JMIR Medical Informatics

Clinical informatics, decision support for health professionals, electronic health records, and eHealth infrastructures
Volume 8 (2020), Issue 9 ISSN: 2291-9694

Contents

Original Papers

Viewpoints

Corrigenda and Addenda

Short Paper

Original Paper

An Iterative Process for Identifying Pediatric Patients With Type 1 Diabetes: Retrospective Observational Study

Heather Lynne Morris[1], BA, MS, PhD; William Troy Donahoo[2], MD; Brittany Bruggeman[2], MD; Chelsea Zimmerman[2], MD; Paul Hiers[2], MD; Victor W Zhong[3], PhD; Desmond Schatz[4], MD

[1]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, United States

[2]Department of Medicine, University of Florida, Gainesville, FL, United States

[3]Division of Nutritional Sciences, Cornell University, Ithaca, NY, United States

[4]Department of Pediatrics, University of Florida, Gainesville, FL, United States

Corresponding Author:
Heather Lynne Morris, BA, MS, PhD
Department of Health Outcomes and Biomedical Informatics
University of Florida
P.O. Box 100177
Gainesville, FL,
United States
Phone: 1 3526279074
Email: hlmorris27@ufl.edu

Abstract

Background: The incidence of both type 1 diabetes (T1DM) and type 2 diabetes (T2DM) in children and youth is increasing. However, the current approach for identifying pediatric diabetes and separating by type is costly, because it requires substantial manual efforts.

Objective: The purpose of this study was to develop a computable phenotype for accurately and efficiently identifying diabetes and separating T1DM from T2DM in pediatric patients.

Methods: This retrospective study utilized a data set from the University of Florida Health Integrated Data Repository to identify 300 patients aged 18 or younger with T1DM, T2DM, or that were healthy based on a developed computable phenotype. Three endocrinology residents/fellows manually reviewed medical records of all probable cases to validate diabetes status and type. This refined computable phenotype was then used to identify all cases of T1DM and T2DM in the OneFlorida Clinical Research Consortium.

Results: A total of 295 electronic health records were manually reviewed; of these, 128 cases were found to have T1DM, 35 T2DM, and 132 no diagnosis. The positive predictive value was 94.7%, the sensitivity was 96.9%, specificity was 95.8%, and the negative predictive value was 97.6%. Overall, the computable phenotype was found to be an accurate and sensitive method to pinpoint pediatric patients with T1DM.

Conclusions: We developed a computable phenotype for identifying T1DM correctly and efficiently. The computable phenotype that was developed will enable researchers to identify a population accurately and cost-effectively. As such, this will vastly improve the ease of identifying patients for future intervention studies.

(JMIR Med Inform 2020;8(9):e18874) doi:10.2196/18874

KEYWORDS

computable phenotype; type 1 diabetes; electronic health record; pediatric

Introduction

Diabetes is one of the most common chronic diseases seen during childhood and adolescence. The incidence and prevalence of diabetes mellitus has continued to increase worldwide for both type 1 diabetes (T1DM) and type 2 diabetes (T2DM), with the rise in T2DM due in large part to the obesity epidemic [1,2]. Uncontrolled T1DM leads to short- and long-term complications and early mortality [3-6].

The vast majority of the population data about the incidence, prevalence, and effects of diabetes in youth in the United States come from select sites, such as the *SEARCH for Diabetes in Youth Study* [7] and the *T1D Exchange* [8]. In the past, outside of highly manicured registries, the thorough and accurate identification of pediatric patients with T1DM versus T2DM could only be accomplished by manual clinical record review, which was both costly and time-consuming, requiring manual medical record reviews. Currently, through the use of algorithms derived from electronic health record data, accurate identification of patients with T1DM versus T2DM may be possible. One such algorithm using a subset of *SEARCH* cohort revealed a 89% positive predictive value (PPV) and a 97% negative predictive value using only ICD-10-CM codes [9]. However, this study was conducted within a self-contained data set overseen by Kaiser Permanente. As such, this does not give a comprehensive insight into patients seen at a variety of health settings using different electronic record systems. There is thus a need for timely real-world population-level monitoring of the incidence, prevalence, and disease course of diabetes in youth that includes the ability to separate T1DM from T2DM.

The overall purpose of this project was to develop and validate an algorithm to identify pediatric patients with T1DM in an efficient and accurate manner that would be valid in a real-world database outside of a closed medical system such as Kaiser Permanente.

Methods

Population

University of Florida Health

Patients eligible for inclusion in this study were aged 0-18 and seen at University of Florida Health (UF Health). The UF Health System is a medical network associated with the University of Florida with the only comprehensive pediatric facility in North Central Florida. The Integrated Data Repository (IDR) is a large-scale database that collects and organizes information across UF Health’s clinical and research enterprises. The IDR is a secure, clinical data warehouse that aggregates data from the university’s clinical and administrative information systems, including the electronic health record system. As of 2018, the IDR housed more than 1 billion observational facts across more than 1 million patients. For query 1, the IDR was utilized to identify 300 patients for the development of the computable phenotype. Similar to other studies, 100 individuals per cohort were selected (T1DM, T2DM, no diagnosis) with the *no diagnosis* classification being used as the reference group.

OneFlorida

The OneFlorida Clinical Research Consortium contains over 12 million unique patient records from as early as of 2012, including Medicaid claims records. This database is maintained and updated on a quarterly basis with information from partners across the state of Florida. The OneFlorida Data Trust’s repository of statewide health care data is regularly updated with the inclusion of new partners and data refreshes from existing partners. All data are cleaned, transformed, curated, and contained in a centralized data warehouse, allowing streamlined inquiries and uniform results based on high-quality data. At present, data on 15 million patients across 22 hospitals are included within the data set going back to 2012, of which approximately 4.3 million are pediatric patients aged 18 or younger across thousands of providers, clinics, practices, and multiple hospital systems throughout the state of Florida. A SAS code that was developed from the algorithm was used to identify eligible members. Previous work has demonstrated that the OneFlorida Data Trust demographics are similar to estimates reported by the US Census Bureau [10,11]. Five OneFlorida sites that did not have prescribing data were excluded. For queries 2 and 3, we limited our results to patients aged 0-18 seen within the OneFlorida Data Trust in the year 2018.

Study Overview

In query 1, the initial algorithm for differentiating T1DM and T2DM was developed and validated with chart reviews using data from the UF Health system. Subsequently, this algorithm was utilized in the OneFlorida database (query 2).

Queries

Query 1: Computable Phenotype Algorithm Development Using UF Health IDR

For the development of the algorithm, we identified individuals in the UF Health System that would meet the criteria of having T1DM or T2DM, and a cohort with no diagnosis of either for comparison. A total of 300 random records were requested from the IDR with 100 of each of the following: T1DM, T2DM, and no diagnosis of either. The criteria for diagnosis of T1DM used diagnosis codes, medication dispensing, and laboratory results. Patients met the T1DM algorithm criteria if they were less than or equal to 18 years of age as of December 31, 2016, and fulfilled the following criteria: (1) inpatient/outpatient with ICD-9/10 for T1DM and insulin medication within 90 days or (2) inpatient/outpatient with ICD-9/10 for T1DM and glucose >200 mg/dL or (3) inpatient/outpatient with ICD-9/10 for T1DM and hemoglobin A1c > 6.5%.

