *bad news* at home suggesting that even complex or challenging discussions may be had over teleconsultations. This is not a consistent finding internationally; patients in an acute medical setting had opposing views on receiving bad news over the video link. One patient in favor of

such an approach stated, "If it was something earth-shattering, you could cry in your own bedroom and not have to worry, I mean driving from downtown and you're upset or what-not...." But others were against this, explaining, "If the doctor were telling me I have a fatal disease or a disease that could be fatal, and I have to go into immediate serious care, probably better in-person" [115].

More is not always necessary or better as sometimes the telephone is sufficient. Therefore, it is important to explore if teleconsultations are needed to provide the intended benefit [61]. It should be noted that patients can be quite satisfied with their teleconsultation but still perceive face-to-face appointments to be the gold standard [107,109].

Among health care professionals, the view is more divided. Preconceptions about its utility from the perspective of the health care professional undoubtedly dictate the enthusiasm with which the service is promoted, in some cases, failing to even consider it as an option [105]. It is unlikely that a consensus will ever be so unanimous as to universally accept teleconsultations, and the wide variation of views on utility and acceptability means that they should only be offered as a choice and not a replacement to traditional models of care [76,107].

Clinical Effectiveness

Teleconsultations appear to be a safe and effective way to assess and manage a variety of clinical situations. Clinical consensus, even within specialties, is not universal, however, and the types of consultations that are suitable are dependent on their complexity and physician comfort with the medium. Although physical examination is limited in teleconsultations, there are many examples in both inpatient and outpatient settings that demonstrate its utility. Neurological conditions and simple ophthalmological presentations such as strabismus could be safely diagnosed and managed.

The inability to perform some aspects of physical examination is likely, in some cases, to restrict its utility to more *routine* outpatient appointments. Among an inpatient, A&E, or acute stroke setting, the presence of a *proxy* examiner appears to be an effective way of overcoming this. Although proxy examiners (often GPs) were used in several outpatient-based studies, more recent outpatient studies assessed patients at home without a proxy present. The NHS is experiencing staff shortages, which are most acute in nursing (1 in 10 posts vacant) and general practice (1.6% decline in numbers) [116]. The use of proxy examiners is unlikely to be viable; therefore, the outcomes reported in such studies may not be replicable in today's health service.

An interesting perspective on physical examination is that it has become a ritual, expected, and performed as tradition rather than clinical usefulness [117]. Novel technological solutions already allow certain physiological parameters—such as peak expiratory flow rate, heart rate/rhythm, and remote blood sugar levels—to be monitored remotely [118]. Digital stethoscopes can allow heart sounds to be transmitted via Bluetooth to a connected device [119], and smartphone ophthalmoscopes may be easier to master than direct ophthalmoscopy [120]. Wearable technology continues to develop, and solutions to other more nuanced aspects of physical examination may be developed in the future, however, for the time being, teleconsultations in outpatient settings are most likely to be confined to dialogue-based consultations where the need for rigorous physical examination is absent.

Cost

The nature and method of assessment of cost were assessed in a heterogeneous way, which makes the comparison between studies speculative; however, it is clear that technology-associated costs have changed. Early studies used bespoke technological solutions with often prohibitive setup costs [26]. Technological advances mean that commercial teleconsultation services are fully scalable to the needs of the health care provider. Whereas many older studies almost universally found costs for the health care provider to be higher than the traditional model of care, more recent evidence from NHS Airedale's experience shows that investment in a large-scale service can save significant costs by reducing unplanned admissions [96]. By reducing missed patient appointments, trusts can also make significant cost savings. Notwithstanding a few examples, real-world data on the financial implications of teleconsultations is lacking. Further in-depth case studies and service evaluations of established services are needed to accurately model the financial implications of teleconsultations.

For patients, the potential cost saving is more clear cut where travel and parking fees are only a part of the cost incurred. The true cost of patient time is likely to be much higher with one estimate putting the actual cost at £17.86 (US $22.89) per hour of travel, compared with just £1 (US $1.28) for a digital interaction [121].

Logistical and Operational Considerations

Clinical trials have failed to replicate real-world operational challenges that such a service would create and is a distinct disadvantage to using a clinical trial methodology to assess the utility of digital technology in health care. The successful adoption of technology may be predicated on demonstrating safety and acceptability, but it will only survive in the real world if it can be integrated into existing health care pathways.

Several examples of real-world evaluations of working teleconsultation services have demonstrated that they can achieve meaningful reductions in DNA rates [102,108]. In one example, this required a complete overhaul of the existing clinical pathway, which was not fit for purpose [108]. Notwithstanding this example, a redesign of most clinical and administrative pathways would be a costly and an enormous logistical undertaking. In this, there is a disconnect between policies aimed at promoting more digital technology use and the real-world practicalities of establishing these services in busy and financially stretched hospital trusts. Embedding teleconsultations into routine clinical practice, in reality, has proven more complex than expected [105].

A key goal of the *Long-Term Plan* is to reduce the number of outpatient appointments, which have doubled to 120 million in the last decade. Through the use of technology, the NHS hopes to reduce this by a third over the next 5 years [4]. To achieve

this, there is the promise of *central funding*, which trusts can access for technological improvements. It is not clear that this is ring-fenced in a way that will allow unencumbered access to service development funding. The expectation may be that these services will fund themselves through the anticipated £1 billion a year saving the outpatient reduction will achieve. As provided for in the National Tariff Program, trusts can opt to fund services on a block contract basis rather than a payment by results (PbR) basis. This would allow trusts more flexibility in how they engage with patients. A concern of this model is that unexpected service costs would not be reimbursed; however, PbR-based reimbursement has been criticized for the apparent incentivization of trusts to simply see more patients. NHS England has introduced a new digital tariff into the National Tariff payment system, which provides for reimbursement at 75% the rate of face-to-face consultations [122]. This should provide financial reassurance for trusts using a PbR model. Redesigning payment systems to be more flexible, perhaps to include elements of both, may be needed to overcome the diverse funding needs of the NHS [123].

Where Do We Go From Here?

Well-funded, goal-orientated implementation of teleconsultations has been shown to be viable on a large scale in the NHS [96]. A number of important factors show that it is not known how replicable these results would be at a national level; NHS IT infrastructure is recognized to be stretched, and reluctance from health care professionals can stifle the growth of such innovations. Further investment is needed to address these issues [70,105].

Teleconsultations may not be suitable for every population. Therefore, teleconsultation services should be introduced gradually in a way that allows proper evaluation, with staff and patient feedback being used to fine-tune the pathway to suit local service needs and expectations. Routine clinical interactions are likely to represent the most pragmatic starting point for most services, but that is not to say that

teleconsultations should be limited to such scenarios. To date, clinical interactions have been limited by the ability to perform examinations, but complex scenarios that involve a verbal exchange only, such as breaking bad news to patients with cancer, can be done effectively over a video link. Such interactions necessitate further investigation.

Although the *gold standard* of research methods, more RCTs are, arguably, not the correct way to find these answers. Finch et al [124], in an ethnographic study of telehealth integration into health care, found that participants felt that the RCT design conflicted with the dynamic nature of the health service environment. Participants saw greater value in pragmatic service evaluations that often produced results, which evaluators could use and—unconstrained by a rigorous trial protocol—they could adapt the service more readily to improve the project's stability.

Limitations

Scoping reviews are not intended to assess the quality of the literature included; therefore, the conclusions of this review are based on the existence of published research rather than the quality of it. Nevertheless, this scoping review provides a comprehensive, contemporary overview of the existing research on teleconsultations in a UK setting.

Conclusions

Teleconsultations appear to be safe and effective in the right clinical situations. Where offered, it is likely that patients will be supportive of such measures, although they should only be offered as an option to support traditional care models rather than replace them outright. Health care staff should be encouraged and supported in using teleconsultations to diversify their practice. Health care organizations should consider developing digital technology strategy and implementation groups to assist health care staff in integrating technologically enabled care into routine practice. The introduction of new technologies should be assessed after a set period with service evaluations, including feedback from key stakeholders.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of UK teleconsultation trials.
[DOCX File , 72 KB - medinform_v8i3e15380_app1.docx]

References

1. Sood S, Mbarika V, Jugoo S, Dookhy R, Doarn CR, Prakash N, et al. What is telemedicine? A collection of 104 peer-reviewed perspectives and theoretical underpinnings. Telemed J E Health 2007 Oct;13(5):573-590. [doi: 10.1089/tmj.2006.0073] [Medline: 17999619]
2. NHS England. 2015 Jan. Technology Enabled Care Services URL: https://www.england.nhs.uk/wp-content/uploads/2014/12/TECS_FinalDraft_0901.pdf [accessed 2018-07-03]
3. Verhoeven F, van Gemert-Pijnen L, Dijkstra K, Nijland N, Seydel E, Steehouder M. The contribution of teleconsultation and videoconferencing to diabetes care: a systematic literature review. J Med Internet Res 2007 Dec 14;9(5):e37 [FREE Full text] [doi: 10.2196/jmir.9.5.e37] [Medline: 18093904]
4. NHS Long Term Plan. URL: https://www.longtermplan.nhs.uk/ [accessed 2019-02-14]
5. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32 [FREE Full text] [doi: 10.1080/1364557032000119616]

6. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]

7. Peters MD, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015 Sep;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]

8. Tricco A, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]

9. Ball CJ, McLaren PM, Summerfield AB, Lipsedge MS, Watson JP. A comparison of communication modes in adult psychiatry. J Telemed Telecare 1995;1(1):22-26. [doi: 10.1177/1357633X9500100105] [Medline: 9375115]

10. McLaren P, Ball CJ, Summerfield AB, Watson JP, Lipsedge M. An evaluation of the use of interactive television in an acute psychiatric service. J Telemed Telecare 1995;1(2):79-85. [doi: 10.1177/1357633X9500100203] [Medline: 9375124]

11. Siderfin CD, Haston W, Milne AH. Telemedicine in the British Antarctic Survey Medical Unit. J Telemed Telecare 1995;1(2):63-68. [Medline: 9375122]

12. Darkins A, Fisk N, Garner P, Wootton R. Point-to-point telemedicine using the ISDN. J Telemed Telecare 1996;2(Suppl 1):82-83. [doi: 10.1258/1357633961929385] [Medline: 9375102]

13. Harrison R, Clayton W, Wallace P. The future role of telemedicine at the interface between primary and secondary care. J Telemed Telecare 1996;2(Suppl 1):87-88. [doi: 10.1258/1357633961929411] [Medline: 9375104]

14. Harrison R, Clayton W, Wallace P. Can telemedicine be used to improve communication between primary and secondary care? Br Med J 1996 Nov 30;313(7069):1377-80; discussion 1380 [FREE Full text] [doi: 10.1136/bmj.313.7069.1377] [Medline: 8956708]

15. Jones DH, Crichton C, Macdonald A, Potts S, Sime D, Toms J, et al. Teledermatology in the highlands of Scotland. J Telemed Telecare 1996;2(Suppl 1):7-9. [doi: 10.1258/1357633961929402] [Medline: 9375077]

16. Kunkler IH, Rafferty P, Hill D, Henry M, Foreman D. A pilot study of tele-oncology in Scotland. J Telemed Telecare 1998;4(2):113-119. [doi: 10.1258/1357633981932055] [Medline: 9744168]

17. Curry GR, Harrop N. The Lancashire telemedicine ambulance. J Telemed Telecare 1998;4(4):231-238. [doi: 10.1258/1357633981932299] [Medline: 10505360]

18. Loane MA, Bloomer SE, Corbett R, Eedy DJ, Gore HE, Mathews C, et al. Patient satisfaction with realtime teledermatology in Northern Ireland. J Telemed Telecare 1998;4(1):36-40. [doi: 10.1258/1357633981931254] [Medline: 9640708]

19. Manchanda M, McLaren P. Cognitive behaviour therapy via interactive video. J Telemed Telecare 1998;4(Suppl 1):53-55. [doi: 10.1258/1357633981931452] [Medline: 9640736]

20. Freir V, Kirkwood K, Peck D, Robertson S, Scott-Lodge L, Zeffert S. Telemedicine for clinical psychology in the Highlands of Scotland. J Telemed Telecare 1999;5(3):157-161. [doi: 10.1258/1357633991933567] [Medline: 10628030]

21. Harrison R, Clayton W, Wallace P. Virtual outreach: a telemedicine pilot study using a cluster-randomized controlled design. J Telemed Telecare 1999;5(2):126-130. [doi: 10.1258/1357633991933440] [Medline: 10628024]

22. Patterson V, Craig J, Pang KA, Wootton R. Successful management of unexplained coma by telemedicine. J Telemed Telecare 1999;5(2):134-136. [doi: 10.1258/1357633991933341] [Medline: 10628026]

23. Wan AC, Gul Y, Darzi A. Realtime remote consultation in the outpatient clinic--experience at a teaching hospital. J Telemed Telecare 1999;5(Suppl 1):S70-S71. [doi: 10.1258/1357633991932621] [Medline: 10534849]

24. Wootton R. Telemedicine and isolated communities: a UK perspective. J Telemed Telecare 1999;5(Suppl 2):S27-S34. [doi: 10.1258/1357633991933495] [Medline: 10628015]

25. Loane MA, Bloomer SE, Corbett R, Eedy DJ, Gore HE, Hicks N, et al. Patient cost-benefit analysis of teledermatology measured in a randomized control trial. J Telemed Telecare 1999;5(Suppl 1):S1-S3. [doi: 10.1258/1357633991932414] [Medline: 10534821]

26. Loane M, Bloomer S, Corbett R, Eedy D, Hicks N, Lotery H, et al. A comparison of real-time and store-and-forward teledermatology: a cost-benefit study. Br J Dermatol 2000 Dec;143(6):1241-1247. [doi: 10.1046/j.1365-2133.2000.03895.x] [Medline: 11122028]

27. Craig J, Chua R, Russell C, Patterson V, Wootton R. The cost-effectiveness of teleneurology consultations for patients admitted to hospitals without neurologists on site. 1: A retrospective comparison of the case-mix and management at two rural hospitals. J Telemed Telecare 2000;6(Suppl 1):S46-S49. [doi: 10.1258/1357633001934122] [Medline: 10793970]

28. Craig J, Patterson V, Russell C, Wootton R. Interactive videoconsultation is a feasible method for neurological in-patient assessment. Eur J Neurol 2000 Nov;7(6):699-702. [doi: 10.1046/j.1468-1331.2000.00133.x] [Medline: 11136358]

29. Craig J, Chua R, Wootton R, Patterson V. A pilot study of telemedicine for new neurological outpatient referrals. J Telemed Telecare 2000;6(4):225-228. [doi: 10.1258/1357633001935400] [Medline: 11027124]

30. Graham LE, McGimpsey S, Wright S, McClean G, Carser J, Stevenson M, et al. Could a low-cost audio-visual link be useful in rheumatology? J Telemed Telecare 2000;6(Suppl 1):S35-S37. [doi: 10.1258/1357633001934078] [Medline: 10793966]

31. Mair F, Whitten P. Systematic review of studies of patient satisfaction with telemedicine. Br Med J 2000 Jun 3;320(7248):1517-1520 [FREE Full text] [doi: 10.1136/bmj.320.7248.1517] [Medline: 10834899]

XSL•FO
RenderX

32. Haslam R, McLaren P. Interactive television for an urban adult mental health service: the Guy's Psychiatric Intensive Care Unit Telepsychiatry Project. J Telemed Telecare 2000;6(Suppl 1):S50-S52. [doi: 10.1258/1357633001934131] [Medline: 10793971]

33. May C, Gask L, Ellis N, Atkinson T, Mair F, Smith C, et al. Telepsychiatry evaluation in the north-west of England: preliminary results of a qualitative study. J Telemed Telecare 2000;6(Suppl 1):S20-S22. [doi: 10.1258/1357633001934618] [Medline: 10793961]

34. Salmon S, Brint G, Marshall D, Bradley A. Telemedicine use in two nurse-led minor injuries units. J Telemed Telecare 2000;6(Suppl 1):S43-S45. [doi: 10.1258/1357633001934113] [Medline: 10793969]

35. Tachakra S, Dutton D, Newson R, Hayes J, Sivakumar A, Jaye P, et al. How do teleconsultations for remote trauma management change over a period of time? J Telemed Telecare 2000;6(Suppl 1):S12-S15. [doi: 10.1258/1357633001934393] [Medline: 10793959]

36. Tachakra S, Jaye P, Bak J, Hayes J, Sivakumar A. Supervising trauma life support by telemedicine. J Telemed Telecare 2000;6(Suppl 1):S7-11. [doi: 10.1258/1357633001934285] [Medline: 10793958]

37. Wootton R, McKelvey A, McNicholl B, Loane M, Hore D, Howarth P, et al. Transfer of telemedical support to Cornwall from a national telemedicine network during a solar eclipse. J Telemed Telecare 2000;6(Suppl 1):S182-S186. [doi: 10.1258/1357633001934609] [Medline: 10794014]

38. Beach M, Miller P, Goodall I. Evaluating telemedicine in an accident and emergency setting. Comput Methods Programs Biomed 2001;64(3):215-223 [FREE Full text] [doi: 10.1016/s0169-2607(00)00141-3]

39. Bose U, McLaren P, Riley A, Mohammedali A. The use of telepsychiatry in the brief counselling of non-psychotic patients from an inner-London general practice. J Telemed Telecare 2001;7(Suppl 1):8-10. [doi: 10.1177/1357633X010070S103] [Medline: 11576473]

40. Chua R, Craig J, Wootton R, Patterson V. Randomised controlled trial of telemedicine for new neurological outpatient referrals. J Neurol Neurosurg Psychiatry 2001 Jul;71(1):63-66 [FREE Full text] [doi: 10.1136/jnnp.71.1.63] [Medline: 11413265]

41. Chua R, Craig J, Wootton R, Patterson V. Cost implications of outpatient teleneurology. J Telemed Telecare 2001;7(Suppl 1):62-64. [doi: 10.1177/1357633X010070S126] [Medline: 11576496]

42. Leggett P, Graham L, Steele K, Gilliland A, Stevenson M, O'Reilly D, et al. Telerheumatology--diagnostic accuracy and acceptability to patient, specialist, and general practitioner. Br J Gen Pract 2001 Sep;51(470):746-748 [FREE Full text] [Medline: 11593837]

43. McCullough A. Viability and effectiveness of teletherapy for pre-school children with special needs. Int J Lang Commun Disord 2001;36 Suppl:321-326. [doi: 10.3109/13682820109177905] [Medline: 11340805]

44. Rayner S, Beaconsfield M, Kennedy C, Collin R, Taylor P, Murdoch I. Subspecialty adnexal ophthalmological examination using telemedicine. J Telemed Telecare 2001;7(Suppl 1):29-31. [doi: 10.1177/1357633X010070S112] [Medline: 11576482]

45. Simpson S. The provision of a telepsychology service to Shetland: client and therapist satisfaction and the ability to develop a therapeutic alliance. J Telemed Telecare 2001;7(Suppl 1):34-36. [doi: 10.1177/1357633X010070S114] [Medline: 11576484]

46. Tachakra S. Depth perception in telemedical consultations. Telemed J E Health 2001;7(2):77-85. [doi: 10.1089/153056201750279575] [Medline: 11421080]

47. Tachakra S, Newson R, Wootton R, Stinson A. Avoiding artificiality in teleconsultations. J Telemed Telecare 2001;7(Suppl 1):39-42. [doi: 10.1177/1357633X010070S116] [Medline: 11576486]

48. Brebner EM, Brebner JA, Ruddick-Bracken H, Wootton R, Ferguson J. Evaluation of a pilot telemedicine network for accident and emergency work. J Telemed Telecare 2002;8(Suppl 2):5-6. [doi: 10.1177/1357633X020080S203] [Medline: 12217114]

49. Dawson E, Kennedy C, Bentley C, Lee J, Murdoch I. The role of telemedicine in the assessment of strabismus. J Telemed Telecare 2002;8(1):52-55. [doi: 10.1258/1357633021937361] [Medline: 11809086]

50. Harley J, McLaren P, Blackwood G, Tierney K, Everett M. The use of videoconferencing to enhance tertiary mental health service provision to the island of Jersey. J Telemed Telecare 2002;8(Suppl 2):36-38. [doi: 10.1177/1357633X020080S216] [Medline: 12217127]

51. MacFarlane A, Harrison R, Wallace P. The benefits of a qualitative approach to telemedicine research. J Telemed Telecare 2002;8(Suppl 2):56-57. [doi: 10.1177/1357633X020080S226] [Medline: 12217137]

52. Martin S, Rankin G. Using commercially available technology to assist in the delivery of person-centred health and social care. J Telemed Telecare 2002;8(Suppl 2):60-62. [doi: 10.1177/1357633X020080S228] [Medline: 12217139]

53. McLaren P, Ahlbom J, Riley A, Mohammedali A, Denis M. The North Lewisham telepsychiatry project: beyond the pilot phase. J Telemed Telecare 2002;8(Suppl 2):98-100. [doi: 10.1177/1357633X020080S245] [Medline: 12217156]

54. Patterson V. Teleneurology in Northern Ireland: a success. J Telemed Telecare 2002 Dec;8 Suppl 3(6):46-47. [doi: 10.1258/13576330260440835] [Medline: 12537903]

55. Scuffham PA, Steed M. An economic evaluation of the Highlands and Islands teledentistry project. J Telemed Telecare 2002;8(3):165-177. [doi: 10.1177/1357633X0200800307] [Medline: 12097178]

56. Simpson S, Morrow E, Jones M, Ferguson J, Brebner E. Video-hypnosis--the provision of specialized therapy via videoconferencing. J Telemed Telecare 2002;8(Suppl 2):78-79. [doi: 10.1177/1357633X020080S236] [Medline: 12217147]

57.   Tachakra S, Rajani R. Social presence in telemedicine. J Telemed Telecare 2002;8(4):226-230. [doi: 10.1258/135763302320272202] [Medline: 12217106]

58.   Tachakra S, Uche CU, Stinson A. Four years' experience of telemedicine support of a minor accident and treatment service. J Telemed Telecare 2002;8(Suppl 2):87-89. [doi: 10.1177/1357633X020080S240] [Medline: 12217151]

59.   Wallace P, Haines A, Harrison R, Barber J, Thompson S, Jacklin P, Virtual Outreach Project Group. Joint teleconsultations (virtual outreach) versus standard outpatient appointments for patients referred by their general practitioner for a specialist opinion: a randomised trial. Lancet 2002 Jun 8;359(9322):1961-1968. [doi: 10.1016/s0140-6736(02)08828-1] [Medline: 12076550]

60.   Jacklin PB, Roberts JA, Wallace P, Haines A, Harrison R, Barber JA, Virtual Outreach Project Group. Virtual outreach: economic evaluation of joint teleconsultations for patients referred by their general practitioner for a specialist opinion. Br Med J 2003 Jul 12;327(7406):84 [FREE Full text] [doi: 10.1136/bmj.327.7406.84] [Medline: 12855528]

61.   Cargill A, Watson AR. Telecare support for patients undergoing chronic peritoneal dialysis. Perit Dial Int 2003;23(1):91-94. [Medline: 12691517]

62.   Ferguson J, Rowlands A, Palombo A, Pedley D, Fraser S, Simpson S. Minor injuries telemedicine. J Telemed Telecare 2003;9(Suppl 1):S14-S16. [doi: 10.1258/135763303322196187] [Medline: 12952707]

63.   Palombo A, Ferguson J, Rowlands A, Pedley D, Fraser S. An evaluation of a telemedicine fracture review clinic. J Telemed Telecare 2003;9(Suppl 1):S31-S33. [doi: 10.1258/135763303322196259] [Medline: 12952714]

64.   Benger JR, Noble SM, Coast J, Kendall JM. The safety and effectiveness of minor injuries telemedicine. Emerg Med J 2004 Jul;21(4):438-445 [FREE Full text] [Medline: 15208226]

65.   Brebner EM, Brebner JA, Ruddick-Bracken H, Wootton R, Ferguson J, Palombo A, et al. Evaluation of an accident and emergency teleconsultation service for north-east Scotland. J Telemed Telecare 2004;10(1):16-20. [doi: 10.1258/135763304322764130] [Medline: 15006210]

66.   Craig J, Chua R, Russell C, Wootton R, Chant D, Patterson V. A cohort study of early neurological consultation by telemedicine on the care of neurological inpatients. J Neurol Neurosurg Psychiatry 2004 Jul;75(7):1031-1035 [FREE Full text] [doi: 10.1136/jnnp.2002.001651] [Medline: 15201365]

67.   Grant IC. Telemedicine in the British Antarctic survey. Int J Circumpolar Health 2004 Dec;63(4):356-364. [doi: 10.3402/ijch.v63i4.17753] [Medline: 15709312]

68.   Leonard S. The development and evaluation of a telepsychiatry service for prisoners. J Psychiatr Ment Health Nurs 2004 Aug;11(4):461-468. [doi: 10.1111/j.1365-2850.2004.00747.x] [Medline: 15255921]

69.   Coelho JJ, Arnold A, Nayler J, Tischkowitz M, MacKay J. An assessment of the efficacy of cancer genetic counselling using real-time videoconferencing technology (telemedicine) compared to face-to-face consultations. Eur J Cancer 2005 Oct;41(15):2257-2261. [doi: 10.1016/j.ejca.2005.06.020] [Medline: 16176873]

70.   Grealish A, Hunter A, Glaze R, Potter L. Telemedicine in a child and adolescent mental health service: participants' acceptance and utilization. J Telemed Telecare 2005;11(Suppl 1):53-55. [doi: 10.1258/1357633054461921] [Medline: 16035994]

71.   Noble SM, Coast J, Benger JR. A cost-consequences analysis of minor injuries telemedicine. J Telemed Telecare 2005;11(1):15-19. [doi: 10.1177/1357633X0501100104] [Medline: 15829038]

72.   Soopramanien A, Pain H, Stainthorpe A, Menarini M, Ventura M. Using telemedicine to provide post-discharge support for patients with spinal cord injuries. J Telemed Telecare 2005;11(Suppl 1):68-70. [doi: 10.1258/1357633054461633] [Medline: 16036001]

73.   Brebner JA, Brebner EM, Ruddick-Bracken H. Accident and emergency teleconsultation for primary care--a systematic review of technical feasibility, clinical effectiveness, cost effectiveness and level of local management. J Telemed Telecare 2006;12(Suppl 1):5-8. [doi: 10.1258/135763306777978542] [Medline: 16884562]

74.   Chambers A, Hennessy E, Powell-Tuck J. Longitudinal trends in quality of life after starting home parenteral nutrition: a randomised controlled study of telemedicine. Clin Nutr 2006 Jun;25(3):505-514. [doi: 10.1016/j.clnu.2006.01.001] [Medline: 16698138]

75.   Norman S. The use of telemedicine in psychiatry. J Psychiatr Ment Health Nurs 2006 Dec;13(6):771-777. [doi: 10.1111/j.1365-2850.2006.01033.x] [Medline: 17087682]

76.   Harrison R, Macfarlane A, Murray E, Wallace P. Patients' perceptions of joint teleconsultations: a qualitative evaluation. Health Expect 2006 Mar;9(1):81-90 [FREE Full text] [doi: 10.1111/j.1369-7625.2006.00368.x] [Medline: 16436164]

77.   Weatherburn G, Dowie R, Mistry H, Young T. An assessment of parental satisfaction with mode of delivery of specialist advice for paediatric cardiology: face-to-face versus videoconference. J Telemed Telecare 2006;12(Suppl 1):57-59. [doi: 10.1258/135763306777978560] [Medline: 16884583]

78.   Dowie R, Mistry H, Young T, Weatherburn G, Gardiner H, Rigby M, et al. Telemedicine in pediatric and perinatal cardiology: economic evaluation of a service in English hospitals. Int J Technol Assess Health Care 2007;23(1):116-125. [doi: 10.1017/S0266462307051653] [Medline: 17234025]

79.   Singh S, Arya D, Peters T. Accuracy of telepsychiatric assessment of new routine outpatient referrals. BMC Psychiatry 2007;7(1):55 [FREE Full text] [doi: 10.1186/1471-244x-7-55]

XSL•FO
RenderX

80. Mair F, Ferguson J. More patients with minor injuries could be seen by telemedicine. J Telemed Telecare 2008;14(3):132-134. [doi: 10.1258/jtt.2008.003009] [Medline: 18430279]

81. Mair F, Fraser S, Ferguson J, Webster K. Telemedicine via satellite to support offshore oil platforms. J Telemed Telecare 2008;14(3):129-131. [doi: 10.1258/jtt.2008.003008] [Medline: 18430278]

82. Mair F, Peterkin G, Laing A, Ferguson J, Fraser S. Feasibility of a telemedicine 'booth' for supporting remote care in Scotland. J Telemed Telecare 2008;14(3):127-129. [doi: 10.1258/jtt.2008.003007] [Medline: 18430277]

83. Miller DR, Alam K, Fraser S, Ferguson J. The delivery of a minor injuries telemedicine service by Emergency Nurse Practitioners. J Telemed Telecare 2008;14(3):143-144. [doi: 10.1258/jtt.2008.003013] [Medline: 18430283]

84. Morgan G, Craig B, Grant B, Sands A, Doherty N, Casey F. Home videoconferencing for patients with severe congential heart disease following discharge. Congenit Heart Dis 2008;3(5):317-324. [doi: 10.1111/j.1747-0803.2008.00205.x] [Medline: 18837809]

85. Styles V. Service users' acceptability of videoconferencing as a form of service delivery. J Telemed Telecare 2008;14(8):415-420. [doi: 10.1258/jtt.2008.071202] [Medline: 19047451]

86. Webster K, Fraser S, Mair F, Ferguson J. Provision of telehealth to the Scottish Police College. J Telemed Telecare 2008;14(3):160-162. [doi: 10.1258/jtt.2008.003020] [Medline: 18430290]

87. Wilkinson OM, Duncan-Skingle F, Pryor JA, Hodson ME. A feasibility study of home telemedicine for patients with cystic fibrosis awaiting transplantation. J Telemed Telecare 2008;14(4):182-185. [doi: 10.1258/jtt.2008.070107] [Medline: 18534951]

88. Howell S, Tripoliti E, Pring T. Delivering the Lee Silverman Voice Treatment (LSVT) by web camera: a feasibility study. Int J Lang Commun Disord 2009;44(3):287-300. [doi: 10.1080/13682820802033968] [Medline: 18821113]

89. Pappas Y, Seale C. The opening phase of telemedicine consultations: an analysis of interaction. Soc Sci Med 2009 Apr;68(7):1229-1237. [doi: 10.1016/j.socscimed.2009.01.011] [Medline: 19201514]

90. Harvey S, Peterkin G, Wootton R. Eleven years of experience with low-bandwidth telemedicine in a nurse-led rural clinic in Scotland. J Telemed Telecare 2010;16(8):417-421 [FREE Full text] [doi: 10.1258/jtt.2010.100310] [Medline: 20876630]

91. Hill A. Telerehabilitation in Scotland: current initiatives and recommendations for future development. Int J Telerehabil 2010;2(1):7-14 [FREE Full text] [doi: 10.5195/ijt.2010.6055] [Medline: 25945168]

92. Chowdhury M, Birns J, Rudd A, Bhalla A. Telemedicine versus face-to-face evaluation in the delivery of thrombolysis for acute ischaemic stroke: a single centre experience. Postgrad Med J 2012 Mar;88(1037):134-137. [doi: 10.1136/postgradmedj-2011-130060] [Medline: 22267526]

93. Gibson. Abstracts of the 22nd European Stroke Conference. London, United Kingdom. May 28-31, 2013. Cerebrovasc Dis 2012;35(Suppl 3):9-854 [FREE Full text] [doi: 10.1159/000353129] [Medline: 23736169]

94. Hill E, Whitehead M, MacInnes B, Ellis G, Talbot A, Brodie F, et al. The first 100 thrombolysis cases in a novel Scottish mesh telestroke system. Scott Med J 2013 Nov;58(4):213-216. [doi: 10.1177/0036933013507868] [Medline: 24215039]

95. Agarwal S, Day DJ, Sibson L, Barry PJ, Collas D, Metcalf K, et al. Thrombolysis delivery by a regional telestroke network--experience from the UK National Health Service. J Am Heart Assoc 2014 Feb 26;3(1):e000408 [FREE Full text] [doi: 10.1161/JAHA.113.000408] [Medline: 24572251]

96. Hex N, Tuggey J, Wright D, Malin R. Telemedicine in care homes in Airedale, Wharfedale and Craven. Clin Gov 2015;20(3):146-154 [FREE Full text] [doi: 10.1108/cgij-07-2015-0022]

97. Murphie P, Mills B, Little S. Telemedicine sleep clinic review - patient acceptability survey. Eur Respir J 2015;46(59):PA3377 [FREE Full text] [doi: 10.1183/13993003.congress-2015.pa3377]

98. Gibson J, Lightbody E, McLoughlin A, McAdam J, Gibson A, Day E, et al. 'It was like he was in the room with us': patients' and carers' perspectives of telemedicine in acute stroke. Health Expect 2016 Feb;19(1):98-111 [FREE Full text] [doi: 10.1111/hex.12333] [Medline: 25581591]

99. Ranta A, Whitehead M, Gunawardana C, Cariga P, Iniesta I, Watson I, et al. International telestroke: the first five cases. J Stroke Cerebrovasc Dis 2016 Apr;25(4):e44-e45. [doi: 10.1016/j.jstrokecerebrovasdis.2015.12.028] [Medline: 26825351]

100. Bagot KL, Cadilhac DA, Bladin CF, Watkins CL, Vu M, Donnan GA, VSTASTUTE investigators. Integrating acute stroke telemedicine consultations into specialists' usual practice: a qualitative analysis comparing the experience of Australia and the United Kingdom. BMC Health Serv Res 2017 Nov 21;17(1):751 [FREE Full text] [doi: 10.1186/s12913-017-2694-1] [Medline: 29157233]

101. Ditchburn J, Marshall A. Renal telemedicine through video-as-a-service delivered to patients on home dialysis: A qualitative study on the renal care team members' experience. J Ren Care 2017 Sep;43(3):175-182. [doi: 10.1111/jorc.12207] [Medline: 28614637]

102. Morris J, Campbell-Richards D, Wherton J, Sudra R, Vijayaraghavan S, Greenhalgh T, et al. Webcam consultations for diabetes: findings from four years of experience in Newham. Pract Diab 2017;34(2):45-50. [doi: 10.1002/pdi.2078]

103. Williams OE, Elghenzai S, Subbe C, Wyatt J, Williams J. The use of telemedicine to enhance secondary care: some lessons from the front line. Future Healthc J 2017 Jun;4(2):109-114 [FREE Full text] [doi: 10.7861/futurehosp.4-2-109] [Medline: 31098445]

104. Freed J, Lowe C, Flodgren G, Binks R, Doughty K, Kolsi J. Telemedicine: Is it really worth it? A perspective from evidence and experience. J Innov Health Inform 2018 Mar 15;25(1):14-18 [FREE Full text] [doi: 10.14236/jhi.v25i1.957] [Medline: 29717950]

105. Greenhalgh T, Shaw S, Wherton J, Vijayaraghavan S, Morris J, Bhattacharya S, et al. Real-world implementation of video outpatient consultations at macro, meso, and micro levels: mixed-method study. J Med Internet Res 2018 Apr 17;20(4):e150 [FREE Full text] [doi: 10.2196/jmir.9897] [Medline: 29625956]

106. Francischetto EO, Damery S, Ferguson J, Combes G, myVideoClinic randomised evaluation steering group. Video clinics versus standard face-to-face appointments for liver transplant patients in routine hospital outpatient care: study protocol for a pragmatic randomised evaluation of myVideoClinic. Trials 2018 Oct 19;19(1):574 [FREE Full text] [doi: 10.1186/s13063-018-2953-4] [Medline: 30340637]

107. Haig-Ferguson A, Loades M, Whittle C, Read R, Higson-Sweeney N, Beasant L, et al. 'It's not one size fits all'; the use of videoconferencing for delivering therapy in a Specialist Paediatric Chronic Fatigue Service. Internet Interv 2019 Mar;15:43-51 [FREE Full text] [doi: 10.1016/j.invent.2018.12.003] [Medline: 30619719]

108. Morey S, Hamoodi A, Jones D, Young T, Thompson C, Dhuny J, et al. Increased diagnosis and treatment of hepatitis C in prison by universal offer of testing and use of telemedicine. J Viral Hepat 2019 Jan;26(1):101-108. [doi: 10.1111/jvh.13017] [Medline: 30315691]

109. Gilbert A, Jaggi A, May C. What is the patient acceptability of real time 1:1 videoconferencing in an orthopaedics setting? A systematic review. Physiotherapy 2018 Jun;104(2):178-186. [doi: 10.1016/j.physio.2017.11.217] [Medline: 29361298]

110. Hex N, Wright D. Airedale NHS Foundation Trust. 2015 May. Airedale NHS Foundation Trust: Evaluation of Telehealth Interventions for Care Homes in Airedale, Wharfedale and Craven. Final Report URL: http://www.airedale-trust.nhs.uk/wp/wp-content/uploads/2014/02/YHEC-Report-Airedale-TM-FINAL-May15-2.pdf [accessed 2018-07-02]

111. Charlesworth A, Firth Z, Gershlick B, Watt T, Johnson P, Kelly E, et al. Securing the Future: Funding Health and Social Care To the 2030s. London: NHS Confederation; May 2018.

112. Office for National Statistics. 2018. Internet Access – Households and Individuals, Great Britain: 2018 URL: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2018 [accessed 2019-02-18]

113. The Office for National Statistics. 2017. Internet Users in the UK: 2017 URL: https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/bulletins/internetusers/2017 [accessed 2019-03-08]

114. Age UK. 2017. Painful Journeys URL: https://www.ageuk.org.uk/globalassets/age-uk/documents/reports-and-publications/reports-and-briefings/active-communities/rb_dec17_painful_journeys_indepth_report.pdf [accessed 2019-02-18]

115. Powell RE, Henstenburg JM, Cooper G, Hollander JE, Rising KL. Patient perceptions of telehealth primary care video visits. Ann Fam Med 2017 May;15(3):225-229 [FREE Full text] [doi: 10.1370/afm.2095] [Medline: 28483887]

116. Buchan J, Charlesworth A, Gershlick B, Seccombe I. A Critical Moment: NHS Staffing Trends, Retention and Attrition. UK: The Health Foundation; 2019.

117. Costanzo C, Verghese A. The physical examination as ritual: social sciences and embodiment in the context of the physical examination. Med Clin North Am 2018 May;102(3):425-431. [doi: 10.1016/j.mcna.2017.12.004] [Medline: 29650064]

118. Hanlon P, Daines L, Campbell C, McKinstry B, Weller D, Pinnock H. Telehealth interventions to support self-management of long-term conditions: a systematic metareview of diabetes, heart failure, asthma, chronic obstructive pulmonary disease, and cancer. J Med Internet Res 2017 May 17;19(5):e172 [FREE Full text] [doi: 10.2196/jmir.6688] [Medline: 28526671]

119. Swarup S, Makaryus AN. Digital stethoscope: technology update. Med Devices (Auckl) 2018;11:29-36 [FREE Full text] [doi: 10.2147/MDER.S135882] [Medline: 29379321]

120. Kim Y, Chao DL. Comparison of smartphone ophthalmoscopy vs conventional direct ophthalmoscopy as a teaching tool for medical students: the COSMOS study. Clin Ophthalmol 2019;13:391-401 [FREE Full text] [doi: 10.2147/OPTH.S190922] [Medline: 30858689]

121. SlideShare. Bernard Quinn - Digital Technology Transformation of Outpatient Services URL: https://www.slideshare.net/InnovationNWC/bernard-quinn-digital-technology-transformation-of-outpatient-services-104962014 [accessed 2019-05-03]

122. NHS Improvement. 2019. National Tariff Payment System 2019/20: A Consultation Notice URL: https://improvement.nhs.uk/resources/national-tariff-1920-consultation/ [accessed 2019-09-04]

123. Appleby J, Harrison T, Hawkins L, Dixon A. Payment by Results: How Can Payment Systems Help to Deliver Better Care?. London: The King's Fund; 2012.

124. Finch T, May C, Mair F, Mort M, Gask L. Integrating service development with evaluation in telehealthcare: an ethnographic study. Br Med J 2003 Nov 22;327(7425):1205-1209 [FREE Full text] [doi: 10.1136/bmj.327.7425.1205] [Medline: 14630758]

Abbreviations

A&E: accident and emergency
DGH: district general hospital
DNA: did not attend
GP: general practitioner

IT: information technologies
NHS: National Health Service
PbR: payment by results
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial

XSL•FO
RenderX

Original Paper

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:

This is a corrected version. See correction statement: http://medinform.jmir.org/2020/8/e23253/

Abstract

Background: Previous research shows that being a "digital native" or growing up in a digital environment does not necessarily lead to increased digital competencies, such as digital health literacy or evaluation of webpage quality.

Objective: This study showed how medical students searched for health information online, specifically the recommended testing for histamine intolerance, by comparing the use of various search engines (Google, Medisuch, and a website of the student's choice) to find out more about search strategies in future health professionals. As Medisuch presents a qualitatively better search engine, we assumed that medical students using this search engine might find valid information faster on more reliable webpages, and might recommend the correct diagnostic steps for histamine intolerance to their patients more often than students using a generic search engine like Google.

Methods: Medical students in their third year of study were asked to find the relevant diagnostic steps of histamine intolerance online. They were randomly assigned to use one search engine: Google, their personal choice, or Medisuch. Their process of seeking information online was video recorded.

Results: In total, 140 medical students participated in this study. The total number of webpages found did not differ among the groups (P=.52). Students using Medisuch (P=.02) correctly identified the elimination diet as a relevant diagnostic step more frequently. The provocation test was reported by almost half of the students independent of the search engine used. In general, medical students commonly identified trustworthy webpages in all three groups (Google: 36/44, 82%; free choice: 31/36; 86%; and Medisuch: 35/45, 78%).

Conclusions: The results indicate that medical students were able to find trustworthy health-related information online independent of the search engine used. Medical students that are digital natives seem to have proper internet skills and a knowledge of how to use them. They entered specific medical terms (evidence-based diagnostic steps) or names of reliable webpages (DocCheck) in the search engines to gain correct information. However, it remains to be seen if this behavior can be called true "digital literacy".

XSL•FO
RenderX

Introduction

The internet is omnipresent and has become the primary source of information for many [1]. It has been reported that 52.2% of Europe's population has used internet searches for health and health-related issues [2]. Çoklar et al [3] reported that online information search strategies represented one of the most important variables in effective and efficient internet use. Previous research has shown that online search strategies might be influenced and explained by the time spent online, Web experience, and individual difference (like domain knowledge or epistemological belief) [4-7]. An investigation of university students' online search strategies in different contexts indicated that the students were able to search online for daily life information. However, they had difficulties in the online search strategies used for learning activities, and the authors suggested that teachers should help students develop online search strategies for academic uses [8]. The need for guidance—particularly for people with little expertise—has also been highlighted by Armstrong and Large [9] in their manual on online search strategies.