The type 2 criteria differed slightly in that it involved ICD-9/10 for patients with T2DM under the age of 18. For each identified member within the 300 total records, we obtained data on age, sex, race, ethnicity, height, weight, BMI, diagnoses, location of services, and the admit date. In order to account for a number of conflicting diagnoses for individual patients, a diagnosis ratio was used to make a final diagnosis categorization (T1DM vs T2DM). Conflicting diagnosis codes occurred when patients were seen by multiple providers, or different settings, and received both a T1DM and T2DM in the electronic health record. In order to receive a designation of T1DM or T2DM, they had to have greater occurrences of one diagnosis. Diagnosis ratio designations were applied prior to the medical record review to allow for further investigation.

The data management for query 1 was managed in a REDCap database [12]. A data abstraction form was developed for use by the medical record reviewers to manually abstract data related to a diabetes mellitus diagnosis and treatment from the medical records. This form was utilized to collect demographic data and diabetes-related clinical information including the most recent record of height, weight, hemoglobin A1c, and if islet autoantibodies were present (and type).

Medical Record Review

A total of 3 pediatric endocrinology fellows (BB, CZ, and PH) evaluated the medical records to determine the *true* diagnosis. A total of 295 cases, with an overlap of 41 cases to assess interrater reliability, were reviewed. For quality assurance, 14% (40/295) of all records were manually abstracted by multiple reviewers (BB, CZ, or PH). Any discrepancies were adjudicated by a senior reviewer (WD). All reviewers were blinded to the diagnosis category patients were assigned to. Each reviewer accessed the patient electronic health records to evaluate the medical record thoroughly to make a final diagnosis. Patients were given a designation of T1DM if they fell into the range of clinical criteria including diagnosis at a younger age, a history of diabetic ketoacidosis, positive antibody status, lower insulin requirements, and lower BMI. Additional data were abstracted so the most up-to-date information for laboratory values was recorded. Reviewers entered all information into a REDCap database. Following the review, data were exported into SPSS and reviewed for interrater reliability. A total of 5 cases were evaluated in greater depth due to missingness, terminology, and a differing diagnosis. The sensitivity, PPV, negative predictive value, and specificity were calculated using the numerators and denominators from the medical record review.

Query 2: Computable Phenotype Algorithm using OneFlorida

Abstraction conducted in query 1 highlighted a number of false-positive diagnoses. In order to correctly categorize patients with other forms of diabetes (eg, cystic fibrosis–related diabetes, maturity-onset diabetes of youth, neonatal hyperglycemia), we separated patients with these diagnostic codes into a third cohort identified as *other diabetes*. We revised the algorithm to include patients with ICD-10 of Neonatal Diabetes Mellitus P70.2 instead of P61.0 for the Other DM categories. This resulted in a reduction of 5397 patients across all years (originally 9727), and 685 patients in the year 2018 alone (previously 1316).

Query 3: Computable Phenotype Algorithm Using OneFlorida Revised

In the initial run of the computable phenotype in the OneFlorida Clinical Research Consortium, there was an inconsistency in the number of cases of patients with T1DM and T2DM. More specifically, there were more cases of patients with T2DM than on average. We revised the algorithm to include additional pharmacy data to identify patients who met the algorithm criteria where patients with a diagnosis code of T2DM were also required to have a prescription of metformin.

Results

Computable Phenotype Algorithm Development Using UF Health IDR

In our first query of 300 medical records drawn from the UF Health IDR, 5 cases had no discerning diagnosis (conflicting diagnosis of T1DM and T2DM) based on the diagnosis ratio, and therefore, these were excluded from the study. A total of 295 records were reviewed. Table 1 shows the demographics of these patients.

After applying a diagnosis ratio between hospital encounters, there were a total of 131 patients with T1DM, 64 with T2DM, and 100 with no diagnosis of either. Of the 131 patients identified using the computable phenotype algorithm, abstractors confirmed a diagnosis of T1DM for 125 patients (true positive; Table 2), which yielded a PPV of 96.8% (Table 2). Upon validation with the medical record review, it was confirmed that 7 patients were incorrectly identified (false positive; Table 2) by the algorithm. These patients instead were found to have either no diagnosis (n=5) or T2DM (n=2). The final computable phenotype algorithm was determined to have a sensitivity of 95.3%. The T2DM algorithm had a lower PPV than T1DM (51.6%) but had a high sensitivity (94.3%) and specificity (97.5%).

Table 1. UF demographics.

Demographic	Overall (N=295)	No diagnosis (N=132)	T1DM[a] (N=128)	T2DM[b] (N=35)
Age, mean (SD)	10.7 (5.44)	7.8 (5.56)	12.3 (4.07)	15.4 (2.87)
Gender				
Male, n (%)	134 (45.4)	63 (47.7)	60 (46.9)	11 (31.4)
Female, n (%)	161 (54.6)	69 (52.3)	68 (53.1)	24 (68.6)
Race				
Caucasian, n (%)	179 (60.7)	79 (59.8)	87 (68.0)	13 (37.1)
African American, n (%)	62 (21.0)	33 (25.0)	10 (7.8)	19 (54.3)
Hispanic, n (%)	31 (10.5)	11 (8.3)	19 (14.8)	1 (2.9)
Asian, n (%)	2 (0.7)	2 (1.5)	0 (0)	0 (0)
Multiple races, n (%)	15 (5.1)	4 (3.0)	9 (7.0)	2 (5.7)
Missing, n (%)	6 (2.0)	3 (2.3)	3 (2.3)	0 (0)
Treatment facility				
UF[c] Health, n (%)	231 (78.3)	81 (61.4)	117 (91.4)	33 (94.3)
Autoantibodies presence, n (%)	67 (22.7)	2 (1.5)	65 (50.8)	0 (0)
Ethnicity				
Hispanic, n (%)	38 (12.9)	13 (9.8)	23 (18.0)	2 (5.7)
Glucose level, mean (SD); range	153.86 (95.45); 7-555	89.43 (30.33); 7-284	207.59 (98.57); 58-555	161.11 (96.83); 64-432
Hemoglobin A1c, mean (SD); range	8.62 (2.31); 4.8-14.00	5.48 (0.58); 4.8-7.5	9.27 (1.80); 5.6-14	7.92 (2.90); 4.9-14.00

[a] T1DM: type 1 diabetes mellitus.

[b] T2DM: type 2 diabetes mellitus.

[c] UF: University of Florida.

Table 2. Results from query 1.

Query 1	Total reviewed, n	Total confirmed, n	Sensitivity, %	Specificity, %	Positive predictive value, %	Negative predictive value, %
T1DM[a] case identified via CP[b] algorithm[c]	131	124	96.9	95.8	94.7	97.6
T2DM[d] case identified via CP algorithm[e]	64	33	94.3	88.1	51.6	99.1

[a] T1DM: type 1 diabetes mellitus.

[b] CP: computable phenotype.

[c] T1DM algorithm: sensitivity=124/124+4; specificity=160/160+7; PPV=124/124+7; NPV=160/160+4.

[d] T2DM: type 2 diabetes mellitus.

[e] T2DM algorithm: sensitivity=33/33+2; specificity=229/229+31; PPV=33/33+31; NPV=229/229+2.

Computable Phenotype Algorithm Performance in OneFlorida

In the second query, the performance of the algorithm was tested in the OneFlorida Data Trust. Although the validity of using only ICD codes for the determination of diabetes type in youth has been demonstrated in the large integrated health system of Kaiser Permanente Southern California [9], and while our algorithm was based largely on ICD codes and did very well in the UF Health IDR, when this was run in the OneFlorida Data Trust, there were issues with appropriate categorization as described in the “Methods” section. As these numbers were not consistent with what we know about the epidemiology and biology of T1DM versus T2DM in youth [13], we undertook a revision of the algorithm.

The revised algorithm included additional pharmacy data to identify patients who met the algorithm criteria. In the revision, patients with a diagnosis code of T2DM were also required to have a prescription of metformin. The results from the final algorithm are presented in Table 3. The majority of patients identified by the algorithm had a diagnosis of T1DM (n=4246) followed by other DM (n=660) and T2DM (n=550). Patients with T1DM had an even distribution of male and female, were predominantly White (2153/4246, 50.71%), between 11 and 15 years of age (1789/4246, 42.13%), and on insulin (3907/4246, 92.02%). Patients identified as having T2DM were more likely to be female (342/550, 62.1%), other race (190/550, 34.5%), Black (241/550, 43.8%), and between 16 and 18 years of age (300/550, 54.5%). Because of the already high sensitivity and specificity of the less robust initial algorithm for T1DM, we did not do additional chart reviews for the revised algorithm.

Table 3. Results of final algorithm in OneFlorida.

Demographic	T1DM[a] (N=4246)	T2DM[b] (N=550)	Other DM (N=660)
Sex			
Female, n (%)	2120 (49.93)	342 (62.18)	326 (49.39)
Male, n (%)	2126 (50.07)	208 (37.82)	334 (50.61)
Race			
White, n (%)	2153 (50.71)	117 (21.27)	195 (29.55)
Black, n (%)	709 (16.70)	241 (43.82)	234 (35.45)
Asian, n (%)	23 (0.54)	N/A[c]	N/A[c]
Other/unknown, n (%)	1361 (32.05)	190 (34.55)	229 (34.70)
Age			
0-5 years, n (%)	253 (5.96)	N/A[c]	512 (77.58)
6-10 years, n (%)	895 (21.08)	N/A[c]	31 (4.70)
11-15 years, n (%)	1789 (42.13)	240 (43.64)	66 (10.00)
16-18 years, n (%)	1309 (30.83)	300 (54.55)	51 (7.73)
Insulin, n (%)	3907 (92.02)	284 (51.64)	63 (9.55)

[a] T1DM: type 1 diabetes mellitus.

[b] T2DM: type 2 diabetes mellitus.

[c] N/A: no available data (ie, no patients identified).

Discussion

Principal Findings

Overall, the computable phenotype we developed to identify pediatric patients with T1DM was effective using data within the electronic health record. The identification of patients with diabetes can be complex and conflicting diagnosis codes make it even more difficult to disentangle an accurate classification. As such, the use of additional clinical parameters to narrow the focus to a specific population refines the specificity of the algorithm. For T1DM, this includes laboratory values (A1c ≥ 6.5, glucose ≥ 200 m/g).

For the purposes of this study, we drew upon the parameters already defined by the SEARCH study which allows researchers to identify adults with T1DM. Referencing this study, we made refinements to account for variations among pediatric patients. The utility of this computable phenotype is that it enables us to identify patients with an accuracy of 97%. Identification of patients solely based on the data found within the electronic health record can be complex, thus accounting for our need of numerous queries. The idiosyncrasies of diagnosis codes and limited recordings of HbA1c for patients added complexities to the methods of identification. In our experience, diagnosis codes for patients often had contradictions. For example, a patient seen multiple times in the measurement year in various settings may have conflicting diagnosis (ie, T1DM and T2DM). To overcome this problem, we applied a diagnosis ratio to include the most prevalent diagnosis. This is an important consideration for other individuals utilizing electronic health records for identification. The identification of pediatric patients solely based on the ICD-9 or ICD-10 code only allows us to look at patients on the surface level rather than as a whole.

The findings from this study were instrumental in developing a computable phenotype to identify pediatric patients with T1DM. Through this process, a number of limitations were of note that should be considered. First, the utilization of the electronic health record presented a few obstacles that were not originally foreseen, particularly the conflicting diagnoses of patients. Inaccuracies and data entry error are plausible within large data sets and need to be accounted for. Being aware of the possibility of inaccurate diagnoses increases the importance of not relying solely on ICD-9 and ICD-10 diagnoses for identifying patients. Similarly, this impacted our proposed methodology of 100 individuals for each of the 3 cohorts (ie, T1DM, T2DM, and no diagnosis). These differences were accounted for in our calculations of predictive value, sensitivity, and specificity, but still need to be noted as a potential limiting factor. Another limitation of this paper is that the medical record review was limited to 1 health care system. While we were able to identify all pediatric patients within the OneFlorida Clinical Research Consortium with T1DM, we were unable to access individualized records within each of the contributing data centers and thus unable to conduct medical record reviews at each site. Additionally, as 5 OneFlorida sites did not have prescribing data, this limits our available data, and generalizability, from the entire state of Florida.

Conclusions

In summary, the computable phenotype that we developed to identify pediatric patients with T1DM is both accurate (PPV=96.8%) and sensitive (95.3%). This computable phenotype will enable future researchers to not only identify a population of interest accurately, but also cost-effectively. As such, this will allow for more precise implementation of interventions to help improve both clinical and psychosocial care, and ultimately improve outcomes important to patients.

Acknowledgments

Research reported in this publication was supported in part by the OneFlorida Clinical Data Network, funded by the Patient-Centered Outcomes Research Institute (No. CDRN-1501-26692), in part by the OneFlorida Cancer Control Alliance, funded by the Florida Department of Health’s James and Esther King Biomedical Research Program (No. 4KB16), and in part by the University of Florida Clinical and Translational Science Institute, which is supported in part by the NIH National Center for Advancing Translational Sciences under award number UL1TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology, the OneFlorida Clinical Research Consortium, the University of Florida’s Clinical and Translational Science Institute, the Florida Department of Health, or the National Institutes of Health.

Conflicts of Interest

None declared.

References

1. Mayer-Davis EJ, Lawrence JM, Dabelea D, Divers J, Isom S, Dolan L, SEARCH for Diabetes in Youth Study. Incidence Trends of Type 1 and Type 2 Diabetes among Youths, 2002-2012. N Engl J Med 2017 Dec 13;376(15):1419-1429 [FREE Full text] [doi: 10.1056/NEJMoa1610187] [Medline: 28402773]
2. Pinhas-Hamiel O, Standiford D, Hamiel D, Dolan LM, Cohen R, Zeitler PS. The type 2 family: a setting for development and treatment of adolescent type 2 diabetes mellitus. Arch Pediatr Adolesc Med 1999 Oct;153(10):1063-1067. [doi: 10.1001/archpedi.153.10.1063] [Medline: 10520614]
3. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care 2014 Jan;37 Suppl 1:S81-S90. [doi: 10.2337/dc14-S081] [Medline: 24357215]
4. Craig ME, Hattersley A, Donaghue KC. Definition, epidemiology and classification of diabetes in children and adolescents. Pediatr Diabetes 2009 Sep;10 Suppl 12:3-12. [doi: 10.1111/j.1399-5448.2009.00568.x] [Medline: 19754613]
5. Galtier F. Definition, epidemiology, risk factors. Diabetes Metab 2010 Dec;36(6 Pt 2):628-651. [doi: 10.1016/j.diabet.2010.11.014] [Medline: 21163426]
6. Massin P, Erginay A, Mercat-Caudal I, Vol S, Robert N, Reach G, et al. Prevalence of diabetic retinopathy in children and adolescents with type-1 diabetes attending summer camps in France. Diabetes Metab 2007 Sep;33(4):284-289. [doi: 10.1016/j.diabet.2007.03.004] [Medline: 17625942]
7. Wake Forest School of Medicine. SEARCH for Diabetes in Youth. 2019. URL: https://www.searchfordiabetes.org/dspHome.cfm [accessed 2020-08-22]
8. Beck RW, Tamborlane WV, Bergenstal RM, Miller KM, DuBose SN, Hall CA, T1D Exchange Clinic Network. The T1D Exchange clinic registry. J Clin Endocrinol Metab 2012 Dec;97(12):4383-4389. [doi: 10.1210/jc.2012-1561] [Medline: 22996145]
9. Chi GC, Li X, Tartof SY, Slezak JM, Koebnick C, Lawrence JM. Validity of ICD-10-CM codes for determination of diabetes type for persons with youth-onset type 1 and type 2 diabetes. BMJ Open Diabetes Res Care 2019;7(1):e000547 [FREE Full text] [doi: 10.1136/bmjdrc-2018-000547] [Medline: 30899525]
10. Filipp SL, Cardel M, Hall J, Essner RZ, Lemas DJ, Janicke DM, et al. Characterization of adult obesity in Florida using the OneFlorida clinical research consortium. Obes Sci Pract 2018 Aug;4(4):308-317 [FREE Full text] [doi: 10.1002/osp4.274] [Medline: 30151226]
11. U.S. Census Bureau. American Community Survey 1-year estimates. Census Reporter Profile page for Florida. 2017. URL: http://censusreporter.org/profiles/04000US12-florida/ [accessed 2020-08-22]
12. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]