Furthermore, it has been shown that falling in the category of "digital natives" [10] or showing "digital nativity" [3] (ie, being born in a digital world and, therefore, frequently using internet or media devices since early childhood) does not automatically correlate with online literacy [1,3]. Therefore, attempts have been made to clarify quality indicators for online information.

One such approach involves introducing specific search engines that prefilter for the user and only show reliable results, such as the German webpage Medisuch [11]. Another approach is to certify webpages with trustworthy, evidence-based medical content and introduce certificates that make it easier for average people and professionals to decide quickly if a webpage can be trusted. At first glance, this approach seems easy; however, people seeking information online still have to be properly trained [12]. Examples of certificates available in the topic of health-related information are the afgis certificate and the Health On the Net Foundation Code of Conduct (HONcode) certificate. For a more comprehensive list, see Pauer et al [13].

To properly advise patients, medical students need to be generally aware of diagnostic steps; if they lack the knowledge, they should know how to find reliable information. As medical students are still learning, specific training can impact their attitude and behavior. This study compared various search engines to investigate how medical students search for health information online. The aim of this study was to track medical students' searches for health information online and identify potentially weak strategies, which could be addressed specifically within the medical curriculum. We wanted to compare three different approaches and focus on the medical students' findings on histamine intolerance. As Medisuch presents a qualitatively better search engine [11], medical students using it may find valid information faster on more reliable webpages, and may recommend the correct diagnostic steps for histamine intolerance to their patients more often than students using a generic search engine like Google.

Methods

Study Design and Participants

This study was performed in an explanatory cross-sectional manner. Third-year students from the Faculty of Medicine of the University Hospital Tuebingen, Germany, were recruited from their curricular courses in the Department of Internal Medicine VI. They had all received theoretical input on functional disorders and differential diagnoses with a focus on intolerances before participating in the study. Group sizes consisted of 8 to 16 students per teaching session. Students were taught by experienced physicians. Participation in the course was mandatory, but participation in the study was on a voluntary basis.

Ethics

The study received ethical approval from the Ethics Committee of the Tuebingen Medical Faculty (443/2018BO2). All participants gave written informed consent. They did not receive a reimbursement for their participation.

Process of Study

Students of the university must attend a 2-week course on psychosomatic medicine in their third year. Every 2 weeks, a new group of approximately 12 to 15 students start the course. The topic is "somatoform disorders" on the third day of the course. Students received a case report involving a patient who consulted her physician due to abdominal pain. The patient in the fictitious case had already researched her health problem beforehand and asked specifically about potential causes such as problems with digestion of histamines. The patient wanted to know which steps had to be taken to rule out or confirm histamine intolerance. Each student had access to one stationary computer with connection to the internet via Internet Explorer or Firefox (both were installed and students were free to choose). Students were instructed to complete a worksheet about histamine intolerance by searching for related information online. They were randomly assigned to one of three groups: Google [14], free choice, or Medisuch [15]. In the free choice group students could choose any search engine they wanted to use. There were no prerequisites for participation, and they were not taught further information about the various search engines. Medisuch is a specific search engine that prefilters the information found and only shows reliable results (ie, medical webpages that are certified or evidence-based) [11]. The students had 10 minutes to research information online and mark their findings on the worksheet, which was collected by a study assistant after the time period. Teaching then continued with an interactive discussion where students were informed about the correct diagnostic steps as well as quality indicators of good webpages and apps. Figure 1 shows an overview of the study. The quality indicators were based on the checklists used by

afgis and HONcode, such as objectivity or accuracy. For a more comprehensive list, see Pauer and colleagues [13].

Figure 1. Process of study including teaching and assessment.
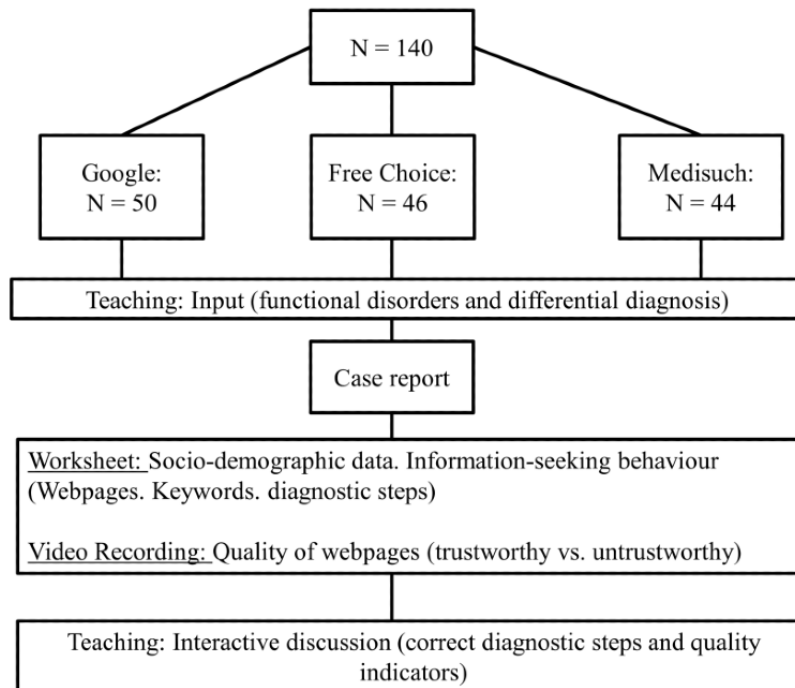
Students were asked about demographic variables such as age and gender, as well as experience with electronics such as possession of electronic devices and time spent online per day. They also rated their competence on the topic of histamine intolerance on a 6-point Likert scale (0 = not competent at all and 5 = highly competent).

Video Recording

Before starting the teaching session, students were asked to start Morae (TechSmith, Michigan) [16] on their computer, which was used to track students' online search for health information via a screen recording of their desktop. Morae has been previously applied within the research area of online search behavior by university students [17-20]. The software generates quantitative data on different variables such as time spent on each webpage and clicks made. To the best of our knowledge, no other standardized procedure exists that would successfully analyze the video-recorded data in a qualitative way. In alignment with the previously described procedure [20], we aimed to develop a categorical system (eg, search engine, number of webpages, search strategies) to guarantee standardized analysis and evaluation of the Morae videos.

All Morae videos were coded according to this categorical system by experienced members of the research team. All webpages students had visited were listed alphabetically, and their quality was classified according to a rating sheet by two independent raters. In cases of a disagreement, a third independent rater was consulted.

Information-Seeking Behavior

Students answered specific questions about the search engine(s) used, keywords entered, and the number of webpages they consulted before making their diagnostic recommendation. Additionally, they rated the number of pages considered as trustworthy as well as the number of pages they would recommend to a patient. Trustworthy was defined as an evidence-based and certified webpage; thus, untrustworthy was defined as a nonevidence-based one such as an advertisement or self-made webpage by an individual. To get an idea on the effectiveness of medical students' searches, the number of pages opened and the keywords needed to get to the desired answer were considered as relevant markers.

Diagnostic Recommendation

Students had to give specific recommendations on the next diagnostic steps. They were asked to choose one or more of the following options: antibody testing, assessment of diaminoxydase, nutritional diary, elimination diet, H_2 breathing test, histamine testing, provocation test, and test of urine and feces. To pass, students needed to mark one or more of the following options: provocation test, nutritional diary, or elimination diet, as these were the diagnostic factors recommended by a clinical expert.

Data Analysis

Statistical analysis was performed using SPSS 25 (IBM Corp, Armonk, New York). Mean values, associated SDs, frequencies, and percentages of relevant factors such as age and gender were calculated. To test possible relationships among the variables, Chi-squared tests were used. Analysis of variance (ANOVAs) were conducted to compare differences between mean values. In addition, the Pearson correlations were calculated. Beforehand, distribution of data was tested by the

Kolmogorov-Smirnov test. *P* values<.05 were reported as significant.

Results

Sample

There were 140 students who took part in the survey, and the response rate was 83.3% (140 /168). The average age was 24.36 (SD 2.84) years. Females constituted 61.7% (85) of the participating students. There were 50 students randomly assigned to Google, 46 to Medisuch, and 44 to the free choice group ($\chi^2_{278}=280.0$, *P*=.46). All participants completed the accompanying worksheet.

There were no significant differences between the groups with regards to age ($F_{2,135}=5.04$, *P*=.008), gender ($\chi^2_4=4.5$, *P*=.34), and previous formal medical or information technology (IT) training ($\chi^2_2=1.5$, *P*=.23). Smartphones were owned by 78.6% (110) of the students; only 2.1% (3) had a smartwatch; 46.4% (65) had a tablet; and 72.1% (111) had a laptop. The vast majority (131, 93.6%) reported spending from 1 to 6 hours online daily.

Video Recording

Videos were successfully recorded from 125 students (44 from the Google group, 45 from the Medisuch group, and 36 from the free choice group). The remaining 15 recordings were lost due to technical reasons such as the recording or storing not working. The following data are based on these video recordings except those on diagnostic recommendation, which were based on the worksheet.

Information-Seeking Behavior

There was no significant difference between groups regarding the total number of webpages students considered before making their diagnostic choice. The same result was shown for the number of pages regarded as useful. 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).

Information seeking-behavior of students regarding the total number of webpages, the number of pages considered useful, and the number of pages considered recommendable for each group, as well as the means and SDs of the webpages are shown in Figure 2. Students in the free choice group opened significantly fewer recommendable pages ($F_{2,133}=5.04$, *P*=.008).

There were no significant differences between groups in regards to the number of keywords entered in the search field or the number of webpages accessed (Figure 3).

There was no significant difference regarding the amount of keywords and webpage names found by students for each group. Means and the corresponding SDs are shown in Figure 3.

There was a highly significant difference between groups in whether or not the students entered specific medical terminology in the search engine ($\chi^2_4=16.6$, *P*=.005). The majority of students in the Google and free choice groups entered either specific medical webpages such as DocCheck or specific medical terminology such as "evidence-based" into the search engine to specify their search; however, in the Medisuch group, more than one-third did not. Of the 44 students in the Google group, 38 (86%) used specific medical terminology for their search. In the free choice group 35 out of 36 (97%) students used specific medical terminology. In the Medisuch group 29 students out of 45 (64%) used specific medical terminology.

Figure 2. Information-seeking behavior of students regarding the total number of webpages.
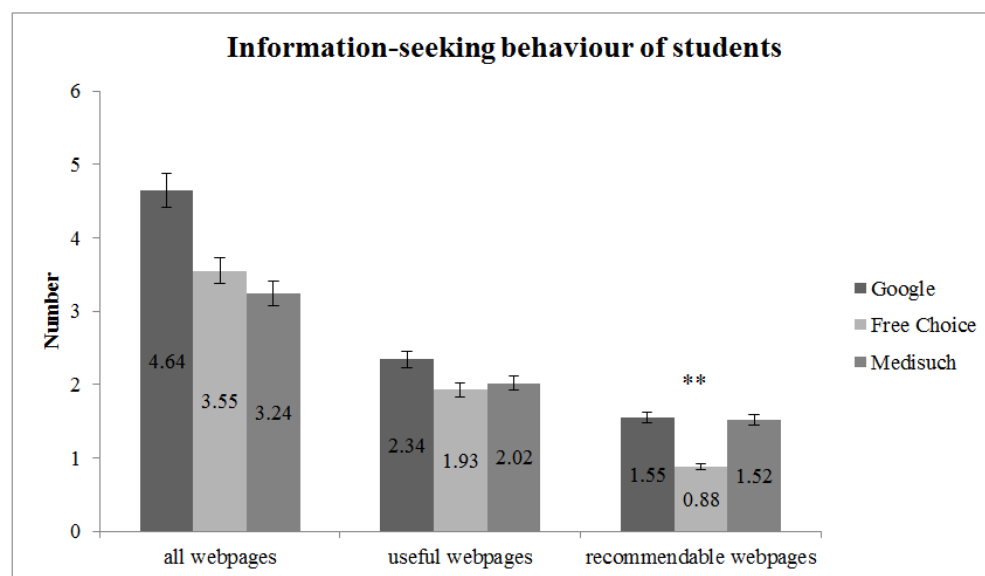
Figure 3. Number of keywords and webpage names found by students for each group.
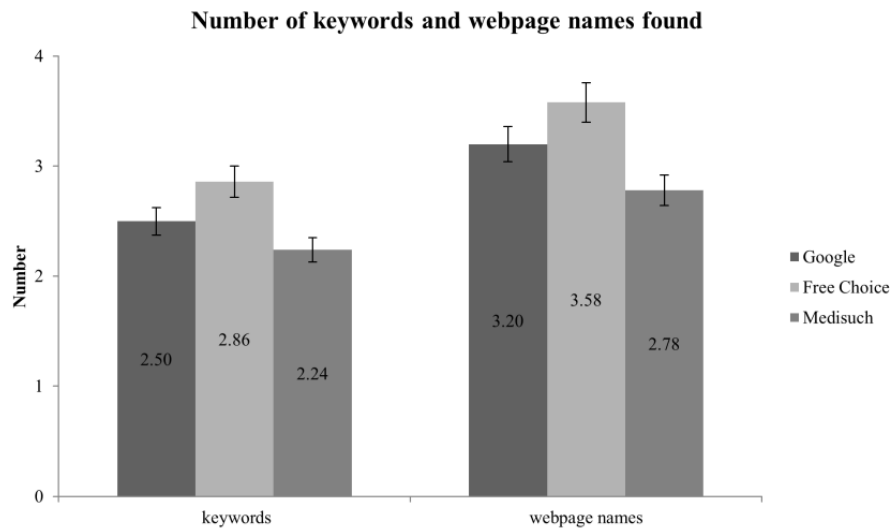
Quality of Webpages

In total, 53 different webpages were accessed by the whole sample. Quality ratings showed that 34 pages fulfilled the criteria of a qualified webpage. Interrater-correlation after Cohen *d* was *r*=.73.

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).

There were no significant differences in the frequencies of trustworthy webpages found among the three groups with χ^2_{14}=16.45, *P*=.29 The webpage of German national treatment guidelines was used by students of all three groups (Google: 15.9%; free choice: 33.3%; and Medisuch: 51.1%). 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. The mean quotients of the single groups were high with Google: 0.82 (SD 0.20); free group: 0.86 (SD 0.22); and Medisuch: 0.77 (SD 0.23) (Figure 4).

Figure 4 presents the means (and corresponding SDs) of the quotients of trustworthy or untrustworthy pages by the total number of pages found, which are separated by each group. There were no significant differences for trustworthy and untrustworthy sources.

Figure 4. Quotient of trustworthy or untrustworthy webpages found by students separated by group.
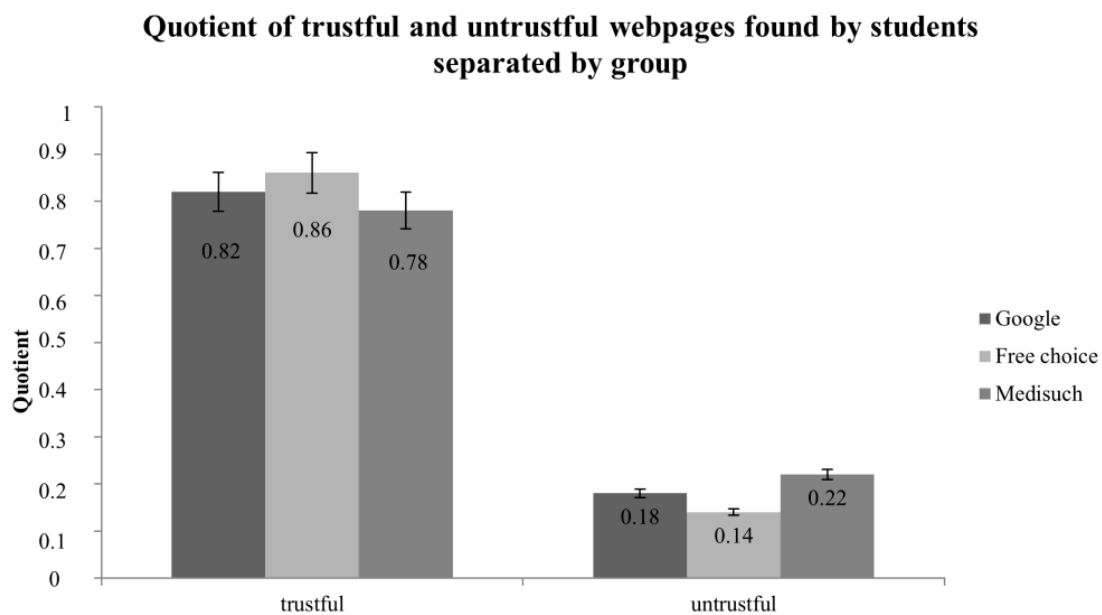
Diagnostic Recommendation

Independent of the search group, approximately 60% of students identified nutrition diary as a correct diagnostic step (Table 1). With regard to the provocation test, almost half the students found this procedure to be correct. The Medisuch student group reported the elimination diet as a correct diagnostic step for histamine intolerance more often than the Google or free choice groups. However, they suggested the wrong answer "antibody detection" significantly more often than the other groups.

Table 1. Diagnostic decision after online search.

Diagnostic step (correct or wrong)	Google (N=50)	Free choice (N=44)	Medisuch (N=46)	Chi-square (*df*)
Antibody detection (wrong)				
yes	2 (4%)	1 (2%)	7 (15%)	6.84 (2), *P*=.03
no	48 (96%)	43 (98%)	39 (85%)	
Histamine testing (wrong)				
yes	19 (38%)	15 (34%)	13 (28%)	1.03 (2),*P*=.60
no	31 (62%)	29 (66%)	33 (72%)	
Assessment of diaminoxydase (wrong)				
yes	29 (58%)	23 (52%)	18 (39%)	3.55 (2),*P*=.17
no	21 (42%)	21 (48%)	28 (61%)	
H_2 breathing test (wrong)				
yes	0 (0%)	0 (0%)	0 (0%)	N/A[a]
no	50 (100%)	44 (100%)	46 (100%)	
Test of urine and feces (wrong)				
yes	7 (14%)	4 (9%)	7 (15 %)	0.84 (2),*P*=.66
no	43 (86%)	40 (91%)	39 (85%)	
Nutrition diary (correct)				
yes	30 (60%)	27 (61%)	28 (61%)	0.02 (2),*P*=.99
no	20 (40%)	17 (39%)	18 (39%)	
Elimination diet (correct)				
yes	26 (52%)	19 (43%)	33 (72%)	7.87 (2),*P*=.02
no	24 (48%)	25 (57%)	13 (28%)	
Provocation test (correct)				
yes	25 (50%)	21 (48%)	22 (48%)	0.06 (2),*P*=.97
no	25 (50%)	23 (52%)	24 (52%)	

[a]Not applicable.

Discussion

In this explanatory study, we examined medical students' information-seeking behavior when assigned to use a generic search engine or a search engine of high quality. The hypothesis that students of the qualitatively better search engine would find the correct diagnostic steps for histamine intolerance more often was not fully supported. The Medisuch group significantly identified the elimination diet as a correct diagnostic step more frequently; however, this was not the case for the nutrition diary or the provocation test. The students of the Google and free choice groups reported the antibody detection as wrong more often than the Medisuch group. Furthermore, students of all three groups identified reliable webpages, which indicates that they do have internet skills that allow them to identify reliable information online. They were able to search for information online without navigation issues by entering specific webpages (like Flexicon) or medical terminology (diagnostic steps) in order to find reliable results [21]. Thus, students have successfully avoided the problem of standard search engine strategies, which often produce a multitude of results that users are forced to scroll through and sort the results [22]. This may be the reason why we did not find any differences between the groups using different search strategies.

A general question arises when viewing these results: Is entering specific terms or webpages into a generic search engine part of "digital literacy"? Many students searched at DocCheck [15], a webpage that is known for its trustworthy medical information. Students know that this webpage is a reliable source for information, as they use it for their medical studies [21,23,24]. Based on these results, medical students were able to comprehend the information found online and develop an

understanding of how to find and evaluate it, which could be interpreted as information literacy [3,25]. Furthermore, medical students showed technical use competency by including specific terms in the search strategies [3,26], which previous studies have described as digital nativity [27,28]. Thus, information literacy and digital nativity were shown by the medical students involved in this study, which are determinants of online information search strategies. This point presents an argument for the medical students having digital literacy [3].

Finding health information online still includes the risk of being misinformed by unreliable information or information providers, which ultimately affects patients [29,30]. Consequently, medical students as future health professionals should have appropriate internet skills and use them to help patients find trustworthy information online. Additionally, online technologies for health information should be implemented in the medical curriculum [1].

One limitation of this study might be that search results and choices of relevant information is only one step out of many. It might be relevant to both track students' search behavior and assess their cognitive processes preceding the search, such as defining the problem, choosing a certain source of information,

or formulating the search strategies [30]. Additionally, in their feedback, students in this study reported that 10 minutes for conducting an information search seemed too short to gather sufficient information. However, as time in daily practice is limited, we consider this a realistic time frame to look for needed information. Furthermore, we need to consider that the number of reliable webpages could differ if other search engines or different topics were used. Generally, it can be argued that the free Web might not offer the same reliable information like customized webpages do (eg, UpToDate or AMBOSS). Future studies, thus, should focus on the patients' online search strategies and use a similar setting to get more insights into general digital health literacy.

This study showed that medical students are able to search and find relevant medical information online regardless of the search engine used, and thus, this study confirms previous findings of medical students having internet skills in a professional sense. Future studies could focus on how to best integrate these internet skills into the medical curriculum. Furthermore, it needs to be determined if the online behavior of the students involved can be considered proper digital literacy. The next step could also focus on patients and their online search strategies.

Acknowledgments

Conflicts of Interest

None declared.

References

1. Kirschner PA, De Bruyckere P. The myths of the digital native and the multitasker. Teaching and Teacher Education 2017 Oct;67:135-142. [doi: 10.1016/j.tate.2017.06.001]

2. Kummervold PE, Chronaki CE, Lausen B, Prokosch H, Rasmussen J, Santana S, et al. eHealth trends in Europe 2005-2007: a population-based survey. J Med Internet Res 2008 Nov 17;10(4):e42 [FREE Full text] [doi: 10.2196/jmir.1023] [Medline: 19017584]

3. Çoklar AN, Yaman ND, Yurdakul IK. Information literacy and digital nativity as determinants of online information search strategies. Computers in Human Behavior 2017 May;70:1-9. [doi: 10.1016/j.chb.2016.12.050]

4. Thatcher A. Web search strategies: The influence of Web experience and task type. Information Processing & Management 2008 May;44(3):1308-1329. [doi: 10.1016/j.ipm.2007.09.004]

5. Tu Y, Shih M, Tsai C. Eighth graders' web searching strategies and outcomes: The role of task types, web experiences and epistemological beliefs. Computers & Education 2008 Nov;51(3):1142-1153. [doi: 10.1016/j.compedu.2007.11.003]

6. Çevik YD. Predicting college students' online information searching strategies based on epistemological, motivational, decision-related, and demographic variables. Computers & Education 2015 Dec;90:54-63. [doi: 10.1016/j.compedu.2015.09.002]

7. Willoughby T, Anderson SA, Wood E, Mueller J, Ross C. Fast searching for information on the Internet to use in a learning context: The impact of domain knowledge. Computers & Education 2009 Apr;52(3):640-648. [doi: 10.1016/j.compedu.2008.11.009]

8. Tsai M, Liang J, Hou H, Tsai C. University students' online information searching strategies in different search contexts. AJET 2012 Jul 26;28(5). [doi: 10.14742/ajet.822]

9. Armstrong CJ, Large JA. Manual of Online Search Strategies. London: Routledge; Dec 20, 2018.

10. Prensky M. Digital Natives, Digital Immigrants Part 1. On the Horizon 2001 Sep;9(5):1-6. [doi: 10.1108/10748120110424816]

11. Medisuch. 2017. Zertifizierungsverfahren Medisuch URL: http://www.medisuch.de/zertifizierungsverfahren.asp [accessed 2019-01-29]

XSL•FO

RenderX

12. LaValley SA, Kiviniemi MT, Gage-Bouchard EA. Where people look for online health information. Health Info Libr J 2017 Jun 21;34(2):146-155. [doi: 10.1111/hir.12143] [Medline: 27207817]

13. Pauer F, Göbel J, Storf H, Litzkendorf S, Babac A, Frank M, et al. Adopting Quality Criteria for Websites Providing Medical Information About Rare Diseases. Interact J Med Res 2016 Aug 25;5(3):e24 [FREE Full text] [doi: 10.2196/ijmr.5822] [Medline: 27562540]

14. Google. URL: https://www.google.de/ [accessed 2019-01-29]

15. DocCheck. URL: https://www.doccheck.com/welcome [accessed 2019-01-29]

16. Morae. Okemos, MI: TechSmith Corporation; 2011. URL: https://www.techsmith.com/tutorial-morae-current.html

17. Currie L, Devlin F, Emde J, Graves K. Undergraduate search strategies and evaluation criteria. New Library World 2010 Mar 30;111(3/4):113-124. [doi: 10.1108/03074801011027628]

18. Holman L. Millennial Students' Mental Models of Search: Implications for Academic Librarians and Database Developers. The Journal of Academic Librarianship 2011 Jan;37(1):19-27. [doi: 10.1016/j.acalib.2010.10.003]

19. Bloom B, Deyrup MM. The SHU research logs: student online search behaviors trans-scripted. The Journal of Academic Librarianship 2015 Sep;41(5):593-601. [doi: 10.1016/j.acalib.2015.07.002]

20. van Deursen AJAM. Internet skill-related problems in accessing online health information. Int J Med Inform 2012 Jan;81(1):61-72. [doi: 10.1016/j.ijmedinf.2011.10.005] [Medline: 22079240]

21. O'Doherty D, Lougheed J, Hannigan A, Last J, Dromey M, O'Tuathaigh C, et al. Internet skills of medical faculty and students: is there a difference? BMC Med Educ 2019 Jan 30;19(1):39 [FREE Full text] [doi: 10.1186/s12909-019-1475-4] [Medline: 30700293]

22. Wang L, Wang J, Wang M, Liang J, Xu D. User experience evaluation of Google search for obtaining medical knowledge: a case study. IJDMB 2011;5(6):626-639. [doi: 10.1504/IJDMB.2011.045414]

23. Mesko B, Győrffy Z, Kollár J. Digital literacy in the medical curriculum: a course with social media tools and gamification. JMIR Med Educ 2015 Oct 01;1(2):e6 [FREE Full text] [doi: 10.2196/mededu.4411] [Medline: 27731856]

24. Scott K, Morris A, Marais B. Medical student use of digital learning resources. Clin Teach 2018 Feb 16;15(1):29-33. [doi: 10.1111/tct.12630] [Medline: 28300343]

25. Ananiadou K, Claro M. 21st century skills and competences for new millennium learners in OECD countries. OECD Education Working Papers 2009. [doi: 10.1787/218525261154]

26. Tsai M, Tsai C. Information searching strategies in web-based science learning: the role of internet self-efficacy. Innovations in Education and Teaching International 2003 Jan;40(1):43-50. [doi: 10.1080/1355800032000038822]

27. Guo RX, Dobson T, Petrina S. Digital natives, digital immigrants: an analysis of age and ict competency in teacher education. Journal of Educational Computing Research 2008 Jun 02;38(3):235-254. [doi: 10.2190/ec.38.3.a]

28. Thornham H, McFarlane A. Discourses of the digital native. Information, Communication & Society 2011 Mar;14(2):258-279. [doi: 10.1080/1369118x.2010.510199]

29. Zimmermann M, Jucks R. How experts' use of medical technical jargon in different types of online health forums affects perceived information credibility: randomized experiment with laypersons. J Med Internet Res 2018 Jan 23;20(1):e30 [FREE Full text] [doi: 10.2196/jmir.8346] [Medline: 29362212]

30. van Deursen AJAM, van Dijk JAGM. Measuring internet skills. International Journal of Human-Computer Interaction 2010 Sep 17;26(10):891-916. [doi: 10.1080/10447318.2010.496338]

Abbreviations

ANOVA: analysis of variance
HONcode: Health On the Net Foundation Code of Conduct
IT: information technology.

XSL•FO

RenderX

Original Paper

A Lightweight Deep Learning Model for Fast Electrocardiographic Beats Classification With a Wearable Cardiac Monitor: Development and Validation Study

Eunjoo Jeon[1*], PhD, RN; Kyusam Oh[1*], MS; Soonhwan Kwon[1], BS; HyeongGwan Son[1], BS; Yongkeun Yun[1], BS; Eun-Soo Jung[1], PhD; Min Soo Kim[1], PhD

Technology Research, Samsung SDS, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Min Soo Kim, PhD
Technology Research
Samsung SDS
56 Seongchon-ro, Seocho-gu
Seoul, 06765
Republic of Korea
Phone: 82 2 6155 3808
Email: minsoo07.kim@samsung.com

Abstract

Background: Electrocardiographic (ECG) monitors have been widely used for diagnosing cardiac arrhythmias for decades. However, accurate analysis of ECG signals is difficult and time-consuming work because large amounts of beats need to be inspected. In order to enhance ECG beat classification, machine learning and deep learning methods have been studied. However, existing studies have limitations in model rigidity, model complexity, and inference speed.

Objective: To classify ECG beats effectively and efficiently, we propose a baseline model with recurrent neural networks (RNNs). Furthermore, we also propose a lightweight model with fused RNN for speeding up the prediction time on central processing units (CPUs).

Methods: We used 48 ECGs from the MIT-BIH (Massachusetts Institute of Technology-Beth Israel Hospital) Arrhythmia Database, and 76 ECGs were collected with S-Patch devices developed by Samsung SDS. We developed both baseline and lightweight models on the MXNet framework. We trained both models on graphics processing units and measured both models' inference times on CPUs.

Results: Our models achieved overall beat classification accuracies of 99.72% for the baseline model with RNN and 99.80% for the lightweight model with fused RNN. Moreover, our lightweight model reduced the inference time on CPUs without any loss of accuracy. The inference time for the lightweight model for 24-hour ECGs was 3 minutes, which is 5 times faster than the baseline model.

Conclusions: Both our baseline and lightweight models achieved cardiologist-level accuracies. Furthermore, our lightweight model is competitive on CPU-based wearable hardware.

Introduction

Background

Arrhythmia refers to any change causing the heart to beat too fast or slow, or erratically [1], and can lead to sudden death or critical adverse outcomes such as embolic stroke [2]. Therefore, early detection and treatment of arrhythmia are very important.

One of the most widely used diagnostic methods for detecting arrhythmia is electrocardiographic (ECG) monitoring. ECG monitoring is a simple and noninvasive method for recording electrical activities of the heart by using electrodes placed on

human skin. However, at least 24 hours of ECG signals should be monitored to confirm arrhythmia since it occurs irregularly [3,4]. Recently, single lead patches that are wireless, compact, and lightweight have been proposed for long-term wear [5-7]. Despite improvements to measuring ECGs and patient comfort, it is still difficult to diagnose arrhythmias because identification of abnormal ECG patterns from large amounts of recorded ECGs is not trivial. For example, an ECG record, measured for 24 hours in patients with a heart rate of 80 bpm, consists of 110,000 beats. It takes at least 2 hours for an expert to analyze this 24-hour ECG signal.

Large-scale machine learning methods have been investigated to reduce the human efforts for ECG beat classification [8-10]. However, most machine learning approaches with static and handcrafted features have performed at lower accuracy rates over new types of ECGs because those features are insufficient for representing the great diversity of ECG patterns from various patients. Therefore, several self-learning approaches based on a deep neural network have been proposed recently [11-13].

Among the deep learning approaches, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are most commonly used for ECG classification. CNNs typically consist of convolution, pooling, and fully connected layers [14]. CNNs extract implicit features of ECGs through each level of convolution layer and use the abstraction from these features to solve problems such as classification and regression [15]. Rajpurkar et al [11] used a CNN to classify 12 ECG rhythms, which are longer units consisting of 2 or more beats. Their model consisted of 33 convolutional layers with shortcut connections followed by a fully connected layer and a softmax layer. The model achieved an F1 score of 0.81 compared with the responses of board-certified cardiologists. Acharya et al [12] proposed an ECG beat classification model using a CNN together with noise removal, wavelet transformation, and segmentation method. Their model consisted of three convolutional layers, three max-pooling layers, three fully connected layers, and finally, a softmax layer with five output neurons. The model resulted in an average accuracy of 94.03% compared with the MIT-BIH (Massachusetts Institute of Technology-Beth Israel Hospital) gold standard.

However, a limitation of CNNs is that the length of inputs must be fixed since the filters of the networks have static sizes. When it comes to ECG classification, the length of ECGs can be varied according to an individual's heart rate. Therefore, adjusting data such as linear interpolation is required to achieve same-size inputs [11-13].

In contrast, RNNs are able to handle this sequential problem because the networks recursively learn data as time progresses [16]. Tan et al [17] proposed the implementation of a long short-term memory network (LSTM), which is the most widely used method among RNN approach, with a CNN to diagnose the presence of coronary artery disease from the ECG signals. Although they focused on specific diseases, they achieved an F1 score of 0.96. Oh et al [18] diagnosed five types of rhythms: normal sinus rhythm, left bundle branch block, right bundle branch block, atrial premature beats, and premature ventricular contraction. Their model consisted of three 1D convolution

layers, one LSTM layer, and three fully connected layers. They achieved a 98.10% accuracy using 10-fold cross-validation. Yildirim [19] also classified five types of rhythms but chose bidirectional LSTM (bi-LSTM) instead of unidirectional LSTM (uni-LSTM). The proposed model was composed of four wavelet transform layers, two bi-LSTMs, and two fully connected layers. This model showed a recognition accuracy of 99.39%.

However, existing studies using RNNs have limitations in application [17-20]. First, subject-specific evaluation to explore differences between patients is generally not conducted. Therefore, it is difficult to trust predictions of RNNs on new patients' ECG signals that were not included in the training data. Second, RNNs have disadvantages related to financial cost and inference time. Most of the papers did not consider the cost of using a graphics processing unit (GPU) instead of a central processing unit (CPU) and did not present the time to inference with their deep learning models. This weakness in computational efficiency is a critical drawback of RNN applications. To accelerate and maximize the computational efficiency of RNN layers, MXNet proposed fused RNN operator by applying several optimization methods: (1) various general matrix multiplication (GEMM) modes such as combining small GEMMs, Batch GEMM, and Pack GEMM; (2) vectorization of elementwise operations using Basic Linear Algebra Subprogram (BLAS) libraries and Intel Math Kernel Library (MKL); and (3) saving and reusing intermediate results during forward computation [21,22].

Objectives

We used ECG signals measured with Samsung S-Patch 2, a small ($120\times29\times4.4$ mm in size) and light (8 g in weight) patch-type ECG monitor [7]. To diagnose arrhythmias using S-Patch devices effectively and efficiently, we propose a baseline model with RNN that can learn sequential patterns. Furthermore, we also propose a lightweight model with fused RNN for conducting the classification process on CPUs with a shorter prediction time.

Methods

Data Collection

We analyzed an open-source ECG database (PhysioBank MIT-BIH Arrhythmia Database [23]) together with our own deidentified dataset collected with the S-Patch device. Overall, the MIT-BIH Arrhythmia Database contains 48 subjects' ECGs, each measured for 24 hours at 360 Hz.

The S-Patch database was obtained according to the following procedures. First, we collected the ECGs using S-Patch at the Samsung Medical Center in Seoul and at Counties Manukau Health in New Zealand from February 2017 to April 2018. A skilled nurse at each hospital attached an ECG monitor and checked whether the ECGs were normally collected over the first 5 minutes. The patches of S-patch cardiac monitor were attached on v2 and v5 positions of the 12-lead placement (Figure 1). Second, three experts, who had more than 5 years of experience working in a territorial hospital, reviewed each record and excluded ones that contained noise levels of more than 80% to enhance the data quality. Subsequently, we anonymized data

XSL•FO

RenderX

by removing personal and location information. Third, the three experts annotated each beat using the Web portal. If consensus could not be reached on the classification of a beat, the experts rediscussed the issue to make a final decision. Consequently, we collected 1828-hours of ECG data from 76 subjects. The average length of the ECGs from S-Patch was 17 hours (from 28 minutes to 45 hours). Each ECG collected with S-Patch was sampled at 256 Hz.

In this study, we used five beat categories defined in the AAMI/IEC (Association for the Advancement of Medical Instrumentation / International Electrotechnical Commission) standard [24] (ie, nonectopic [N], supraventricular ectopic [S], ventricular ectopic [V], fusion [F], and paced or unknown [Q]). Overall, 5,575,512 ECG beats were used in this study, as shown in Table 1.

Figure 1. Usage of S-Patch for Samsung SDS Cardio.

Table 1. The five subtype classes and the number of samples.

AAMI[a]/IEC[b] categories	Number of beats	
	MIT-BIH[c] dataset, n (%)	S-Patch dataset, n (%)
Nonectopic	90,386 (82.19)	5,303,245 (97.03)
Supraventricular ectopic	3026 (2.75)	27,288 (0.5)
Ventricular ectopic	7708 (7.01)	135,013 (2.47)
Fusion	803 (0.74)	0 (0)
Paced or unknown	8043 (7.31)	0 (0)
Total	109,966 (100.00)	5,465,546 (100.00)

[a]AAMI: Association for the Advancement of Medical Instrumentation.

[b]IEC: International Electrotechnical Commission.

[c]MIT-BIH: Massachusetts Institute of Technology-Beth Israel Hospital.

Data Preprocessing

We performed ECG preprocessing as follows (Figure 2A): downsampling, noise removal, segmentation, and short-time Fourier transform (STFT). The examples of preprocessed signals for the beat classes are also depicted (Figure 2B): A normal beat has a regular beat interval with a small wave (P-wave) before

a larger and sharper wave (QRS wave); a supraventricular beat has an irregular beat interval; a ventricular beat has a wide QRS wave with a vague or no P-wave; a fusion beat is a combined pattern of normal and ventricular beat; and a paced or unknown beat has none of the abovementioned features and can be observed in diverse patterns.

Figure 2. Data preprocessing of electrocardiograms: (A) full steps from downsampling to short-time Fourier transform; (B) an example of a 3-beat electrocardiographic segment for each class. ECG: electrocardiogram; RNN: recurrent neural network.
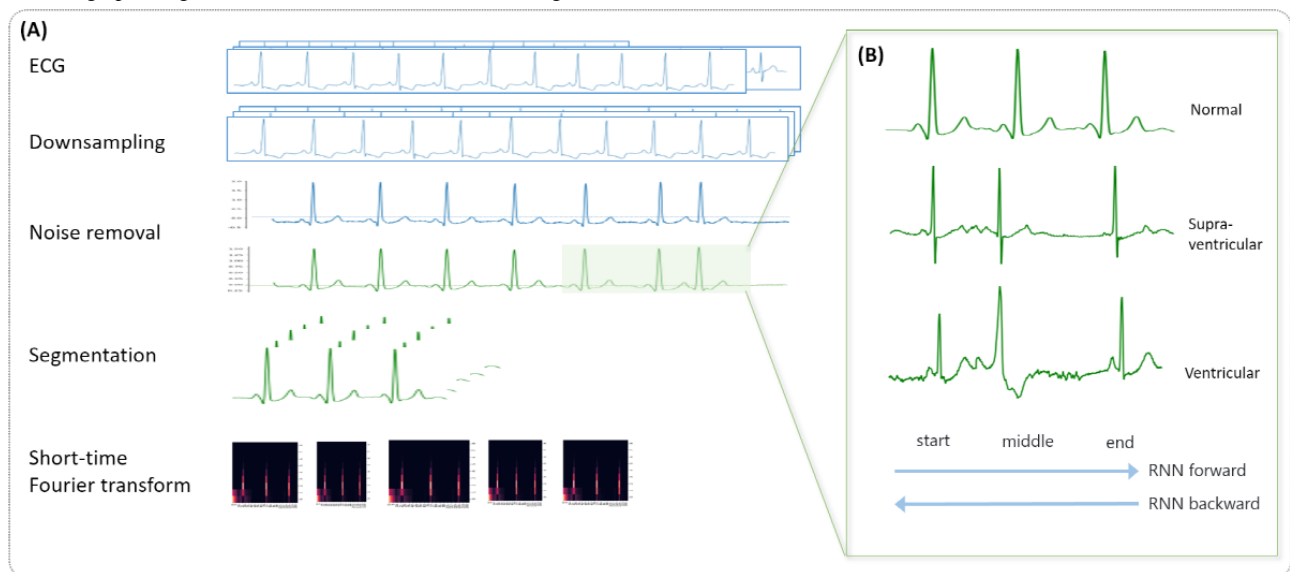
First, we downsampled the ECG signals to handle the different sampling rates between data from difference databases. For consistency, the MIT-BIH records were downsampled to 256 Hz, which is the same as the sampling rate of the S-Patch dataset. Second, we tried to reduce artifacts in the data. The ECGs collected with S-Patch are real-world data. Therefore, they contained all kinds of noise such as loose contacts, motion artifact, muscular activation interference, baseline wandering, and AC (alternating current) interference (Figure 3). We excluded noise caused by loose contact that falls below 0 mV. Thereafter, we applied a bandpass filter with a high-frequency cutoff at 40 Hz and a low-frequency cutoff at 0.5 Hz to handle other types of noise. Third, in order to deal with 24-hour ECGs effectively, we segmented the ECGs into beat units. Since the duration of a beat is different according to the heart rate of each patient, we extracted 3-beat ECG signals with R-peaks identified using the algorithm developed by Kathirvel et al [25] instead of an arbitrary time duration. Specifically, we selected a window with a length of 3 beats because information on the middle beat, which is the target to be classified, is affected by the preceding and following beats. Finally, we applied STFT to all ECG segments.