13. Jaiswal M, Divers J, Urbina EM, Dabelea D, Bell RA, Pettitt DJ, SEARCH for Diabetes in Youth Study Group. Cardiovascular autonomic neuropathy in adolescents and young adults with type 1 and type 2 diabetes: The SEARCH for Diabetes in Youth Cohort Study. Pediatr Diabetes 2018 Jun;19(4):680-689 [FREE Full text] [doi: 10.1111/pedi.12633] [Medline: 29292558]

Abbreviations

IDR: Integrated Data Repository
PPV: positive predictive value
T1DM: type 1 diabetes mellitus
T2DM: type 2 diabetes mellitus
UF Health: University of Florida Health system

Edited by M Focsa; submitted 25.03.20; peer-reviewed by S Oser, R Bush, C Wu; comments to author 05.05.20; revised version received 30.06.20; accepted 16.07.20; published 04.09.20.

Please cite as:
Morris HL, Donahoo WT, Bruggeman B, Zimmerman C, Hiers P, Zhong VW, Schatz D
An Iterative Process for Identifying Pediatric Patients With Type 1 Diabetes: Retrospective Observational Study
JMIR Med Inform 2020;8(9):e18874
URL: http://medinform.jmir.org/2020/9/e18874/
doi:10.2196/18874
PMID:32886067

Original Paper

Integrating and Evaluating the Data Quality and Utility of Smart Pump Information in Detecting Medication Administration Errors: Evaluation Study

Yizhao Ni[1,2], PhD; Todd Lingren[1,2], MSc; Hannah Huth[3,4], BA; Kristen Timmons[2], BA; Krisin Melton[2,5], MD; Eric Kirkendall[1,2,3,6], MBI, MD

[1]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

[2]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States

[3]Wake Forest Center for Healthcare Innovation, Wake Forest School of Medicine, Winston Salem, NC, United States

[4]Indiana University, Bloomington, IN, United States

[5]Division of Neonatology and Pulmonary Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

[6]Department of Pediatrics, Wake Forest School of Medicine, Winston Salem, NC, United States

Corresponding Author:
Yizhao Ni, PhD
Division of Biomedical Informatics
Cincinnati Children's Hospital Medical Center
3333 Burnet Avenue
MLC 7024
Cincinnati, OH, 45229
United States
Phone: 1 5138034269
Email: yizhao.ni@cchmc.org

Abstract

Background: At present, electronic health records (EHRs) are the central focus of clinical informatics given their role as the primary source of clinical data. Despite their granularity, the EHR data heavily rely on manual input and are prone to human errors. Many other sources of data exist in the clinical setting, including digital medical devices such as smart infusion pumps. When incorporated with prescribing data from EHRs, smart pump records (SPRs) are capable of shedding light on actions that take place during the medication use process. However, harmoniz-ing the 2 sources is hindered by multiple technical challenges, and the data quality and utility of SPRs have not been fully realized.

Objective: This study aims to evaluate the quality and utility of SPRs incorporated with EHR data in detecting medication administration errors. Our overarching hypothesis is that SPRs would contribute unique information in the med-ication use process, enabling more comprehensive detection of discrepancies and potential errors in medication administration.

Methods: We evaluated the medication use process of 9 high-risk medications for patients admitted to the neonatal inten-sive care unit during a 1-year period. An automated algorithm was developed to align SPRs with their medica-tion orders in the EHRs using patient ID, medication name, and timestamp. The aligned data were manually re-viewed by a clinical research coordinator and 2 pediatric physicians to identify discrepancies in medication ad-ministration. The data quality of SPRs was assessed with the proportion of information that was linked to valid EHR orders. To evaluate their utility, we compared the frequency and severity of discrepancies captured by the SPR and EHR data, respectively. A novel concordance assessment was also developed to understand the detec-tion power and capabilities of SPR and EHR data.

Results: Approximately 70% of the SPRs contained valid patient IDs and medication names, making them feasible for data integration. After combining the 2 sources, the investigative team reviewed 2307 medication orders with 10,575 medication administration records (MARs) and 23,397 SPRs. A total of 321 MAR and 682 SPR dis-crepancies were identified, with vasopressors showing the highest discrepancy rates, followed by narcotics and total parenteral nutrition. Compared with EHR MARs, substantial dosing discrepancies were more commonly detectable using the SPRs. The concordance analysis showed little overlap between MAR and SPR discrepan-cies, with most discrepancies captured by the SPR data.

Conclusions: We integrated smart infusion pump information with EHR data to analyze the most error-prone phases of the medication lifecycle. The findings suggested that SPRs could be a more reliable data source for medication error detection. Ultimately, it is imperative to integrate SPR information with EHR data to fully detect and mitigate medication administration errors in the clinical setting.

(JMIR Med Inform 2020;8(9):e19774) doi:10.2196/19774

KEYWORDS

medication administration errors; smart infusion pumps; electronic health records; concordance

Introduction

Background

Electronic health records (EHRs) are the central focus of many efforts in clinical research and quality improvement given their role as the primary clinical data source [1-4]. Despite their granularity, the data heavily rely on manual input and are prone to human errors [3,4]. Many digital devices have been used in clinical environments, and they provide additional sources of data for understanding health care processes, a form of real-world data from clinical settings. One example is digital medication infusion pumps, more commonly known as *smart pumps*. These pumps, which are now commonplace in modern health care settings, record copious amounts of rich, granular data about medication administration. Smart pumps have been shown to prevent some errors while propagating others. One systematic review found that smart pumps can intercept multiple error types, such as wrong dose and wrong rate errors, as well as reduce adverse drug events [5]. This effect, however, is heavily dependent on user compliance and utilization of specific functionalities vendor products afford, including dose error reduction software [6]. As with EHRs, infusion pump alerts are another source of alert burden and are subject to alert fatigue, which raises a trade-off between potential safety benefits and increased workload for providers [7]. Smart pumps, compared with their analogue counterparts, generate a lot of data to log user interaction with the pumps (eg, pausing of pump infusions and pump alert overrides) and the pump status (eg, infusion start and infusion complete), which are associated with granular timestamps. The data create useful information such as user compliance with alerts, pump states at different time points, and mechanical alarm records. These smart pump records (SPRs) can be harnessed to help understand actions that take place during medication administration.