Figure 3. Examples of electrocardiographic signal noise: (A) loose contact, (B) motion artifact, (C) muscular activation interference, (D) baseline wandering, (E) alternating current (AC) interference (low signal-to-noise ratio), (F) AC interference (high signal-to-noise ratio).
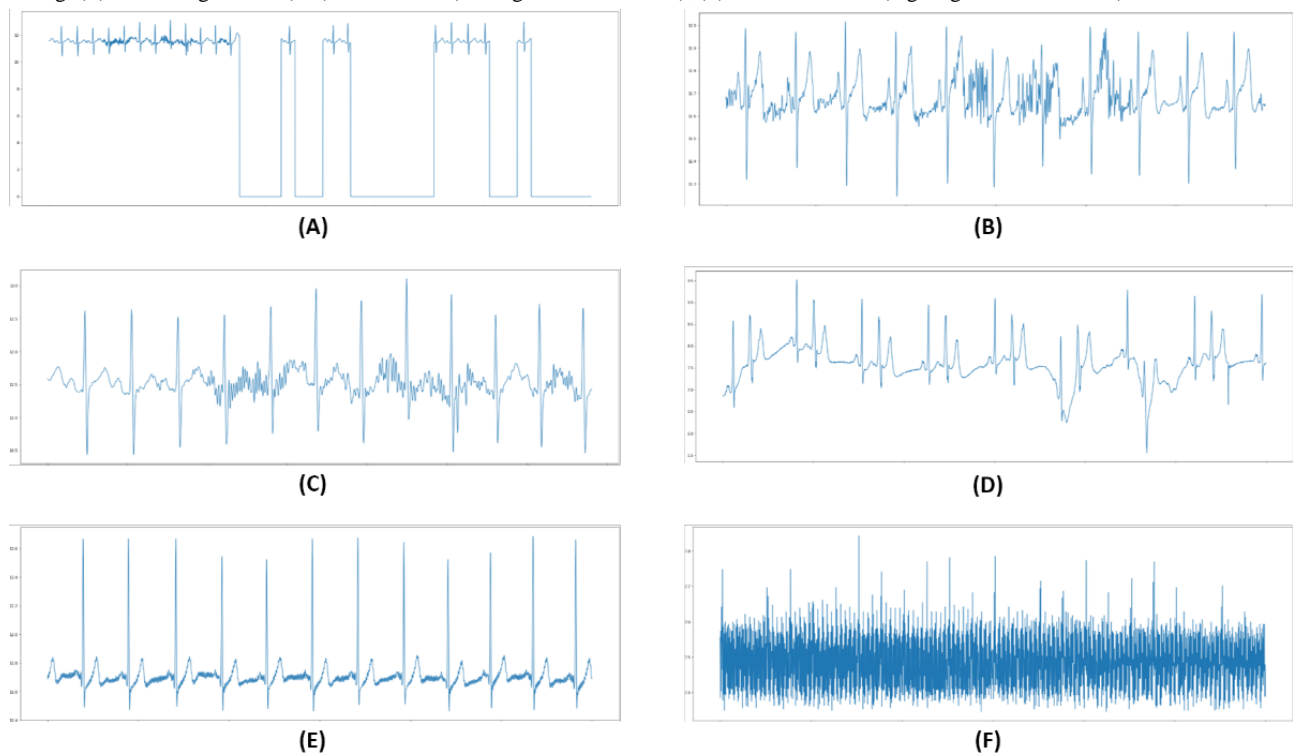
Baseline Model With Recurrent Neural Network

We segmented ECG signals into 3-beat units during the preprocessing; thus, their lengths varied according to the subjects' heart rates. As baselines for ECG pattern classification, we implemented Vanilla RNNs (Figure 4), which can handle input sequences with various lengths. However, the variable length of input data can reduce the learning efficiency of deep learning models. Therefore, we used a bucketing method to handle the variable length of inputs. Bucketing is suggested to improve the parallelization capabilities of the recurrent training process. We set up several buckets and assigned each instance to the bucket with the closest size. Within a bucket, each instance was padded with zeroes up to the length of the bucket. Although the buckets had different internal models, their parameters were shared in time.

Figure 4. Model architecture (general).
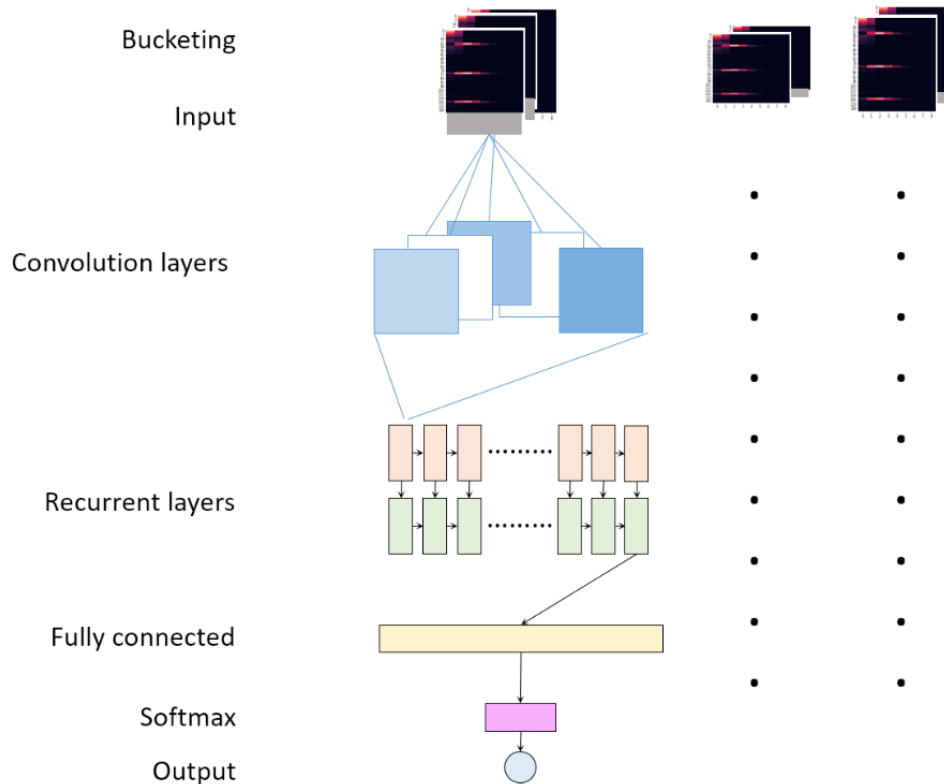
After bucketing, the training process for instances of each bucket was as follows: first, the input data was convolved with 11 filters (3×11 in size) with a stride of 1 in the first convolution layer, which was followed by a convolution layer with 11 filters (3×3 in size) with a stride of 1. Second, the outputs of the convolution layer proceeded through consecutive two Vanilla RNN layers with hidden states of 1760 for each. Finally, the outputs of the RNN layers passed to a fully connected layer, and a softmax function with 5 output nodes was used in the final layer. Batch normalization was used in each layer of the architecture.

Lightweight Model With Fused Recurrent Neural Network

Baseline model inference was performed on CPUs (Intel Xeon Platinum 8000 v4). However, it was about 6 times slower than the GPU-based inference (Tesla K80). In order to improve the inference speeds on CPUs, we propose a lightweight model by reducing the input size and adopting fused RNN (Table 2).

To reduce the input size, we selected a minimum sampling rate by halving the sampling rate from 256, provided there was no degradation in accuracy. Finally, we downsampled the MIT-BIH (360 Hz) and S-Patch data (256 Hz) to 64 Hz. As the input size decreased, the filter size changed from 3×11 to 3×5 and the number of convolution layers changed from two to one, compared to the baseline model. Additionally, we changed the RNN layers to fused RNN instead of Vanilla RNN to maximize the computation efficiency of the RNN layers in CPUs. Additionally, we used Intel Math Kernel Library 2018 update 3 for matrix-multiplication operation.

Table 2. Comparison of the baseline and lightweight models.

Characteristic	Baseline model	Lightweight model
Sampling rate (Hz)	256	64
Convolution layer	2	1
Convolution filter size	3×11×11, stride 1; 3×3×11, stride 1	3×5×11, stride 1
Recurrent layer	Vanilla RNN[a]	Fused RNN

[a]RNN: recurrent neural network.

Experimental Setup

We divided a total of 124 subjects into two groups—112 and 12 subjects for the train (including validation) and test sets, respectively. The train set consisted of 43 subjects from MIT-BIH and 69 from S-Patch, and the test set consisted of 5 subjects from MIT-BIH and 7 from S-Patch. Specifically, the 12 subjects in the test set were carefully selected by cardiologists to evaluate various types of beats.

Normal beats comprised more than 90% of the total data; therefore, we randomly sampled beats in the normal class equal to the total number of beats in other classes every epoch to avoid this data imbalance problem. Moreover, most of the samples for abnormal classes were from MIT-BIH; thus, data imbalance between MIT-BIH and S-Patch were also handled by balancing the number of samples for normal and other classes.

We used MXNet to create the baseline model with RNN and the lightweight model with fused RNN [26]. We trained both models on GPUs (OS: Linux, CPU: Intel Xeon E5-2686 v4 processor, memory: 488 GB, GPU: four NVIDIA K80 GPU) with Xavier initialization and Adam optimizer. The baseline model's learning rate was 5E-06 with a batch size of 1000 over 400 epochs, and the lightweight model's learning rate was 1E-05 with a batch size of 900 over 300 epochs. After training, we selected the best model with the highest validation accuracies for the three classes (N, S, and V Classes).

Evaluation Metric

Classification performance was measured by four standard metrics (ie, accuracies, sensitivities, specificities, and positive predictive values) that have been used in the literature [9,12,13]. These were calculated using the four values from the confusion matrix, true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The accuracy is the ratio of the number of correctly classified patterns to the total numbers of patterns classified, ▱. The sensitivity is the rate of correctly classified events among all events, ▱. The specificity is the rate of correctly classified nonevents among all nonevents, ▱. The positive predictive value is the rate of correctly classified events in all detected events, ▱. In addition to the metrics, we calculated an overall accuracy using the equation proposed by Landis and Koch [27]. To measure the inference speed, 24-hour ECGs with 64,976 beats were inferenced on both the GPUs and CPUs (OS: Linux, CPU: eight 3.0 GHz Intel Xeon Platinum processors, memory: 16 GB).

Results

Accuracy of Baseline and Lightweight Models

Figure 5 shows the overall performances in the classification of ECGs for the baseline (RNN) and lightweight (fused RNN) models. There was no significant difference between the baseline and lightweight models because the overall accuracies were close to each other: 99.72% for baseline and 99.80% for the lightweight model.

Figure 5. Normalized confusion matrix of the recurrent neural network (RNN) and fused RNN models.
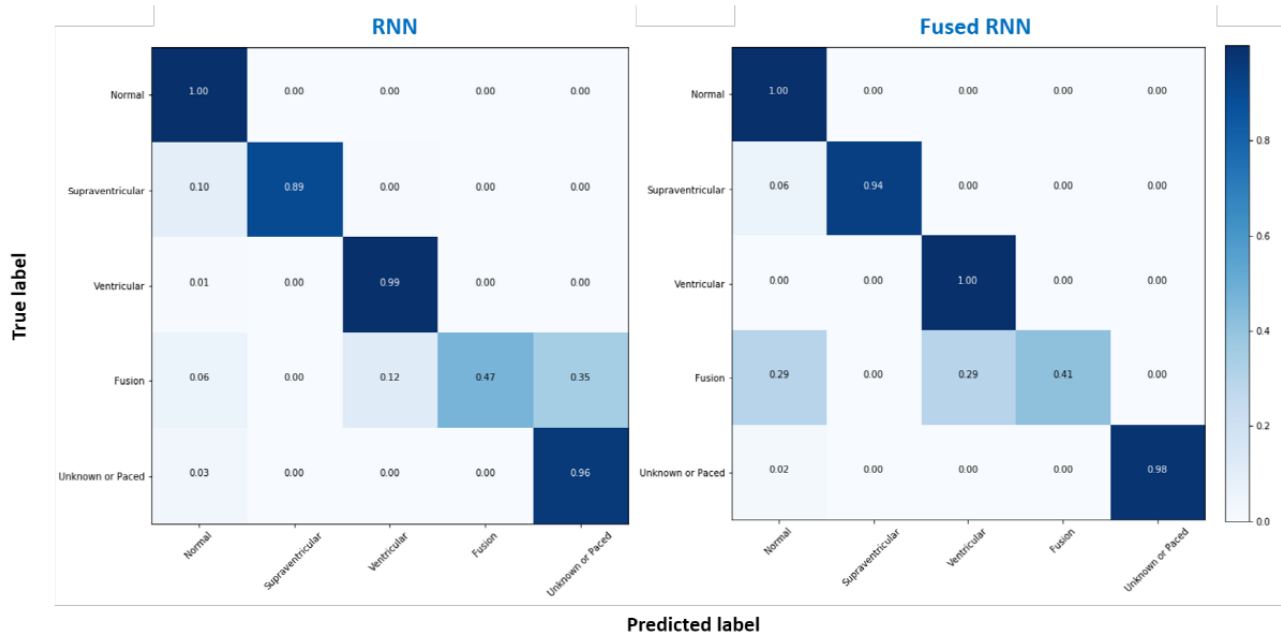
Tables 3 and 4 demonstrate the accuracies of the baseline and lightweight models for Class S (99.82% and 99.90%, respectively) and for Class V (99.91% and 99.89%, respectively). These results were far superior to those of previous works, wherein the results ranged from 92.4% to 97.6% for Class S and from 96.7% to 99.0% for Class V [13,28,29]. Note that we excluded the results for Class F because the number of beats in Class F was only 17.

Furthermore, both baseline and lightweight models performed higher than 97% in subject-specific accuracies, similar to the overall accuracy (Table 5). These results ensured the internal reliability of the models. The accuracy fluctuation range was 2.42% for baseline (from 97.56% to 99.88%) and 0.84% for lightweight model (from 99.12% to 99.96%). Our lightweight model with fused RNN improved overall accuracy and internal reliability even though it is relatively lighter.

Table 3. A confusion matrix of the baseline model for the test set.

Class (ground truth)	Classification results					ACC[f] (%)	PPV[g] (%)	SEN[h] (%)	SPEC[i] (%)
	Predicted class								
	n[a]	s[b]	v[c]	f[d]	q[e]				
N	481,491	362	215	2	46	99.75	99.87	99.86	98.31
S	438	3935	20	0	5	99.82	89.47	89.17	99.91
V	172	116	33,980	5	11	99.90	99.11	99.29	99.95
F	1	0	2	8	6	N/A[j]	N/A	N/A	N/A
Q	71	0	5	0	1984	99.97	96.31	96.69	99.99

[a]n/N: nonectopic.

[b]s/S: supraventricular ectopic.

[c]v/V: ventricular ectopic.

[d]f/F: fusion beat.

[e]q/Q: paced or unknown beat.

[f]ACC: accuracy.

[g]PPV: positive predictive value.

[h]SEN: sensitivity.

[i]SPEC: specificity.

[j]Not applicable.

Table 4. A confusion matrix of the lightweight model for the test set.

Class (ground truth)	Classification results								
	Predicted class					ACC[f] (%)	PPV[g] (%)	SEN[h] (%)	SPEC[i] (%)
	n[a]	s[b]	v[c]	f[d]	q[e]				
N	481,587	130	403	1	7	99.83	99.89	99.92	99.11
S	264	4123	6	0	0	99.91	93.85	94.80	99.96
V	51	96	34,157	7	0	99.89	99.55	98.80	99.92
F	5	0	5	7	0	N/A[j]	N/A	N/A	N/A
Q	43	0	0	0	2017	99.99	97.91	99.65	100.00

[a]n/N: nonectopic.

[b]s/S: supraventricular ectopic.

[c]v/V: ventricular ectopic.

[d]f/F: fusion beat.

[e]q/Q: paced or unknown beat.

[f]ACC: accuracy.

[g]PPV: positive predictive value.

[h]SEN: sensitivity.

[i]SPEC: specificity.

[j]Not applicable.

Table 5. Overall accuracies of the baseline and lightweight models according to subjects in the test set. (Subject numbers 100 to 223 are from MIT-BIH, and the rest are from S-patch.)

Subject #	Rhythm	Beats, n	RNN[a]				Fused RNN			
			ACC[b] (%)	PPV[c] (%)	SEN[d] (%)	SPEC[e] (%)	ACC (%)	PPV (%)	SEN (%)	SPEC (%)
100	Normal	2273	99.74	99.34	99.34	99.84	99.91	99.78	99.78	99.95
104	Paced	2225	98.04	95.10	95.10	98.78	99.12	97.80	97.80	99.45
108	Normal	1756	97.56	93.91	93.91	98.48	99.66	99.15	99.15	99.79
202	Afib[f]	2133	98.87	97.19	97.19	99.30	99.51	98.78	98.78	99.70
223	VT[g]	2604	97.45	93.63	93.63	98.41	98.54	96.35	96.35	99.09
12006	Normal	60,867	99.87	99.66	99.66	99.92	99.90	99.76	99.76	99.94
12007	Normal	75,805	99.87	99.66	99.66	99.92	99.99	99.97	99.97	99.99
12008	Normal	78,866	99.94	99.84	99.84	99.96	99.98	99.96	99.96	99.99
12010	SVT[h] and VT	75,914	99.89	99.72	99.72	99.93	99.97	99.94	99.94	99.98
12011	Bigeminy	66,229	99.92	99.79	99.79	99.95	99.71	99.27	99.27	99.82
12012	Normal	63,820	99.98	99.94	99.94	99.99	99.99	99.99	99.99	99.99
12358	Noise	85,091	99.97	99.92	99.92	99.98	99.96	99.90	99.90	99.97

[a]RNN: recurrent neural network.

[b]ACC: accuracy.

[c]PPV: positive predictive value.

[d]SEN: sensitivity.

[e]SPEC: specificity.

[f]Afib: atrial fibrillation.

[g]VT: ventricular tachycardia.

[h]SVT: supraventricular tachycardia.

Inference Speed

The inference time of the baseline model on GPUs took 15 minutes and 12 seconds. On the other hand, the lightweight model on CPUs took 3 minutes and 1 second (Table 6). Namely, our lightweight model took only 2 milliseconds to process one beat, and this implies that our model is competitive on a CPU-based wearable hardware [30]. The inference speed according to each parameter can be found in the AWS Re:Invent [31]; controlling the sampling rate and adopting fused RNNs each reduced the inference time. This result also demonstrated that replacing Vanilla RNNs to fused RNNs does not change the processes or parameters constituting a network but only accelerates the processes; consequently, the inference speed was improved without loss of accuracy.

Table 6. Comparison of accuracies and latencies.

Model	Latency (min)	Accuracy (%)
RNN[a] in GPUs[b]	15.12	99.72
RNN in CPUs[c]	120	99.80
Fused RNN in CPUs	3.01	99.80

[a]RNN: recurrent neural network.

[b]GPU: graphics processing unit.

[c]CPU: central processing unit.

Discussion

Principal Findings

The results showed that both baseline and lightweight models achieved high prediction performances (ie, accuracies of over 99%). The final model, fused RNNs, showed superior performance in both subclasses: supraventricular and ventricular beat. In addition, the reliability of the lightweight model with fused RNNs was supported with a prediction accuracy of over 99% in each subject as well as overall model performance.

Limitations

The accuracy of the lightweight model with fused RNN for beat classification was high, but there were still incorrect cases. These false cases were caused with specific rhythms, such as supraventricular tachycardia, bigeminy, and paroxysmal atrial fibrillation.

We reviewed 1018 beats that were falsely predicted by the lightweight model and interpreted the errors. As a result, the beats composing the abnormal rhythm were often misunderstood (Figure 6). For example, in the case of subject #12010, supraventricular beats in the supraventricular tachycardia rhythm were predicted incorrectly as normal beats. This is because the supraventricular beats in supraventricular tachycardia rhythm have very short intervals between two consecutive beats but have similar morphology to normal beats in normal rhythm (Figure 6A). Supraventricular beats can be easily misjudged as normal beats when the model reviewed only one segment, which has 3 beats. In another example, subject #12011, normal beats in the bigeminy rhythm were predicted as ventricular rhythm. This is because the intervals between normal beats in the bigeminy rhythm are relatively longer due to the leading and trailing ventricular beats. Therefore, normal beats in bigeminy rhythm have a different morphology from those in other rhythms (Figure 6B). The differences in beat morphology were also confirmed in the atrial fibrillation rhythm (Figure 6C).

Figure 6. Beats in rhythms with different shapes. SVT: supraventricular tachycardia; S: supraventricular ectopic; N: nonectopic.
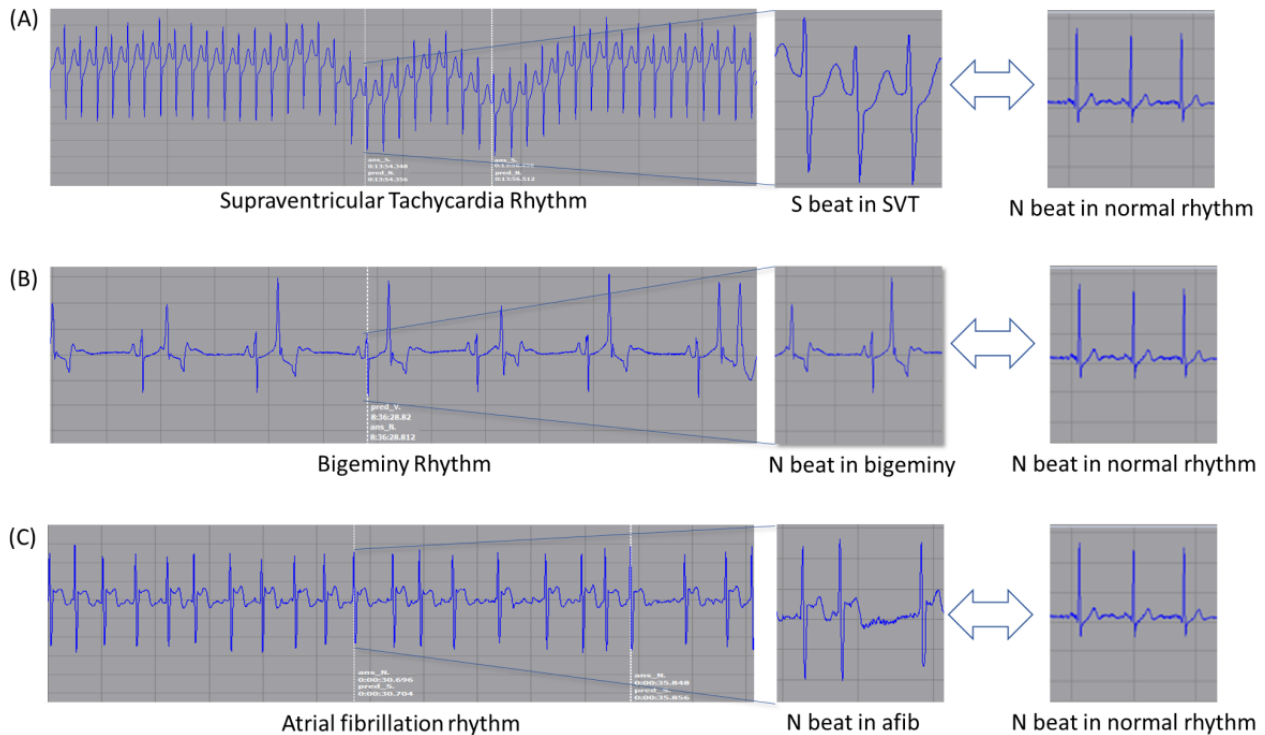
The case review confirmed that the rhythm affects each ECG beat. Therefore, it is necessary to develop a rhythm model using a wider ECG window. Although Rajpurkar et al [11] recently conducted a classification task for 14 ECG rhythms, most existing studies on rhythm predictions were limited to a specific rhythm such as atrial fibrillation [32,33]. We are developing a universal rhythm prediction model that integrates the results of the beat model developed in this study and other features such as R-R interval and R-peak amplitude.

In our case, the subject-specific evaluation was conducted to demonstrate the reliability of the deep learning models. However, it is necessary to perform an evaluation based on real-world data to support the model's reliability, which can provide generalized predictions for new data. Currently, this is in progress at a tertiary hospital in South Korea.

Future Works

The lightweight deep learning model for ECG classification proposed in this paper was adopted as an analysis module for Samsung SDS Cardio. Cardio is a service that collects ECG signals with S-Patch, sends the signals to a cloud-based Web portal for ECG diagnosis, and reports the diagnosis to users through their mobile or gear applications. This service also can be extended to health monitoring for elderly people, who are vulnerable to cardiovascular disease, and for first responders such as firemen. Furthermore, our ECG classifier can be embedded into a health care system together with patient-generated biomedical information analysis (eg, mobile search log, geotagged data) [34] and provide wider and deeper information to users.

Conclusion

We proposed lightweight deep neural network models that were effective and efficient for ECG beat classification. The proposed models were trained using both the standard Pysionet MIT-BIH database and Samsung S-Patch 2 dataset collected by two major hospitals in New Zealand and South Korea. Our lightweight model with fused RNN achieved a cardiologist-level accuracy of 99.80%. Furthermore, the lightweight model conducted ECG beat predictions on a CPU five times faster than the baseline model with Vanilla RNNs without accuracy loss.

Conflicts of Interest

None declared.

References

1. American Heart Association. 2016. About Arrhythmia URL: https://www.heart.org/en/health-topics/arrhythmia/about-arrhythmia [accessed 2016-08-30]
2. Sayantan G, Kien PT, Kadambari KV. Classification of ECG beats using deep belief network and active learning. Med Biol Eng Comput 2018 Oct;56(10):1887-1898. [doi: 10.1007/s11517-018-1815-2] [Medline: 29651694]
3. Lobodzinski SS, Laks MM. New devices for very long-term ECG monitoring. Cardiol J 2012;19(2):210-214 [FREE Full text] [Medline: 22461060]

4.   Guirguis E. Holter monitoring. Can Fam Physician 1987 Apr;33:985-992 [FREE Full text] [Medline: 21263911]

5.   Turakhia MP, Hoang DD, Zimetbaum P, Miller JD, Froelicher VF, Kumar UN, et al. Diagnostic utility of a novel leadless arrhythmia monitoring device. Am J Cardiol 2013 Aug 15;112(4):520-524 [FREE Full text] [doi: 10.1016/j.amjcard.2013.04.017] [Medline: 23672988]

6.   Engel J, Mehta V, Fogoros R, Chavan A. Study of arrhythmia prevalence in NUVANT Mobile Cardiac Telemetry system patients. : IEEE; 2012 Aug Presented at: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 28 Aug.-1 Sept. 2012; San Diego, CA, USA p. 2440-2443. [doi: 10.1109/embc.2012.6346457]

7.   Korea Food Drug Administration, Medical Device Safety Bureau. Approval information (No 16-4537). URL: https://emed.mfds.go.kr [accessed 2016-06-16]

8.   Martis R, Acharya U, Lim C, Suri J. Characterization of ECG beats from cardiac arrhythmia using discrete cosine transform in PCA framework. Knowledge-Based Systems 2013 Jun;45:76-82 [FREE Full text] [doi: 10.1016/j.knosys.2013.02.007]

9.   Martis RJ, Acharya UR, Lim CM, Mandana KM, Ray AK, Chakraborty C. Application of higher order cumulant features for cardiac health diagnosis using ECG signals. Int J Neural Syst 2013 Aug;23(4):1350014. [doi: 10.1142/S0129065713500147] [Medline: 23746287]

10. Ho T, Huang C, Lin C, Lai F, Ding J, Ho Y, et al. A telesurveillance system with automatic electrocardiogram interpretation based on support vector machine and rule-based processing. JMIR Med Inform 2015 May 07;3(2):e21 [FREE Full text] [doi: 10.2196/medinform.4397] [Medline: 25953306]

11. Rajpurkar P, Hannun A, Haghpanahi M, Bourn C, Ng A. Cardiologist-level arrhythmia detection with convolutional neural networks (preprint). arXiv.org 2017:6 [FREE Full text]

12. Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adam M, Gertych A, et al. A deep convolutional neural network model to classify heartbeats. Comput Biol Med 2017 Oct 01;89:389-396. [doi: 10.1016/j.compbiomed.2017.08.022] [Medline: 28869899]

13. Kiranyaz S, Ince T, Gabbouj M. Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks. IEEE Trans Biomed Eng 2016 Mar;63(3):664-675. [doi: 10.1109/TBME.2015.2468589] [Medline: 26285054]

14. Ferreira A, Giraldi G. Convolutional Neural Network approaches to granite tiles classification. Expert Systems with Applications 2017 Oct;84:1-11 [FREE Full text] [doi: 10.1016/j.eswa.2017.04.053]

15. Bengio Y. Learning Deep Architectures for AI. FNT in Machine Learning 2009;2(1):1-127. [doi: 10.1561/2200000006]

16. Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, et al. Deep speech 2nd-to-end speech recognition in english and mandarin. 2016 Presented at: International Conference on Machine Learning; June 19-24 2016; New York City, NY, USA p. 173-182 URL: http://proceedings.mlr.press/v48/amodei16.html [doi: 10.1007/978-3-030-14596-5_12]

17. Tan JH, Hagiwara Y, Pang W, Lim I, Oh SL, Adam M, et al. Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. Comput Biol Med 2018 Mar 01;94:19-26. [doi: 10.1016/j.compbiomed.2017.12.023] [Medline: 29358103]

18. Oh SL, Ng EYK, Tan RS, Acharya UR. Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. Comput Biol Med 2018 Nov 01;102:278-287. [doi: 10.1016/j.compbiomed.2018.06.002] [Medline: 29903630]

19. Yildirim O. A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. Comput Biol Med 2018 May 01;96:189-202. [doi: 10.1016/j.compbiomed.2018.03.016] [Medline: 29614430]

20. Kwon S, Hong J, Choi E, Lee E, Hostallero DE, Kang WJ, et al. Deep Learning Approaches to Detect Atrial Fibrillation Using Photoplethysmographic Signals: Algorithms Development Study. JMIR Mhealth Uhealth 2019 Jun 06;7(6):e12770 [FREE Full text] [doi: 10.2196/12770] [Medline: 31199302]

21. Apache MXNet (incubating). LSTMCell. 2019. URL: https://github.com/apache/incubator-mxnet/blob/master/python/mxnet/rnn/rnn_cell.py [accessed 2019-12-27]

22. Apache MXNet (achieved proposal). Fused RNN Operators for CPU. 2018. URL: https://cwiki.apache.org/confluence/display/MXNET/Fused+RNN+Operators+for+CPU [accessed 2019-12-27]

23. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):E215-E220 [FREE Full text] [Medline: 10851218]

24. International Electrotechnical Commission. Medical electrical equipment Part 2-27: Particular requirements for the safety, including essential performance, of electrocardiographic monitoring equipment. 2011. URL: https://webstore.iec.ch/publication/2638 [accessed 2011-12-27]

25. Kathirvel P, Sabarimalai Manikandan M, Prasanna S, Soman K. An Efficient R-peak Detection Based on New Nonlinear Transformation and First-Order Gaussian Differentiator. Cardiovasc Eng Tech 2011 Oct 12;2(4):408-425 [FREE Full text] [doi: 10.1007/s13239-011-0065-3]

26. Chen T, Li M, Li Y, Lin M, Wang N, Wang M, et al. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems (preprint). arXiv.org 2015:3 [FREE Full text]

27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]

XSL•FO

RenderX

28.   Li T, Zhou M. ECG Classification Using Wavelet Packet Entropy and Random Forests. Entropy 2016 Aug 05;18(8):285
        [FREE Full text] [doi: 10.3390/e18080285]

29.   Zubair M, Kim J, Yoon C. An automated ECG beat classification system using convolutional neural networks. : IEEE;
        2016 Presented at: 2016 6th International Conference on IT Convergence and Security (ICITCS); 2016 Sep 26; Prague,
        Czech Republic p. 1-5 URL: https://doi.org/10.1109/ICITCS.2016.7740310 [doi: 10.1109/icitcs.2016.7740310]

30.   Saadatnejad S, Oveisi M, Hashemi M. LSTM-Based ECG Classification for Continuous Monitoring on Personal Wearable
        Devices. IEEE J Biomed Health Inform 2019 Apr 15. [doi: 10.1109/JBHI.2019.2911367] [Medline: 30990452]

31.   Hagay L, Seungjai M. AWS Re:Invent 2018. 2018 Nov 26. Deep Learning for Developers: An Introduction, Featuring
        Samsung SDS URL: https://tinyurl.com/tz6m9dt [accessed 2020-03-06]

32.   Xia Y, Wulan N, Wang K, Zhang H. Detecting atrial fibrillation by deep convolutional neural networks. Comput Biol Med
        2018 Feb 01;93:84-92. [doi: 10.1016/j.compbiomed.2017.12.007] [Medline: 29291535]

33.   Clifford G, Liu C, Moody B, Lehman LWH, Silva I, Li Q, et al. AF Classification from a Short Single Lead ECG Recording:
        the PhysioNet/Computing in Cardiology Challenge 2017. Comput Cardiol (2010) 2017 Sep;44 [FREE Full text] [doi:
        10.22489/CinC.2017.065-469] [Medline: 29862307]

34.   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]

Abbreviations

AAMI: Association for the Advancement of Medical Instrumentation
ACC: accuracy
BLAS: Basic Linear Algebra Subprogram
CNN: convolutional neural network
CPU: central processing unit
ECG: electrocardiogram (electrocardiographic)
F: fusion beat
FN: false negative
FP: false positive
GEMM: general matrix multiplication
GPU: graphics processing unit
IEC: International Electrotechnical Commission
MIT-BIH: Massachusetts Institute of Technology-Beth Israel Hospital
MKL: Intel Math Kernel Library
N: nonectopic
PPV: positive predictive value
Q: paced or unknown beat
RNN: recurrent neural network
S: supraventricular ectopic
SEN: sensitivity
SPEC: specificity
STFT: short-time Fourier transform
TN: true negative
TP: true positive
V: ventricular ectopic

Original Paper

Development of a Real-Time Risk Prediction Model for In-Hospital Cardiac Arrest in Critically Ill Patients Using Deep Learning: Retrospective Study

Junetae Kim[1,2,3], PhD; Yu Rang Park[4], PhD; Jeong Hoon Lee[4], PhD; Jae-Ho Lee[5,6], MD, PhD; Young-Hak Kim[5,7,8,9], MD, PhD; Jin Won Huh[10], MD, PhD

[1]Graduate School of Cancer Science and Policy, National Cancer Center, Goyang-si, Republic of Korea

[2]Cancer Data Center, National Cancer Control Institute, National Cancer Center, Goyang-si, Republic of Korea

[3]Healthcare AI Team, Healthcare Platform Center, National Cancer Center, Goyang-si, Republic of Korea

[4]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea

[5]Department of Biomedical Informatics, Asan Medical Center, Seoul, Republic of Korea

[6]Department of Emergency Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[7]Health Innovation Big Data Center, Asan Institute for Life Science, Asan Medical Center, Seoul, Republic of Korea

[8]Medical Information Office, Asan Medical Center, Seoul, Republic of Korea

[9]Department of Cardiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[10]Department of Pulmonary and Critical Care Medicine, Asan Medical Center, College of Medicine, University of Ulsan, Seoul, Republic of Korea

Corresponding Author:
Yu Rang Park, PhD
Department of Biomedical Systems Informatics
Yonsei University College of Medicine
50-1 Yonsei-ro, Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 2 2228 2493
Email: yurangpark@yuhs.ac

Abstract

Background: Cardiac arrest is the most serious death-related event in intensive care units (ICUs), but it is not easily predicted because of the complex and time-dependent data characteristics of intensive care patients. Given the complexity and time dependence of ICU data, deep learning–based methods are expected to provide a good foundation for developing risk prediction models based on large clinical records.

Objective: This study aimed to implement a deep learning model that estimates the distribution of cardiac arrest risk probability over time based on clinical data and assesses its potential.

Methods: A retrospective study of 759 ICU patients was conducted between January 2013 and July 2015. A character-level gated recurrent unit with a Weibull distribution algorithm was used to develop a real-time prediction model. Fivefold cross-validation testing (training set: 80% and validation set: 20%) determined the consistency of model accuracy. The time-dependent area under the curve (TAUC) was analyzed based on the aggregation of 5 validation sets.

Results: The TAUCs of the implemented model were 0.963, 0.942, 0.917, 0.875, 0.850, 0.842, and 0.761 before cardiac arrest at 1, 8, 16, 24, 32, 40, and 48 hours, respectively. The sensitivity was between 0.846 and 0.909, and specificity was between 0.923 and 0.946. The distribution of risk between the cardiac arrest group and the non–cardiac arrest group was generally different, and the difference rapidly increased as the time left until cardiac arrest reduced.

Conclusions: A deep learning model for forecasting cardiac arrest was implemented and tested by considering the cumulative and fluctuating effects of time-dependent clinical data gathered from a large medical center. This real-time prediction model is expected to improve patient's care by allowing early intervention in patients at high risk of unexpected cardiac arrests.

KEYWORDS

deep learning; cardiac arrest; Weibull distribution; forecasting; intensive care units; gated recurrent unit

Introduction

Background

In-hospital cardiac arrest (IHCA) is etiologically different from out-of-hospital cardiac arrest owing to the variety of underlying illnesses in hospitalized patients. Unfortunately, despite efforts to improve survival following IHCA, outcomes have not significantly improved over the last few decades [1,2]. In particular, unexpected cardiac arrest is the most serious adverse event related to death in intensive care units (ICUs). The common reasons for cardiac arrest in critically ill patients are severe respiratory insufficiency and hypotension.

Several studies have reported that mortality after IHCA was associated with the timing of cardiac arrest (day vs night shift), type of institution (teaching vs nonteaching hospital), and subsets of patients (ie, age and sex of patients) [3-5]. However, these are not preventable factors. To reduce mortality, we need to be able to predict which critically ill patients are at high risk for IHCA before arrest and the actionable factor to reduce the risk of cardiac arrest. Although many arrests are preceded by clinical deterioration that is either unrecognized or not managed effectively, the complex data of ICU patients make it difficult to model and monitor early warning signs [6]. An additional problem with arrest prediction is the nature of time to event (TTE); we would like to be able to predict not only whether or not cardiac arrest will occur but also when that event will occur [7].

Given the complexity and time dependency of ICU data, machine learning–based methods including the deep learning–based early warning system and gradient boosting machine have provided a good basis to develop risk prediction models using large clinical data contained within electronic medical records [8-12]. Specifically, several deep neural network algorithms have been applied to develop an early warning system for cardiac arrest to predict IHCA a few hours before the event [12,13]. However, in the time series data, constructing the label of the data by assuming the dependent variable is binary has various risks because the time for the onset of symptoms associated with cardiac arrest varies from patient to patient. In this study, rather than simply predicting the probability of cardiac arrest at the current time by binary classification, parameters of the Weibull distribution were used to predict the distribution of the probability of occurrence over time. This allows us to predict when cardiac arrest will occur at this time point, which will enable clinicians to alter the clinical trajectory to prevent cardiac arrest.

Objectives

This study aimed to develop a real-time deep learning model to predict the risk of cardiac arrest in critically ill patients in a medical intensive care unit (MICU). Then, we evaluated the performance of this system depending on the remaining time from the event occurrence.

Methods

Study Design and Subjects

We conducted a retrospective study with patients admitted to the MICU at the Asan Medical Center in Seoul, South Korea, between January 1, 2013, and July 31, 2015. For the development of a deep learning–based prediction model of cardiac arrest in critical ill patients of the MICU, we identified 759 distinct patients aged 18 years or older who stayed in the MICU for 1 day or more (Figure 1). All clinical data were extracted from our deidentified clinical data warehouse [14]. The extracted clinical data were categorized by patient demographics, diagnosis, medication, vital signs, medication, and inputs and outputs (Multimedia Appendix 1). As most clinical data were time series data, they were automatically recorded using patient monitoring devices. Vital signs and laboratory test data were collected at equally spaced intervals of 1 hour and 1 day, respectively.

The data were preprocessed in 2 ways. First, we selected features that patients have in common (see Feature selection in Figure 1). The features were divided into 6 broad categories (vitals, Sequential Organ Failure Assessment scores, laboratory results, demographics, diagnosis, and medications), and 45 common variables of 981 patients were selected. Second, we created a data pipeline to fit a gated recurrent unit (GRU) algorithm structure (see Data structure settings in Figure 2). Patient observations that had too many missing values were excluded to prevent biased model estimation. To filter out observations with many missing variables, 2 criteria were used: P_{id} and P_{ir} (Figure 1). P_{id} refers to the amount of missing observations after 2 tables (data pipeline fitted for the GRU structure and features) were joined. The threshold for P_{id} was set to 1000. P_{ir} is the ratio approach and is calculated as P_{id} divided by the total observations of patients in outcome variables. The threshold for P_{ir} was set to 20%. Furthermore, a threshold (τ) is defined to prevent large values of remaining hours in the uncensored group (ie, patients who did not experience cardiac arrest) from causing biased model estimation. For instance, if the uncensored group has many values greater than many hours (ie, 72 hours), then the likelihood of cardiac arrest may be estimated to be lower than that in the censored group. In addition, to accurately predict the occurrence of cardiac arrest, it is important to allow the model to learn the relationship between the time remaining just before cardiac arrest (eg, 1-3 hours) and the variables. Hence, only patients whose data were observed at least 1 hour before cardiac arrest were included in the study (Figure 2). Finally, the 2 tables (data structure that fits the GRU algorithm and features) were joined.

This study was approved by the institutional review board of the Asan Medical Center, Korea (institutional review board number 2015-1015). The need for informed consent was waived by the ethics committee as this study involved routinely collected medical data that were anonymously managed at all stages, including data cleaning and statistical analyses.

XSL•FO

RenderX

Figure 1. Data preprocessing flowchart. Obs: observation; TTE: time to event.