The ability to link and leverage different data sources across the full medication lifecycle has the potential to make medication errors recognizable and rectifiable. Theoretically, when combining smart infusion pump information with prescribing data from EHRs, one can see the bookends of the medication use process, from medication origin (order) to terminus (administration). Although there are gaps in the intermediate steps (traditionally the transcription and dispensing stages), given that errors are more frequent in the ordering and administration phases [8], integrating the EHR and SPR data permits visibility in the most error-prone phases. As such, the harmonization of these 2 data sources can provide insight and information about safe and unsafe practices.

Our research is specifically directed at developing accurate and scalable informatics technologies to monitor the medication use process and detect medication administration errors. In our previous studies, we developed artificial intelligence–based algorithms for monitoring the use of high-risk medications including vasopressors, narcotics, insulin, total parenteral nutrition (TPN), and fluids [1,9,10]. By analyzing order and medication administration record (MAR) data residing in EHRs, the algorithms identified discrepancies and potential errors in how medications were being ordered and documented as administered. Despite their viability in discrepancy detection, the algorithms relied on a single data source that resulted in a number of false positives and false negatives. For instance, the algorithms might miss an error in administration (a false negative) if an order adjustment was not placed in the EHR or was incorrectly documented in the MAR [10]. To improve the accuracy of error detection, we sought to integrate smart pump information into the computerized algorithms.

Integration of SPRs with EHR data requires advanced informatics technologies and is not without significant challenges [4]. Most health care institutions that use smart pumps have not fully integrated them into a *closed-loop* system, which would permit automatic linkage of order data in the EHRs to administration information from the pumps. One barrier to integration is the cost and complexity of the implementation process, including its impact on clinical workflows. Another barrier is the maturity of the technology and its associated challenges [11]. Single-site studies have reported increased work efficiencies and revenue benefits, but widespread integration is not yet ubiquitous [12]. As such, although great potential exists, the insight gained from combining the data has not yet been realized.

Objectives

To fill these gaps in knowledge, we integrated SPRs with EHR data and evaluated their quality and utility in detecting medication administration errors. Our overarching hypothesis was that SPRs would contribute unique information in the medication use process, enabling more comprehensive detection of discrepancies and potential errors in medication administration. The specific aims of this study were to (1) develop an automated algorithm that aligns SPRs with EHR data to facilitate manual review of medication administration, (2) characterize discrepancies identified from EHRs and SPRs, and (3) develop a novel assessment that measures the concordance between the ability of EHR and SPR data in detecting medication administration discrepancies. This study is among the first to integrate multiple clinical data sources to understand medication safety events. Our long-term objective is to establish a more effective and generalizable program that assembles comprehensive data sources in clinical environments to improve patient safety across health care institutions.

Methods

Setting and Study Population

We evaluated medication administration for patients admitted to the neonatal intensive care unit (NICU) at the Cincinnati Children's Hospital Medical Center (CCHMC). Approval for this study was provided by the CCHMC institutional review board (study ID: 2015-3824), and a waiver of consent was authorized.

The CCHMC NICU is a level 4 NICU that provides the highest level of neonatal intensive care to complex and critically ill newborns. The unit has an average daily census of 70 patients and an average of 750 admissions per year. The institution utilizes a fully computerized commercial EHR system (Epic Systems Corporation). Additional NICU safety interventions include the use of computerized provider order entry with embedded clinical decision support, a bar code medication administration (BCMA) system, smart infusion pump technology with a customized neonatal library of medications, daily prescription review by dedicated NICU pharmacists, and clinical guidelines for high-risk medications.

Study Medications and Study Periods

We focused on reconciling 9 high-risk, continuous intravenous infusion medications prescribed to NICU inpatients, including vasopressors (dopamine, dobutamine, epinephrine, milrinone, and vasopressin), narcotics (fentanyl and morphine), TPN, and lipids. Continuous intravenous infusions have a higher risk and severity of error than other medication administrations [13,14]. In particular, its administration usually spans multiple nursing shifts and involves complex dosage adjustments that are not captured by in-place interventions such as BCMA. Medication administrations for vasopressors and narcotics were studied over the period of January 1, 2014, to December 31, 2014. Due to changes to our ordering system, TPN and lipid administrations were studied over the period of January 1, 2016, to December 31, 2016. All vasopressors, narcotics, and lipid orders were included in the analysis. Owing to the large volume of TPN orders, we randomly selected 8.05% (669/8308) of the TPN orders for analysis.

Clinical Data Extraction and Federation

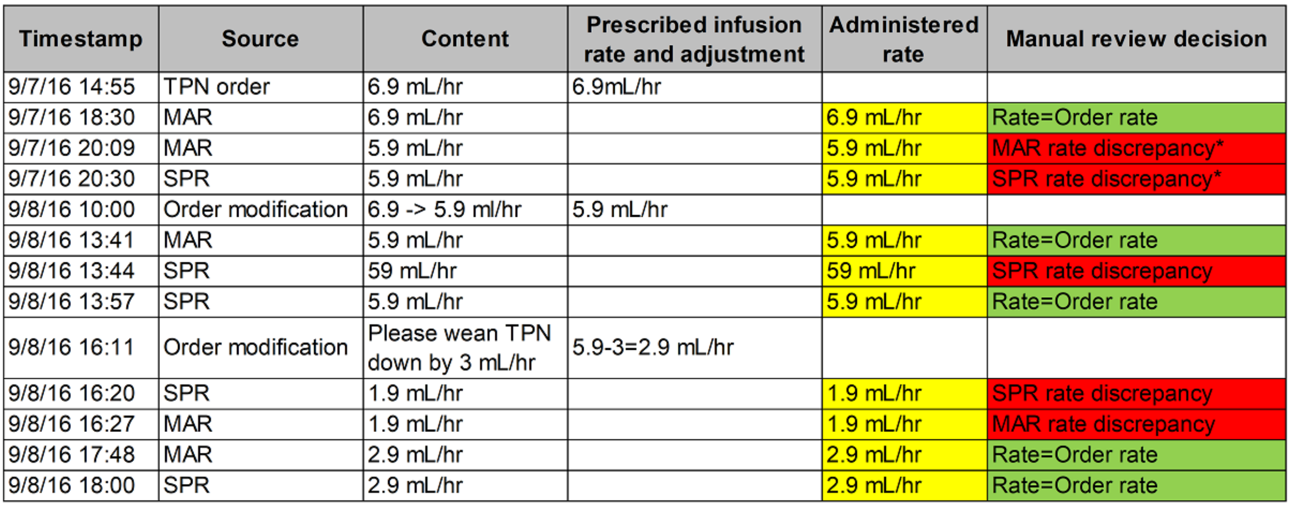
Medication use information was extracted retrospectively from the institution's EHR system. The information included (1) medication orders that documented infusion doses (or infusion rates) prescribed to the patients, (2) structured order modifications that adjusted the original doses and rates via computerized physician order entry, (3) MARs documented by clinical professionals that describe doses or rates administered to patients, and (4) free-text physician to nurse communication orders that specified complex medication dose or rate adjustments during patient care. The infusion pump information was extracted separately from the vendor-provided reporting system (CareFusion). The information included (1) patient IDs, (2) medication names, and (3) SPRs that documented actual doses or rates administered to patients. The SPRs contained multiple pump state categories including infusion started or restarted, stopped, completed, paused, canceled, and delayed. Only SPRs that indicated infusion started or restarted were used for this analysis because they represented the initiation of medication delivery and the point at which one would want to intercept potential erroneous infusions.

As the SPRs were not integrated into the institution's EHR system, there was no explicit association between an SPR and its corresponding medication order. As such, we developed a computerized algorithm to merge the 2 data sources and align SPRs with their potential medication orders. The EHR and SPR data were first grouped by patient IDs and medication names. The algorithm then chronologically aligned the EHR and SPR data for each patient medication group, where each SPR was linked to the closest medication order with order placement, modification, or MAR documented within 24 hours of its administration. The SPRs with invalid patient IDs or unknown medication names could not be definitely linked to any order. As such, they were excluded from subsequent manual review and discrepancy analysis.

Manual Review for Gold Standard Creation

A clinical research coordinator (CRC) and 2 board-certified pediatric physicians on the research team (including 1 neonatologist) manually reviewed the aligned data for each patient medication group to identify medication administration discrepancies in MARs and SPRs. Figure 1 illustrates an example of the chronological ordering of EHR and SPR data and discrepancies identified by manual review. A discrepancy was defined as a mismatch between the prescribed dose or rate of a medication and the electronic documentation of its administration in MARs or SPRs [10]. A discrepancy may be a medication administration error, or it may be a false positive subject to further investigation. For the purposes of this study, we defined an a priori 30-min window to allow for verbal orders to be transcribed into the EHR, in line with our institutional policy and expectations. As such, a discrepancy occurred if an order was placed more than 30 min after an administration, even if the correct dose or rate was administered (the starred discrepancy in Figure 1). If a discrepancy was detected, the reviewers additionally identified the correct dose or rate prescribed. Differences between the reviewers' decisions were resolved during the adjudication sessions. Inter-rater reliability was calculated using Cohen kappa to define the agreement [15].

Figure 1. An example of chronological ordering of medication use data and medication administration records or smart pump record discrepancies identified by manual review. The discrepancy occurred because the order modification was placed over 30 min after the MAR or SPR, which did not meet the institutional expectations even if the administration was correct. MAR: medication administration records; SPR: smart pump record; TPN: total par-enteral nutrition.



Timestamp	Source	Content	Prescribed infusion rate and adjustment	Administered rate	Manual review decision
9/7/16 14:55	TPN order	6.9 mL/hr	6.9mL/hr		
9/7/16 18:30	MAR	6.9 mL/hr		6.9 mL/hr	Rate=Order rate
9/7/16 20:09	MAR	5.9 mL/hr		5.9 mL/hr	MAR rate discrepancy*
9/7/16 20:30	SPR	5.9 mL/hr		5.9 mL/hr	SPR rate discrepancy*
9/8/16 10:00	Order modification	6.9 -> 5.9 ml/hr	5.9 mL/hr		
9/8/16 13:41	MAR	5.9 mL/hr		5.9 mL/hr	Rate=Order rate
9/8/16 13:44	SPR	59 mL/hr		59 mL/hr	SPR rate discrepancy
9/8/16 13:57	SPR	5.9 mL/hr		5.9 mL/hr	Rate=Order rate
9/8/16 16:11	Order modification	Please wean TPN down by 3 mL/hr	5.9-3=2.9 mL/hr		
9/8/16 16:20	SPR	1.9 mL/hr		1.9 mL/hr	SPR rate discrepancy
9/8/16 16:27	MAR	1.9 mL/hr		1.9 mL/hr	MAR rate discrepancy
9/8/16 17:48	MAR	2.9 mL/hr		2.9 mL/hr	Rate=Order rate
9/8/16 18:00	SPR	2.9 mL/hr		2.9 mL/hr	Rate=Order rate

Analysis of Discrepancies

The consensus annotations served as a gold standard to understand medication use processes and discrepancies. To assess the data quality, we analyzed the proportion of valid information in the SPR source. An SPR was valid if it contained both a valid patient ID and a medication name. An ID was considered valid if its value mapped to an existing patient ID in the NICU. Clinical staff manually entered patient IDs into the infusion pumps; hence, invalid IDs may represent entry or programming errors. Medication names were present in the SPRs if the staff selected their profile from the pump drug library. Infusions programmed under a generic *basic* infusion status did not have a medication associated with the records. We then investigated the number of discrepancies identified by MARs and SPRs to characterize the scale of discrepancies captured by the 2 sources. The magnitude of discrepancy (MoD), as defined by the percentage of discrepancy over a correct dose or rate, was also calculated to quantify the severity of a discrepancy. Finally, we developed a concordance assessment to understand the detection power and capabilities of the MAR data alone, the SPR data alone, and their overlap. We hypothesized that MAR discrepancies often represented documentation errors. As such, the use of a concordance measure could help differentiate documentation issues versus true administration errors, reflected by the concordance of the MAR and SPR discrepancies or SPR discrepancies alone. The assessment first divided each order sequence into multiple event blocks separated by order modifications (either order initiations or modifications and audits). It then identified whether an event block contained 1 of 4 categories: no discrepancies, MAR-only discrepancies, SPR-only discrepancies, or both MAR and SPR discrepancies. For example, the TPN order in Figure 1 contained 3 event blocks, 1 with an SPR-only discrepancy and 2 with both MAR and SPR discrepancies. Medication orders containing both MARs and SPRs were included in the analysis. The descriptive statistics of the 4 categories were calculated for each medication and in aggregation to study the concordance. Cohen kappa was calculated to assess the agreement between MAR and SPR discrepancies.

Results

Table 1 presents the distribution of SPRs with and without valid patient IDs and medication names in the SPR data source. A total of 543,791 out of 764,624 SPRs (71.11%) in 2014 and 521,113 out of 787,692 SPRs (66.16%) in 2016 contained valid patient IDs and medication names and were therefore feasible for data federation. Of the 220833 invalid SPRs in 2014, 66.7% (147,304) were because of invalid patient IDs, 52,680 (23.%) were because of missing medication names, and 20,849 (9.4%) were because of missing identifiers. A similar distribution of invalid SPRs was observed in the 2016 data. Table 2 shows the distribution and categorization of valid and invalid patient IDs documented in the SPRs.

Table 1. The distribution of smart pump records with and without valid patient IDs and medication names.

Data sources	Patient ID+[a], n (%)	Patient ID−[b], n (%)	Total, n (%)
2014 data			
Medication name+	543,791 (71.12)	147,304 (19.26)	691,095 (90.38)
Medication name−	52,680 (6.89)	20,849 (2.73)	73,529 (9.62)
Total	596,471 (78.01)	168,153 (21.99)	764,624 (100.00)
2016 data			
Medication name+	521,113 (66.16)	175,830 (22.32)	696,943 (88.48)
Medication name−	63,326 (8.04)	27,423 (3.48)	90,749 (11.52)
Total	584,439 (74.20)	203,253 (25.80)	787,692 (100.00)

[a] The patient ID or medication name was valid.

[b] The patient ID or medication name was invalid or missing.

Table 2. Descriptive statistics of patient IDs in smart pump records and categorization of invalid patient IDs.