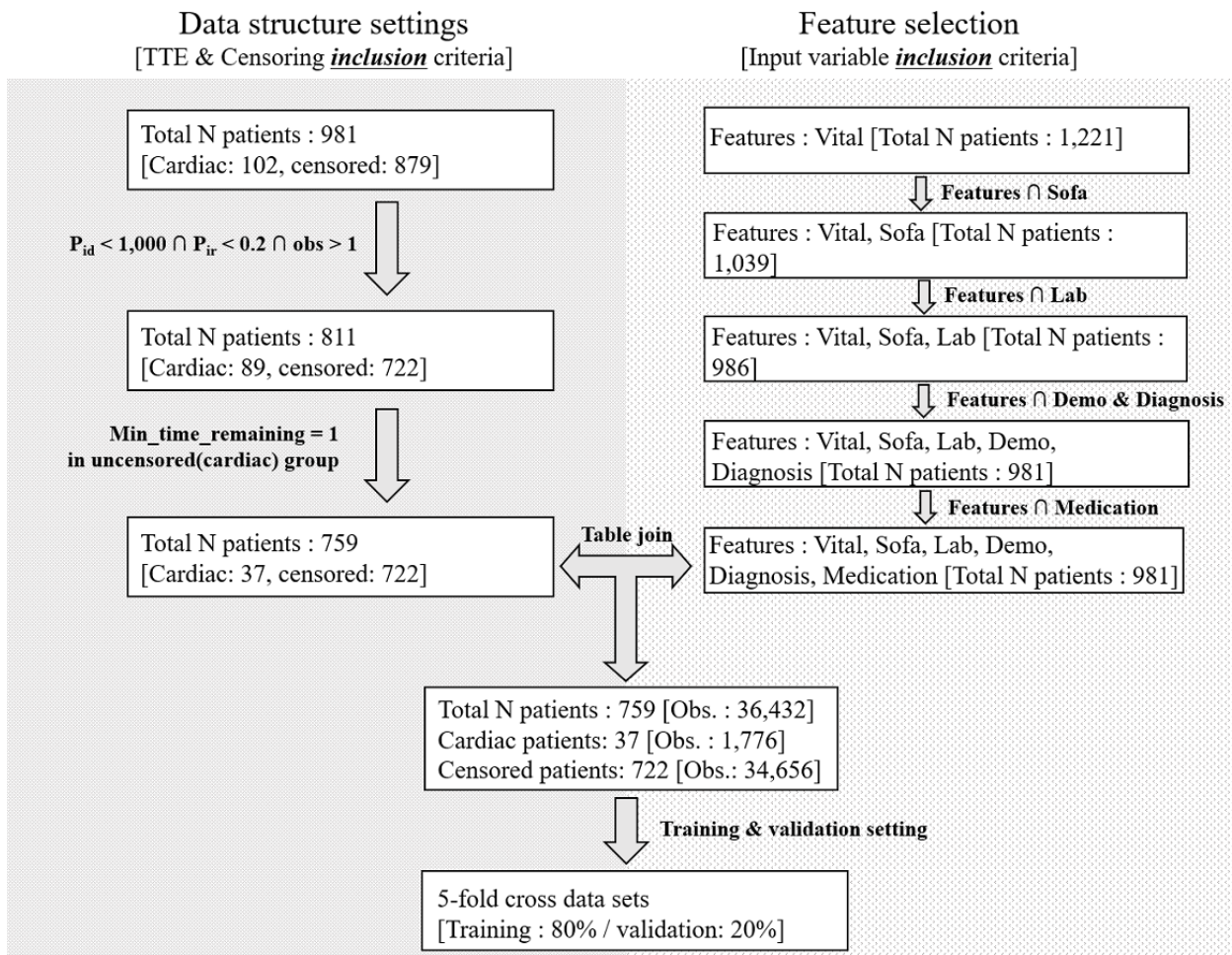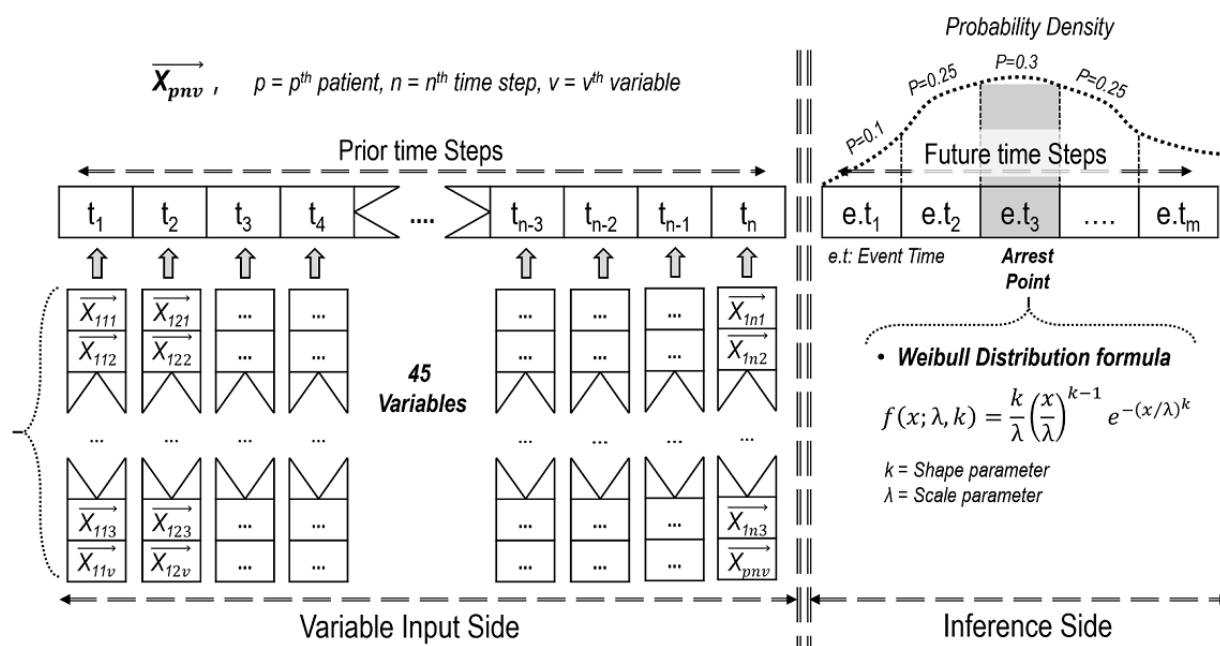**Figure 2.** Character-level gated recurrent unit structure combined with the Weibull distribution.

Development of Risk Prediction Model

The Weibull distribution, a continuous probability distribution, is a parametric model that can calculate the distribution form of survival time. Given the advantage of parametric models in survival analysis, the Weibull model is often used to estimate

failure rate over time [15,16]. The probability density function of a Weibull random variable is shown in Figure 3.

Figure 3. The probability density function of a Weibull random variable. k: shape parameter; λ: scale parameter; x: the quantity of time to failure.

$$f(x; \lambda, k) = \frac{k}{\lambda} \left(\frac{x}{\lambda} \right)^{k-1} e^{-(x/\lambda)^k}$$

The distribution consists of 2 parameters—the shape parameter k and the scale parameter λ. The variable x is the quantity of time to failure. Failure rates decrease over time when the shape parameter is less than 1. Conversely, failure rates increase over time when the shape parameter is greater than 1. The scale parameter is a location parameter that affects the width of the distribution. The larger the scale parameter value, the larger the width of the distribution.

Character-level gated recurrent unit (Char-GRU) is often used to predict the next token given a sequence of previous tokens [17]. In this research setting, each patient time history (45 historical variables for 1 patient) was preprocessed to become a set of overlapping time histories (see the list of 45 variables in Multimedia Appendix 1). Thus, the data structure consists of a 3-dimensional array: the number of observations × the number of time steps × the number of variables.

Figure 2 represents the structure of the input tensor o, n, v stands for o^{th} observation, n^{th} time step, v^{th} variable, respectively. For outcome variables, the algorithm estimates 2 parameters of the Weibull distribution by accumulating input variables for up to 48 hours (ie, 2 days). Thus, the 3-dimensional structure of tensor values at the input stage is changed to a 2-dimensional structure through the GRU network, and the parameters of shape and scale are estimated by the last point of the network.

The challenging point of learning the model was related to the censoring feature of the data structure (ie, 1=cardiac arrest occurred or 0=censored). The TTE of cardiac arrest is actually observed data, unlike in the case where the data point is not censored. However, the TTE of cardiac arrest is unknown when the data point is censored. In this study, τ was defined as a threshold value indicating the time to safety in the censored group. We set 72 hours as a threshold based on the median number of hours that patients stayed in the MICU.

Cost Function and Model Structure

The outcomes of Char-GRU with the Weibull distribution algorithm are 2 parameters corresponding to the shape and scale of the Weibull model. These 2 parameters enable calculation of likelihood through the function proposed in Figure 3. The goal is to obtain the optimal parameter of the Weibull distribution from the sequential patient data; therefore, the negative log of the likelihood is set to the objective function to maximize the likelihood of the training batch. With the objective function, the Char-GRU network parameters were optimized using the Adam stochastic optimization [18].

The total number of patients was 759, consisting of 37 cardiac arrest patients and 722 non–cardiac arrest patients. As 45

variables for 1 patient are repeatedly observed 48 times, the number of observations for cardiac arrest patients is 1776 and that for censored patients is 34,656. Thus, the shape of the input data delivered to the GRU algorithm is a 3-dimensional array of 36,432 × 48 × 45. If 45 variables of a timewise vector are missing, we apply a masking layer that skips the vector and the learning. It is then delivered sequentially to a layer of 50 GRU units. The activation function of this layer is an all hyperbolic tangent function. Next, a fully connected layer of 20 units is connected with the hyperbolic activation function. Finally, the 2 fully connected layers are connected to estimate the shape and scale, the parameters of the Weibull distribution with a softplus activation and exponential function, respectively.

Cross-Validation Procedure

A fivefold cross-validation test (training set: 80% and validation set: 20%) was implemented to determine the consistency of the model's accuracy. Overall, 5 models were learned independently from each dataset each time. Time-dependent receiver operating characteristic (ROC) analysis was performed from the aggregated set of the probability of an individual having cardiac arrest in each time step, which was estimated from 5 validation sets [19]. The mechanism for applying the result of the deep learning model to time-dependent ROC analysis is as follows: tensors, which were 3-dimensional in the input level, were passed through the GRU network to estimate the Weibull distribution by learning the shape and scale parameters. The Weibull distribution in this study setting indicates the likelihood (from 0%, low risk, to 100%, high risk) of a heart attack within the next hours from the current point in time. In other words, the time-dependent risk of cardiac arrest for each patient was estimated based on the deep learning model. Thus, the time-dependent risk probability of having a cardiac arrest was passed to the time-dependent ROC analysis.

Open Source Software

All procedures for data preprocessing and model implementation were conducted through the open source programming languages R and Python. To handle data in the format of a data frame (ie, data table) and an array, 2 open source libraries—Pandas and Numpy—were used. Char-GRU with a Weibull distribution was implemented in Keras (version 2.2.2), a wrapper library from Tensorflow (version 1.10.0), and a representative open source tool supporting the implementation of deep learning algorithms. Detailed concepts and mechanisms at the code level of this algorithm have been well documented in a previous study [16,20]. The ROC analysis was performed based on the R package pROC.

Results

Patient Characteristics

A total of 759 patients admitted in the ICU of the Asan Medical Center from March 2015 to March 2017 were enrolled in the study. Descriptive analysis was performed in 2 broad categories: demographics with 3 variables and diagnostic status with 8 variables. The Student t test was used for continuous variables such as age; the chi-square test was used for categorical variables such as Diab (ie, 1=diabetes or 0=no diabetes). The

results of a descriptive analysis are reported in Table 1. Both age and body weight in the cardiac arrest group were statistically higher than those in the non–cardiac arrest group (ie, censored group; *P*<.001). However, there were no statistical differences between the 2 groups for the remaining variables, including gender and underlying diseases.

Table 1. Descriptive statistics of the demographics and underlying diseases of the patients.

Variables	Cardiac group (n=37)	Censored group (n=722)	*P* value (test type)
Demographics			
Age (years), mean (SD)	62.509 (12.311)	60.526 (13.991)	<.001 (*t* test)
Weight (kg), mean (SD)	59.734 (13.166)	57.816 (13.435)	<.001 (*t* test)
Gender, n			**.15 (chi-square test)**
Male	28	451	
Female	9	271	
Diagnosis, n			
Hematologic malignancy			**.35 (chi-square test)**
Yes	8	105	
No	29	617	
Liver disease			**.43 (chi-square test)**
Yes	8	111	
No	29	611	
Oxygenation index			**.97 (chi-square test)**
Yes	2	28	
No	35	694	
Respiratory index			**.99 (chi-square test)**
Yes	0	10	
No	37	712	
Heart failure			**.84 (chi-square test)**
Yes	4	61	
No	33	661	
Diabetes			**.92 (chi-square test)**
Yes	12	218	
No	25	504	
Coronary Sinus Pressure			**.68 (chi-square test)**
Yes	0	18	
No	37	704	
Dialysis			**.85 (chi-square test)**
Yes	3	76	
No	34	646	

[a]The digits outside the parentheses mean *P* value.

Model Learning Results

As 5 cross-validation procedures were performed in this study, each of the 5 models was trained independently. Multimedia Appendix 1 shows the cost values over 1000 epochs for the training set and the validation set of each model. Although there were many points where the cost changed rapidly over the course of learning, the cost value decreased continuously over the epochs. From the first fold to the fifth fold, the cost value for the validation set was 0.217, 0.242, 0.271, 0.329, and 0.251 (Multimedia Appendix 1). As there is no clear criterion on when to stop training parameters during model training, we stopped model training with a heuristic approach based on the shape of the cost function. Specifically, the overall cost value hardly decreased when the epochs exceed 300. However, the cost values of both the training and validation sets suddenly increased after 500 epochs and settled after 700 epochs (Multimedia Appendix 1). To check if the cost instability was recaptured,

we trained the model up to 1000 epochs and then stopped the learning.

Time-Dependent Model Performance

Overall, 5 time-dependent areas under the curve (TAUCs) were calculated using the aggregated set of 5 validation sets (Figure 4). In this study, these TAUCs were segmented according to 5 time points. The TTE equated to 1, 8, 16, 24, 32, 40, or 48 hours remaining to cardiac arrest in the cardiac arrest group and to be censored in the non–cardiac arrest group. The number of cardiac arrest cases according to the 5 folds was 41 (6.63%), 41 (6.63%), 42 (6.77%), 42 (6.77%), and 42 (6.77%). We show the performance of the 5 folds through the median for all time points

(Multimedia Appendix 1). TAUCs for TTEs 1, 8, 16, 24, 32, 40, and 48 hours were calculated as 0.963, 0.942, 0.917, 0.875, 0.850, 0.842, and 0.761, respectively, indicating that model performance decreases as TTE increases. For all time points, the area under the curve of performance for the 5 folds is increased linearly (Multimedia Appendix 1). The average correlation coefficient between TAUC and time point in 5 folds was 0.910. Despite the smaller number of patients with cardiac arrest as compared with censored patients, the sensitivity ranged from 0.846 to 0.909. The specificity was generally high, ranging from 0.923 to 0.946, except for 48 hours, when there was a lack of prior information.

Figure 4. Results of time-dependent receiver operating characteristic analysis according to the fold change. AUC: area under the curve.
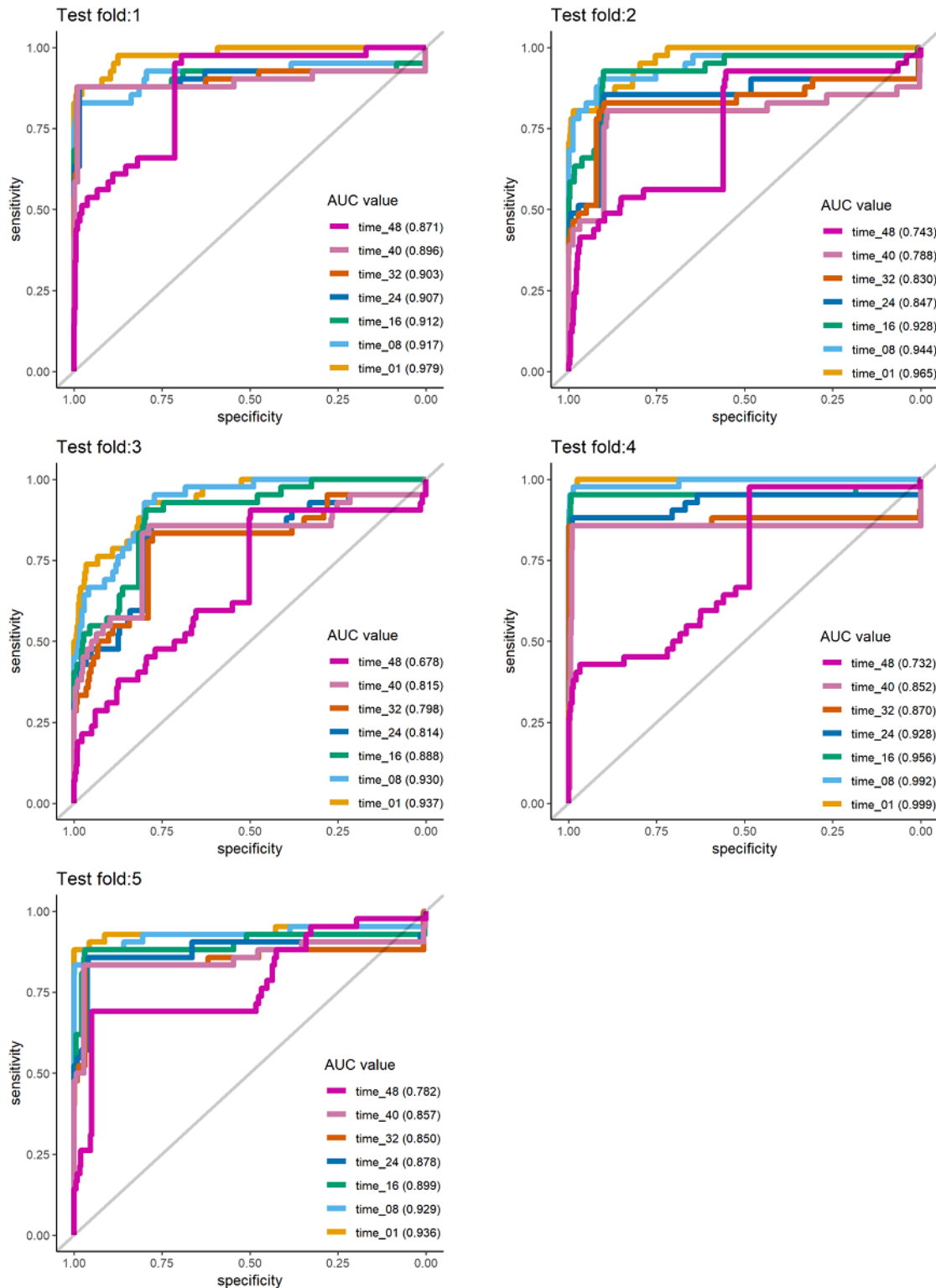
Risk Score Comparison

Figure 5 shows how the risk probability in the cardiac arrest and the non–cardiac arrest groups changed over time. For the group with a cardiac arrest, the risk probability increases as the time for a cardiac arrest approaches. Conversely, the group without a cardiac arrest did not show an increase in the risk of cardiac arrest when the data were closer to the censored time. From 48 to 16 hours before cardiac arrest, the interquartile range (IQR) values overlap for the cardiac and noncardiac groups. However, the IQR of the risk probability of the 2 groups is separated from 15 hours ago. The median risk probability value from 15 hours ago also differs more than 10 times, and the difference continues to increase 1 hour before cardiac arrest.

Figure 5. Risk probability comparison between cardiac arrest and the non–cardiac arrest groups. The x-axis represents the time point, and the y-axis represents the distribution of probability density values for cardiac arrest obtained for each patient corresponding to each time point.
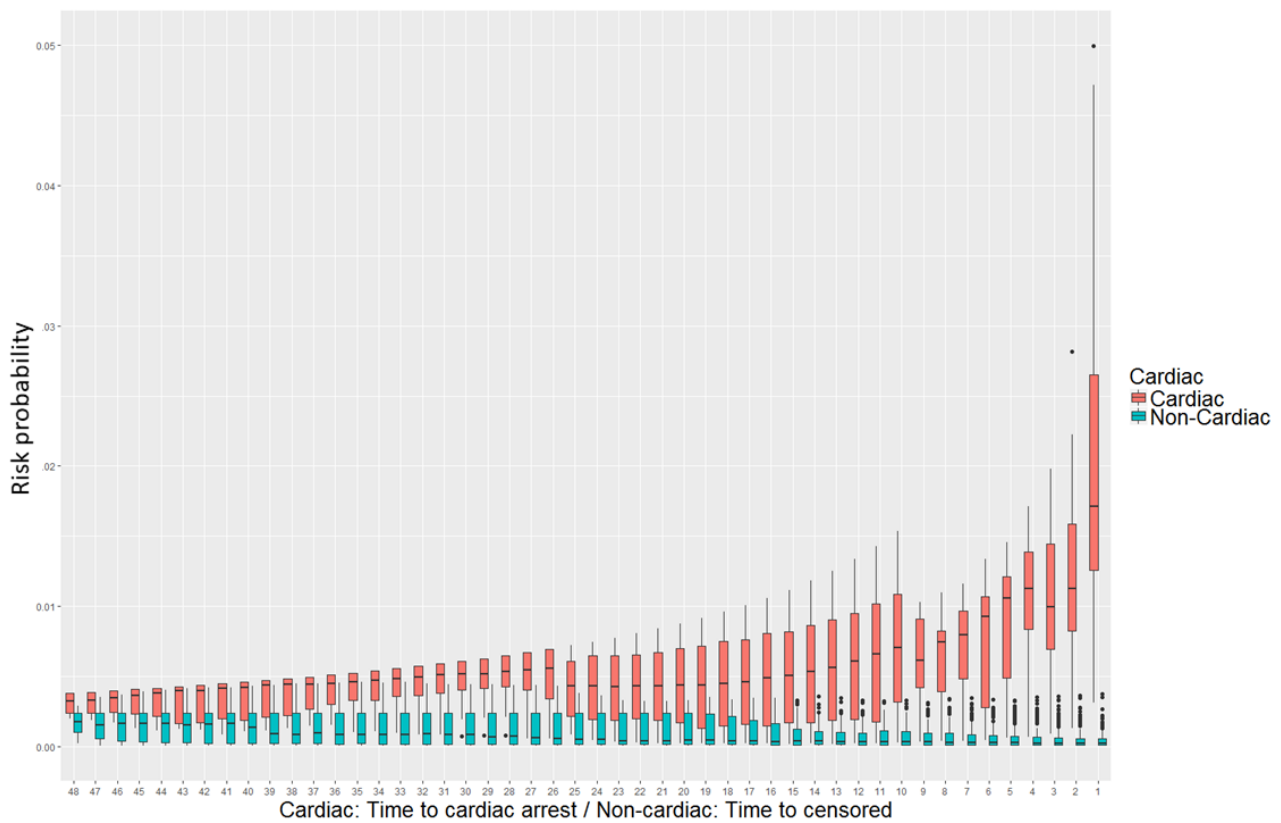
Predicted Cumulative Distribution Function at Time Point

An additional problem with arrest prediction is predicting when a cardiac arrest event will occur. A cumulative distribution function was derived through the shape and scale inferred by the model from each time point. Using the Weibull distribution parameter derived for the 48 time points, curves corresponding to cumulative distribution functions were drawn (A in Figure 6). Each line represents the probability of a cardiac arrest occurring from the start time point of the parameters until the remaining time. The closer the time to the cardiac arrest, the higher the beginning of the cumulative distribution function. This shows that even when the time point is far from cardiac arrest, a patient can be predicted to be high risk. Furthermore, the predicted time remaining until the patient has a cardiac arrest is presented (B in Figure 6). As the time approaches the cardiac arrest, the time remaining before the cardiac arrest occurs is estimated to be very small.

Figure 6. (A) Cumulative distribution function lines from the predicted time point to censoring time point for a patient with cardiac arrest at 48 time points; Each function line is color-coded. (B) Predicted hours remaining until a patient has cardiac arrest; the y-axis was limited to less than 25 hours for readability. pTime: predicted time.
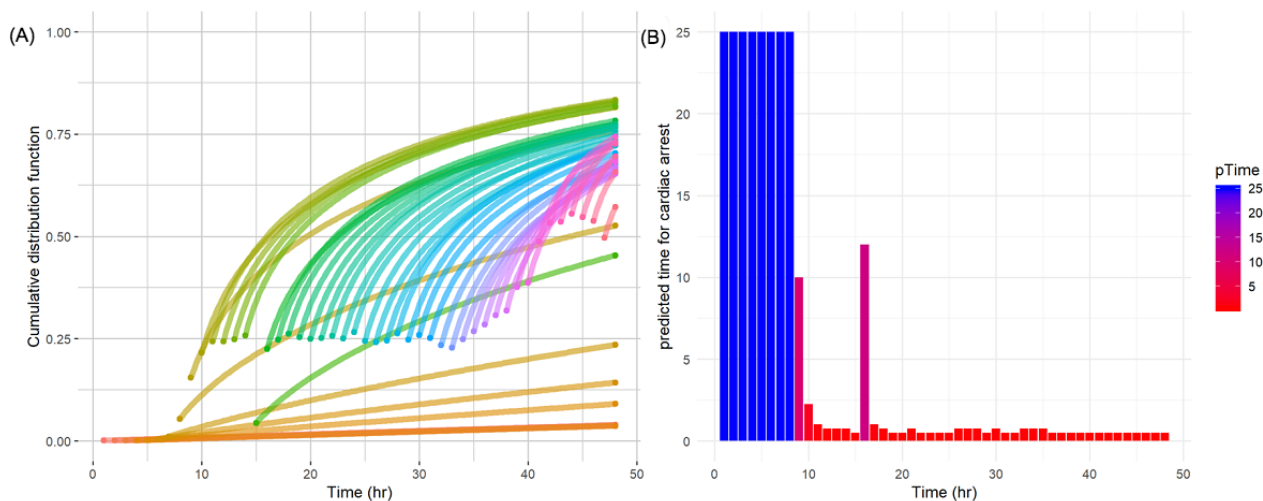
Conversely, the distribution of the cumulative distribution function of a certain patient without cardiac arrest shows that, at all time points, the probability does not increase over time (A in Figure 7). Likewise, the time remaining until the patient has cardiac arrest is predicted to be very high (ie, more than 25 hours) over the entire time (B in Figure 7).

Figure 7. (A) Cumulative distribution function lines from the predicted time point to censoring time point for a patient without cardiac arrest at 48 time points; Each function line is color-coded. (B) Predicted hours remaining until a patient has cardiac arrest; the y-axis was limited to less than 25 hours for readability. pTime: predicted time.

Discussion

Principal Findings

In this study, we developed the prediction model for cardiac arrest in critically ill patients through machine learning using electronic medical records. Besides vital sign, we used the underlying disease, laboratory data, medication, and organ failure as parameters to improve the accuracy of the prediction model. The TAUCs for TTE of 8, 16, and 24 hours were 0.942, 0.91, and 0.811, respectively, and the model performance decreased in accordance with increasing TTE.
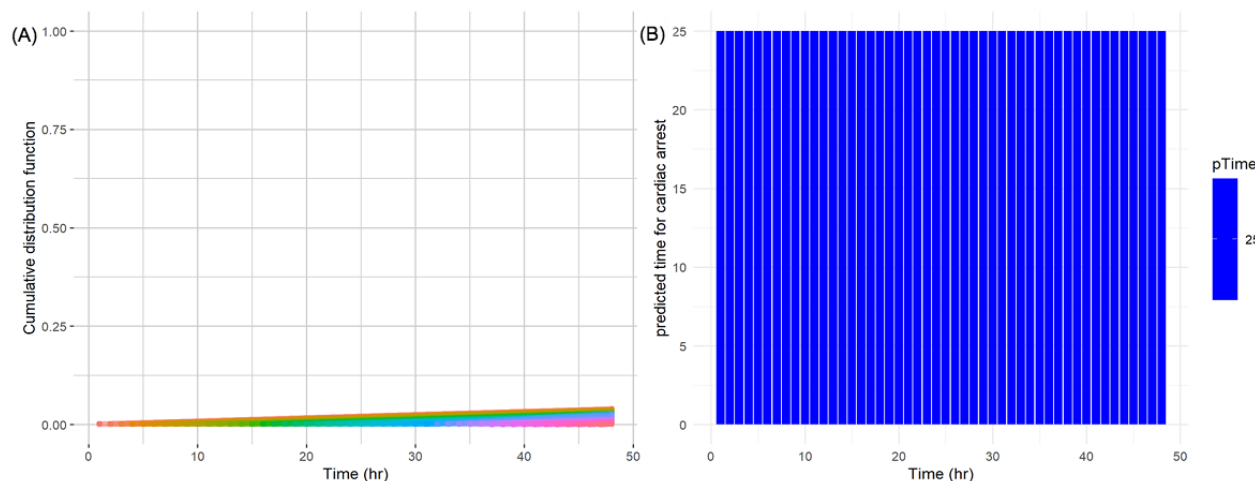
Informative Outcomes

In previous studies related to cardiac arrest predictions, modeling techniques that predict whether an event occurs within a predefined predicted time window have often been implemented [4,9,12]. Although these approaches are efficient in terms of model implementation, these approaches have limitations because it is impossible to forecast whether cardiac arrest occurs or not outside a defined window or when it will occur. To overcome the limitations, we attempted to combine the Weibull distribution estimation technique with a Char-GRU. This modeling approach provides information about the cardiac arrest risk probability over the future time. Therefore, it is possible to provide an answer to how many hours are left until cardiac arrest occurs without a predefined time threshold that may limit the information. Thus, it is obviously much more informative to predict cardiac arrest in clinical practice.

Early Warning in Real Time

The early recognition of cardiac arrest and its prompt correction are critical to reducing the mortality of critically ill patients. To decide clinically who is unstable or who is going to deteriorate, many intensivists often scrutinize the vital signs of intensive care patients, such as blood pressure, heart rate, respiratory rate, and peripheral capillary oxygen saturation [21,22]. However,

several studies have shown that these signs may not be effective in forecasting the risk early (ie, several hours before) [21,23]. This may be because of insufficient information on vital signs in predicting cardiac arrest in advance. However, considering that critical patient data are continuously generated in real time from numerous sources, including vital signs and information from organ support devices [12,13], the use of big data may provide models with sufficient information for the early prediction of cardiac arrest. Furthermore, the use of deep learning models, taking into account cumulative historical patterns of large clinical data, is expected to be very effective in predicting cardiac arrest in advance. In this regard, we implemented a deep learning–based model using a large dataset of 45 variables and found that the model could potentially be used for the early prediction of cardiac arrest.

Flexibility and Operational Reality

As cumulative and fluctuating effects of clinical variables over time can be reflected in deep learning algorithms, the use of long time series data to predict cardiac arrest is ideal. However, it is not appropriate to take no action until the patient has accumulated sufficient time series data. Waiting for sufficient time (ie, 48 hours) to accumulate patient time series data in clinical settings is undesirable for both patients and intensivists. Even if variables have not yet accumulated for a sufficient amount of time, a model should be available. In this situation, the Char-GRU structure allows the model to use the clinical variables. Specifically, the Char-GRU model can predict the risk of a patient's cardiac arrest using clinical variables accumulated up to the present time (ie, 3 hours after entering the ICU) [17].

Estimation Efficiency

The early detection of disease onset is challenging in terms of the configuration of deep learning algorithm structures and data pipelines, as there is no reference for *early* time. Previous studies have been limited in predicting the onset of an illness just 1

time step ahead (a week or a month before) [24,25]. However, disease onset could be forecasted at various time points. For instance, the probability of 3 hours before, 2 hours before, and 1 hours before the onset of a disease may be predicted at the same time. In this setting, 3 dimensions—the number of observations, time steps, and disease indication (1=onset or 0=nononset)—should be considered in determining cardiac arrest. However, it is inefficient to predict cardiac arrest onsets over all time steps because it is unknown how many hours earlier the onsets should be predicted. However, when estimating the occurrence of a disease by estimating a Weibull distribution, the dimension of the outcome variables (ie, shape and scale parameters) is 2 (observation ID and Weibull distribution parameters—shape and scale). In other words, this setting estimates the time remaining until the onset of the disease in the form of a continuous variable so that the time dimension (ie, the cardiac arrest onset over time remaining) is removed from the outcome variable area. Therefore, this method is much more efficient because it can significantly reduce the need for various experiments.

Limitations

This study has limitations that need to be addressed in further studies before applying Char-GRU with the Weibull distribution algorithm to clinics. In this study, rigorous validation was not performed while focusing on algorithm implementation using clinical data. As clinical data from only 1 medical institution were used, various additional validations are needed to generalize the results. To conduct rigorous validation, it is recommended to validate deep learning–based Weibull models using published data such as the modified early warning score [12].

Another limitation is the inability to fully control the reflection of certain effects in the collected data, which may affect the model results. For instance, data from a treated patient who is perceived to be in a very dangerous condition may cause a bias against the time series characteristics in the high-risk group [26]. In other words, the patient's data reflecting the time series characteristics of the non–cardiac arrest group would ultimately reflect the time series characteristics of the cardiac arrest group if the patient had not been treated [26,27]. In the previous studies, such data were just removed or corrected based on statistical methods [26,28]. Therefore, further research that validates this algorithm requires in-depth consideration of data selection and preprocessing.

Conclusions

The cardiac arrest survival rate in hospitals is about 24%, and even after survival, patients suffer from fatal problems such as brain damage [29,30]. However, because of the difficulty in forecasting cardiac arrest in advance, adequate prior interventions were rarely provided. We hope that the early prediction of cardiac arrest is linked to early intervention for the prevention of cardiac arrest. For that purpose, further research is essential to discuss how to operate deep learning models linked with a database and what forms of model outcomes should be provided to medical providers in practice.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary figures and tables.
[PDF File (Adobe PDF File), 696 KB - medinform_v8i3e16349_app1.pdf]

References

1. Peberdy MA, Kaye W, Ornato JP, Larkin GL, Nadkarni V, Mancini ME, et al. Cardiopulmonary resuscitation of adults in the hospital: a report of 14720 cardiac arrests from the National Registry of Cardiopulmonary Resuscitation. Resuscitation 2003 Sep;58(3):297-308. [doi: 10.1016/s0300-9572(03)00215-6] [Medline: 12969608]
2. Ehlenbach WJ, Barnato AE, Curtis JR, Kreuter W, Koepsell TD, Deyo RA, et al. Epidemiologic study of in-hospital cardiopulmonary resuscitation in the elderly. N Engl J Med 2009 Jul 2;361(1):22-31 [FREE Full text] [doi: 10.1056/NEJMoa0810245] [Medline: 19571280]
3. Carr BG, Goyal M, Band RA, Gaieski DF, Abella BS, Merchant RM, et al. A national analysis of the relationship between hospital factors and post-cardiac arrest mortality. Intensive Care Med 2009 Mar;35(3):505-511. [doi: 10.1007/s00134-008-1335-x] [Medline: 18936907]
4. Peberdy MA, Ornato JP, Larkin GL, Braithwaite RS, Kashner TM, Carey SM, National Registry of Cardiopulmonary Resuscitation Investigators. Survival from in-hospital cardiac arrest during nights and weekends. J Am Med Assoc 2008 Feb 20;299(7):785-792. [doi: 10.1001/jama.299.7.785] [Medline: 18285590]

5. Al-Dury N, Rawshani A, Israelsson J, Strömsöe A, Aune S, Agerström J, et al. Characteristics and outcome among 14,933 adult cases of in-hospital cardiac arrest: a nationwide study with the emphasis on gender and age. Am J Emerg Med 2017 Dec;35(12):1839-1844. [doi: 10.1016/j.ajem.2017.06.012] [Medline: 28624147]

6. Sandroni C, D'Arrigo S, Antonelli M. Rapid response systems: are they really effective? Crit Care 2015 Mar 16;19:104 [FREE Full text] [doi: 10.1186/s13054-015-0807-y] [Medline: 25887250]

7. Sun J. The Statistical Analysis of Interval-censored Failure Time Data. Cham: Springer; 2007.

8. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]

9. Zaremba W, Sutskever I, Vinyals O. arXiv e-Print archive. 2015. Recurrent Neural Network Regularization URL: https://arxiv.org/pdf/1409.2329.pdf [accessed 2020-01-14]

10. Nanayakkara S, Fogarty S, Tremeer M, Ross K, Richards B, Bergmeir C, et al. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: a retrospective international registry study. PLoS Med 2018 Nov;15(11):e1002709 [FREE Full text] [doi: 10.1371/journal.pmed.1002709] [Medline: 30500816]

11. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Crit Care Med 2016 Feb;44(2):368-374 [FREE Full text] [doi: 10.1097/CCM.0000000000001571] [Medline: 26771782]

12. Kwon J, Lee Y, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in-hospital cardiac arrest. J Am Heart Assoc 2018 Jun 26;7(13):pii: e008678 [FREE Full text] [doi: 10.1161/JAHA.118.008678] [Medline: 29945914]

13. Lee Y, Kwon J, Lee Y, Park H, Cho H, Park J. Deep learning in the medical domain: predicting cardiac arrest using deep learning. Acute Crit Care 2018 Aug;33(3):117-120 [FREE Full text] [doi: 10.4266/acc.2018.00290] [Medline: 31723874]

14. Shin SY, Park YR, Shin Y, Choi HJ, Park J, Lyu Y, et al. A de-identification method for bilingual clinical texts of various note types. J Korean Med Sci 2015 Jan;30(1):7-15 [FREE Full text] [doi: 10.3346/jkms.2015.30.1.7] [Medline: 25552878]

15. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. Hoboken, NJ: John Wiley & Sons; 2011.

16. Martinsson E. Focus on the objective · Egils blog. 2016. WTTE-RNN: Weibull Time to Event Recurrent Neural Network URL: https://ragulpr.github.io/assets/draft_master_thesis_martinsson_egil_wtte_rnn_2016.pdf [accessed 2020-01-14]

17. Hwang K, Sung W. Character-Level Language Modeling With Hierarchical Recurrent Neural Networks. In: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. 2017 Presented at: ICASSP'17; March 5-9, 2017; New Orleans, LA, USA. [doi: 10.1109/icassp.2017.7953252]

18. Kingma DP, Ba JL. arXiv e-Print archive. 2015. Adam: A Method for Stochastic Optimization URL: https://arxiv.org/pdf/1412.6980.pdf [accessed 2020-01-14]

19. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000 Jun;56(2):337-344. [doi: 10.1111/j.0006-341x.2000.00337.x] [Medline: 10877287]

20. Martinsson E. GitHub. 2019. WTTE-RNN a framework for churn and time to event prediction URL: https://github.com/ragulpr/wtte-rnn [accessed 2020-01-10]

21. Smith GB. Vital signs: vital for surviving in-hospital cardiac arrest? Resuscitation 2016 Jan;98:A3-A4. [doi: 10.1016/j.resuscitation.2015.10.010] [Medline: 26597106]

22. Andersen LW, Kim WY, Chase M, Berg KM, Mortensen SJ, Moskowitz A, American Heart Association's Get With the Guidelines – Resuscitation Investigators. The prevalence and significance of abnormal vital signs prior to in-hospital cardiac arrest. Resuscitation 2016 Jan;98:112-117 [FREE Full text] [doi: 10.1016/j.resuscitation.2015.08.016] [Medline: 26362486]

23. Brekke IJ, Puntervoll LH, Pedersen PB, Kellett J, Brabrand M. The value of vital sign trends in predicting and monitoring clinical deterioration: A systematic review. PLoS One 2019;14(1):e0210875 [FREE Full text] [doi: 10.1371/journal.pone.0210875] [Medline: 30645637]

24. Kam HJ, Kim HY. Learning representations for the early detection of sepsis with deep neural networks. Comput Biol Med 2017 Oct 1;89:248-255. [doi: 10.1016/j.compbiomed.2017.08.015] [Medline: 28843829]

25. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J Am Med Assoc 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]

26. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015 Aug 5;7(299):299ra122. [doi: 10.1126/scitranslmed.aab3719] [Medline: 26246167]

27. Florens JP, Heckman JJ, Meghir C, Vytlacil E. Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. Econometrica 2008;76(5):1191-1206. [doi: 10.3982/ECTA5317]

28. Pan W. A multiple imputation approach to Cox regression with interval-censored data. Biometrics 2000 Mar;56(1):199-203. [doi: 10.1111/j.0006-341x.2000.00199.x] [Medline: 10783796]

29. Javan SL, Sepehri MM, Javan ML, Khatibi T. An intelligent warning model for early prediction of cardiac arrest in sepsis patients. Comput Methods Programs Biomed 2019 Sep;178:47-58. [doi: 10.1016/j.cmpb.2019.06.010] [Medline: 31416562]

30. Daya MR, Schmicker RH, May S, Morrison LJ. The National Academies. 2015. Current Burden of Cardiac Arrest in the United States: Report From the Resuscitation Outcomes Consortium URL: http://www.nationalacademies.org/hmd/~/media/Files/Report%20Files/2015/ROC.pdf [accessed 2020-01-14]

Abbreviations

Char-GRU: character-level gated recurrent unit
GRU: gated recurrent unit
ICU: intensive care unit
IHCA: in-hospital cardiac arrest
IQR: interquartile range
MICU: medical intensive care unit
ROC: receiver operating characteristic
TAUC: time-dependent area under the curve
TTE: time to event

XSL•FO
RenderX

Original Paper

Predicting Metabolic Syndrome With Machine Learning Models Using a Decision Tree Algorithm: Retrospective Cohort Study

Cheng-Sheng Yu[1,2*], PhD; Yu-Jiun Lin[1,2*], MD; Chang-Hsien Lin[1,2], MD; Sen-Te Wang[1,2], MD, PhD; Shiyng-Yu Lin[1,2], MD; Sanders H Lin[1], MD; Jenny L Wu[1,2], BSc; Shy-Shin Chang[1,2], MD, PhD

[1]Department of Family Medicine, Taipei Medical University Hospital, Taipei, Taiwan

[2]Department of Family Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan

[*]these authors contributed equally

Corresponding Author:
Shy-Shin Chang, MD, PhD
Department of Family Medicine
School of Medicine, College of Medicine
Taipei Medical University
No. 250, Wuxing St
Taipei
Taiwan
Phone: 886 2 23565926
Email: sschang0529@gmail.com

Abstract

Background: Metabolic syndrome is a cluster of disorders that significantly influence the development and deterioration of numerous diseases. FibroScan is an ultrasound device that was recently shown to predict metabolic syndrome with moderate accuracy. However, previous research regarding prediction of metabolic syndrome in subjects examined with FibroScan has been mainly based on conventional statistical models. Alternatively, machine learning, whereby a computer algorithm learns from prior experience, has better predictive performance over conventional statistical modeling.

Objective: We aimed to evaluate the accuracy of different decision tree machine learning algorithms to predict the state of metabolic syndrome in self-paid health examination subjects who were examined with FibroScan.

Methods: Multivariate logistic regression was conducted for every known risk factor of metabolic syndrome. Principal components analysis was used to visualize the distribution of metabolic syndrome patients. We further applied various statistical machine learning techniques to visualize and investigate the pattern and relationship between metabolic syndrome and several risk variables.

Results: Obesity, serum glutamic-oxalocetic transaminase, serum glutamic pyruvic transaminase, controlled attenuation parameter score, and glycated hemoglobin emerged as significant risk factors in multivariate logistic regression. The area under the receiver operating characteristic curve values for classification and regression trees and for the random forest were 0.831 and 0.904, respectively.

Conclusions: Machine learning technology facilitates the identification of metabolic syndrome in self-paid health examination subjects with high accuracy.

Introduction

Metabolic syndrome is a cluster of disorders, including insulin resistance or hyperglycemia, visceral adiposity (identified by a large waistline or overweight), atherogenic dyslipidemia (eg, raised triglycerides or reduced high-density lipoprotein [LDL]), and endothelial dysfunction (characterized by elevated blood pressure) [1]. Metabolic syndrome has significant impacts on the development and deterioration of several diseases and is a critical predictor of cardiovascular diseases [2,3]. Numerous modifiable risk factors and practical intervention strategies regarding metabolic syndrome have been proposed [4-14]. Identifying high-risk patients to prevent the incidence and

XSL•FO
RenderX

deterioration of metabolic dysregulation and relevant diseases is therefore vital.

A recent study showed that nonalcoholic fatty liver disease (NAFLD) is closely correlated to metabolic syndrome. Patients with metabolic syndrome frequently show an increase in fat accumulation in the liver (steatosis) and hepatic insulin resistance [15]. Nevertheless, the gold-standard method for NAFLD diagnosis is liver biopsy, which is a highly invasive procedure for patients. Several reports have demonstrated that ultrasound using FibroScan, also known as transient elastometry, can accurately assess the staging of NAFLD in a noninvasive manner with comparable results to liver biopsy [16,17]. The new models of FibroScan (marketed after 2013) can assess the staging of NAFLD using a liver stiffness score (E score) and a liver steatosis score (controlled attenuation parameter [CAP] score). Interestingly, the CAP score alone was found to be a useful indicator of the presence and severity of metabolic syndrome [18]. Using traditional statistical modeling, we previously validated this finding, confirming that the CAP score alone can be used to detect metabolic syndrome with moderate accuracy (area under the receiver operating characteristic curve [ROC] of 0.79), and the accuracy was improved to 0.88 when combined with other biomarkers [19].