Groups	Year	
	2014	2016
Valid patient IDs, n	569	476
Invalid patient IDs, n	173	148
Date of birth out of range (1965-2014), n (%)	42 (24.3)	25 (16.9)
Entering patient names instead of IDs, n (%)	33 (19.1)	10 (6.8)
Missing digits in patient IDs, n (%)	30 (17.2)	23 (15.5)
Random numbers, n (%)	23 (13.3)	10 (6.8)
Entering encounter IDs instead of patient IDs, n (%)	20 (11.6)	49 (33.1)
Invalid letters in patient IDs, n (%)	13 (7.5)	8 (5.4)
Expired patient IDs due to merged charts, n (%)	4 (2.3)	10 (6.8)
Extra digits in patient IDs, n (%)	4 (2.3)	10 (6.8)
Potential typographical errors in patient IDs, n (%)	4 (2.3)	3 (2.0)

Table 3 presents the descriptive statistics of the medication use data. The CRC and physicians reviewed 2307 medication orders with 10,575 MARs and 23,397 SPRs during the study period. A total of 321 discrepancies were identified from MARs (discrepancy rate 321/10,575, 3.0%), and 682 discrepancies were identified from SPRs (discrepancy rate 682/23,397, 2.9%). The overall inter-rater reliabilities were 0.92/0.90 (MAR/SPR), indicating almost perfect agreement on decision making [15]. Among the targeted medications, vasopressors including epinephrine, dopamine, and vasopressin had the highest discrepancy rates, followed by narcotics (fentanyl) and TPN. The SPR discrepancy rates were higher than that of MARs for all medications except epinephrine.

Table 3. Descriptive statistics of the gold standard medication use data.

Medication	Patients, n (%)	Orders, n (%)	MAR[a], n (%)	MAR discrepancy, n (% total[b])	MAR discrepancy rate, %	SPRs[c], n (%)	SPR discrepancy, n (% total)	SPR discrepancy rate, %
Dobutamine	1 (0.2)	3 (0.1)	7 (0.1)	0 (0.0)	0.0	18 (0.1)	0 (0.0)	0.0
Dopamine	10 (1.6)	18 (0.8)	60 (0.6)	4 (1.2)	6.7	152 (0.6)	13 (1.9)	8.6
Epinephrine	87 (13.6)	325 (14.1)	1937 (18.3)	233 (72.6)	12.0	2994 (12.8)	290 (42.5)	9.7
Fentanyl	38 (6.0)	134 (5.8)	723 (6.8)	9 (2.8)	1.2	2332 (10.0)	87 (12.8)	3.7
Lipid	179 (28.1)	604 (26.2)	1725 (16.3)	0 (0.0)	0.0	1884 (8.1)	6 (0.9)	0.3
Milrinone	33 (5.2)	71 (3.1)	744 (7.0)	3 (0.9)	0.4	1188 (5.1)	14 (2.1)	1.2
Morphine	110 (17.2)	434 (18.8)	2850 (27.0)	13 (4.0)	0.5	10,051 (43.0)	150 (22.0)	1.5
TPN[d]	160 (25.1)	669 (29.0)	2281 (21.6)	43 (13.4)	1.9	4524 (19.3)	105 (15.4)	2.3
Vasopressin	20 (3.1)	49 (2.1)	248 (2.3)	16 (5.0)	6.5	254 (1.1)	17 (2.5)	6.7
Overall	638 (100.0)	2307 (100.0)	10,575 (100.0)	321 (100.0)	3.0	23,397 (100.0)	682 (100.0)	2.9

[a] MAR: medication administration record.

[b] The numbers in parentheses represent the percentage of total discrepancies attributable to a medication.

[c] SPR: smart pump record.

[d] TPN: total parenteral nutrition.

Table 4 presents the MoD for MARs and SPRs across all targeted medications. A total of 58.2% (187/321) of the MAR discrepancies were overdoses, of which 21.9% (41/187) were substantial overdoses (administered dose was 100% greater than the prescribed dose). A total of 66.9% (456/682) of the SPR discrepancies were overdoses, of which 27.6% (126/456) were substantial overdoses. The few discrepancies with 0% magnitude represent documentation issues where the administrated doses or rates were correct but the orders or order modifications were placed more than 30 min after administration. Figure 2 depicts the MoD distributions for MARs and SPRs over discrepancy categories. Epinephrine, fentanyl, morphine, and TPN were responsible for most MAR and SPR discrepancies, particularly for substantial overdoses. Figure 3 depicts the MoD distributions for MARs and SPRs for each medication. Dopamine, epinephrine, and vasopressin showed similar distributions between MARs and SPRs. Other medications such as fentanyl, morphine, and TPN had low numbers of substantial overdoses on MARs but higher numbers on SPRs.

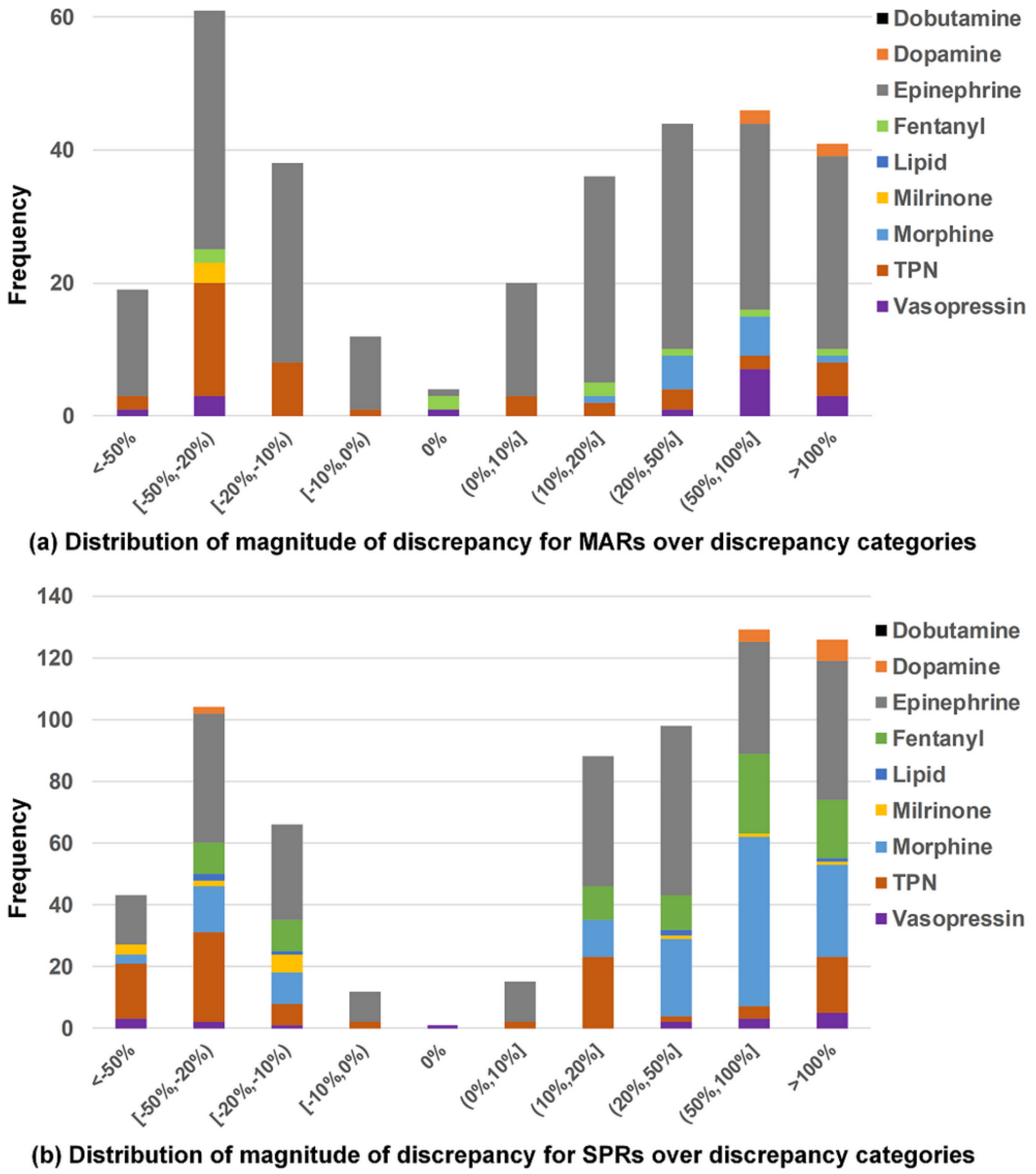
Table 4. Magnitude of discrepancy for medication administration records and smart pump records across all medications.