Machine learning, whereby a computer algorithm learns from prior experience, was recently shown to have better performance over traditional statistical modeling approaches [20-22]. Various supervised machine learning models based on decision trees have been successfully applied to medical data [23-29] for accurate prediction of a wide range of clinical conditions such as myocardial infarction [30], atrial fibrillation [31], trauma [32], breast cancer [33-35], Alzheimer disease [36-38], cardiac surgery [27,39], and others [27,28,40-42]. However, each decision tree machine learning algorithm has its own strength and weakness. Therefore, comparing different decision tree algorithms can reduce the bias in the results and provide a more robust outcome. Accordingly, the aim of this study was to determine whether decision tree algorithms can predict the state of metabolic syndrome among self-paid health examination subjects who were examined with FibroScan.

Methods

Study Design

This was a single-center retrospective cohort study. The cohort comprised self-paid health examination subjects at the Health Management Center of Taipei Medical University Hospital who were examined with FibroScan from September 2015 to December 2018.

Setting

The electronic healthcare records of subjects examined with FibroScan were reviewed at Taipei Medical University Hospital, which is a private, tertiary-care, 800-bed teaching hospital in Taiwan. The Institutional Review Board of Taipei Medical University Hospital approved the study design for data collection (TMU-JIRB No.: N201903080) in accordance with the original and amended Declaration of Helsinki. The requirement for informed consent was waived owing to the retrospective nature of the study.

Population and Data Collection

The study included all Taiwanese adult patients aged >18 years who had undergone a self-paid health examination comprising an abdominal transient elastography inspection using FibroScan 502 Touch (Echosens, Paris, France). Individuals who underwent FibroScan examination on physician's orders were excluded. The routine protocols of the Health Management Center were applied to all participants. The subjects were first interviewed by thoroughly trained personnel who verified the correctness of self-completed questionnaires on demographics, existing medical conditions, and medication use. In addition, the personnel confirmed adherence to health examination prerequisites (eg, overnight fasting for at least 8 hours) for the package chosen by the study participant. Those found to have not fulfilled the necessary prerequisites were advised to reschedule their appointment. Anthropometrics, including weight, height, waist circumference, and arterial pressure, were measured. Instruments were regularly calibrated per the manufacturer's specifications. According to the chosen package, the required samples of blood, urine, and specimens were collected for laboratory tests. Regular laboratory test items included alpha-fetoprotein, glycated hemoglobin (HbA_{1c}), serum glutamic oxaloacetic transaminase (GOT), serum glutamic pyruvic transaminase (GPT), uric acid, creatinine, blood urine nitrogen, red blood cell count, hemoglobin, hematocrit, mean corpuscular hemoglobin, mean corpuscular volume, mean corpuscular hemoglobin concentration, platelet count, white blood cell count, percentage of neutrophils, lymphocytes, monocytes, eosinophils and basophils, total protein, albumin, globulin, albumin/globulin ratio, total bilirubin, direct bilirubin, alkaline phosphatase, gamma-glutamyl transpeptidase (γ-GT), total cholesterol, LDL cholesterol, high-density lipoprotein (HDL) cholesterol, LDL/HDL ratio, triglycerides, fasting blood sugar, and thyroid-stimulating hormone. The estimated glomerular filtration rate (eGFR) was calculated using equations for the Modification of Diet in Renal Disease for Chinese patients [43], with chronic kidney disease (CKD) measured as follows: $175 \times (\text{Scr})^{-1.234} \times (\text{Age})^{-0.179} \times 0.79$ (if female). CKD was defined as an eGFR of <60 mL/min per 1.73 m^2 of body surface (mL/min/1.73 m^2), according to the definition from the Kidney Disease Outcomes Quality Initiative for CKD ≥ stage 3 [39,44]. Body mass index categories were defined as follows: obesity, ≥27 kg/m^2; overweight, 24-26.9 kg/m^2; and normal weight, <23.9 kg/m^2, according to the ranges established for Asian populations by the Ministry of Health and Welfare of Taiwan [45].

Outcome

According to the National Cholesterol Education Program Adult Treatment Panel III definition of metabolic syndrome consensus, metabolic syndrome was identified if at least three out of the following five symptoms were present: large waistline (80 cm for women and 90 cm for men), high triglycerides (150 mg/dL) or use of medication to control triglycerides, reduced HDL levels (<50 mg/dL for women and <40 mg/dL for men) or use

of medication to control HDL, elevated blood pressure (systolic blood pressure 130 mmHg or diastolic blood pressure 85 mmHg) or use of relevant medication to control blood pressure, and increased fasting blood sugar (100 mg/dL) or use of relevant medication to control blood sugar. The classification of cutoff points was adopted from the National Cholesterol Education Program Adult Treatment Panel III definition consensus with ethnicity-specific cutoff points for waist circumference [46,47] and an equality principle for the five disorders.

FibroScan

FibroScan is a noninvasive device that assesses the hardness of the liver using ultrasound-based elastography. Liver hardness is evaluated by measuring the velocity of a vibration wave, which is determined by measuring the time that the vibration wave takes to travel to a particular depth inside the liver from the skin (Figure 1). For each FibroScan inspection, two scores are reported: the CAP score and E score. The dashboard of FibroScan provides a CAP score only when an E score derived from identical signals is validated as successfully computed; higher E scores indicate higher transmission velocity and liver stiffness levels, and higher CAP scores indicate faster wave amplitude attenuation and higher levels of liver steatosis. Notably, the adoption of probe size (medium or extra large) is based on the recommendation of the instrumental autodetection function.

Figure 1. Illustration of the FibroScan device: liver diagnosis by ultrasound-based elastography. FibroScan measures fibrosis and steatosis in the liver. Measurements are performed by scanning the right liver lobe through the right intercostal space. The fibrosis result is measured in kiloPascals (kPa), and is normally between 2.5 and 6 kPa; the highest possible result is 75 kPa. Fibrosis score: F0 to F1, no liver scarring or mild liver scarring; F2, moderate liver scarring; F3, severe liver scarring; F4, advanced liver scarring (cirrhosis). The steatosis result is measured in decibels per meter (dB/m), and is normally between 100 and 400 dB/m. Steatosis can be graded from S0 to S3, corresponding to the severity of fatty liver from "0-10%" to "67% or more".
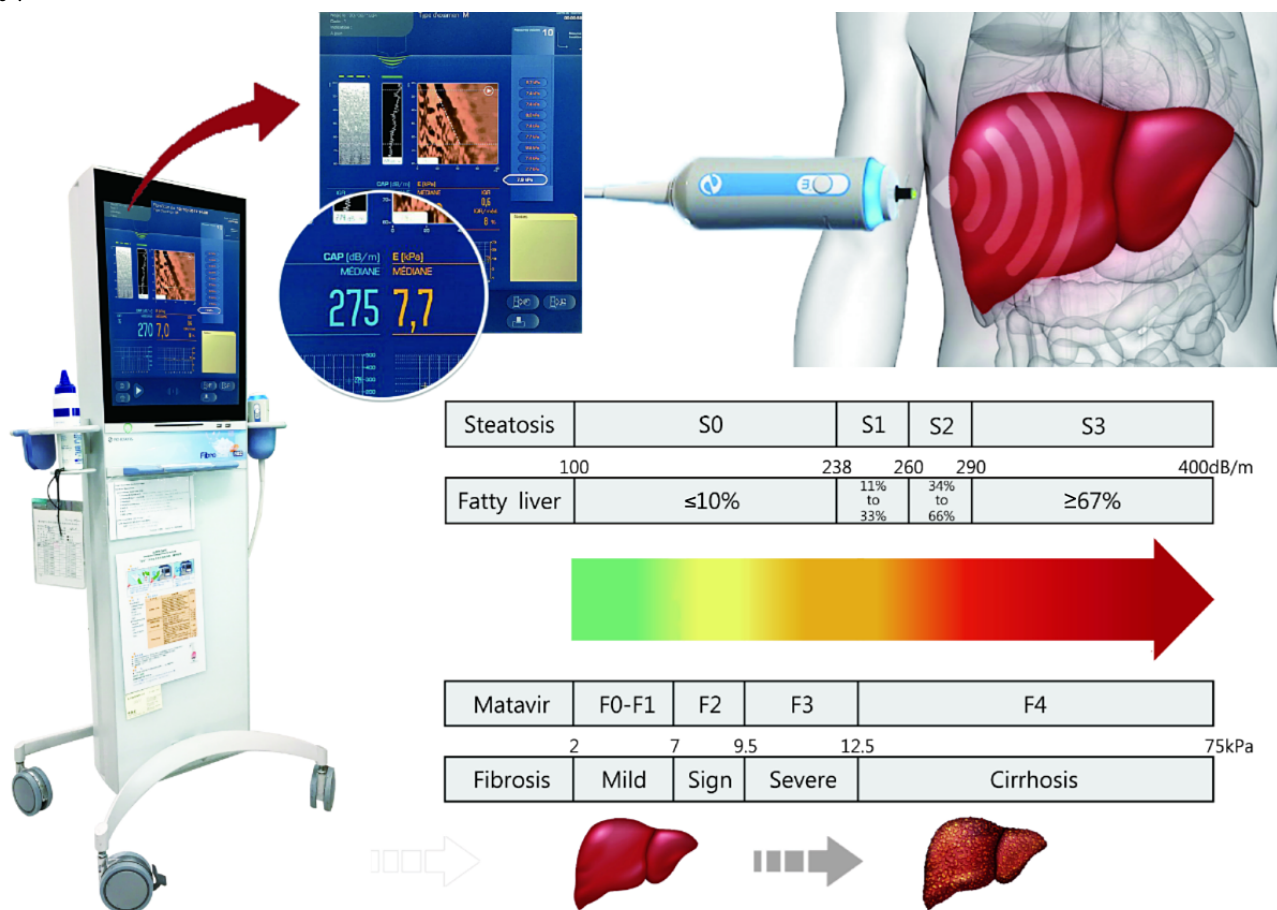
Machine Learning Technique

Overview

A decision tree is a widely used effective nonparametric machine learning modeling technique for regression or classification purposes. To obtain solutions, a decision tree makes a sequential, hierarchical decision regarding outcome variables based on the predictor [48].

Classification and Regression Trees

Classification and regression trees (CART), the typical tree-based models, explore the structure of data, while evolving to visualize decision rules for predicting a categorical (classification tree) or continuous (regression tree) outcome [49]. The decision at each internal node is assessed by information gain or entropy to compare the value of attributes in the data from the root to each of the leaves. CART was generated through the "rpart" package in R [50].

XSL•FO
RenderX

C5.0

C5.0 is derived from C4.5 and ID3 with improvements according to the disadvantages of the predecessor trees. The "C50" package was applied to implement the C5.0 tree [51,52].

Chi-Square Automatic Interaction Detection

Chi-square automatic interaction detection (CHAID) is a specific decision tree using adjusted significance testing (Bonferroni testing) for prediction. An algorithm for recursive partitioning is implemented by maximizing the significance of a chi-square statistic for crosstabulations between the categorical dependent variable and the categorical predictors at each partition. Moreover, CHAID can create nonbinary trees since nominal, ordinal, and continuous data are used. CHAID tree is available from the "CHAID" package in R [53].

Conditional Interference Trees

Conditional inference trees (ctrees) embed tree-structured regression models into a well-defined theory of conditional inference procedures. They use a significance test procedure to select variables instead of selecting the variable that maximizes any information measure. In addition, ctree is applicable to all types of regression issues, including nominal, ordinal, numeric, censored, and multivariate response variables, as well as arbitrary measurement scales of covariates. A flexible and extensible computational tool in the "partykit" package of R is suitable for fitting and visualizing ctrees [54,55].

Evolutionary Learning of Globally Optimal Trees

Evolutionary learning of globally optimal trees (evtree) describes recursive partitioning methods that create models using a forward stepwise search. An evtree is learned using an evolutionary algorithm. Notably, a set of trees is initialized with random split rules in the root nodes. Mutation and crossover operators are then applied to modify the tree's structure and tests that are applied in the internal nodes. After each modification step, a survivor selection mechanism identifies the best candidate models for the next iteration, terminating when the quality of the best trees ceases to improve. The "evtree" package in R applies an evolutionary algorithm for learning globally optimal classification and regression trees [56].

Generalized Linear Model Trees

Generalized linear model trees (glmtree) involve model-based recursive partitioning based on generalized linear models. They are convenient for fitting model-based recursive partitions using "mob" functions in R. A glmtree internally sets up a model-fit function for mob using the negative log likelihood as the objective function. It is also implemented by the "partykit" package in R [54,57].

Random Forest

Random decision forests are an ensemble learning method for classification, regression or other applications based on decision tree structures at the time of training. The idea of random forest is to create multiple decision trees (CART) and then combine the output generated by each of the decision trees. In the decision tree algorithm, the Gini index is a measure of the frequency of a randomly chosen element from the set that would be incorrectly labeled. The Gini index is calculated by subtracting the sum of the squared probabilities of each class from 1. This approach removes the bias that a decision tree model might introduce to a system while considerably improving the predictive power. In addition, random forests can be used to rank the importance of variables in a regression or classification problem in a natural manner, which can be conducted in the R package "randomForest" [58].

Statistical Analysis

Basic Statistics

Statistical analysis was conducted using R (version 3.6.1; R Foundation for Statistical Computing, Vienna, Austria) or SPSS (version 17.0; SPSS Inc, Chicago, IL, USA) software.

Categorical variables were tested using the chi-square test or Fisher exact test. The nonparametric Mann-Whitney U-test was applied to determine differences in the median of continuous variables between the two groups. Multivariate logistic regression was employed to assess the significance of clinical data, and the variance inflation factor was also used to check for multicollinearity. $P<.05$ was considered statistically significant [59,60].

Principal Components Analysis

High-dimension data were processed by principal components analysis (PCA), using an orthogonal transformation to convert a set of observations of correlated variables to provide a two-dimensional or three-dimensional visualization with its leading principal components.

Receiver Operating Characteristic Curve

ROC curves were used to illustrate the diagnostic ability of classification trees in the machine learning methodology. The area under the ROC curve (AUC), true positive rate (also called sensitivity or recall), and false positive rate (specificity) are represented in a graphical plot [61]. The F1 score, which constitutes the harmonic mean of precision and recall, was also evaluated. The F1 score has been widely used in the natural language processing literature and for machine learning [62,63].

Missing Values

Data with missing values were statistically regulated by the expectation-maximization algorithm, which is an iterative procedure that preserves the relationship with other variables. Only 9 factors had missing values, and most of them accounted for less than 5% of the sample size. Direct bilirubin, which had the largest proportion of missing values (298/1333, 22.36%), was at high risk of multicollinearity; thus, it was not a crucial element in the model [64].

Comparison of Decision Trees

To compare the performance of the aforementioned decision trees, the same setting for the training set and testing set was considered. In addition, the boundary for each tree's height was limited between 4 and 5 instead of pruning each decision tree according to its own criteria. Finally, outcomes from each decision tree were summarized to investigate common and reliable results supporting the conclusions.

Results

After data cleaning, a total of 1333 individuals undergoing self-paid annual health examination were enrolled in this study. The baseline characteristics of the 193 patients diagnosed with metabolic syndrome and 1140 participants without metabolic syndrome are compared in Table 1. All categorical elements were found to be extremely significant in the chi-square test. Among the continuous variables, most of the risk factors were highly significantly different between groups in the nonparametric test, although not enough evidence was found for age, alpha-fetoprotein, bilirubin, and thyroid stimulating hormone to support rejection of the null hypothesis. However, large samples and *P* value problems had to be considered owing to the numerous and complex data in this analysis [65]. The foremost factors were then validated by a series of additional evaluations.

The visualization of the two groups was achieved by PCA with the advantage of dimensionality reduction (Figure 2). All factors with significant outcomes by the tests mentioned above and shown in Table 1 depicted an intermixing view because the two groups of patients overlapped (Figure 2), with weak explanatory power for the first two principal components PC1 and PC2 at 27.7% and 13.2%, respectively. A variety of views in three-dimensional PCA plots are also displayed in Multimedia Appendix 1. The two groups could not be clearly discriminated, even if the coordinates were rotated in the three-dimensional graph.

Table 1. Descriptive statistics and testing of risk factors in health examination data with potential metabolic syndrome as the dependent variable.

Risk factors	No metabolic syndrome (N=1140)	Metabolic syndrome (N=193)	P value
Chronic kidney disease[a], n (%)			
Stage 1	585 (51.32)	71 (36.8)	.001
Stage 2	530 (46.49)	115 (59.6)	
Stage 3	24 (2.11)	6 (3.1)	
Stage 4	1 (0.09)	1 (0.5)	
Sex, n (%)			
female	564 (49.47)	44 (22.8)	<.001
male	576 (50.53)	149 (77.2)	
Obesity[a], n (%)			
underweight	49 (4.30)	0 (0.0)	<.001
normal weight	667 (58.51)	22 (11.4)	
overweight I	293 (25.70)	65 (33.7)	
overweight II	131 (11.49)	106 (54.9)	
Age (years), median (IQR)	44 (38-50)	45 (40-51)	.12
Hepatic indices, median (IQR)	4.6 (4.4-4.8)	4.7 (4.5-4.8)	<.001
Albumin (g/dL)	2.26 (1.637-3.12)	2.43 (1.71-3.14)	.15
AFP[b] (ng/mL)	58 (48-69)	62 (54-75)	<.001
ALKp[c] (IU/L)	20 (17-24)	24 (19-31)	<.001
GOT[d] (IU/L)	19 (13-27)	33 (22-51)	<.001
GPT[e] (IU/L)	0.6 (0.5-0.8)	0.7 (0.5-0.9)	.08
Total bilirubin (mg/dL)	0.2 (0.2-0.3)	0.2087 (0.2-0.3)	.66
Direct bilirubin (mg/dL)	16 (12-25)	29 (20-45)	<.001
γ-GT[f] (U/L)	239 (209-274)	311 (271-340)	<.001
CAP[g] score (dB/m)	4 (3.4-4.8)	4.9 (4.3-5.8)	<.001
E score (kPa)	12 (10-15)	13 (11-15)	.03
Nephritic indices, median (IQR)			
BUN[h] (mg/dL)			
Creatinine (mg/dL)	0.8 (0.6-0.9)	0.9 (0.7-1.0)	<.001
MDRD[i]	91.07 (81.3-105.17)	86.23 (75.05-98.82)	<.001
UA[j] (mg/dL)	5.2 (4.3-6.5)	6.3 (5.5-7.3)	<.001
Blood lipid and thyroid markers, median (IQR)			
Cholesterol (mg/dL)	187 (165-208)	194 (165-220)	.03
LDL[k] (mg/dL)	121 (101-142)	136 (106-158)	<.001
HbA$_{1c}$[l] (%)	5.4 (5.2-5.7)	5.7 (5.4-6.1)	<.001
TSH[m] (μIU/mL)	1.93 (1.30-2.52)	1.89 (1.38-2.51)	.83

[a]Progressive discrete variables.

[b]AFP: alpha-fetoprotein.

[c]ALKp: alkaline phosphatase.

[d]GOT: glutamic-oxalocetic transaminase.

[e]GPT: glutamic-pyruvic transaminase.

[f]γ-GT: gamma-glutamyl transpeptidase.

[g]CAP: controlled attenuation parameter.

[h]BUN: blood urea nitrogen.

[i]MDRD: Modification of Diet in Renal Disease.

[j]UA: uric acid.

[k]LDL: low-density lipoprotein cholesterol.

[l]HbA$_{1c}$: glycated hemoglobin.

[m]TSH: thyroid-stimulating hormone.

Figure 2. Principal components analysis (PCA) of metabolic and nonmetabolic groups by two-dimensional and three-dimensional visualization. (a) PCA with 95% CI shown as ellipses for all risk factors in Table 1. (b) PCA with 95% CI shown as ellipses for specific major variables in Table 3. Light blue nodes represent people diagnosed with metabolic syndrome, and dark blue nodes represent people without metabolic syndrome. PC1 explains the most variability among the samples, followed by PC2, PC3, and so on. (c) Three-dimensional PCA including PC1, PC2, and PC3 for all risk factor data from Table 1. (d) Three-dimensional PCA for specific major variables from Table 3.
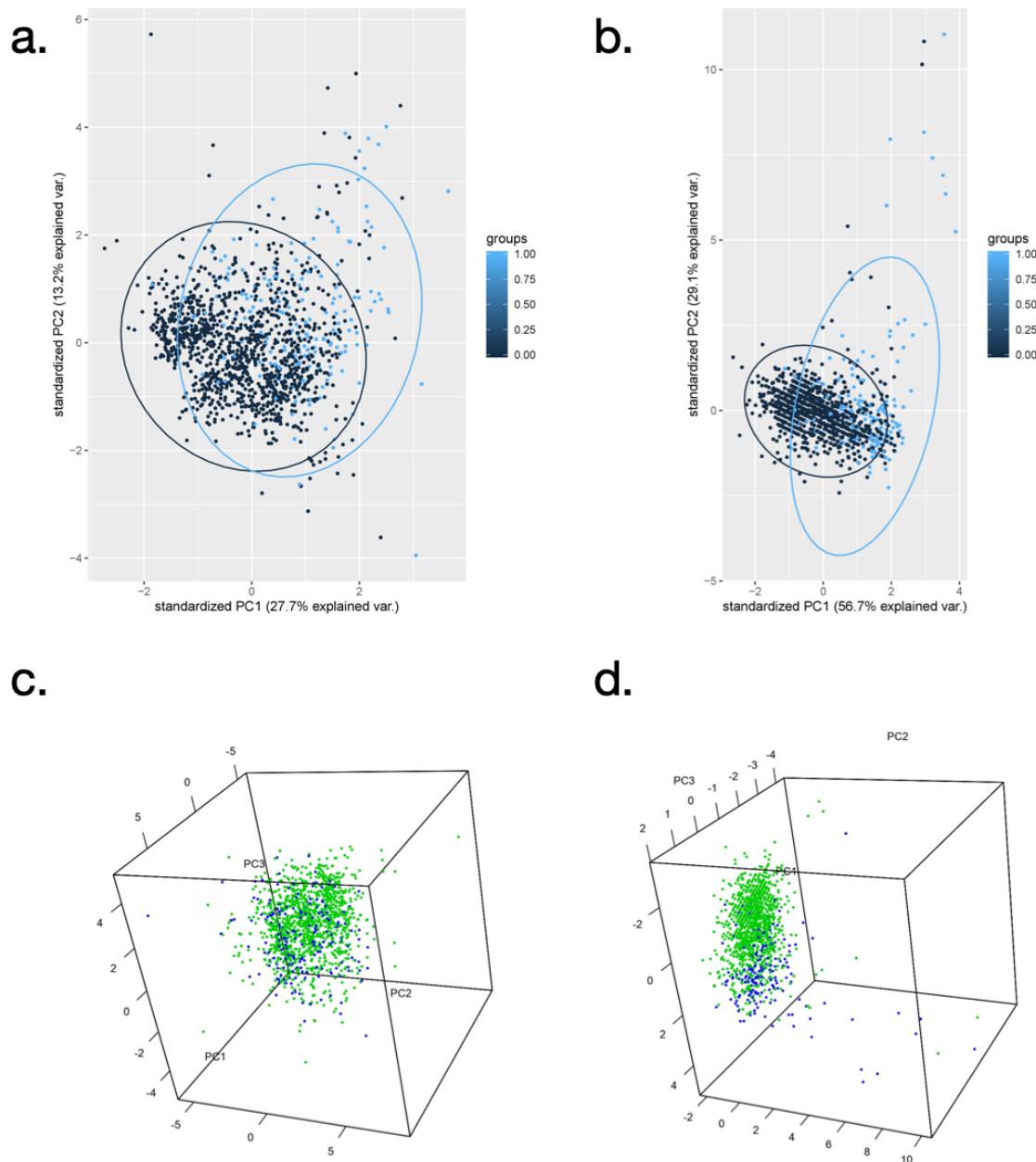
Next, we applied multivariate logistic regression to assess factors influencing metabolic syndrome. As shown in Table 2, the number of significant variables was reduced to 3, and included obesity, CAP score, and HbA$_{1c}$. Among these, HbA$_{1c}$ was obtained from blood tests, whereas information on obesity and the CAP score was obtained through noninvasive means.

Notably, obesity and HbA_{1c} exhibited high odds ratios, exceeding 2. In addition, the variance inflation factor was taken into account for multicollinearity.

Table 2. Multivariate logistic regression analysis of risk factors related to metabolic syndrome.

Factor	Odds ratio[a] (95% CI)	VIF[b]	ΔVIF[c]	P value
Sex, Male/Female	0.742 (0.335-1.641)	3.590	1.630	.99
Age, years	1.025 (0.999-1.051)	1.622	1.506	.15
Obesity	2.915 (2.175-3.907)	1.429	1.406	<.001
Albumin, g/dL	1.866 (0.821-4.239)	1.218	1.176	.12
AFP[d], ng/mL	1.045 (0.915-1.193)	1.162	1.153	.48
ALKp[e], IU/L	0.995 (0.983-1.007)	1.158	1.140	.52
GOT[f], IU/L	0.959 (0.923-0.997)	*7.226*	-	-
GPT[g], IU/L	1.023 (1.003-1.045)	*7.747*	*1.555*	.51
Total bilirubin, mg/dL	2.599 (0.562-12.015)	*8.334*	*1.246*	.39
Direct bilirubin, mg/dL	0.011 (0-2.507)	*8.413*	-	-
γ-GT[h], U/L	1.002 (0.994-1.009)	1.414	1.379	.77
CAP[i] score, dB/m	1.011 (1.007-1.016)	1.455	1.398	<.001
E score, kPa	1.046 (0.926-1.182)	1.284	1.256	.61
CKD[j]	1.135 (0.615-2.097)	3.387	2.550	.27
BUN[k], mg/dL	0.952 (0.893-1.016)	1.397	1.338	.13
Creatinine, mg/dL	4.288 (0.196-94.014)	*10.957*	-	-
MDRD[l]	1.012 (0.994-1.031)	4.860	2.863	.39
UA[m], mg/dL	1.127 (0.967-1.314)	1.642	1.596	.08
Cholesterol, mg/dL	1.003 (0.986-1.021)	*9.855*	-	-
LDL[n] mg/dL	0.994 (0.976-1.012)	*9.701*	*1.069*	.82
HbA_{1c}[o], %	2.170 (1.631-2.888)	1.236	1.230	<.001
TSH[p], μIU/mL	0.876 (0.727-1.054)	1.086	1.078	.13

[a]The odds ratio represents the exp(β), which is the exponential of the estimator in logistic regression.

[b]VIF: variance inflation factor (to check multicollinearity); factors with high VIF values are italicized.

[c]ΔVIF: variance inflation factor after removal of predictor variables with high VIF values; VIF values with a sharp decline are italicized.

[d]AFP: alpha-fetoprotein.

[e]ALKp: alkaline phosphatase.

[f]GOT: glutamic-oxalocetic transaminase.

[g]GPT: glutamic-pyruvic transaminase.

[h]γ-GT: gamma-glutamyl transpeptidase.

[i]CAP: controlled attenuation parameter.

[j]CKD: chronic kidney disease.

[k]BUN: blood urea nitrogen.

[l]MDRD: Modification of Diet in Renal Disease.

[m]UA: uric acid.

[n]LDL: low-density lipoprotein cholesterol.

[o]HbA_{1c}: glycated hemoglobin.

[p]TSH: thyroid-stimulating hormone.

To inspect the potential indices used for metabolic syndrome, several types of decision trees were applied to health examination data for the classification of metabolic syndrome (Figure 3). In general, obesity, CAP score, and HbA_{1c} were found to be important predictive variables in the decision trees.

Moreover, important variables appearing in each node of the decision trees were recorded 100 times (Table 3). CAP score, obesity, and HbA$_{1c}$ were regarded as outstanding variables in the root, and E score, γ-GT, LDL, and GPT were secondary variables in the decision trees. The thresholds for factors classified as nodes are listed on the branches of each decision tree. In addition, a right skew pattern at the leaves was apparent and expected because the classification of metabolic syndrome was achieved efficiently and hierarchically by the decision trees (Figure 3).

Figure 3. Metabolic syndrome prediction by various decision tree models. The decision tree takes on a flowchart-like structure. The six most commonly used decision trees are shown: (a) classification and regression tree (CART), (b) C5.0 classification tree modified from C4.5 and ID3 tree, (c) chi-square automatic interaction detection (CHAID), (d) conditional inference tree (ctree), (e) evolutionary learning of globally optimal tree (evtree), and (f) generalized linear model tree (glmtree). Each decision tree is applied for the prediction of metabolic syndrome to explore the factors with the greatest influence as an index to distinguish metabolic syndrome.
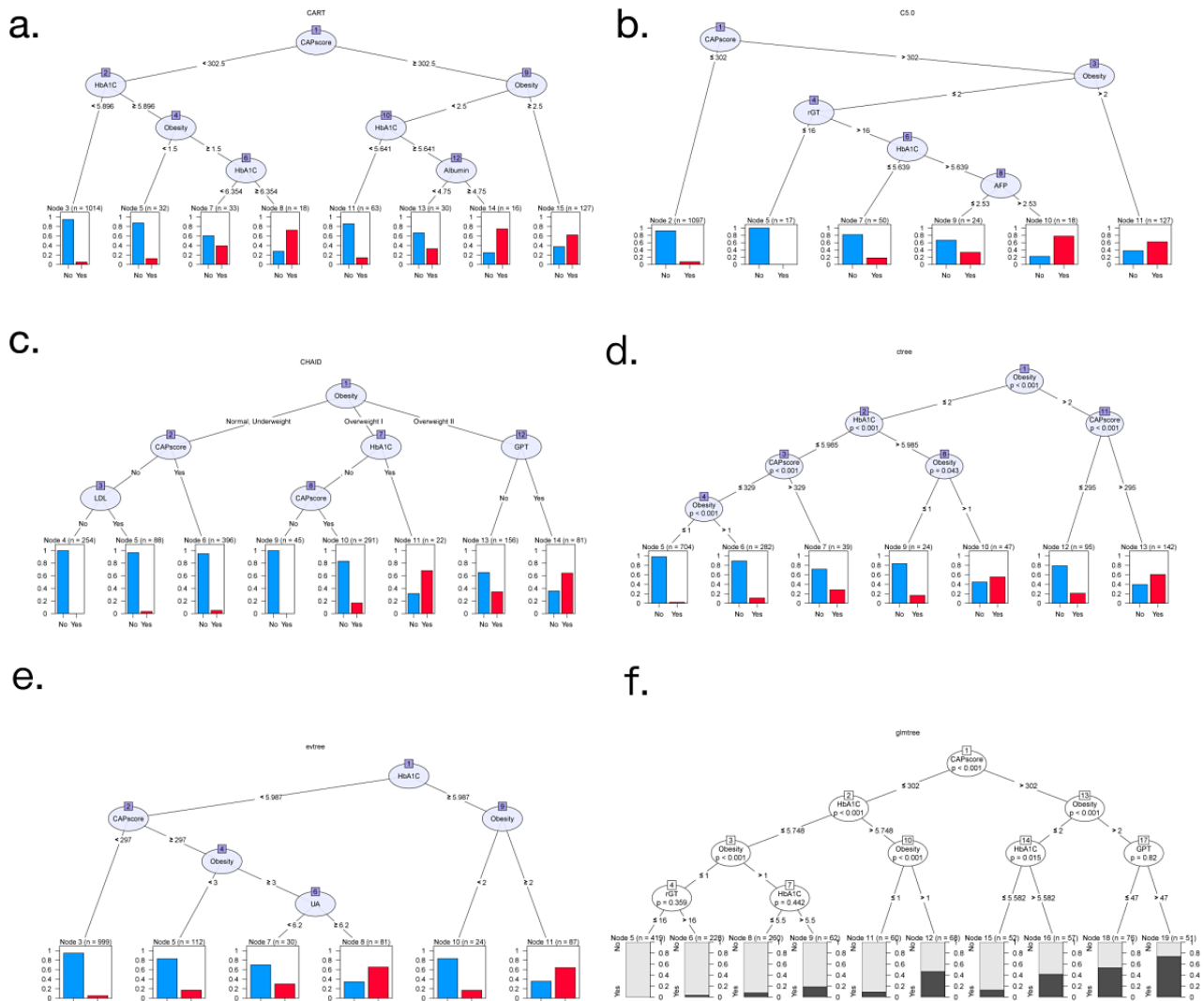
Table 3. Major factors as classified nodes in decision trees[a].

Decision tree	Root[b]	Primary node[c] (root included)	Secondary node[d]	
CART[e]	*CAP [f] score (0.91)*	*CAP score (0.99)*	*HbA₁c (0.93)*	Total bilirubin (0.19)
	Obesity (0.09)	*HbA₁c[g] (0.95)*	*Obesity (0.54)*	Albumin (0.18)
		Obesity (0.94)	γ-GT[h] (0.40)[v]	ALKp[l] (0.17)
		E score (0.05)[v]	Age (0.36)	*CAP score (0.17)*
			GOT[i] (0.31)	GPT[m] (0.17)[v]
			UA[j] (0.28)	LDL[n] (0.15)[v]
			AFP[k] (0.25)	MDRD[o] (0.13)
			E score (0.21)[v]	TSH[p] (0.10)
C5.0	*CAP score (0.90)*	*Obesity (1.06)*	*HbA₁c (0.52)*	Sex (0.28)
	Obesity (0.06)	*CAP score (0.94)*	Total bilirubin (0.41)	*Obesity (0.23)*
	HbA₁c (0.04)	*HbA₁c (0.25)*	GOT (0.35)	AFP (0.19)
			γ-GT (0.34)[v]	CKD[q] (0.16)
			E score (0.29)[v]	UA (0.11)
CHAID[r]	*Obesity (1.00)*	*Obesity (1.00)*	*CAP score (1.21)*	AFP (0.12)
		GPT (0.99)[v]	GPT (0.41)[v]	*HbA₁c (0.12)*
		HbA₁c (0.74)	E score (0.32)[v]	
		CAP score (0.70)	LDL (0.32)[v]	
		LDL (0.42)[v]	TSH (0.19)	
		γ-GT (0.06)[v]	Sex (0.16)	
ctree[s]	*Obesity (0.96)*	*HbA₁c (0.89)*	*Obesity (1.23)*	AFP (0.16)
	CAP score (0.04)	*CAP score (0.85)*	*CAP score (0.67)*	
		Obesity (0.51)	*HbA₁c (0.66)*	
		γ-GT (0.14)[v]	LDL (0.17)[v]	
evtree[t]	*Obesity (0.33)*	*Obesity (0.82)*	*HbA₁c (0.48)*	GOT (0.19)
	HbA₁c (0.30)	*HbA₁c (0.58)*	*CAP score (0.44)*	γ-GT (0.13)[v]
	CAP score (0.17)	*CAP score (0.53)*	*Obesity (0.42)*	GPT (0.11)[v]
	Escore (0.05)[v]	Escore (0.10)[v]	UA (0.26)	
	UA (0.05)	UA (0.10)	E score (0.25)[v]	
		rGT (0.05)[v]	ALKp (0.19)	
glmtree[u]	*CAP score (0.85)*	*Obesity (1.16)*	*Obesity (1.13)*	LDL (0.23)[v]
	Obesity (0.15)	*CAP score (0.92)*	*HbA₁c (0.85)*	*CAP score (0.22)*
		HbA₁c (0.79)	γ-GT (0.79)[v]	MDRD (0.18)
		LDL (0.06)[v]	Escore (0.35)[v]	GOT (0.15)
			GPT (0.26)[v]	

[a]Major variables are listed with their weights as candidate nodes in each decision tree; since some variables may be considered candidate nodes in the decision tree more than once, the proportion of variables can be larger than 1.

[b]The root shows factors appearing as the first classified node and their proportions.

[c]The primary node (italicized) includes variables selected as the top three nodes (root included) with their proportions (>0.05); variables with lower weights as candidate nodes in the primary nodes are excluded.

[d]The secondary node includes all remaining candidate nodes in each decision tree with their proportions; only candidate nodes with proportions >0.1 with a certain influence in the classification of metabolic syndrome are shown.

[e]CART: classification and regression trees.

[f]CAP: controlled attenuation parameter.

[g]HbA₁c: glycated hemoglobin.

[h]γ-GT: gamma-glutamyl transpeptidase.

[i]GOT: glutamic-oxalocetic transaminase.

[j]UA: uric acid.

[k]AFP: alpha-fetoprotein.

[l]ALKp: alkaline phosphatase.

[m]GPT: glutamic-pyruvic transaminase.

[n]LDL: low-density lipoprotein cholesterol.

[o]MDRD: Modification of Diet in Renal Disease.

[p]TSH: thyroid-stimulating hormone.

[q]CKD: chronic kidney disease.

[r]CHAID: chi-square automatic interaction detection.

[s]ctree: conditional inference tree.

[t]evtree: evolutionary learning of globally optimal tree.

[u]glmtree: generalized linear model tree.

[v]Secondary variables for classification of metabolic syndrome in several decision tree algorithms.

PCA was then applied again to visualize the nonmetabolic syndrome and metabolic syndrome groups according to the prominence of factors from the decision trees, which comprised the CAP score, obesity, and HbA_{1c} (Figure 2b). PC1 and PC2 explained greater variability of 56.7% and 29.1%, respectively. With this analysis, discrimination between the two groups was evident, although the junction of the two groups was explicit in the union (Multimedia Appendix 2).

Finally, the accuracies of various decision trees were determined using 500 rounds of random sampling from the entire health examination dataset with fixed-size divisions of training and testing sets (Table 4). Independent training and testing sets were used for each evaluation to confirm the performance and reliability of each model. The AUC of the ROC curve was determined to evaluate the performance of each decision tree and random forest (Table 4, Figure 4, and Multimedia Appendix 3). Prominent variables obtained with random forest are shown in Figure 4. In general, CAP score, obesity, HbA_{1c}, GPT, and γ-GT were the leading variables in accuracy, whereas CAP score, HbA_{1c}, obesity, GPT, γ-GT, and E score played essential roles in random forest for classification.

Table 4. Accuracy[a] and area under the curve (AUC) values of various decision trees in receiver operating characteristic curve analysis.

Decision tree	Accuracy			F1-score			AUC
	minimum	mean	maximum	minimum	mean	maximum	
CART[b]	0.797	0.857	0.914	0.888	0.919	0.948	0.831
C5.0	0.805	0.861	0.921	0.884	0.922	0.951	0.769
CHAID[c]	0.823	0.873	0.917	0.894	0.930	0.956	0.867
ctree[d]	0.801	0.864	0.914	0.883	0.923	0.954	0.896
evtree[e]	0.805	0.857	0.906	0.880	0.920	0.953	0.815
glmtree[f]	–[g]	–	–	–	–	–	0.889
Random forest	0.812	0.870	0.940	0.888	0.928	0.959	0.904

[a]Accuracy and F1-score were calculated from 500 machine learning trials with different training sets for comparison with the number of candidate trees from random forest. Accuracy is the probability of true positives and true negatives for all data, whereas F1-score is a measure of performance, which is the harmonic mean of precision and recall. The dataset was divided 80% as the training set and 20% as the testing set independently for each analysis with randomized sampling.

[b]CART: classification and regression trees.

[c]CHAID: chi-square automatic interaction detection.

[d]ctree: conditional inference tree.

[e]evtree: evolutionary learning of globally optimal tree.

[f]glmtree: generalized linear model tree.

[g]The terminal nodes of the R package glmtree are not a simple classification form to calculate the confusion matrix for accuracy; therefore, the area under the curve was used to reach a balance in comparison between the seven decision tree techniques on the same training and testing set.

XSL•FO

RenderX

Figure 4. Random forest model for predicting classification performance and variable importance. Receiver operating characteristic (ROC) curve with area under the curve values for (a) classification and regression tree and (b) random forest. The color bar indicates the value of specificity in the false positive rate. (c) Variable importance ordered by accuracy of a mean decrease in random forest. (d) Variable importance ordered by the gini index of a mean decrease in random forest. The leading variables obtained by random forest are listed in darker blue, and less important variables are in lighter blue.
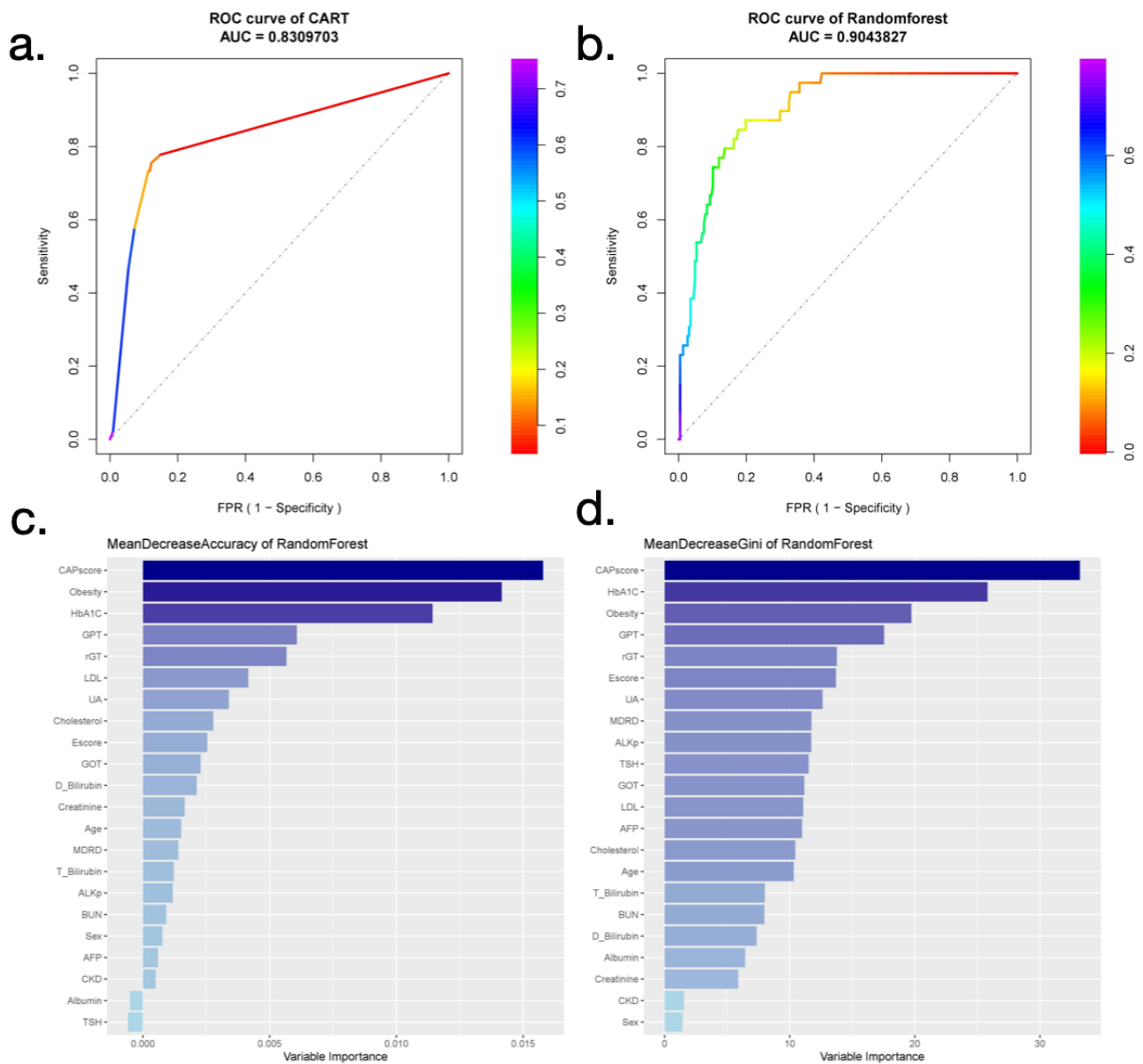
Discussion

Principal Findings

The use of artificial intelligence in health care, particularly machine learning methods, can help to discover underlying patterns and correlations through the learning of data-driven prediction models. We applied various machine learning techniques to visualize and investigate predictive variables leading to metabolic syndrome, which revealed that obesity, serum GOT, serum GPT, CAP score, and HbA_{1c} are the most important predictive variables.

Among these predictive variables, the predictive power of the CAP score was similar to that of other key indices such as obesity. Despite the significance of the CAP score, these factors make sense cumulatively rather than as exclusive alternatives.

In other words, more research is required to determine whether the CAP score can be used as a standalone test method to screen for metabolic syndrome, and whether a minimum set of nonblood test variables can be combined with the CAP score to improve the accuracy of predicting metabolic syndrome. Such future research may help subjects who are resistant to the inconvenience of overnight starvation or painful blood assays.

Metabolic syndrome demonstrates a spectrum of physiological manifestations with groups of pathologies that are complicated and progressive. Traditional diagnostic criteria often dichotomize the population into those with metabolic syndrome and those without. However, based on the results of our PCA, such a sharp distinction may be inappropriate. We found that CAP score, obesity, and HbA_{1c} were the principal factors predicting metabolic syndrome, although E score, γ-GT, LDL,

and GPT also considerably affected the predictions. Notably, GPT had more predictive power than GOT. We consider this difference to be related to aspartate aminotransferase as a relatively less specific indicator of liver damage than alanine aminotransferase, which is common in patients with fatty liver. Our study suggests that current diagnostic criteria for metabolic syndrome fail to capture its wide range of presentations, and should thus be expanded to include hepatic and nephritic indices.

Liver-related indices such as γ-GT, GPT, and E score ranked among the highest predictors in our models. A previous study also showed a strong correlation between liver function tests and metabolic syndrome based on Pearson correlation coefficients [66]. HbA_{1c} is reported to be more closely associated with several chronic diseases than fasting plasma glucose. In addition, although fasting glucose levels are commonly believed to be reproducible across days, acute perturbations of glucose homeostasis due to stress and other factors have been described. By contrast, HbA_{1c} is not influenced by acute perturbations or insufficient fasting; thus, it can be measured at any time. Accordingly, HbA_{1c} might prove to be a more suitable predictor of metabolic syndrome [67].

Multivariate logistic regression has been extensively utilized in medical research, and its many biases have been well documented. One of the drawbacks we observed in our models was the multicollinearity problem. To avoid multicollinearity (Table 2), GOT, direct bilirubin, creatinine, and cholesterol were eliminated from the regression model. By contrast, the decision trees had few such disadvantages and offered more intuitive visualizations. The trained decision tree models could also be more easily interpreted by human experts, which is vital for establishing various important pathways to metabolic syndrome. In general, our result that random forest has the best

accuracy in detecting metabolic syndrome agrees with previous research [68]. One of the reasons for the better accuracy of a random forest model is that it creates multiple decision trees and then combines the output generated by each tree; each tree is built from a sample drawn with replacement from the training set. This approach therefore removes the bias that a decision tree model might introduce in the system, thus substantially improving the predictive power.

Limitations

This study has several limitations. First, this was a retrospective study, and therefore a sufficiently powered prospective cohort study is needed to conclusively address the usefulness of supervised machine learning models to diagnose metabolic syndrome. Second, this study included only health-conscious Taiwanese participants that underwent a self-paid health examination; therefore, this study should be replicated and validated in other populations. Third, this study failed to include some new obesity biomarkers (such as leptin and adiponectin) that may further improve the prediction of metabolic syndrome [69].

Conclusion

To the best of our knowledge, this is the first study to apply machine learning algorithms to identify metabolic syndrome in subjects examined with FibroScan. We found that decision tree learning algorithms identified metabolic syndrome in self-paid health examination subjects with high accuracy, and obesity, serum GOT, serum GPT, CAP score, and HbA_{1c} emerged as important predictive variables. More research is required to validate the CAP score as a standalone test method to screen for metabolic syndrome, and to determine whether a minimum set of nonblood tests variables can be combined with the CAP score to improve the accuracy of predicting metabolic syndrome.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Three-dimensional principal components analysis (PCA) of all risk factors. The three-dimensional PCA plots provide different visual points to observe the scatter of both the metabolic syndrome and nonmetabolic syndrome groups. All factors in Table 1 are considered in this analysis. The leading principal components PC1, PC2, and PC3—which explain more variability among the samples—are shown in all three-dimensional graphs. The aggregation of two groups is obvious when rotating the coordinate in the three-dimensional graph.
[PNG File , 330 KB - medinform_v8i3e17110_app1.png]

Multimedia Appendix 2
Three-dimensional principal components analysis of only the major risk factors. In this case, the distinction between the metabolic syndrome and nonmetabolic syndrome groups is apparent because only the major variables obtained from Table 3 are included, although the borders of the two groups still overlap.
[PNG File , 317 KB - medinform_v8i3e17110_app2.png]

XSL•FO
RenderX

Multimedia Appendix 3

Receiving operator characteristic curves and area under the curve (AUC) values of six decision trees. The specificity is revealed by the color bar, and the diagonal line is presented as a dashed line. Most AUC values exceed 0.80 except for that of the C5.0 tree.

[PNG File , 388 KB - medinform_v8i3e17110_app3.png]

References

1. Huang PL. A comprehensive definition for metabolic syndrome. Dis Model Mech 2009;2(5-6):231-237 [FREE Full text] [doi: 10.1242/dmm.001180] [Medline: 19407331]

2. Gurusamy J, Gandhi S, Damodharan D, Ganesan V, Palaniappan M. Exercise, diet and educational interventions for metabolic syndrome in persons with schizophrenia: A systematic review. Asian J Psychiatr 2018 Aug;36:73-85. [doi: 10.1016/j.ajp.2018.06.018] [Medline: 29990631]

3. Bassi N, Karagodin I, Wang S, Vassallo P, Priyanath A, Massaro E, et al. Lifestyle modification for metabolic syndrome: a systematic review. Am J Med 2014 Dec;127(12):1242. [doi: 10.1016/j.amjmed.2014.06.035] [Medline: 25004456]

4. de Lédinghen V, Vergniol J. Transient elastography (FibroScan). Gastroentérol Clin Biol 2008 Sep;32(6):58-67. [doi: 10.1016/s0399-8320(08)73994-0]

5. Chang PE, Goh GB, Ngu JH, Tan HK, Tan CK. Clinical applications, limitations and future role of transient elastography in the management of liver disease. World J Gastrointest Pharmacol Ther 2016 Feb 06;7(1):91-106 [FREE Full text] [doi: 10.4292/wjgpt.v7.i1.91] [Medline: 26855815]

6. Chivinge A, Harris R, Guha N, Aithal G, James M, Ryder S, et al. Risk-stratified screening for chronic liver disease using vibration-controlled transient elastography (Fibroscan). Gastrointest Nurs 2018 Jun 02;16(5):S15-S22. [doi: 10.12968/gasn.2018.16.5.s15]

7. Wong GL. Transient elastography: Kill two birds with one stone? World J Hepatol 2013 May 27;5(5):264-274 [FREE Full text] [doi: 10.4254/wjh.v5.i5.264] [Medline: 23717737]

8. Recio E, Cifuentes C, Macías J, Mira JA, Parra-Sánchez M, Rivero-Juárez A, et al. Interobserver concordance in controlled attenuation parameter measurement, a novel tool for the assessment of hepatic steatosis on the basis of transient elastography. Eur J Gastroenterol Hepatol 2013 Aug;25(8):905-911. [doi: 10.1097/MEG.0b013e32835f4c3d] [Medline: 23459105]

9. Fraquelli M, Rigamonti C, Casazza G, Conte D, Donato MF, Ronchi G, et al. Reproducibility of transient elastography in the evaluation of liver fibrosis in patients with chronic liver disease. Gut 2007 Jul;56(7):968-973 [FREE Full text] [doi: 10.1136/gut.2006.111302] [Medline: 17255218]

10. Boursier J, Konaté A, Gorea G, Reaud S, Quemener E, Oberti F, et al. Reproducibility of liver stiffness measurement by ultrasonographic elastometry. Clin Gastroenterol Hepatol 2008 Nov;6(11):1263-1269. [doi: 10.1016/j.cgh.2008.07.006] [Medline: 18995217]

11. Li Y, Huang Y, Wang Z, Yang Z, Sun F, Zhan S, et al. Systematic review with meta-analysis: the diagnostic accuracy of transient elastography for the staging of liver fibrosis in patients with chronic hepatitis B. Aliment Pharmacol Ther 2016 Feb;43(4):458-469. [doi: 10.1111/apt.13488] [Medline: 26669632]

12. Pavlov CS, Casazza G, Nikolova D, Tsochatzis E, Burroughs AK, Ivashkin VT, et al. Transient elastography for diagnosis of stages of hepatic fibrosis and cirrhosis in people with alcoholic liver disease. Cochrane Database Syst Rev 2015 Jan 22;1:CD010542. [doi: 10.1002/14651858.CD010542.pub2] [Medline: 25612182]

13. Hartl J, Denzer U, Ehlken H, Zenouzi R, Peiseler M, Sebode M, et al. Transient elastography in autoimmune hepatitis: Timing determines the impact of inflammation and fibrosis. J Hepatol 2016 Oct;65(4):769-775. [doi: 10.1016/j.jhep.2016.05.023] [Medline: 27238753]

14. Roulot D, Costes J, Buyck J, Warzocha U, Gambier N, Czernichow S, et al. Transient elastography as a screening tool for liver fibrosis and cirrhosis in a community-based population aged over 45 years. Gut 2011 Jul;60(7):977-984. [doi: 10.1136/gut.2010.221382] [Medline: 21068129]

15. Kotronen A, Westerbacka J, Bergholm R, Pietiläinen KH, Yki-Järvinen H. Liver fat in the metabolic syndrome. J Clin Endocrinol Metab 2007 Sep;92(9):3490-3497. [doi: 10.1210/jc.2007-0482] [Medline: 17595248]

16. Vuppalanchi R, Siddiqui MS, Van Natta ML, Hallinan E, Brandman D, Kowdley K, NASH Clinical Research Network. Performance characteristics of vibration-controlled transient elastography for evaluation of nonalcoholic fatty liver disease. Hepatology 2018 Jan;67(1):134-144 [FREE Full text] [doi: 10.1002/hep.29489] [Medline: 28859228]

17. Eddowes PJ, Sasso M, Allison M, Tsochatzis E, Anstee QM, Sheridan D, et al. Accuracy of FibroScan Controlled Attenuation Parameter and Liver Stiffness Measurement in Assessing Steatosis and Fibrosis in Patients With Nonalcoholic Fatty Liver Disease. Gastroenterology 2019 May;156(6):1717-1730. [doi: 10.1053/j.gastro.2019.01.042] [Medline: 30689971]

18. Hu Y, Dong N, Qu Q, Zhao X, Yang H. The correlation between controlled attenuation parameter and metabolic syndrome and its components in middle-aged and elderly nonalcoholic fatty liver disease patients. Medicine (Baltimore) 2018 Oct;97(43):e12931. [doi: 10.1097/MD.0000000000012931] [Medline: 30412101]

19. Lin YJ, Lin CH, Wang ST, Lin SY, Chang SS. Noninvasive and Convenient Screening of Metabolic Syndrome Using the Controlled Attenuation Parameter Technology: An Evaluation Based on Self-Paid Health Examination Participants. J Clin Med 2019 Oct 24;8(11):E1775 [FREE Full text] [doi: 10.3390/jcm8111775] [Medline: 31653028]

XSL•FO
RenderX

20. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]

21. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. N Engl J Med 2017 Jun 29;376(26):2507-2509. [doi: 10.1056/NEJMp1702071] [Medline: 28657867]

22. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2017 Jun 14;38(23):1805-1814. [doi: 10.1093/eurheartj/ehw302] [Medline: 27436868]

23. Koh HC, Tan G. Data mining applications in healthcare. J Healthc Inf Manag 2005;19(2):64-72. [Medline: 15869215]

24. Lisboa PJ. A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Networks 2002 Jan;15(1):11-39. [doi: 10.1016/s0893-6080(01)00111-3]

25. Bertsimas D, Bjarnadóttir MV, Kane MA, Kryder JC, Pandey R, Vempala S, et al. Algorithmic Prediction of Health-Care Costs. Operations Res 2008 Dec;56(6):1382-1392. [doi: 10.1287/opre.1080.0619]

26. Bell LM, Grundmeier R, Localio R, Zorc J, Fiks AG, Zhang X, et al. Electronic health record-based decision support to improve asthma care: a cluster-randomized trial. Pediatrics 2010 Apr;125(4):e770-e777. [doi: 10.1542/peds.2009-1385] [Medline: 20231191]

27. Tomar D, Agarwal S. A survey on Data Mining approaches for Healthcare. Int J Bioci Biotechnol 2013 Oct 31;5(5):241-266. [doi: 10.14257/ijbsbt.2013.5.5.25]

28. Fei Y, Hu J, Gao K, Tu J, Li W, Wang W. Predicting risk for portal vein thrombosis in acute pancreatitis patients: A comparison of radical basis function artificial neural network and logistic regression models. J Crit Care 2017 Jun;39:115-123. [doi: 10.1016/j.jcrc.2017.02.032] [Medline: 28246056]

29. Santelices LC, Wang Y, Severyn D, Druzdzel MJ, Kormos RL, Antaki JF. Development of a hybrid decision support model for optimal ventricular assist device weaning. Ann Thorac Surg 2010 Sep;90(3):713-720 [FREE Full text] [doi: 10.1016/j.athoracsur.2010.03.073] [Medline: 20732482]

30. Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med 1991 Dec 01;115(11):843-848. [doi: 10.7326/0003-4819-115-11-843] [Medline: 1952470]

31. Artis SG, Mark RG, Moody GB. Detection of atrial fibrillation using artificial neural networks. In: (1991) Proceedings Computers in Cardiology.: IEEE; 1991 Presented at: Computers in Cardiology (CinC); September 23-26, 1991; Venice, Italy p. 23-26 URL: https://ieeexplore.ieee.org/document/169073 [doi: 10.1109/CIC.1991.169073]

32. Eftekhar B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC Med Inform Decis Mak 2005 Feb 15;5:3 [FREE Full text] [doi: 10.1186/1472-6947-5-3] [Medline: 15713231]

33. Belciug S, Salem A, Gorunescu F, Gorunescu M. Clustering-based approach for detecting breast cancer recurrence. : IEEE; 2010 Presented at: 10th International Conference on Intelligent Systems Design and Applications; 2010; Cairo, Egypt p. 533-538. [doi: 10.1109/ISDA.2010.5687211]

34. Salama GI, Abdelhalim MB, Zeid M. Breast Cancer Diagnosis on Three Different Datasets using Multi-classifiers. Int J Comput Inf Technol 2012;1(1):36-43 [FREE Full text]

35. Ayer T, Chhatwal J, Alagoz O, Kahn CE, Woods RW, Burnside ES. Informatics in radiology: comparison of logistic regression and artificial neural network models in breast cancer risk estimation. Radiographics 2010 Jan;30(1):13-22 [FREE Full text] [doi: 10.1148/rg.301095057] [Medline: 19901087]

36. Joshi S, Shenoy DG, G.G VS, Rrashmi PL, Venugopal KR, Patnaik LM. Classification of Alzheimer's Disease and Parkinson's Disease by Using Machine Learning and Neural Network Methods. IEEE: IEEE Computer Society; 2010 Presented at: ICMLC 2010: The 2nd International Conference on Machine Learning and Computing; February 9-11, 2010; Bangalore, India p. 218-222. [doi: 10.1109/ICMLC.2010.45]

37. Escudero J, Zajicek JP, Ifeachor E. Early detection and characterization of Alzheimer's disease in clinical scenarios using Bioprofile concepts and K-means. In: Conf Proc IEEE Eng Med Biol Soc.: IEEE; 2011 Presented at: 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2011; Boston. [doi: 10.1109/IEMBS.2011.6091597]

38. Ramani RG, Sivagami G. Parkinson Disease Classification using Data Mining Algorithms. International Journal of Computer Applications October. Int J Comput Appl 2011;32(9):17-22 [FREE Full text] [doi: 10.5120/3932-5571]

39. Nilsson J, Ohlsson M, Thulin L, Höglund P, Nashef SA, Brandt J. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. J Thorac Cardiovasc Surg 2006 Jul;132(1):12-19 [FREE Full text] [doi: 10.1016/j.jtcvs.2005.12.055] [Medline: 16798296]

40. Kazemi Y, Mirroshandel SA. A novel method for predicting kidney stone type using ensemble learning. Artif Intell Med 2018 Jan;84:117-126. [doi: 10.1016/j.artmed.2017.12.001] [Medline: 29241659]

41. Edelstein P. Emerging directions in analytics. Predictive analytics will play an indispensable role in healthcare transformation and reform. Health Manag Technol 2013 Jan;34(1):16-17. [Medline: 23420986]

42. Moradi M, Ghadiri N. Different approaches for identifying important concepts in probabilistic biomedical text summarization. Artif Intell Med 2018 Jan;84:101-116. [doi: 10.1016/j.artmed.2017.11.004] [Medline: 29208328]

43. Ma YC, Zuo L, Chen JH, Luo Q, Yu XQ, Li Y, et al. Modified glomerular filtration rate estimating equation for Chinese patients with chronic kidney disease. J Am Soc Nephrol 2006 Oct;17(10):2937-2944 [FREE Full text] [doi: 10.1681/ASN.2006040368] [Medline: 16988059]

44. National Kidney Foundation. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. Am J Kidney Dis 2002 Feb;39(2 Suppl 1):S1-266. [Medline: 11904577]

45. Choo V. WHO reassesses appropriate body-mass index for Asian populations. Lancet 2002 Jul 20;360(9328):235. [doi: 10.1016/S0140-6736(02)09512-0] [Medline: 12133671]

46. Alberti KG, Zimmet P, Shaw J. Metabolic syndrome--a new world-wide definition. A Consensus Statement from the International Diabetes Federation. Diabet Med 2006 May;23(5):469-480. [doi: 10.1111/j.1464-5491.2006.01858.x] [Medline: 16681555]

47. Alberti KG, Zimmet P, Shaw J, IDF Epidemiology Task Force Consensus Group. The metabolic syndrome--a new worldwide definition. Lancet 2005;366(9491):1059-1062. [doi: 10.1016/S0140-6736(05)67402-8] [Medline: 16182882]

48. Quinlan JR. Simplifying decision trees. Int J Man-Machine Stud 1987 Sep;27(3):221-234. [doi: 10.1016/s0020-7373(87)80053-6]

49. Breiman L, Friedman J, Stone CJ, Olshen R. Classification and Regression Trees (Wadsworth Statistics/Probability). Boca Raton, Florida: Chapman and Hall; 1984.

50. Therneau T, Atkinson B, Ripley B. The rpart Package. 2010. URL: https://cran.r-project.org/web/packages/rpart/rpart.pdf [accessed 2019-10-15]

51. Quinlan JR. C4.5: Programs For Machine Learning (Morgan Kaufmann Series In Machine Learning). San Francisco, CA: Morgan Kaufmann; 1993.

52. Kuhn M, Weston S, Culp M, Coulter N, Quinlan R. Package 'C50'. 2018. URL: https://cran.r-project.org/web/packages/C50/C50.pdf [accessed 2019-10-01]

53. Kass GV. An Exploratory Technique for Investigating Large Quantities of Categorical Data. Appl Stat 1980;29(2):119. [doi: 10.2307/2986296]

54. Hothorn T, Zeileis A. partykit: A modular toolkit for recursive partytioning in R. J Machine Learn Res 2015;16(1):3905-3909 [FREE Full text]

55. Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. J Comput Graph Stat 2006 Sep;15(3):651-674. [doi: 10.1198/106186006X133933]

56. Grubinger T, Zeileis A, Pfeiffer K. evtree: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R. J Stat Soft 2014;61(1):1-29. [doi: 10.18637/jss.v061.i01]

57. Zeileis A, Hothorn T, Hornik K. Model-Based Recursive Partitioning. J Comput Graph Stat 2008 Jun;17(2):492-514. [doi: 10.1198/106186008x319331]

58. Breiman L. Random forests. Machine learning 2001;45(1):5-32 [FREE Full text] [doi: 10.1023/A:1010933404324]

59. Fox J, Monette G. Generalized Collinearity Diagnostics. J Am Stat Assoc 1992 Mar;87(417):178. [doi: 10.2307/2290467]

60. Fox J, Bates D, Firth D, Friendly M, Gorjanc G, Graves S. The car Package. 2007. URL: ftp://ftp.uni-bayreuth.de/pub/math/statlib/R/CRAN/doc/packages/car.pdf [accessed 2019-10-20]

61. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett 2006 Jun;27(8):861-874. [doi: 10.1016/j.patrec.2005.10.010]

62. Sasaki Y. The truth of the F-measure. 2007 Jun. URL: https://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf [accessed 2019-10-15]

63. Powers DMW. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol 2011;2(1):37-63 [FREE Full text] [doi: 10.9735/2229-3981]

64. von Hippel PT. Biases in SPSS 12.0 Missing Value Analysis. Am Stat 2004 May;58(2):160-164. [doi: 10.1198/0003130043204]

65. Lin M, Lucas HC, Shmueli G. Research commentary—Too big to fail: large Samples and the p-value problem. Inf Syst Res 2013 Dec;24(4):906-917. [doi: 10.1287/isre.2013.0480]

66. Wang S, Zhang J, Zhu L, Song L, Meng Z, Jia Q, et al. Association between liver function and metabolic syndrome in Chinese men and women. Sci Rep 2017 Mar 20;7(1):44844. [doi: 10.1038/srep44844] [Medline: 28317840]

67. Vijayakumar P, Nelson RG, Hanson RL, Knowler WC, Sinha M. HbA1c and the Prediction of Type 2 Diabetes in Children and Adults. Diabetes Care 2017 Jan;40(1):16-21 [FREE Full text] [doi: 10.2337/dc16-1358] [Medline: 27810987]

68. Alam MZ, Rahman MS, Rahman MS. A Random Forest based predictor for medical data classification using feature ranking. Informat Med Unlocked 2019;15:100180. [doi: 10.1016/j.imu.2019.100180]

69. Srikanthan K, Feyh A, Visweshwar H, Shapiro JI, Sodhi K. Systematic Review of Metabolic Syndrome Biomarkers: A Panel for Early Detection, Management, and Risk Stratification in the West Virginian Population. Int J Med Sci 2016;13(1):25-38 [FREE Full text] [doi: 10.7150/ijms.13800] [Medline: 26816492]

Abbreviations

AFP: alpha-fetoprotein
ALKp: alkaline phosphatase

XSL•FO

RenderX

AUC: area under the curve
BUN: blood urea nitrogen
CAP score: controlled attenuation parameter score
CART: classification and regression tree
CHAID: Chi-square automatic interaction detection
CKD: chronic kidney disease
ctree: conditional interference tree
eGFR: estimated glomerular filtration rate
E score: liver stiffness score
evtree: evolutionary learning of globally optimal trees
γ-GT: gamma-glutamyl transpeptidase
GOT: serum glutamic oxaloacetic transaminase
GPT: serum glutamic pyruvic transaminase
HbA$_{1c}$: glycated hemoglobin
HDL: high-density lipoprotein
LDL: low-density lipoprotein
MDRD: Modification of Diet in Renal Disease
NAFLD: nonalcoholic fatty liver disease
PCA: principal components analysis
ROC: receiver operating characteristic
TSH: thyroid-stimulating hormone
UA: uric acid
VIF: variance inflation factor

Original Paper

Predicting Adverse Outcomes for Febrile Patients in the Emergency Department Using Sparse Laboratory Data: Development of a Time Adaptive Model

Sungjoo Lee[1], BS; Sungjun Hong[1], MS; 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,
Republic of Korea
Phone: 82 10 5386 6597
Email: wc.cha@samsung.com

Abstract

Background: A timely decision in the initial stages for patients with an acute illness is important. However, only a few studies have determined the prognosis of patients based on insufficient laboratory data during the initial stages of treatment.

Objective: This study aimed to develop and validate time adaptive prediction models to predict the severity of illness in the emergency department (ED) using highly sparse laboratory test data (test order status and test results) and a machine learning approach.

Methods: This retrospective study used ED data from a tertiary academic hospital in Seoul, Korea. Two different models were developed based on laboratory test data: order status only (OSO) and order status and results (OSR) models. A binary composite adverse outcome was used, including mortality or hospitalization in the intensive care unit. Both models were evaluated using various performance criteria, including the area under the receiver operating characteristic curve (AUC) and balanced accuracy (BA). Clinical usefulness was examined by determining the positive likelihood ratio (PLR) and negative likelihood ratio (NLR).

Results: Of 9491 eligible patients in the ED (mean age, 55.2 years, SD 17.7 years; 4839/9491, 51.0% women), the model development cohort and validation cohort included 6645 and 2846 patients, respectively. The OSR model generally exhibited better performance (AUC=0.88, BA=0.81) than the OSO model (AUC=0.80, BA=0.74). The OSR model was more informative than the OSO model to predict patients at low or high risk of adverse outcomes (P<.001 for differences in both PLR and NLR).

Conclusions: Early-stage adverse outcomes for febrile patients could be predicted using machine learning models of highly sparse data including test order status and laboratory test results. This prediction tool could help medical professionals who are simultaneously treating the same patient share information, lead dynamic communication, and consequently prevent medical errors.

XSL•FO
RenderX

Introduction

For time-sensitive diseases, timely decisions are essential; however, the availability of data is extremely limited in the early stages of medicine [1,2]. Data obtained in the long term after the patient's visit provide sufficient information, and the results of analysis to predict the patient's outcome are highly accurate. However, the timing and effectiveness of this long-term data are limited in early decision making because the results do not reflect the patient's initial status. Therefore, it is necessary to develop a time adaptive model that reflects the decision-making process by utilizing the pattern of interim information in uncertain situations during the initial stages of the patient's visit.

Biomarkers, especially those obtained via laboratory data, play a key role in clinical decisions in emergency settings [3-5]. Laboratory data are important for predicting the patient's prognosis but can lead to delays in decision making since many test results are not available during the initial stages [6-9]. This further exacerbates the inherent problems with laboratory data including a high level of sparsity due to the many test types and variation in individual orders [8,10]. Therefore, we try to maximize the utilization of laboratory information through patterns and the use of order status, which can infer the patient's initial status before obtaining test results.

Previous studies have focused on utilizing a sufficient amount of laboratory test data. Most predictive models have been developed based on long intervals such as those to predict mortality occurring within 24 or 48 hours rather than earlier periods; with these longer periods, researchers can be guaranteed of adequate information from test results [11,12]. Previous studies have also used a limited number of the frequently measured test variables, rather than including all possible tests, to predict mortality, and they found that sufficient data were available to develop prediction models [13,14].

This study aimed to develop time adaptive models that predict adverse outcomes for febrile patients in the emergency department (ED) based on a machine learning approach and highly sparse data.

Methods

Study Setting

This retrospective study was conducted with ED data from a tertiary academic hospital in Seoul, Korea. The hospital has approximately 2000 beds. The outpatient department has an average of approximately 9000 patients per day, while the ED has approximately 220 patients per day. Since the opening of

its comprehensive cancer center in 2003, the hospital has a large portion of oncology patients undergoing both surgical and medical procedures. This study was approved by the institutional review board of the study site (IRB File No: SMC 2018-08-125). This report follows the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline.

Source of Data

Data were obtained from a clinical data warehouse containing medical data for research, which enables de-identification and retrieval of patient information from electronic medical records for research purposes. It uses global standard terminology and provides near realtime data through daily updates. In addition to basic patient demographics, it provides information on tests, medications, diagnoses, and operations.

Participants

Patients who visited the ED from March 2017 through February 2019 were included in the study. Then, only febrile (body temperature >38°C) [15] and adult patients (aged ≥18 years) were included. Patients were excluded if the main reason for their visit was determined as trauma.

Outcome and Predictors

We used a binary composite outcome for severity. Severity was considered as death or admission to the intensive care unit after transfer from the ED.

Only laboratory test data were used as predictors, and the list of laboratory tests was selected by experts. Predictors were selected based on the typical ED process in which all possible laboratory tests could be performed after the initial assessment by physicians [16]. Figure 1 shows the general process from the typical initial process to patient discharge from the ED. This study focused on the initial process and the ordering of laboratory tests and test results in particular. Two models were developed from the viewpoint of the initial process, using laboratory test order status and laboratory test results that became available later.

The laboratory test data provide the order status and result for each laboratory test, and all the variables were categorized. Order status indicates whether a patient has an order for a laboratory test, and the test result reflects whether it was normal, abnormal, or not reported. When the test was repeatedly performed, only the first test data were included. We developed two predictive models using these laboratory test data: order status only (OSO) for the first model and order status and results (OSR) for model 2 (Figure 2).

Figure 1. Process flow in the emergency department.
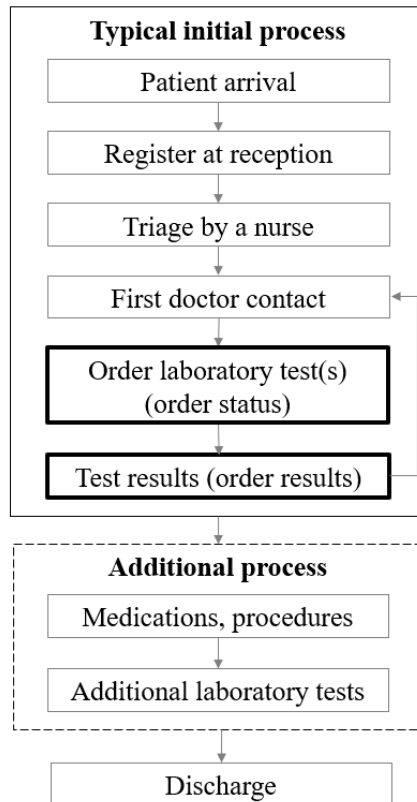
Figure 2. Representative example of the range of predictors for each model, where each row indicates a patient's record of laboratory tests. Additional laboratory tests and patient records can be added. Order status, which indicates whether the test was ordered, was used in the OSO model. The OSR model was developed using order status and test results, which had three levels: normal, abnormal, and NA (not reported). CRP: C-reactive Protein; Hb: hemoglobin.

For a group of laboratory tests that are not frequently ordered but are conducted for only a few patients, the order status information causes severe data sparsity. Rather than using the order status information for each of those tests, new variables were introduced. First, rarely ordered tests (ROTs) were identified as tests that had an ordered rate <5%. The new variable ROT was defined as the number of ROTs ordered for each patient. Likewise, for tests that generate rarely detected abnormal results (RARs; <5% of the results are abnormal), a new variable RAR was defined as the number of RARs obtained among those tests for each patient.

Analytical Methods

Patients were randomly assigned to two cohorts for model development (70%) and validation (30%), which had similar distributions with respect to the outcome. We applied and compared various machine learning methods, including random forest (RF), support vector machine, logistic regression with least absolute shrinkage and selection operator, ridge, and elastic net (EN) regularization [17-21]. For optimization, the grid search was used for all algorithms, and the stochastic gradient descent method was used for penalized regression algorithms. The Gini index was selected to measure the split quality in RF. Linear and sigmoid kernels were considered for the support vector

machine. Moreover, hyperparameters for each algorithm were tuned with the accuracy measure based on 10-fold cross-validations that were repeatedly conducted 5 times to reduce the partition bias in model development.

The predictive models were evaluated with the validation cohort using various performance measures, such as the area under the receiver operating characteristic curve (AUC), area under the precision recall curve (AUPRC), balanced accuracy (BA), sensitivity, specificity, *F*1 score, positive likelihood ratio (PLR), and negative likelihood ratio (NLR) [22]. For each measure, we provide 95% CIs, which were estimated by bootstrapping 2000 resamples, and selected the optimal thresholds with which BA was maximized. Moreover, net reclassification improvement [23] was used to measure the incremental value of adding test results to the order status in the prediction. Comparison of the performance between the models was conducted using the bootstrap-t method [24]. The resulting predictive models were further compared with the Modified Early Warning Score (MEWS), a reference algorithm currently used to predict the severity of a patient's condition in clinical practice [25]. The clinical usefulness of the models was demonstrated in two ways. First, we conducted Spearman correlation analysis to evaluate whether clinically meaningful variables were selected consistently across the algorithms. Second, the Kaplan-Meier method and log-rank test were used to estimate and compare survival curves between high-risk and low-risk groups as predicted from the OSR model. We used 2-sided *P*<.05 for statistical significance.

Class imbalance existed in our outcome data. This can lead to the classifier having poor performance because it can create bias against a class and may not able to distinguish between noise and the individuals from the minority class [26,27]. We investigated its effects on prediction performance using various scenarios in which different techniques and class ratio were considered for imbalance reduction. Three oversampling and three undersampling approaches were considered. The oversampling methods included random, synthetic minority oversampling technique, and adaptive synthetic sampling. The undersampling methods included random, NearMiss-2, and edited nearest neighbors [28]. Furthermore, we increased the ratio between the minority and majority classes from 1:1 to 1:10 and tried to find the best performance. For each scenario, we used EN for model development, and the performance was assessed using 100 bootstrap resamples from the original dataset.

The preprocess was conducted using R version 3.4.4 [29], and the analytic process was performed with Python version 3.6.2 (Python Software Foundation, Wilmington, DE) using pandas, numpy, sklearn, and imblearn library. More details related to the model development are available in Multimedia Appendix 1.

Results

Patient Demographics

A total of 154,402 patients visited the ED between March 1, 2017, and February 28, 2019. Based on the inclusion and exclusion criteria, 9491 patients remained in the final dataset used for modeling (Multimedia Appendix 2). The randomly divided model development and validation cohorts included 6645 and 2846 patients, respectively, with a composite adverse outcome frequency of 4.6% in the 2 datasets. The baseline characteristics reflect only the initial patient status. The mean age (SD) was 55.2 years (17.7 years); 4839 of the 9491 patients (51.0%) were female; and 432 of the 9491 patients (4.6%) experienced the composite adverse outcome. Patients in the development and validation cohorts were similarly distributed (Table 1).

The three most frequently observed laboratory tests were C reactive protein, chlorine, and sodium. Among a total of 286 laboratory tests after preprocessing, 201 ROTs (order rate <5%) and 231 tests with RARs (abnormal rate <5%) were identified. The OSO model had 85 order status variables as well as the ROT variable. Similarly, the OSR model had 55 result variables, the RAR variable, and the variables in the OSO model.

Table 1. Baseline characteristics of the total sample and comparisons between the two patient cohorts used to develop and validate the two models.

Characteristic	Total sample	Model development cohort (n=6645)	Model validation cohort (n=2846)	P value
Sex, n (%)				
Female	4839 (51.0)	3399 (51.2)	1440 (50.6)	.64
Male	4652 (49.0)	3246 (48.8)	1406 (49.4)	
Age (years), mean (SD)	55.2 (17.7)	55.0 (17.8)	55.6 (17.6)	.17
Transportation, n (%)				
Other	7483 (78.8)	5224 (78.6)	2259 (79.4)	.42
Ambulance	2008 (21.2)	1421 (21.4)	587 (20.6)	
Route, n (%)				
Indirect	1217 (12.8)	830 (12.5)	387 (13.6)	.15
Direct	8274 (87.2)	5815 (87.5)	2459 (86.4)	
Mentality, n (%)				
Alert	9206 (97.0)	6449 (97.1)	2757 (96.9)	.69
Not alert	285 (3.0)	196 (2.9)	89 (3.1)	
Pulse rate, n (%)				
Normal (60-120 beats per minute)	7108 (75.1)	4956 (74.8)	2152 (75.7)	.35
Abnormal (<60 or >120 beats per minute)	2361 (24.9)	1671 (25.2)	690 (24.3)	
Respiratory rate, n (%)				
Normal (10-30 breaths per minute)	9399 (99.2)	6577 (99.2)	2822 (99.2)	1.00
Abnormal (<10 or >30 breaths per minute)	73 (0.8)	51 (0.8)	22 (0.8)	
Systolic blood pressure, n (%)				
Normal (90-140 mmHg)	7212 (76.0)	5044 (75.9)	2168 (76.2)	.80
Abnormal (<90 or >140 mmHg)	2279 (24.0)	1601 (24.1)	678 (23.8)	
Diastolic blood pressure, n (%)				
Normal (60-90 mmHg)	6790 (71.5)	4738 (71.3)	2052 (72.1)	.44
Abnormal (<60 or >90 mmHg)	2701 (28.5)	1907 (28.7)	794 (27.9)	
SpO_2[a], n (%)				
Normal (>90)	9070 (97.4)	6356 (97.5)	2714 (97.2)	.48
Abnormal (<90)	245 (2.6)	166 (2.5)	79 (2.8)	
Outcome, n (%)				
Normal	9059 (95.4)	6342 (95.4)	2717 (95.5)	.99
Composite adverse outcome[b]	432 (4.6)	303 (4.6)	129 (4.5)	

[a]SpO_2 : peripheral oxygen saturation.

[b]Defined as death or admission to the intensive care unit.

Model Performance and Specification

The OSO and OSR models were each developed based on 5 different algorithms. The RF-based models were selected as the final predictive OSO and OSR models because they had better performance overall in terms of the most evaluation measures, including specificity, precision, $F1$ score, NLR, and PLR (Multimedia Appendix 3). Note that the EN-based OSR model was comparable to the RF-based OSR model. Compared to the MEWS (AUC = 0.68), the final OSO and OSR models showed significant AUC improvement, at 12% and 20%, respectively (Table 2, Multimedia Appendix 4). Both models had better performance than the MEWS according to most of the other measures, including the AUPRC, maximum BA, and $F1$ score.

Table 2. Internal validation of the models using different laboratory information, reported as the score and 95% CI.

Measure[a]	MEWS[b]	OSO[c]	OSR[d]	Difference (MEWS vs OSO)[e]	Difference (OSO vs OSR)[f]
AUC[g]	0.68 (0.63 to 0.73)	0.80 (0.76 to 0.84)	0.88 (0.85 to 0.91)	0.12 (0.12 to 0.12)	0.08 (0.08 to 0.08)
AUPRC[h]	0.14 (0.10 to 0.20)	0.25 (0.18 to 0.33)	0.39 (0.30 to 0.47)	0.11 (0.11 to 0.11)	0.14 (0.14 to 0.14)
Sensitivity	0.49 (0.42 to 0.61)	0.70 (0.62 to 0.82)	0.81 (0.76 to 0.89)	0.22 (0.21 to 0.22)	0.10 (0.10 to 0.10)
Specificity	0.82 (0.66 to 0.83)	0.78 (0.66 to 0.83)	0.81 (0.75 to 0.83)	–0.04 (–0.04 to –0.04)	0.04 (0.04 to 0.04)
Balanced accuracy	0.65 (0.62 to 0.69)	0.74 (0.71 to 0.77)	0.81 (0.78 to 0.84)	0.09 (0.09 to 0.09)	0.07 (0.07 to 0.07)
Precision	0.11 (0.08 to 0.14)	0.13 (0.10 to 0.16)	0.17 (0.13 to 0.20)	0.02 (0.02 to 0.02)	0.04 (0.04 to 0.04)
*F*1 score	0.18 (0.14 to 0.22)	0.22 (0.17 to 0.26)	0.28 (0.23 to 0.32)	0.04 (0.04 to 0.04)	0.06 (0.06 to 0.06)
PLR[i]	2.68 (1.76 to 3.27)	3.10 (2.25 to 4.29)	4.22 (2.92 to 4.94)	0.49 (0.48 to 0.5)	1.07 (1.06 to 1.08)
NLR[j]	0.63 (0.49 to 0.73)	0.39 (0.24 to 0.49)	0.23 (0.12 to 0.31)	–0.25 (–0.25 to –0.25)	–0.14 (–0.15 to –0.14)

[a]Calculations were completed with the validation set, and 95% CIs were computed using 2000 bootstrap replicates for each performance measure.

[b]MEWS: Modified Early Warning Score.

[c]OSO: model with order status only.

[d]OSR: model with both order status and test result.

[e]Difference in each performance measure between the MEWS and OSO model.

[f]Difference in each performance measure between the OSO and OSR models.

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

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

[i]PLR: positive likelihood ratio.

[j]NLR: negative likelihood ratio.

Compared with the OSO model, the OSR model showed significant improvement in the AUC, at 8%, and maximum BA, at 7%. Additionally, the OSR model was more informative than the OSO model in predicting low-risk and high-risk patients in terms of outcome (*P*<.001 for difference in both PLR and NLR). A significant additional increment in reclassification was also observed between the OSO and OSR models (net reclassification improvement=0.15). Despite the lack of information from laboratory test results, the OSO model showed considerable performance (AUC = 0.80). Therefore, the order pattern itself can be important information for prediction, and it is better to use the OSO model in the early stages than to wait until test results are obtained. However, because it utilizes laboratory test results obtained later in time, the OSR model has higher accuracy and better performance than the OSO model. According to the Kaplan-Meier survival curves (log-rank test, *P*<.001) comparing the predicted outcome groups, the OSR model can classify patients well (Figure 3). The complementary use of these two models can be beneficial both before and after all laboratory test results are available in clinical situations such as those in the ED.

Important variables selected from the RF-based and EN-based models were moderately correlated in terms of their value importance and odds ratios, respectively (r_s=0.603 and 0.626 for the OSO and OSR models, respectively; Multimedia Appendix 5). Among the top 10 variables selected in each of the RF-based and EN-based models, 80% and 60%, respectively, were shared by the OSO and OSR models. Therefore, the important variables were very similar between the RF and EN models, potentially suggesting our models are robust, regardless of the algorithm used. The order statuses of cardiac troponin I, creatine kinase, and creatine kinase-MB were the top 3 variables in terms of importance in both the RF-based and EN-based OSO models. The order status of creatine kinase remained in the top 10 important variables in the OSR models. The lactic acid test result was the most important variable in both the RF-based and EN-based OSR models.

The data had severe outcome imbalance: 95.4% (9059) for the majority class and 4.6% (432) for the minority class. However, the sensitivity analysis to calibrate the imbalance with various reduction scenarios did not reveal any considerable improvement in the prediction performance. Therefore, our models are not affected much by the imbalance problem (Multimedia Appendix 6).

Figure 3. The curves indicate how the actual outcome developed over time when the patients were divided into high-risk and low-risk categories, as predicted from the OSR model. The graph was plotted using the Kaplan-Meier survival curve, and the P-value shows the log-rank test result.
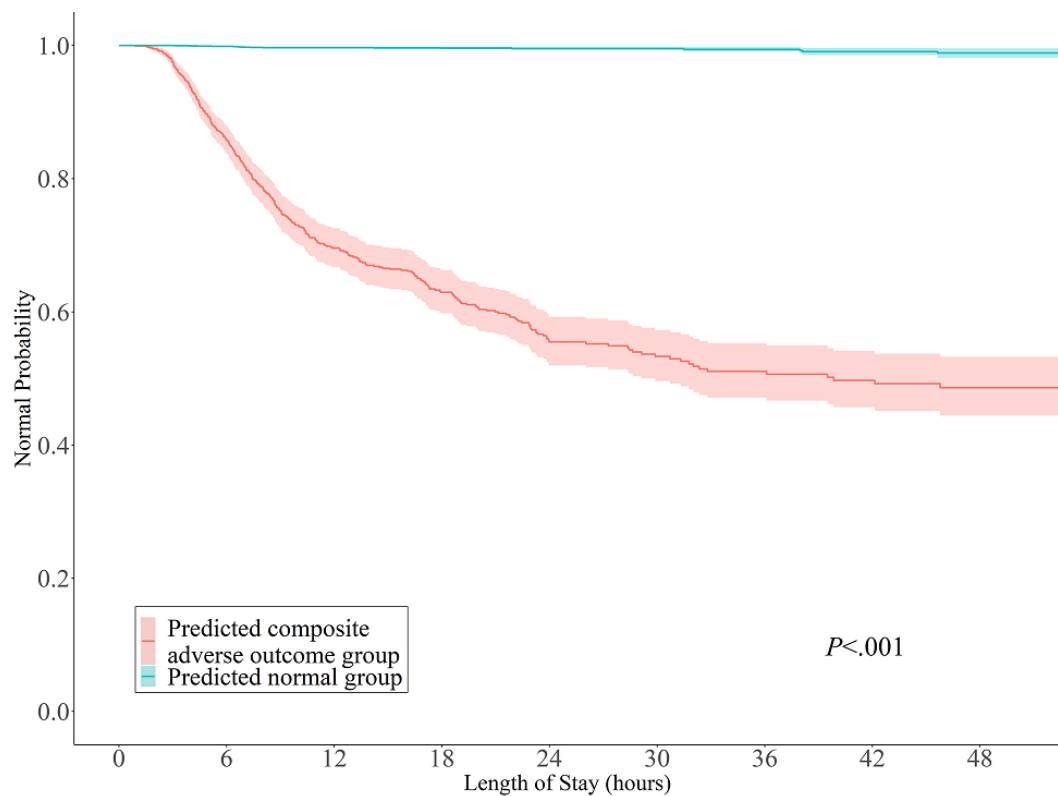
Patients at risk	Length of Stay (hours)								
	0	6	12	18	24	30	36	42	48
Predicted composite adverse outcome group	1186	884	527	387	227	154	124	101	70
Predicted normal group	8305	4672	2908	2151	1195	830	664	540	389

Discussion

Principal Findings

In this study, we developed a time adaptive model to predict adverse outcomes for patients in the ED. These patients are likely to have insufficient and unconfirmed clinical information, especially in the early stages of the ED process. The OSO model, which only utilizes test order status, supports our hypothesis that it is feasible to predict patient prognosis based only on the fact that a laboratory test has been ordered and without the test results. Patient demographics or vital signs were also not required for the prediction.

Febrile patients have a considerable number of laboratory tests to consider. The ED receives patients with different illnesses and febrile patients with various diseases in particular. Fever is also the most common sign of potential sepsis [30,31], requiring more laboratory tests. Febrile patients were selected as the target population to investigate the sparsity and large number of tests typically required initially and to reflect as many tests as possible.

The OSO model mimics the ED physician's clinical reasoning process in practical settings, while prediction models developed previously are limited by using only confirmed results [11-14]. For time-sensitive conditions, multiple tests are performed simultaneously, resulting in a combination of confirmed and unconfirmed results, which necessitates models that can be applied in real practice. This study is the first step to overcome these limitations. Furthermore, it is possible to predict the initial outcome of patients with a severe condition using the OSO model and then update the prediction using the OSR model when additional information becomes available in a time adaptive manner.

In modern medicine, a multidisciplinary approach is a cornerstone of better quality [32,33]. In the ED, multiple providers work as a team to simultaneously treat various conditions [7]. Analyzing laboratory test orders and results could benefit the whole treatment team by interpreting the initial impression and intention of physicians who give the orders for tests. It is possible that each physician may not be aware of the intention of others while simultaneously treating the same patient [7,34,35]. In this context, the prediction tool can help

share information, lead dynamic communication, and consequently prevent medical errors.

This study could be expanded further by including vital signs, procedures, and medications for better prediction. In addition, the application could be broadened to include diagnosis as well as adverse outcomes, especially for diseases where the patient's response over time after a particular treatment is important. Additionally, it can be extended to anticipate clinical decisions, which may be integrated as clinical decision support. The time variable is the most essential component for these predictions, and this model has successfully shown its feasibility.

Limitations

This study has some limitations. First, the models were developed and internally validated using data from a single large hospital. Although cross-validation was performed with repetition, optimization, and several candidates of hyperparameters, along with survival analysis to increase their clinical impact, further studies are required for external and prospective validation.

Second, the primary parameters such as laboratory results, which may vary across individuals and clinical fields, were from febrile patients. Therefore, it could be difficult to apply these to other populations, although we attempted to include as many tests as possible. However, we believe the important variables that were selected related to laboratory tests from the 2 models are clinically relevant for the outcome variables, so there is potential to extend the models to other target populations in future studies.

Third, the OSO and OSR models were not developed with a continuous time sequence. Instead of creating a continuous model, we tried to build representative models to reflect test order status and results. Further research will be required to create a continuous model for practical use, which can be applied to various time thresholds.

Last, the imbalance of data could have affected the performance of models developed using raw data. Although various methods to deal with the issues related with imbalanced data were applied to develop and validate the model, only a few of algorithms among the methods for calibrating the class imbalance were used. It is possible that the use of other algorithms would have changed the results even though the results in this study were not significantly improved after addressing the issue of imbalanced data. Therefore, various additional algorithms should be used to address the imbalanced data in future studies.

Conclusions

Adverse outcomes during the early stages for febrile patients could be predicted using a time adaptive model and machine learning approach based on the highly sparse data from test order status and laboratory test results.

Authors' Contributions

WCC and KK designed the concept of study, and WCC acquired the data. SL and KK conducted all data analysis, and all authors interpreted the results. SL, WCC, and KK wrote the draft, and all authors contributed to critical revision of the manuscript. WCC and KK contributed equally as corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental code for developing models (OSO).
[PDF File (Adobe PDF File), 83 KB - medinform_v8i3e16117_app1.pdf]

Multimedia Appendix 2
Number of laboratory log per person.
[PDF File (Adobe PDF File), 255 KB - medinform_v8i3e16117_app2.pdf]

Multimedia Appendix 3
Performance of the OSO and OSR models.
[PDF File (Adobe PDF File), 143 KB - medinform_v8i3e16117_app3.pdf]

Multimedia Appendix 4
Receiver operating characteristic (ROC) curves for models.
[PDF File (Adobe PDF File), 322 KB - medinform_v8i3e16117_app4.pdf]

Multimedia Appendix 5
Important variables selected from developed models.
[PDF File (Adobe PDF File), 423 KB - medinform_v8i3e16117_app5.pdf]

Multimedia Appendix 6
Model performance without and with applying imbalance-easing techniques.
[PDF File (Adobe PDF File), 211 KB - medinform_v8i3e16117_app6.pdf]

References

1. International Federation for Emergency Medicine. Framework for quality and safety in the emergency department. West Melbourne: International Federation for Emergency Medicine; 2012.
2. Sun B, Ma W, Zhao H. An approach to emergency decision making based on decision-theoretic rough set over two universes. Soft Comput 2015 May 31;20(9):3617-3628. [doi: 10.1007/s00500-015-1721-6]
3. Sikaris K. Enhancing the Clinical Value of Medical Laboratory Testing. Clin Biochem Rev 2017 Nov;38(3):107-114 [FREE Full text] [Medline: 29332975]
4. Vroonhof K, van Solinge WW, Rovers M, Huisman A. Differences in mortality on the basis of laboratory parameters in an unselected population at the Emergency Department. Clin Chem Lab Med 2005;43(5):536-541. [doi: 10.1515/CCLM.2005.093] [Medline: 15899676]
5. Froom P, Shimoni Z. Prediction of hospital mortality rates by admission laboratory tests. Clin Chem 2006 Feb 23;52(2):325-328. [doi: 10.1373/clinchem.2005.059030] [Medline: 16449218]
6. Jalili M, Shalileh K, Mojtahed A, Mojtahed M, Moradi-Lakeh M. Identifying causes of laboratory turnaround time delay in the emergency department. Arch Iran Med 2012 Dec;15(12):759-763 [FREE Full text] [Medline: 23199248]
7. Eisenberg EM, Murphy AG, Sutcliffe K, Wears R, Schenkel S, Perry S, et al. Communication in Emergency Medicine: Implications for Patient Safety. Communication Monographs 2005 Dec;72(4):390-413. [doi: 10.1080/03637750500322602]
8. Yadav P, Steinbach M, Kumar V, Simon G. Mining Electronic Health Records (EHRs). ACM Comput. Surv 2018 Jan 12;50(6):1-40. [doi: 10.1145/3127881]
9. Awad A, Bader-El-Den M, McNicholas J, Briggs J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. Int J Med Inform 2017 Dec;108:185-195. [doi: 10.1016/j.ijmedinf.2017.10.002] [Medline: 29132626]
10. Masud M, Al HA. Mortality prediction of ICU patients using lab test data by feature vector compaction & classification. 2017 Feb 06 Presented at: IEEE International Conference on Big Data (Big Data). IEEE Dec; 5-8 Dec. 2016; Washington, DC, USA p. 3404-3411. [doi: 10.1109/bigdata.2016.7841001]
11. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. Acad Emerg Med 2016 Mar 13;23(3):269-278 [FREE Full text] [doi: 10.1111/acem.12876] [Medline: 26679719]
12. Mohammed MA, Rudge G, Watson D, Wood G, Smith GB, Prytherch DR, et al. Index blood tests and national early warning scores within 24 hours of emergency admission can predict the risk of in-hospital mortality: a model development and validation study. PLoS One 2013 May 29;8(5):e64340 [FREE Full text] [doi: 10.1371/journal.pone.0064340] [Medline: 23734195]
13. Redfern OC, Pimentel MA, Prytherch D, Meredith P, Clifton DA, Tarassenko L, et al. Predicting in-hospital mortality and unanticipated admissions to the intensive care unit using routinely collected blood tests and vital signs: Development and validation of a multivariable model. Resuscitation 2018 Dec;133:75-81 [FREE Full text] [doi: 10.1016/j.resuscitation.2018.09.021] [Medline: 30253229]
14. Kristensen M, Iversen AKS, Gerds TA, Østervig R, Linnet JD, Barfod C, et al. Routine blood tests are associated with short term mortality and can improve emergency department triage: a cohort study of >12,000 patients. Scand J Trauma Resusc Emerg Med 2017 Nov 28;25(1):115 [FREE Full text] [doi: 10.1186/s13049-017-0458-x] [Medline: 29179764]
15. Póvoa P, Almeida E, Moreira P, Fernandes A, Mealha R, Aragão A, et al. C-reactive protein as an indicator of sepsis. Intensive Care Med 1998 Oct 20;24(10):1052-1056. [doi: 10.1007/s001340050715] [Medline: 9840239]
16. Ruohonen T, Neittaanmaki P, Teittinen J. Simulation model for improving the operation of the emergency department of special health care. 2007 Mar 05 Presented at: Proceedings of the Winter Simulation Conference. IEEE; 3-6 Dec. 2006; Monterey, CA, USA p. 453-458. [doi: 10.1109/wsc.2006.323115]
17. Liaw A, Wiener M. Classification and regression by randomForest. R News 2002 Dec;2(3):18-22 [FREE Full text]
18. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/bf00994018]
19. 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]
20. Tikhonov A, Goncharsky A, Stepanov V, Yagola A. Numerical methods for the solution of ill-posed problems. In: Berlinpringer. Berlin/Heidelberg, Germany: Springer Science & Business Media, 2013; 2013:978-994.

21. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Statistical Soc B 2005 Apr;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]

22. 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):7-9 [FREE Full text] [doi: 10.1258/shorts.2012.012089] [Medline: 23413409]

23. Leening MJ, Vedder MM, Witteman JC, Pencina MJ, Steyerberg EW. Net Reclassification Improvement: Computation, Interpretation, and Controversies. Ann Intern Med 2014 Jan 21;160(2):122-131. [doi: 10.7326/m13-1522]

24. Efron B. The Jackknife, the Bootstrap, and Other Resampling Plans. Philadelphia: Society for Industrial and Applied Mathematics (SIAM); 1982.

25. Subbe C, 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]

26. Liu J, Chen XX, Fang L, Li JX, Yang T, Zhan Q, et al. Mortality prediction based on imbalanced high-dimensional ICU big data. Computers in Industry 2018 Jun;98:218-225. [doi: 10.1016/j.compind.2018.01.017]

27. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. IDA 2002 Nov 15;6(5):429-449. [doi: 10.3233/ida-2002-6504]

28. Haibo He, Garcia E. Learning from Imbalanced Data. IEEE Trans. Knowl. Data Eng 2009 Sep;21(9):1263-1284. [doi: 10.1109/tkde.2008.239]

29. R Core Team. R Foundation for Statistical Computing. Vienna, Austria; 2019. R: A Language and Environment for Statistical Computing URL: https://www.R-project.org/ [accessed 2020-03-02]

30. Bossink AW, Groeneveld AJ, Hack CE, Thijs LG. Prediction of mortality in febrile medical patients: How useful are systemic inflammatory response syndrome and sepsis criteria? Chest 1998 Jun;113(6):1533-1541. [doi: 10.1378/chest.113.6.1533] [Medline: 9631790]

31. Young PJ, Bellomo R. Fever in sepsis: is it cool to be hot? Crit Care 2014 Feb 13;18(1):109 [FREE Full text] [doi: 10.1186/cc13726] [Medline: 24521542]

32. Christensen D, Maaløe R, Jensen NM, Rudolph SS, Perrild H. Quality of care using a multidisciplinary team in the emergency room. Dan Med Bull 2011 Jun;58(6):A4227. [Medline: 21651874]

33. Epstein N. Multidisciplinary in-hospital teams improve patient outcomes: A review. Surg Neurol Int 2014;5(Suppl 7):S295-S303 [FREE Full text] [doi: 10.4103/2152-7806.139612] [Medline: 25289149]

34. Jenkins V, Fallowfield L, Poole K. Are members of multidisciplinary teams in breast cancer aware of each other's informational roles? Qual Health Care 2001 Jun;10(2):70-75 [FREE Full text] [doi: 10.1136/qhc.10.2.70] [Medline: 11389314]

35. Firth-Cozens J. Multidisciplinary teamwork: the good, bad, and everything in between. Qual Health Care 2001 Jun;10(2):65-66 [FREE Full text] [doi: 10.1136/qhc.10.2.65] [Medline: 11389310]

Abbreviations

AUC: area under the receiver operating characteristic curve.
AUPRC: area under the precision recall curve.
BA: balanced accuracy.
CRP: C reactive protein.
ED: emergency department.
EN: elastic net.
Hb: hemoglobin.
MEWS: Modified Early Warning Score.
NLR: negative likelihood ratio.
OSO: order status only.
OSR: order status and result.
PLR: positive likelihood ratio.
RAR: number of rarely detected results.
RF: random forest.
ROT: number of rarely ordered tests.
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis.

XSL•FO
RenderX

Original Paper

Peak Outpatient and Emergency Department Visit Forecasting for Patients With Chronic Respiratory Diseases Using Machine Learning Methods: Retrospective Cohort Study

Junfeng Peng[1*], BSc; Chuan Chen[1*], PhD; Mi Zhou[2], MD; Xiaohua Xie[1], PhD; Yuqi Zhou[3], MD; Ching-Hsing Luo[1], MD, PhD

[1]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

[2]Surgical Intensive Care Unit, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

[3]Department of Respiratory and Critical Care Medicine, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Ching-Hsing Luo, MD, PhD
School of Data and Computer Science
Sun Yat-sen University
National Supercomputing Guangzhou Center Bldg, 5th Fl
Guangzhou, 510006
China
Phone: 86 13265384134
Email: luojinx5@mail.sysu.edu.cn

Abstract

Background: The overcrowding of hospital outpatient and emergency departments (OEDs) due to chronic respiratory diseases in certain weather or under certain environmental pollution conditions results in the degradation in quality of medical care, and even limits its availability.

Objective: To help OED managers to schedule medical resource allocation during times of excessive health care demands after short-term fluctuations in air pollution and weather, we employed machine learning (ML) methods to predict the peak OED arrivals of patients with chronic respiratory diseases.

Methods: In this paper, we first identified 13,218 visits from patients with chronic respiratory diseases to OEDs in hospitals from January 1, 2016, to December 31, 2017. Then, we divided the data into three datasets: weather-based visits, air quality-based visits, and weather air quality-based visits. Finally, we developed ML methods to predict the peak event (peak demand days) of patients with chronic respiratory diseases (eg, asthma, respiratory infection, and chronic obstructive pulmonary disease) visiting OEDs on the three weather data and environmental pollution datasets in Guangzhou, China.

Results: The adaptive boosting-based neural networks, tree bag, and random forest achieved the biggest receiver operating characteristic area under the curve, 0.698, 0.714, and 0.809, on the air quality dataset, the weather dataset, and weather air quality dataset, respectively. Overall, random forests reached the best classification prediction performance.

Conclusions: The proposed ML methods may act as a useful tool to adapt medical services in advance by predicting the peak of OED arrivals. Further, the developed ML methods are generic enough to cope with similar medical scenarios, provided that the data is available.

Introduction

Worldwide, one of the fundamental issues in hospital management is the sudden inflow of outpatient and emergency department (OED) patients [1]. Influenza season (epidemic period) is one of the causes for OED overcrowding and generates a large flow of patients [2]. In particular, weather and air quality are important factors that affect the health status of

individuals and populations with chronic respiratory diseases [3]. Chronic respiratory diseases such as asthma and chronic obstructive pulmonary disease (COPD) often require regular OED medication as the condition changes, which can cause further OED overcrowding [4]. Nevertheless, the crowding could be alleviated and mitigated considerably by forecasting levels of demand for OED care and giving health care staff an opportunity to prepare for this demand [5]. Efficient patient flow has been proven to potentially increase the capacity of the existing system, minimize patient care delays, and improve overall quality of health care [6-10].

There have been many attempts to predict daily patient volumes visiting emergency departments (EDs) using machine learning (ML) and deep learning models based on weather and air quality [11,12].

Bibi et al [13] created a computer-based model called an artificial neural network (ANN) using a backpropagation to predict volumes of ED visits of patients with asthma, COPD, or acute or chronic bronchitis 7 days in advance. The study included a dataset (1020 days of ED activity) extracted from an ED admittance database at the Barzilai Medical Center (Ashkelon, Israel). The mode integrated 5 indicators (ie, temperature, relative humidity, barometric pressure, sulfur dioxide, and nitrogen oxide) and achieved the prediction accuracy with an average error of 12%. However, indicators and data collections are relatively inadequate.

Moustris et al [14] developed three different ANN models to forecast the childhood asthma admissions 7 days in advance for the subgroups of 0 to 4 years of age and 5 to 14 years of age, as well as for the whole study population. The study used 6 indicators, that is ozone, carbon monoxide, PM10 (particulate matter of 10 μm in diameter or smaller), PM25 (particulate matter less than 2.5 μm in diameter), and sulfur dioxide, from Athens, Greece to train the ANN model. The evaluation of the three ANN models' forecasting abilities on the root mean square error (mean bias error) were 6.8 (1.4), 3.2 (1.3), and 5.2 (0.3) for 0 to 4 years of age, 5 to 14 years of age, and the whole study population, respectively. However, the study only took into account air quality indicators and ignored weather factors.

Soyiri [15] explored the base and reduced predictive quantile regression models (QRMs) to detect peak numbers of daily asthma admissions in London with sensitivity levels of 76% and 62%; as well as specificities of 66% and 76%, respectively. The research used 10 indicators (ie, air temperature, vapor pressure, humidity, ozone, carbon monoxide, nitrogen dioxide, nitrogen oxide, PM10, and formaldehyde) to build the QRMs. The findings also reaffirmed the known associations between asthma and temperature, and ozone and carbon monoxide levels. Nevertheless, the accuracy of the model is not very high.

Khatri et al [16] employed an ANN–based classifier using multilayered perceptions with a backpropagation algorithm that predicts peak events, that is days of peak demand, for patients with respiratory diseases. The study used 8 predictors (ie, outdoor temperature, relative humidity, wind speed, carbon monoxide, ozone, sulphur dioxide, nitrogen dioxide, and PM25) to construct the model. The proposed ANN model achieved a good forecasting performance with the overall accuracy of the

system at 81.0%. Even so, the study population only included visits for respiratory diseases data in EDs. Further, the research did not consider dividing data into weather and air pollution separately.

Yucesan et al [17] developed a multi-method patient arrival forecasting outline for EDs using a private hospital ED case in Turkey. The methods followed within this study include the single methods linear regression (LR), autoregressive integrated moving average (ARIMA), ANN, exponential smoothing, and the hybrid methods ARIMA-ANN and ARIMA-LR. The ARIMA-ANN hybrid model is shown to outperform in terms of forecasting accuracy. This study explored a novel attempt of applying these methods to model ED patient arrivals and making an overall assessment among them.

Muhammet et al [18] analyzed variations in annual, monthly, and daily ED arrivals based on regression and neural network models with the aid of collected data from a public hospital ED in Istanbul. Both of the methods have been proven to be useful and readily available tools for forecasting ED patient arrivals. The results show that ANN–based models have higher model accuracy values and lower values of absolute error in terms of forecasting ED patient arrivals over the long- and medium-term. The value of the standard error of regression for the ANN modeling, which is 30.022306, refers to the difference between the real ED patient arrivals and the forecasted ED patient arrivals per day covering the total of the three patient groups.

Although ED forecasting has attracted many researchers, we found few studies designed to predict OED visits of patients with chronic respiratory diseases using multiple ML methods. In a real medical scene, patients with chronic respiratory diseases often go to outpatient clinics. Therefore, it would be of great significance to forecast the peak OED visits for chronic respiratory diseases.

In this paper, we employed bagging [19], adaptive boosting [20] and random forest [21] algorithms to predict the peak arrival of chronic respiratory disease OED visits based on the weather and air quality data. Meanwhile, we compared the ensemble models with the general linear model (GLM) [22] and the polynomial nuclear support vector machine (SVM) [23]. The results show that ensemble models outperform the GLMs and SVM. Further, we found that the predictive performance of ML algorithms gradually improves with the increase of input features. By the ML approaches, the OED managers can plan resources to meet the excessive demand of patients with respiratory diseases after short-term fluctuations in air pollution or weather.

Methods

Data Acquisition

Figure 1 shows the flowchart of participants in our research. We identified 13,208 OED visits to the Second Affiliated Hospital of Guangzhou Medical University that had a major diagnosis of a chronic respiratory disease defined by the International Classification of Diseases, Tenth Revision, Clinical Modification code (J45.900, J44.001, J44.101, J44.803, and J98.801). The duration of the collected data lasted from January

1, 2016, to December 31, 2017, which is 731 days of continuous data. For statistical purposes, the days where the daily volume was less than 24 were labeled as nonpeak events, and the rest were labeled as peak events.

Table 1 describes the Pearson correlation coefficient between OED visit numbers and input indicators. We found that OED

visit numbers showed positive correlations with wind speed, atmospheric pressure, carbon monoxide, sulphur dioxide, nitrogen dioxide, and PM25. However, OED visit numbers showed negative correlations with outdoor temperature, relative humidity, and ozone. The weather and air quality data distribution of patients with acute exacerbations of COPD from peak and nonpeak groups was shown in Table 2.

Figure 1. Flowchart of participants. ICD-10-CM: International Classification of Diseases, 10th revision, Clinical Modification.
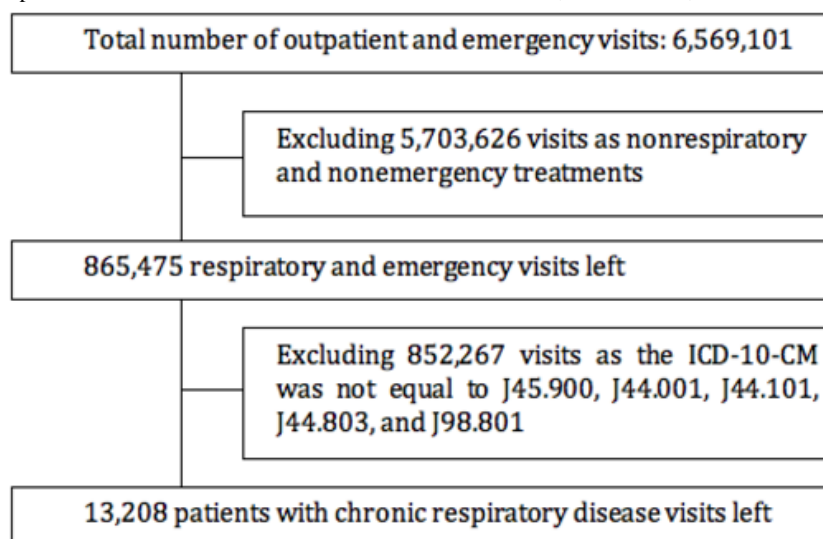

Table 1. The Pearson correlation coefficients between outpatient and emergency department visit numbers and input indicators.

Variable	WS[a], r	TP[b], r	AP[c], r	RH[d], r	PM25[e], r	SO$_2$[f], r	CO[g], r	NO$_2$[h], r	O$_3$_8h[i], r	Number of visits, r
WS	1	–0.32	0.27	–0.4	–0.34	–0.33	–0.26	–0.42	–0.24	0.15
TP	–0.32	1	–0.88	0.35	–0.23	0.03	–0.24	–0.25	0.39	–0.38
AP	0.27	–0.88	1	–0.5	0.31	0.09	0.21	0.29	–0.18	0.39
RH	–0.4	0.35	–0.5	1	–0.18	–0.27	0.2	0.03	–0.28	–0.2
PM25	–0.34	–0.23	0.31	–0.18	1	0.73	0.65	0.81	0.29	0.29
SO$_2$	–0.33	0.03	0.09	–0.27	0.73	1	0.35	0.66	0.43	0.22
CO	–0.26	–0.24	0.21	0.21	0.65	0.35	1	0.68	–0.07	0.35
NO$_2$	–0.42	–0.25	0.29	0.03	0.81	0.66	0.68	1	0.13	0.35
O$_3$_8h	–0.24	0.39	–0.18	–0.28	0.29	0.43	–0.07	0.13	1	–0.14
Number of visits	0.15	–0.38	0.39	–0.2	0.29	0.22	0.35	0.35	–0.14	1

[a]WS: wind speed.

[b]TP: outside temperature.

[c]AP: atmospheric pressure.

[d]RH: relative humidity.

[e]PM25: particulate matter less than 2.5 μm in diameter.

[f]SO$_2$: sulphur dioxide.

[g]CO: carbon monoxide.

[h]NO$_2$: nitrogen dioxide.

[i]O$_3$_8h: 8-hour average ozone slip in a day.

Table 2. Weather and air quality data distribution of peak and nonpeak groups visiting outpatient and emergency departments.

Variables	Peak group, mean (SD)	Nonpeak group, mean (SD)
Wind speed (m/sec)	2.49 (1.10)	2.15 (0.91)
Outside temperature (°C)	17.81 (5.59)	23.11 (5.81)
Atmosphere pressure (mb)	1009.99 (5.26)	1003.73 (6.57)
Relative humidity (%)	77 (12.51)	82.15 (9.65)
Particulate matter less than 2.5 μm in diameter	43.74 (23.69)	32.83 (16.49)
Sulphur dioxide	13.16 (4.65)	11.45 (3.73)
Carbon monoxide	1.06 (0.25)	0.92 (0.17)
Nitrogen dioxide	60.05 (26.09)	46.43 (17.67)
8-hour average ozone slip in a day	74.28 (54.90)	90.24 (52.46)

Data Analysis

Since the effect of weather and air quality on respiratory conditions in humans was not instantaneous, representative lags were applied to variables based on the work done previously in this area [3,24-26]. To simplify the delayed impact of respiratory conditions, we considered a 3-day lag for all variables.

We removed records with less than 10 people on weekends to eliminate weekend effects, bringing the total number of samples collected to 559. To create a meaningful feature vector for training and cross-validation, the date field was removed to obtain a (X, y), where X was a matrix with the dimensions ($m \times n = 559 \times 9$) representing values of variables, and y was a vector of length (m=559) representing the output class of the examples (ie, events). Analysis of the data suggested that the output class was highly imbalanced with 413 examples of nonpeak and 146 examples of peak events.

Machine Learning Approaches

In this section the ML algorithms are presented and discussed; details of the updating and classification processes are described in the following algorithms.

Generalized Linear Models

1. Construct the common linear model from the original training set: $f = w^T x + b$, where w is the weight vector and b is the bias, both of which are only determined by the training samples
2. Identify the contact function f^{-1}
3. Build the GLMs: $= f^{-1}(w^T x + b)$
4. Calculate the classification on the test set

Support Vector Machine

1. Convert the sample space into linearly separable space with polynomial core functions $K(x_i, y_i)$
2. Calculate the support vectors with the following formula:

3. Then identify the hyperplane. The regular parameter C is a penalty factor, which can balance the model complexity and empirical risk. In addition, ε_i indicates the positive parameters called slack variables, which represent the

distance between the misclassified sample and the optimal hyperplane.
4. Forecast the classification of the test dataset using hyperplane and support vectors

Bagging

1. Generate a new training set by sampling from the original training set
2. Repeat step 1 N times to get the N new training sets, and train N trees in N different training sets
3. Calculate the classification results by averaging the predicted value of each tree or use the majority
4. Out-of-bag error estimation: The data not sampled in step 1 is used as the test set of the corresponding generated tree to evaluate the predicted results

Random Forest

1. Create a new training set from a sample of a training set
2. Repeat step 1 N times to get N new training sets, and train N trees on the training sets
3. Identify the optimal candidate node as the prediction space from the randomly selected m feature set when building the tree model

Boosting

1. Initialize the weight vector of the training data
2. Construct m weak classifiers
3. Calculate the classification error rate of the m weak classifiers
4. If one sample is misclassified, its weight will be increased, and the next weak classifier pays more attention to this sample; otherwise, its weight will be decreased.
5. After all the weak classifiers finish the training, the stronger classifier is constructed.

Results

Metrics

Precision, recall, and F measure are the metrics that are used to evaluate our proposed ML methods. Based on the classification of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), we have the following formulas.

We then define the F measure, a metric that balances precision and recall.

Evaluation

We calculate the overall accuracy, precision, recall, and F measure for nonpeak events and peak events, respectively. Evaluation of the ML approaches on the weather and air quality data are shown in Table 3. It showed that the developed random forest gave the best predictive performance. This was mainly due to the data collection fitting better with the random forest.

Table 3. Evaluation of machine learning approaches on weather and air quality.

Machine learning approaches	F1 measure	Accuracy, % (n/N)
Generalized linear model		85.6 (479/559)
Peak	0.667	
Nonpeak	0.908	
Support vector machine		80.2 (448/559)
Peak	0.289	
Nonpeak	0.882	
Adaptive boosting neural networks		84.7 (473/559)
Peak	0.667	
Nonpeak	0.900	
Tree bag		83.8 (468/559)
Peak	0.640	
Nonpeak	0.895	
Random forest		88.3 (494/559)
Peak	0.745	
Nonpeak	0.924	

In addition, we used the receiver operating characteristic (ROC) curve to evaluate the multiple ML approaches on the same dataset (Table 4). We found that adaptive boosting neural networks achieved the biggest ROC area under the curve on the air quality data, tree bag on the climate data, and random forest on weather and air quality data. In general, we discovered that the predictive performance of the ML approaches improves as data variables increase.

Table 4. Evaluation of machine learning approaches using receiver operating characteristic.

| Machine learning approaches	Weather, AUC[a]	Air quality, AUC	Weather and air quality, AUC
Generalized linear model	0.538	0.682	0.758
Support vector machine	0.500	0.494	0.621
Adaptive boosting neural network	0.611	0.698	0.734
Tree bag	0.714	0.680	0.780
Random forest	0.669	0.692	0.809

[a]AUC: area under the curve.

Discussion

Clinical Significance

Recent studies have shown that weather and air pollution have been a major problem leading to an increase in daily deaths and hospital admissions for chronic respiratory illnesses [3-5,27,28].

We focused the distribution of daily patient visits for 2 years (ie, 2016 and 2017) (Figure 2). It is worth noting that peak days are more dominant from October to March, which indicates that the haze is a strong predictor, as these months are mostly colder in Guangzhou. Thus, it is important to recognize the peak OED visits for respiratory conditions.

Figure 2. Histogram of patients visiting outpatient and emergency rooms.
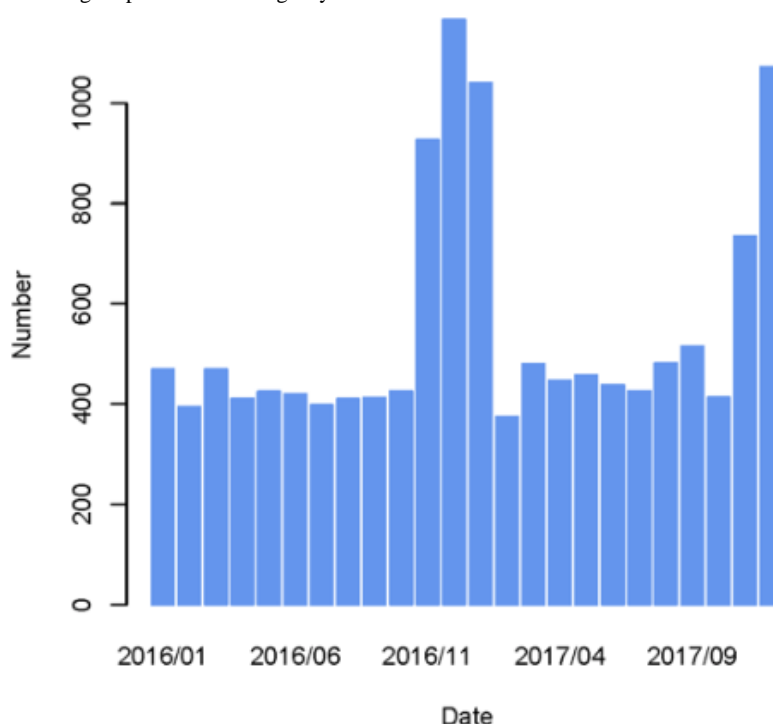
Previous studies mainly focused on the peak event forecasting ED visits for patients with one or more diseases. We expanded the study population to include outpatient visits for patients with chronic respiratory diseases. In fact, many patients with chronic respiratory diseases also seek treatment from outpatient departments. Thus, predicting the OED peak visits for chronic respiratory disease plays an important role in clinical management.

We developed a variety of learning methods to forecast the OED peak visits, from simple models to complex ensemble learning ones. In particular, the ensemble learning models achieved good prediction results. In terms of indicators, most of the previous studies used air pollution indicators to predict the peak events of ED visits; however, we used weather and air quality indicators to build a more complete set of features.

Limitations

There are a few limitations to this study. In this study, we used nine variables, namely, wind speed, atmospheric pressure, outdoor temperature, relative humidity, carbon monoxide, ozone, sulphur dioxide, nitrogen dioxide, and PM25, as these variables have been associated with exacerbation of respiratory diseases. However, there are some other variables that also contribute to the exacerbation of these diseases, such as formaldehyde and nitrogen oxide [29]. The Environmental Protection Agency of Guangzhou does not disclose the daily data for variables such as formaldehyde. Other pollutants are either not measured or

had too many missing values. Therefore, we were not able to include these variables in our study.

In terms of weather, Guangzhou as a coastal city in southern China has a higher air humidity than other northern cities. In terms of air pollution, some studies have shown that patients with lower levels of economics and education are more susceptible to air pollution [30]. Guangzhou has a significantly higher economic and educational level than the national average. However, the pollution of haze and the harmful emissions of Guangzhou are also serious [31]. In particular, the lighter particulate matter is higher than other northern cities due to automobile exhaust and industrial emissions. Therefore, the prediction result of this study may not be directly applicable to other regions due to the regional differences in climate and air pollution.

Conclusion

In this paper, we investigated ML methods to forecast the peak events of patients with chronic respiratory diseases visiting OEDs combined with nine weather and air quality predictors. Overall, random forest outperforms the other methods in the accuracy, F measure, and ROC on the validation dataset. Compared with similar studies before, we used more indicators and ML methods to study the subject and achieved good results. The ML methods may act as a useful tool to adapt medical services in advance by predicting the peak number of OED arrivals.

Conflicts of Interest

None declared.

References

1. Kadri F, Pach C, Chaabane S, Berger T, Trentesaux D, Tahon C, et al. Modelling and management of strain situations in hospital systems using an ORCA approach. 2013 Mar 13 Presented at: Proceedings of 2013 International Conference on Industrial Engineering & Systems Management; 2013; Morocco.

2. Schull MJ, Mamdani MM, Fang J. Influenza and emergency department utilization by elders. Acad Emerg Med 2005 Apr;12(4):338-344 [FREE Full text] [doi: 10.1197/j.aem.2004.11.020] [Medline: 15805325]

3. Soyiri IN, Reidpath DD. Evolving forecasting classifications and applications in health forecasting. Int J Gen Med 2012;5:381-389 [FREE Full text] [doi: 10.2147/IJGM.S31079] [Medline: 22615533]

4. Wargon M, Guidet B, Hoang TD, Hejblum G. A systematic review of models for forecasting the number of emergency department visits. Emerg Med J 2009 Jun 22;26(6):395-399. [doi: 10.1136/emj.2008.062380] [Medline: 19465606]

5. Davidson SJ, Koenig KL, Cone DC. Daily patient flow is not surge: "management is prediction". Acad Emerg Med 2006 Nov;13(11):1095-1096 [FREE Full text] [doi: 10.1197/j.aem.2006.07.021] [Medline: 17085737]

6. Wargon M, Casalino E, Guidet B. From model to forecasting: a multicenter study in emergency departments. Acad Emerg Med 2010 Sep 22;17(9):970-978 [FREE Full text] [doi: 10.1111/j.1553-2712.2010.00847.x] [Medline: 20836778]

7. Gul M, Celik E. An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. Health Systems 2018 Nov 19:1-22. [doi: 10.1080/20476965.2018.1547348]

8. Jones S, Thomas A, Evans R, Welch S, Haug P, Snow GL. Forecasting daily patient volumes in the emergency department. Acad Emerg Med 2008 Feb;15(2):159-170 [FREE Full text] [doi: 10.1111/j.1553-2712.2007.00032.x] [Medline: 18275446]

9. Gul M, Guneri AF. A comprehensive review of emergency department simulation applications for normal and disaster conditions. Computers & Industrial Engineering 2015 May;83:327-344. [doi: 10.1016/j.cie.2015.02.018]

10. Gul M, Guneri AF. Forecasting patient length of stay in an emergency department by artificial neural networks. J Aeronaut Space Technol 2015 Oct 7;8(2). [doi: 10.7603/s40690-015-0015-7]

11. McCarthy ML, Zeger SL, Ding R, Aronsky D, Hoot NR, Kelen GD. The challenge of predicting demand for emergency department services. Acad Emerg Med 2008 Apr;15(4):337-346 [FREE Full text] [doi: 10.1111/j.1553-2712.2008.00083.x] [Medline: 18370987]

12. Boyle J, Jessup M, Crilly J, Green D, Lind J, Wallis M, et al. Predicting emergency department admissions. Emerg Med J 2012 May;29(5):358-365. [doi: 10.1136/emj.2010.103531] [Medline: 21705374]

13. Bibi H, Nutman A, Shoseyov D, Shalom M, Peled R, Kivity S, et al. Prediction of emergency department visits for respiratory symptoms using an artificial neural network. Chest 2002 Nov;122(5):1627-1632. [doi: 10.1378/chest.122.5.1627] [Medline: 12426263]

14. Moustris KP, Douros K, Nastos PT, Larissi IK, Anthracopoulos MB, Paliatsos AG, et al. Seven-days-ahead forecasting of childhood asthma admissions using artificial neural networks in Athens, Greece. Int J Environ Health Res 2012;22(2):93-104. [doi: 10.1080/09603123.2011.605876] [Medline: 21854178]

15. Soyiri IN, Reidpath DD, Sarran C. Forecasting peak asthma admissions in London: an application of quantile regression models. Int J Biometeorol 2013 Jul;57(4):569-578. [doi: 10.1007/s00484-012-0584-0] [Medline: 22886344]

16. Khatri KL, Tamil LS. Early detection of peak demand days of chronic respiratory diseases emergency department visits using artificial neural networks. IEEE J Biomed Health Inform 2018 Jan;22(1):285-290. [doi: 10.1109/JBHI.2017.2698418] [Medline: 28459697]

17. Yucesan M, Gul M, Celik E. A multi-method patient arrival forecasting outline for hospital emergency departments. International Journal of Healthcare Management 2018 Oct 10:1-13. [doi: 10.1080/20479700.2018.1531608]

18. Gul M, Guneri AF. Planning the future of emergency departments: forecasting ED patient arrivals by using regression and neural network models. International Journal of Industrial Engineering 2016;23(2):137-154.

19. Breiman L. Bagging predictors. Mach Learn 1996 Aug;24(2):123-140. [doi: 10.1007/BF00058655]

20. Freund Y, Schapire R. Experiments with a new boosting algorithm. 1996 Jul 3 Presented at: Proceedings of the Thirteenth International Conference on Machine Learning; 1996; Bari.

21. Ho T. Random decision forests. 1995 Aug Presented at: Proceedings of 3rd International Conference on Document Analysis and Recognition; 1995; Montreal. [doi: 10.1109/ICDAR.1995.598994]

22. Nelder JA, Wedderburn RWM. Generalized Linear Models. Journal of the Royal Statistical Society 1972;135(3):370. [doi: 10.2307/2344614]

23. Vapnik VN. Statistical Learning Theory. New York, NY: Wiley; 1998.

24. Schwartz J. Short term fluctuations in air pollution and hospital admissions of the elderly for respiratory disease. Thorax 1995 May;50(5):531-538 [FREE Full text] [doi: 10.1136/thx.50.5.531] [Medline: 7597667]

25. Giovannini M, Sala M, Riva E, Radaelli G. Hospital admissions for respiratory conditions in children and outdoor air pollution in Southwest Milan, Italy. Acta Paediatr 2010 Aug;99(8):1180-1185. [doi: 10.1111/j.1651-2227.2010.01786.x] [Medline: 20219045]

26. Soyiri IN, Reidpath DD, Sarran C. Forecasting peak asthma admissions in London: an application of quantile regression models. Int J Biometeorol 2013 Jul;57(4):569-578. [doi: 10.1007/s00484-012-0584-0] [Medline: 22886344]

27. World Health Organization, Regional Office of Europe. Air quality guidelines: global update 2005. Particulate matter, ozone, nitrogen dioxide and sulfur dioxide. Indian Journal of Medical Research 2007;4(4):493.

XSL•FO
RenderX

28.   Launay F. 7 million deaths annually linked to air pollution. Cent Eur J Public Health 2014 Mar;22(1):53-59. [Medline: 24844108]

29.   Hulin M, Simoni M, Viegi G, Annesi-Maesano I. Respiratory health and indoor air pollutants based on quantitative exposure assessments. Eur Respir J 2012 Oct;40(4):1033-1045 [FREE Full text] [doi: 10.1183/09031936.00159011] [Medline: 22790916]

30.   Li L, Yang J, Song Y, Chen P, Ou C. The burden of COPD mortality due to ambient air pollution in Guangzhou, China. Sci Rep 2016 May 19;6:25900 [FREE Full text] [doi: 10.1038/srep25900] [Medline: 27195597]

31.   Lin H, Wang X, Liu T, Li X, Xiao J, Zeng W, et al. Air Pollution and Mortality in China. Adv Exp Med Biol 2017;1017:103-121. [doi: 10.1007/978-981-10-5657-4_5] [Medline: 29177960]

Abbreviations

ANN: artificial neural network

ARIMA: autoregressive integrated moving average

COPD: chronic obstructive pulmonary disease

ED: emergency department

FN: false negatives

FP: false positives

GLM: general linear model

LR: linear regression

ML: machine learning

OED: outpatient and emergency department

PM10: particulate matter of 10 μm in diameter or smaller

PM25: particulate matter less than 2.5 μm in diameter

QRM: quantile regression model

ROC: receiver operating characteristic curve

SVM: support vector machine

TN: true negatives

TP: true positives

XSL•FO
RenderX

Original Paper

Detecting Potential Medication Selection Errors During Outpatient Pharmacy Processing of Electronic Prescriptions With the RxNorm Application Programming Interface: Retrospective Observational Cohort Study

Corey A Lester[1], PhD, PharmD; Liyun Tu[2]; Yuting Ding[1], MS; Allen J Flynn[3], PhD, PharmD

[1]Department of Clinical Pharmacy, College of Pharmacy, University of Michigan, Ann Arbor, MI, United States

[2]Department of Electrical Engineering and Computer Science, College of Engineering, University of Michigan, Ann Arbor, MI, United States

[3]Department of Learning Health Sciences, School of Medicine, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
Corey A Lester, PhD, PharmD
Department of Clinical Pharmacy
College of Pharmacy
University of Michigan
428 Church Street
Ann Arbor, MI, 48109
United States
Phone: 1 734 647 8849
Email: lesterca@umich.edu

Abstract

Background: Medication errors are pervasive. Electronic prescriptions (e-prescriptions) convey secure and computer-readable prescriptions from clinics to outpatient pharmacies for dispensing. Once received, pharmacy staff perform a transcription task to select the medications needed to process e-prescriptions within their dispensing software. Later, pharmacists manually double-check medications selected to fulfill e-prescriptions before dispensing to the patient. Although pharmacist double-checks are mostly effective for catching medication selection mistakes, the cognitive process of medication selection in the computer is still prone to error because of heavy workload, inattention, and fatigue. Leveraging health information technology to identify and recover from medication selection errors can improve patient safety.

Objective: This study aimed to determine the performance of an automated double-check of pharmacy prescription records to identify potential medication selection errors made in outpatient pharmacies with the RxNorm application programming interface (API).

Methods: We conducted a retrospective observational analysis of 537,710 pairs of e-prescription and dispensing records from a mail-order pharmacy for the period January 2017 to October 2018. National Drug Codes (NDCs) for each pair were obtained from the National Library of Medicine's (NLM's) RxNorm API. The API returned RxNorm concept unique identifier (RxCUI) semantic clinical drug (SCD) identifiers associated with every NDC. The SCD identifiers returned for the e-prescription NDC were matched against the corresponding SCD identifiers from the pharmacy dispensing record NDC. An error matrix was created based on the hand-labeling of mismatched SCD pairs. Performance metrics were calculated for the e-prescription-to-dispensing record matching algorithm for both total pairs and unique pairs of NDCs in these data.

Results: We analyzed 527,881 e-prescription and pharmacy dispensing record pairs. Four clinically significant cases of mismatched RxCUI identifiers were detected (ie, three different ingredient selections and one different strength selection). A total of 546 less significant cases of mismatched RxCUIs were found. Nearly all of the NDC pairs had matching RxCUIs (28,787/28,817, 99.90%-525,270/527,009, 99.67%). The RxNorm API had a sensitivity of 1, a false-positive rate of 0.00104 to 0.00312, specificity of 0.99896 to 0.99688, precision of 0.00727 to 0.04255, and F1 score of 0.01444 to 0.08163. We found 872 pairs of records without an RxCUI.

Conclusions: The NLM's RxNorm API can perform an independent and automatic double-check of correct medication selection to verify e-prescription processing at outpatient pharmacies. RxNorm has near-comprehensive coverage of prescribed medications and can be used to recover from medication selection errors. In the future, tools such as this may be able to perform automated

XSL•FO
RenderX

verification of medication selection accurately enough to free pharmacists from having to perform manual double-checks of the medications selected within pharmacy dispensing software to fulfill e-prescriptions.

Introduction

Background

Medical error is the third leading cause of death in the United States, and medication errors are the most common type of these errors [1]. On a global basis, medication errors cost US $42 billion annually [2]. One of the medication errors with the greatest potential for harm happens when patients receive a different medication than that prescribed [3,4]. Outpatient pharmacies can dispense an incorrect medication in several ways [5]. For one, pharmacy staff members essentially transcribe e-prescription information by using software to select the medication product for dispensing based on the prescribed medication conveyed in an electronic prescription (e-prescription) [6-8]. Pharmacy software aids in the on-screen selection of a medication product from the pharmacy's medication inventory by linking the medication data transmitted with an e-prescription to closely related drug product options. To identify the prescribed drug product, e-prescriptions carry a representative National Drug Code (NDC) along with a standard drug description from a commercial drug compendium (eg, DailyMed). Sometimes, when one exists, a corresponding RxNorm concept unique identifier (RxCUI) is also included in transmitted e-prescriptions for drug identification.

There are instances when pharmacy staff enter e-prescription medication information manually, bypassing available product identification safety features [8]. In these cases, pharmacy staff can type the name of any medication held in inventory into a new prescription record. This increases the risk of entering an incorrect medication into the pharmacy's software and dispensing the wrong medication to the patient. An independent double-check by a pharmacist helps identify these medication selection errors. However, the pharmacist can still miss these errors as well because of fatigue, workload, or stress [9-12]. To support pharmacists, an independent, automated double-check of the medications selected to fulfill e-prescriptions could identify medication selection errors post hoc. The electronic transmission of e-prescriptions, coupled with electronic representations of prescriptions in pharmacy dispensing systems, enables machines to perform novel safety checks not possible in the past with handwritten or verbal prescriptions.

RxNorm, the US National Library of Medicine (NLM) drug terminology system, has the potential to detect medication selection mistakes and prevent patient harm. Bell et al [13] demonstrated the comprehensiveness of RxNorm by finding an RxCUI for all but 1 of 19,743 sample e-prescriptions. They also reported an RxCUI mismatch rate of 3.4% between e-prescription and pharmacy dispensing records. Most mismatches between the medication prescribed and dispensed were deemed not of clinical significance. One limitation of previous work is their use of downloadable .csv files of RxNorm content. RxNorm content is also available through an application programming interface (API). This API is important because it can support continuous detection of medication selection errors in pharmacy practice, making it possible to identify and resolve errors before they reach the patient. To demonstrate this, we make use of the publicly available RxNorm API to perform checks of past dispensing records.

Objective

The objective of this study was to establish an automated RxNorm API double-check and evaluate its performance as a method for detecting potential medication selection errors occurring in outpatient pharmacies. This study contributes a blueprint for how to do this with the widely available API resource from NLM.

Methods

Overview

We analyzed e-prescription and corresponding dispensing records that were transmitted through a mail-order pharmacy in the United States. NDCs for each medication were matched mainly to its associated semantic clinical drug (SCD) RxCUI identifiers using the NLM's RxNorm API. Semantic branded drug (SBD) and generic pack (GPCK) RxCUIs were also used. A direct match was performed between e-prescription and pharmacy system dispensing record SCDs. Afterward, all mismatched pairs found were hand-labeled into categories. Finally, performance metrics were applied to evaluate the e-prescription-dispensing record matching algorithm for both total pairs and unique pairs of NDCs in the data. We use a prototypical example with a specific NDC pair in this section to communicate how the system functions (ie, NDC=00093-5117-98; Diltiazem HCl ER coated beads 180 mg and NDC=00008-0841-81; Protonix 40 mg oral tablet).

Data Source

We obtained 537,710 pairs of e-prescription and pharmacy dispensing records over the period of January 2017 to October 2018. For e-prescriptions, the variables included the free-text medication name, an alpha-numeric NDC (ie, these are representative NDCs required to transmit an e-prescription), and national provider identifier (NPI). A corresponding dispensing record for data appearing on the prescription label included the free-text medication name and an alpha-numeric NDC (ie, the specific product dispensed to the patient). To determine the origin of the e-prescription, we linked our prescriber NPI to the Centers for Medicare and Medicaid's National Plan and Provider Enumeration System NPI download file (November 2018 release) [14]. We used no personally

identifiable patient-specific data. The University of Michigan Institutional Review Board reviewed our study protocol and assigned exemption status.

Data Filtering and Cleaning

To ensure we included only valid e-prescription-dispensing record pairs, we performed a series of data filtering and cleaning tasks. We removed e-prescription-dispensing record pairs containing certain signal words: *duplicate, cancel, wrong, and denied* from the data. We also removed pairs containing missing NDC data. When necessary, we padded NDC codes with leading zeros to make them conform to a standard 11-digit format. This step normalized NDC codes and prepared them for the NLM's RxNorm API. Figure 1 shows the results of our data cleaning steps.

Figure 1. Data cleaning process. e-prescription: electronic prescription; NDC: National Drug Code.
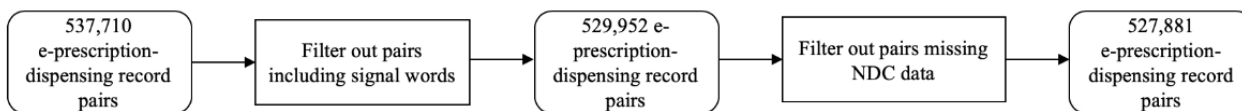

Data Analysis

We conducted a three-step analysis. The three steps of the analysis were (1) linking NDCs to SCD RxCUIs via the RxNorm API, (2) matching SCD RxCUIs in each e-prescription-dispensing record pair, and (3) calculating performance metrics. Besides reporting the results of the three-step analysis, we also report the most common medications dispensed by the mail-order pharmacy as well as the most common medication classes represented in our dataset. As the dataset came from a mail-order pharmacy, the dataset does not include injectable products exclusively used in inpatient or long-term care settings.

Moving from left to right in Figure 2, the first step in our analysis mapped the NDC conveyed by the e-prescription and the NDC on the pharmacy's dispensing record to a corresponding SCD (or SBD or GPCK or branded pack) RxCUI. The automated checking system we created called the RxNorm API for "ndcstatus" to do this mapping [15]. When called in this manner using an NDC number, the RxNorm API returns a corresponding RxCUI for all current and retired NDC numbers passed to the API. According to the NLM RxNorm Technical Documentation, there is a one-to-many relationship between RxCUIs and NDC numbers [16]. Thus, no ambiguity should exist in this mapping process.

Figure 2 shows an overview of the analytic process and proposed future automated checking system. The figure also includes an embedded example of e-prescription record that failed to match its corresponding dispensing record.

Figure 2. Proposed system for detecting medication selection errors of e-prescriptions with RxNorm application programming interface. A prototypical example is included. API: application programming interface; e-prescription: electronic prescription; JSON: JavaScript object notation; NDC: National Drug Code; NLM: National Library of Medicine; RXCUI: RxNorm concept unique identifier; SBD: semantic branded drug; SCD: semantic clinical drug.
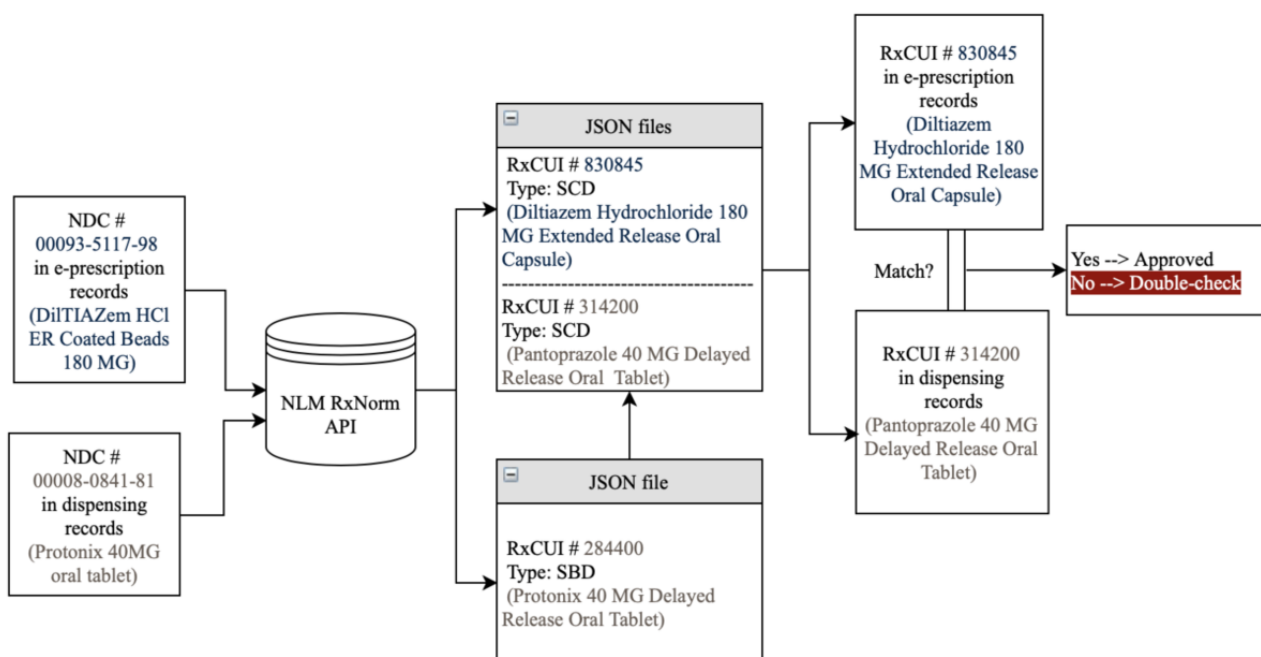
We performed additional identification of RxCUI term types to ensure RxCUIs in e-prescription and dispensing records were comparable. As SCD RxCUIs uniquely identify instances of drug name, strength, and form for the majority of drug products

with NDCs, we chose to perform the comparison between corresponding SCD terms. If a term type other than SCD (eg, an SBD or GPCK) was returned by the RxNorm API (API method *rxcui/{rxcui}/properties*), we would call the API a second time (API method *rxcui/{rxcui}/related?tty=SCD*) to identify all SCDs corresponding to the SBD or GPCK in those cases [15]. For example, in Figure 2, the NDC 00008-0841-81 found in a dispensing record is mapped to SBD RxCUI 284400, which is linked using the RxNorm API to SCD RxCUI 314200.

The second step of analysis used a matching algorithm to compare RxCUI returned for each e-prescription and dispensing record pair. The algorithm identified pairs with matched and unmatched RxCUIs. The output divided pairs into three categories: pairs with missing RxCUIs, pairs with matching RxCUIs, and pairs with nonmatching RxCUIs.

We report examples of actual NDC numbers for those pairs missing a corresponding RxCUI from RxNorm. It is important to note that manufacturers can generate new NDCs on their own. The majority, but not all NDCs, are curated by NLM and appear in RxNorm.

For the other two categories of pairs of e-prescription and dispensing records, we analyzed incorrect medication selection using an *error matrix*. We labeled the outcomes of our matching algorithm in the following way:

- True positives (TPs): SCD mismatch with incorrect drug/strength
- False negatives: Nonequivalent drugs with matching SCD
- False positives (FPs): SCD mismatch with incorrect quantity/form/qualitative distinction/releasing mechanism or outside RxNorm
- True negatives: Equivalent drugs with matching SCD

In our running example that is also embedded in Figure 2, a mismatch between SCD RxCUI 830845 for the drug diltiazem and SCD RxCUI 314200 for the drug pantoprazole is found. Diltiazem and pantoprazole are different medications and so a medication selection error has been detected by the system. Therefore, in our analysis, this example would be classified as a TP in the scheme above.

As a final step, to evaluate the automated checking system's overall performance, we analyzed unique pairs of NDCs and total NDC pairs. This allowed us to consider how the system might learn over time to reduce nuisance alerts in the pharmacy. For each specific mismatched e-prescription-dispensing record

pair, we classified the mismatched pair as clinically significant or not. In a learning automated checking system, once a mismatch is tagged as nonclinically significant, the system should ignore that mismatch in the future and not fire an alert to pharmacists on subsequent mismatches of that pair [17].

For each unique NDC pair with mismatched RxCUIs, we identified the type and frequency of the mismatch. These types of mismatches included incorrect ingredient, incorrect strength, incorrect quantity, incorrect dosage form (eg, capsule vs cream), and problematic RxCUIs. We defined a true positive as a different ingredient selection or a different strength selection (eg, 25 mg vs 50 mg). These errors signaled a TP error because the different ingredient or strength could lead to patient harm [18,19]. We defined an FP as an incorrect medication selection that would typically not lead to patient harm (ie, different quantity, different dosage form, or problematic RxCUI). Using the error matrix, we calculated the following performance metrics: accuracy, sensitivity/recall, false-positive rate, specificity, precision, and F1 score.

We report these performance metrics for both the unique pairs of NDCs and for the total pairs of NDCs in the dataset. This allows us to consider the automated checking system's performance if we deactivated alerts for specific mismatched RxCUI pairs after a pharmacist overrides them once (ie, as if the system operated as a learning system). For total pairs, we report performance of the automated checking system if the same mismatched RxCUI pair creates an alert every time the pharmacy dispensed the mismatched RxCUI pair (ie, as if the system operated as a static system that could not learn).

Results

The processed dataset included 527,881 (527,881/537,710, 98.17%) of the original dataset) e-prescription-dispensing record pairs from 64,805 prescribers in all 50 US states. There were 17,123 unique NDCs and 3,838 unique SCDs in the prescription-dispensing record pair data. The most frequently dispensed medications were atorvastatin, amlodipine, hydrochlorothiazide, and omeprazole (Table 1). The most common therapeutic drug classes in the dataset were cholesterol-lowering agents, which accounted for 118,877 pairs (118,877/527,881, 22.52%). The next most commonly analyzed therapeutic drug classes were blood glucose–lowering agents for 46,553 pairs (46,553/527,881, 8.82%) and antidepressants for 44,586 pairs (44,586/527,881, 8.45%).

Table 1. Five most frequently found clinical drug packs in the analyzed dataset (N=527,881).

RxCUI	Semantic clinical drug name	Matched pairs, n (%)
617311	Atorvastatin 40 mg oral tablet	7688 (1.46)
617310	Atorvastatin 20 mg oral tablet	7683 (1.46)
198051	Omeprazole 20 mg delayed release oral capsule	7674 (1.45)
197361	Amlodipine 5 mg oral tablet	6631 (1.26)
310798	Hydrochlorothiazide 25 mg oral tablet	6252 (1.18)

The first step of our analysis mapped each pair of NDC codes arising from an e-prescription-dispensing record pair to their

related SCD RxCUI. We found 872 pairs with 1731 NDCs where the RxNorm API did not return an RxCUI. Table 2

contains a list of examples with NDC codes that did not map to an RxCUI. One NDC code sometimes corresponded to more than one medication product description within the e-prescription data. Examples of medications with these

unmatched NDC numbers are *multivitamin tablet*, *blood sugar diagnostic strip*, and *Lancets Miscellaneous*. We excluded these 872 pairs from further analysis steps.

Table 2. Ten most frequent National Drug Codes without a corresponding RxNorm concept unique identifier (n=1731).

National Drug Code	Frequency, n (%)	Sample medication name[a]
0000-20002-02	440 (25.42)	Cyanocobalamin (Vitamin B-12) 100 mcg tablet
0000-00000-07	344 (1987)	Blood glucose test strips
0000-00000-08	220 (12.71)	Lancets 28 gauge
0000-00000-09	108 (6.24)	Blood glucose meter
2743-40010-21	38 (2.20)	Magnesium oxide 400 mg capsule
9289-60000-08	36 (2.08)	Blood glucose meter kit
0888-16096-00	34 (1.96)	Insulin syringe-needle U-100 1 ml 30 gauge $\times$ 5/16 syringe
3841-50003-08	34 (1.96)	Insulin pen needle 30 g $\times$ 8 mm
9403-00002-02	30 (1.73)	Embrace blood glucose system strips
0862-70014-01	22 (1.27)	Dexcom G5 mobile transmitter

[a]Some National Drug Codes correspond to multiple medication product descriptions.

As a result of the matching process performed by the automated checking system, 0.10% (550/527,009) pairs had different SCDs for their e-prescription and corresponding dispensing records. Table 3 reports on these 550 mismatched pairs by issue category. Three mismatched pairs showed different ingredients. One mismatched pair contained different strengths of the correct medication. Of the other 546 mismatched pairs found, 347 had

concept names different in one or more essential term, including quantity (ie, number of units in a pack) or form (eg, solution for injection vs prefilled syringe). The other 199 were either special cases (eg, sugar-free vs not sugar-free) or were cases with NDCs that were not curated by RxNorm, meaning that the NDCs appeared in RxNorm but came from a vocabulary source other than DailyMed or First Data Bank [16,20].

Table 3. Performance evaluation of the algorithm for detecting different medication selection, with examples (n=527,009).

Issues	Frequency, n (%)	Electronic Rx (E-Rx) RxNorm concept unique identifier (RxCUI)	Medication conveyed by E-Rx	Pharmacy prescription label RxCUI	Medication dispensed by pharmacy
Same medication	526,457 (99.90)	751620	Bystolic 2.5 mg oral tablet	751620	Bystolic 2.5 mg oral tablet
Different ingredient	3 (0.00)	313585	24 HR venlafaxine 75 mg extended release oral capsule	966225	Levothyroxine sodium 0.15 mg oral tablet
Different strength	1 (0.00)	861700	Amylases 82000 UNT/endopeptidases 51000 UNT/lipase 15000 UNT delayed release oral capsule	1595476	Amylases 84000 UNT/endopeptidases 63000 UNT/lipase 20000 UNT delayed release oral capsule
Different pack quantity	285 (0.05)	905100	12 (risedronate sodium 35 mg oral tablet) pack	905092	4 (risedronate sodium 35 mg oral tablet) pack
Different form	67 (0.01)	835840	testosterone cypionate 200 mg/mL injectable solution	2047882	1 mL testosterone cypionate 200 mg/mL injection
Concept outside RxNorm	40 (0.01)	1371671	Proprietary	1371861	Proprietary
Others (eg, qualitative distinctions)	159 (0.03)	198034	24 HR nifedipine 30 mg extended release oral tablet	1812011	Osmotic 24 HR nifedipine 30 mg extended release oral tablet

After removing duplicates, 0.33% (94/28,817) unique e-prescription-dispensing record pairs contained a mismatched RxCUI. Figure 3 shows the error matrix for clinically significant

incorrect medication selection detection from the data. Table 4 reports performance with and without duplicate mismatched e-prescription-dispensing record pairs.

Figure 3. Results of medication safety event error matrix with and without duplicate e-prescription-dispensing record pairs.

Table 4. Algorithm performance with and without duplicate electronic prescription–dispensing record pairs.

Metrics	Unique electronic prescription (e-prescription)–dispensing record pairs (n=28,817)	Total e-prescription-dispensing record pairs (n=527,009)[a]
Accuracy	0.99688	0.99896
Sensitivity/recall	1.00000	1.00000
False-positive rate	0.00312	0.00104
Specificity	0.99688	0.99896
Precision	0.04255	0.00727
F1 Score	0.08163	0.01444

[a]Pairs including duplicate dispensing of the same National Drug Code pair.

Discussion

Principal Findings

This study demonstrates a straightforward method of using the publicly available US NLM RxNorm API to identify potential medication selection errors made during transcription and processing of e-prescriptions in outpatient pharmacies. We evaluated the performance of an automated comparison for e-prescription-dispensing record pairs of medications ordered and dispensed using a dataset from a mail-order pharmacy. The automated checking system we developed identified 550 cases (550/527,009, 0.10%) of mismatched e-prescription-dispensing record pairs with issues ranging from incorrect medications to problems with unresolved NDCs. This rate is low when compared with a study reporting a 3.4% mismatch rate using RxNorm [13]. One potential reason for this is that the pharmacies sampled in the previous study were largely community pharmacies and may have different dispensing software, physical environments, and workloads.

Clinical Implications

To promote correct dispensing of medication, the RxNorm SCD is recommended by the National Council for Prescription Drug Programs (NCPDP) SCRIPT standards (currently 10.6), when one exists, to be included with the transmission of all e-prescriptions [21]. Unlike NDCs, RxNorm codes are centrally managed, making them easier to resolve to an actual drug product and potentially more suitable as a drug product identifier data standard. Including SCDs with all e-prescriptions can help to ensure that the meaning of the clinical drug product information is communicated correctly to the pharmacy so that pharmacy staff can select the drug product that they will dispense to the patient accurately. Besides, as more health information technology applications consume and process e-prescription information, having SCDs included in e-prescriptions will help protect the public from unintended software errors.

A 2014 national sample of e-prescriptions revealed that only 33.0% (n=49,997) were transmitted with an SCD in the HL7 message, even though the e-prescribing network allows for transmitting these identifiers. Since then, the adoption of SCD

RxCUIs has increased significantly as a result of Office of National Coordinator supporting the NCPDP SCRIPT standards [22] and more recently, due to the Centers for Medicare and Medicaid Services requiring the use of NCPDP SCRIPT standards for conveying e-prescriptions [21]. The use of a representative NDC to convey a prescribed drug product via an e-prescription is too specific, and NDCs are also potentially stale and no longer accurate [23]. In the pharmacy, staff are responsible for selecting the exact NDC of the drug product that is dispensed to the patient. Inclusion of the RxNorm SCD can aid the transcription of e-prescription information into pharmacy systems, resulting in fewer dispensing errors. As an added benefit, our automated checking system can be extended to the international community based on the recent expansion of RxNorm to integrate with other medication terminology standards, such as the Identification of Medicinal Products used by the European Medicines Agency and DrugBank used in Canada [24].

Consistent with previous literature, the evident rate of errors that reached the patient because of incorrect medication selection was approximately 0.1% [25]. We detected different ingredient errors, different medication strength errors, dosage form errors, and errors in the quantity to be dispensed. These are all common types of dispensing errors reported by community pharmacies around the world [26-31]. These errors differ in their potential to cause harm. In the worst case, when the wrong drug is dispensed, two kinds of harm may follow. First, in such a case, an individual does not benefit from getting the correct drug until the error is discovered. Second, the individual is exposed to a different drug. In other cases, for example, when the wrong package size of the correct drug is dispensed, the harm that may result is less, although such an error can lead to confusion for the patient. Owing to these important differences in the potential of errors to cause harm, in the future, we plan to explore methods of quantifying harm averted as well as errors avoided.

This study demonstrates the potential for the RxNorm API to enable an automated double-check of e-prescription and dispensing record information that could substitute for many of the manual double-checks provided by pharmacists today. Instead of having pharmacists double-check the fidelity of every e-prescription entered into the pharmacy system, the number requiring a manual double-check might be reduced 1000-fold or more, saving pharmacists' worktime and cognitive effort [9]. In addition, the efficiency and reliability of the RxNorm-enabled matching algorithm suggest it may be feasible to apply this type of automated checking at different stages in the medication dispensing workflow. For example, using the RxNorm API in real time to fire an alert if the technician makes an incorrect drug selection. Alternatively, automated checking could trigger alerts to an off-site staff member to determine the clinical significance of the problem before interrupting the workflow in the pharmacy.

However, there is a potential for alert fatigue by using this type of automated checking. A previous study found that pharmacy staff were annoyed by false-positive alerts but also that their satisfaction with alerting goes up when given opportunities to prevent dispensing errors [13]. We found that just 4 out of the 550 cases of mismatched pairs were of clinical significance in

this study. This means there were 137 false alerts for every clinically significant one. To overcome alert fatigue, we are interested in developing a learning capability to complement the automated checking capability demonstrated here. If the current automated checking system could learn the difference between mismatched pairs that are clinically significant and those that are not, then we could suppress alerts for insignificant mismatches. Our tests show that in this case, given the sequence of prescribing events, there is the potential to lower the number of alerts needed to surface a clinically significant issue from 138 (137:1) to 24 (23:1). Over time, continuing to learn from the mismatched pairs identified would further lower the ratio of false-positive to true-positive alerts. We believe this type of learning approach could decrease the chance of alert fatigue [32].

Previous studies have demonstrated a variety of uses for the RxNorm drug terminology system, including improving medication history taking, resolving free-text medication naming conventions from electronic health records (EHRs), and matching clinical drug names across medication formularies [33-35]. RxNorm provides a distinct advantage over using NDCs in that the RxNorm's RxCUIs provide normalization of drug naming concepts across systems with different configurations. This benefit from RxNorm is particularly relevant for EHRs and pharmacy dispensing software. EHRs are configured with different medication naming conventions. When different naming conventions are associated with the same NDCs, ambiguity arises in the e-prescription data ecosystem upon which pharmacies depend. RxNorm provides a means to reconcile the different ways that medication information is configured in EHRs, helping to address and remove unwanted ambiguity from the e-prescription data ecosystem. Using RxNorm RxCUIs with e-prescriptions and dispensing records improves the quality of data and helps ensure accurate transmission of e-prescription information.

Future Work

To validate our findings, a logical next step is to expand the use of automated checking to process a larger dataset drawn from multiple pharmacies. Given that 1.91 billion new e-prescriptions are transmitted annually in the United States, other future work needs to establish a reliable technical infrastructure for outpatient pharmacies to analyze e-prescription-dispensing record pairs routinely with the existing capabilities of the RxNorm API. One solution to prevent the worst medication selection errors would be to use the RxNorm API to link an NDC from the e-prescription to a list of NDCs mapped to the same SCD in the pharmacy dispensing software. Once these capabilities are further demonstrated and made widely available, additional research should focus on other areas where automated checking can have this kind of positive impact.

Limitations

These findings are limited primarily because the dataset came from a single mail-order pharmacy in the United States. Although the e-prescriptions received by this pharmacy came from all over the country, the dataset is not a representative national sample of e-prescription and dispensing records. These data also only contain medications used in the outpatient setting

(ie, not injectable or infusion products using in inpatient or long-term care settings). A second limitation is that although our algorithm detected seeming clinically significant medication safety incidents, it is possible that there were other plausible explanations for these mismatches (eg, the case of a pharmacist who receives verbal approval from a prescriber to change the medication product). Another concern is that drug concepts and their codes are only valid for a limited period of time. During our analysis, we have checked and remapped obsolete NDCs and RxCUIs to active concept unique identifiers. But the reuse of an NDC number can lead to problems when combining datasets. Today, the Food and Drug Administration (FDA) allows firms to reassign an NDC 5 years after the expiration date of a discontinued drug [36]. As the FDA does not restrict the creation of NDCs, some codes fall outside the intended scope of RxNorm (eg, proprietary prescription compounds or diabetic testing supplies). This means that some NDCs will never be analyzable with the proposed RxCUI matching system.

Conclusions

In this study, we used the NLM's RxNorm API to enable accurate automated checks of e-prescription-dispensing record pairs. We identified a small but critical number of e-prescription processing errors made at the pharmacy. Our method can detect potential dispensing errors before they cause harm and, if combined with other machine-checking interventions, has the potential to eliminate the need for the majority of the manual pharmacist double-check to compare the e-prescription medication product with the medication product selected for dispensing. This type of automation can help reduce the risk of dispensing the incorrect medication to the patient while sparing the pharmacist's worktime for higher value patient care tasks. Using these results, validation in other pharmacies is necessary before the widespread adoption of this system as one part of a safety management system that relies on the retrospective analysis of e-prescription and dispensing record data.

Authors' Contributions

CL conceptualized the research question and led the analysis and writing/editing of manuscript. LT performed data analysis and drafted/edited sections of manuscript. YD designed and edited tables/figures, assisted in study design, and assisted in editing the manuscript. AF assisted in research question development, and edited, wrote, and contributed clinical significance to the manuscript.

Conflicts of Interest

In the past 12 months, AF has received an honorarium from the Omnicell Corporation and the American Society of Health-system Pharmacists for participating in the Autonomous Pharmacy Advisory Board and Midyear Clinical Meeting, respectively. All other authors report no conflicts of interest.

References

1. Makary MA, Daniel M. Medical error-the third leading cause of death in the US. Br Med J 2016 May 3;353:i2139. [doi: 10.1136/bmj.i2139] [Medline: 27143499]
2. World Health Organization. WHO Global Patient Safety Challenge: Medication Without Harm. Geneva, Switzerland: World Health Organization; 2017. URL: https://www.who.int/patientsafety/medication-safety/medication-without-harm-brochure/en/ [accessed 2019-08-01]
3. Muhammad KW, Carson-Stevens A, Avery AJ, Boyd MJ. What medication errors occur in community pharmacies and why do they occur? A national analysis of patient safety incidents. Research in Social and Administrative Pharmacy 2018;14(8):e37. [doi: 10.1016/j.sapharm.2018.05.061]
4. Darbishire PL, Zhao JC, Sodhi A, Anderson CM. Student observations of medication error reporting practices in community pharmacy settings. Res Social Adm Pharm 2019 Jul;15(7):902-906. [doi: 10.1016/j.sapharm.2019.02.009] [Medline: 30852086]
5. Pervanas HC, Revell N, Alotaibi AF. Evaluation of medication errors in community pharmacy settings. J Pharm Technol 2015;32(2):71-74. [doi: 10.1177/8755122515617199]
6. Odukoya OK, Stone JA, Chui MA. E-prescribing errors in community pharmacies: exploring consequences and contributing factors. Int J Med Inform 2014 Jun;83(6):427-437 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.02.004] [Medline: 24657055]
7. Odukoya OK, Schleiden LJ, Chui MA. The hidden role of community pharmacy technicians in ensuring patient safety with the use of e-prescribing. Pharmacy (Basel) 2015 Dec;3(4):330-343 [FREE Full text] [doi: 10.3390/pharmacy3040330] [Medline: 27525221]
8. Odukoya OK, Stone JA, Chui MA. How do community pharmacies recover from e-prescription errors? Res Social Adm Pharm 2014;10(6):837-852 [FREE Full text] [doi: 10.1016/j.sapharm.2013.11.009] [Medline: 24373898]
9. Gaither C, Schommer J, Doucette W, Kreling D, Mott D. American Association of Colleges of Pharmacy.: American Association of Colleges of Pharmacy; 2015 Apr 8. Executive Summary of the Final Report of the 2014 National Sample

XSL•FO

RenderX

Survey of the Pharmacist Workforce to Determine Contemporary Demographic Practice Characteristics and Quality of Work-Life URL: https://www.aacp.org/sites/default/files/executivesummaryfromthenationalpharmacistworkforcestudy2014.pdf [accessed 2019-08-01]

10. Chui MA, Look KA, Mott DA. The association of subjective workload dimensions on quality of care and pharmacist quality of work life. Res Social Adm Pharm 2014;10(2):328-340 [FREE Full text] [doi: 10.1016/j.sapharm.2013.05.007] [Medline: 23791360]

11. Chui MA, Mott DA. Community pharmacists' subjective workload and perceived task performance: a human factors approach. J Am Pharm Assoc (2003) 2012;52(6):e153-e160 [FREE Full text] [doi: 10.1331/JAPhA.2012.11135] [Medline: 23229977]

12. Hewitt T, Chreim S, Forster A. Double checking: a second look. J Eval Clin Pract 2016 Apr;22(2):267-274 [FREE Full text] [doi: 10.1111/jep.12468] [Medline: 26568537]

13. Bell DS, O'Neill SM, Reynolds K, Schoeff D. Evaluation of RxNorm in ambulatory electronic prescribing. Rand Health Q 2011;1(3):8 [FREE Full text] [Medline: 28083195]

14. Centers for Medicare and Medicaid Services. 2018. National Plan and Provider Enumeration System (NPPES) URL: https://download.cms.gov/nppes/NPI_Files.html [accessed 2018-11-20]

15. RxNav, . RxNorm API URL: https://rxnav.nlm.nih.gov/RxNormAPIs.html [accessed 2019-08-16]

16. US National Library of Medicine. RxNorm Technical Documentation URL: https://www.nlm.nih.gov/research/umls/rxnorm/docs/techdoc.html [accessed 2019-08-16]

17. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med 2010 Nov 10;2(57):57cm29. [doi: 10.1126/scitranslmed.3001456] [Medline: 21068440]

18. Ukens C. Deadly dispensing: an exclusive survey of Rx errors by pharmacists. Drug Topics 1997;141(5) [FREE Full text]

19. Ukens C. Breaking the trust: Exclusive survey of dispensing errors. Drug Topics 1992;23:58-69 Found as citation here: https://www.researchgate.net/publication/41892753_Incidence_type_and_causes_of_dispensing_errors_A_review_of_the_literature.

20. Peters LB, Bodenreider O. Approaches to supporting the analysis of historical medication datasets with RxNorm. AMIA Annu Symp Proc 2015;2015:1034-1041 [FREE Full text] [Medline: 26958241]

21. National Council for Prescription Drug Programs. 2019. SCRIPT Implementation Recommendations URL: https://www.ncpdp.org/NCPDP/media/pdf/SCRIPT-Implementation-Recommendations.pdf [accessed 2019-08-01]

22. Health IT, Office of the National Coordinator for Health Information Technology. Certification Companion Guide: Electronic Prescribing URL: https://www.healthit.gov/test-method/electronic-prescribing [accessed 2019-08-01]

23. Surescripts, Inc. E-Prescribing Quality Guidelines URL: https://surescripts.com/docs/default-source/pressrelease-library/e-prescribing-quality-guidelines.pdf?sfvrsn=fb807dca_38 [accessed 2019-08-01]

24. NLM Technical Bulletin, US National Library of Medicine. 2018 Sep 6. RxNorm Aligns with International Drug Models by Adding New Attributes URL: https://www.nlm.nih.gov/pubs/techbull/so18/brief/so18_rxnorm_boss.html [accessed 2019-10-11]

25. Teagarden JR, Nagle B, Aubert RE, Wasdyke C, Courtney P, Epstein RS. Dispensing error rate in a highly automated mail-service pharmacy practice. Pharmacotherapy 2005 Nov;25(11):1629-1635. [doi: 10.1592/phco.2005.25.11.1629] [Medline: 16232024]

26. James KL, Barlow D, McArtney R, Hiom S, Roberts D, Whittlesea C. Incidence, type and causes of dispensing errors: a review of the literature. Int J Pharm Pract 2009 Feb;17(1):9-30. [Medline: 20218026]

27. Varadarajan R, Barker KN, Flynn EA, Thomas RE. Comparison of two error-detection methods in a mail service pharmacy serving health facilities. J Am Pharm Assoc (2003) 2008;48(3):371-378. [doi: 10.1331/JAPhA.2008.07005] [Medline: 18595822]

28. Flynn EA, Barker KN. Effect of an automated dispensing system on errors in two pharmacies. J Am Pharm Assoc (2003) 2006;46(5):613-615. [doi: 10.1331/1544-3191.46.5.613.flynn] [Medline: 17036648]

29. Flynn EA, Barker KN, Carnahan BJ. National observational study of prescription dispensing accuracy and safety in 50 pharmacies. J Am Pharm Assoc (Wash) 2003;43(2):191-200. [doi: 10.1331/108658003321480731] [Medline: 12688437]

30. Allan EL, Barker KN, Malloy MJ, Heller WM. Dispensing errors and counseling in community practice. Am Pharm 1995 Dec;NS35(12):25-33. [Medline: 8546071]

31. Timonen J, Kangas S, Kauppinen H, Ahonen R. Electronic prescription anomalies: a study of frequencies, clarification and effects in Finnish community pharmacies. J Pharm Health Serv Res 2018;9(3):183-189. [doi: 10.1111/jphs.12224]

32. Saverno K, Hines L, Warholak T, Grizzle AJ, Babits L, Clark C, et al. Ability of pharmacy clinical decision-support software to alert users about clinically important drug-drug interactions. J Am Med Inform Assoc 2011;18(1):32-37 [FREE Full text] [doi: 10.1136/jamia.2010.007609] [Medline: 21131607]

33. Bennett CC. Utilizing RxNorm to support practical computing applications: capturing medication history in live electronic health records. J Biomed Inform 2012 Aug;45(4):634-641 [FREE Full text] [doi: 10.1016/j.jbi.2012.02.011] [Medline: 22426081]

34. Freimuth R, Wix K, Zhu Q, Siska M, Chute C. Evaluation of RxNorm for medication clinical decision support. AMIA Annu Symp Proc 2014;2014:554-563 [FREE Full text] [Medline: 25954360]

35. Peters L, Kapusnik-Uner JE, Nguyen T, Bodenreider O. An approximate matching method for clinical drug names. AMIA Annu Symp Proc 2011;2011:1117-1126 [FREE Full text] [Medline: 22195172]

36. US Food and Drug Administration. CFR - Code of Federal Regulations Title 21 URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=207.35 [accessed 2019-08-01]

Abbreviations

API: application programming interface
EHR: electronic health record
e-prescription: electronic prescription
FDA: Food and Drug Administration
FP: false positive
GPCK: generic pack
NCPDP: National Council for Prescription Drug Programs
NDC: National Drug Code
NLM: National Library of Medicine
NPI: National Provider Identifier
RxCUI: RxNorm concept unique identifier
SBD: semantic branded drug
SCD: semantic clinical drug
TP: true positive

XSL•FO
RenderX