Data sources	Magnitude of discrepancy, n (%)									
	<−50%	[−50%,−20%)	[−20%,−10%)	[−10%,0%)	0%	(0%,10%]	(10%,20%]	(20%,50%]	(50%,100%]	>100%
MAR[a]	19 (5.9)	61 (19.0)	38 (11.8)	12 (3.7)	4 (1.2)	20 (6.2)	36 (11.2)	44 (13.7)	46 (14.3)	41 (12.8)
SPR[b]	43 (6.3)	104 (15.2)	66 (9.7)	12 (1.8)	1 (0.1)	15 (2.2)	88 (12.9)	98 (14.4)	129 (18.9)	126 (18.5)

[a] MAR: medication administration record.

[b] SPR: smart pump record.

Figure 2. Distribution of magnitude of discrepancy for (a) medication administration records and (b) smart pump records over discrepancy categories. MARs: medication administration records; SPRs: smart pump records; TPN: total parenteral nutrition.



(a) Distribution of magnitude of discrepancy for MARs over discrepancy categories

(b) Distribution of magnitude of discrepancy for SPRs over discrepancy categories

Figure 3. Distribution of magnitude of discrepancy for (a) medication administration records and (b) smart pump records over medications. MARs: medication administration records; SPRs: smart pump records; TPN: total parenteral nutrition.

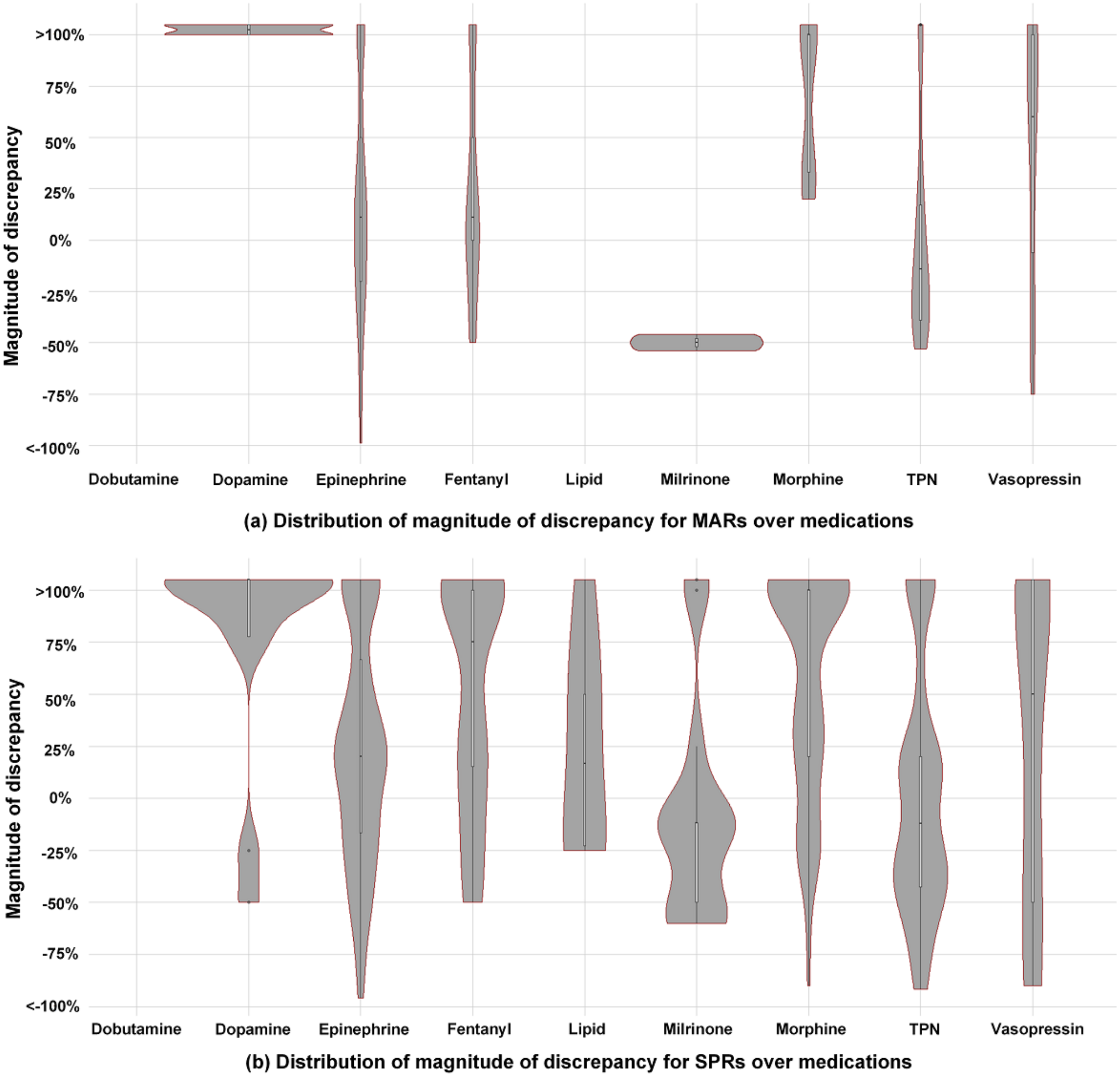


Table 5 presents the concordance between MAR and SPR discrepancies. The analysis included 60.58% (1397/2306) medication orders that contained both MARs and SPRs. The orders were segmented into 2638 event blocks, of which 308 (11.67%) had discrepancies. Of these 308 blocks, 197 (64.0%) contained only SPR discrepancies, 44 (14.3%) contained only MAR discrepancies, and 67 (21.7%) contained both. The Cohen kappa was 0.32, suggesting minimal agreement between MAR and SPR discrepancies [15]. The event blocks with SPR discrepancies were higher than those with MAR discrepancies across all targeted medications.

Table 5. Concordance assessment between medication administration record and smart pump record discrepancies.

Medications	Orders[a], n		Analysis block[b], n	MAR[c] discrepancy[d], n	SPR[e] discrepancy[d], n	Concordance category, n			
	All	Included				None	MAR	SPR	Both
Dobutamine	3	2	3	0	0	3	0	0	0
Dopamine	17	12	40	4	13	35	1	3	1
Epinephrine	325	189	901	196	288	793	25	37	46
Fentanyl	134	82	182	6	79	145	3	32	2
Lipid	604	352	353	0	6	349	0	4	0
Milrinone	71	42	60	1	14	51	1	8	0
Morphine	434	315	631	12	143	529	5	91	6
TPN[f]	669	380	380	27	105	347	8	15	10
Vasopressin	49	23	88	4	17	78	1	7	2
Overall	2306	1397	2638	250	665	2330	44	197	67

[a] All represents the orders in the data set, whereas *Included* represents the orders included in the concordance analysis (ie, the orders having both MAR and SPR data associated with them).

[b] Analysis block represents the event blocks included in the analysis.

[c] MAR: medication administration record.

[d] MAR discrepancy and SPR discrepancy represent the MAR or SPR discrepancies, respectively, found in the analysis blocks and included in the analysis.

[e] SPR: smart pump record.

[f] TPN: total parenteral nutrition.

Discussion

Principal Findings

This study is among the first to integrate smart infusion pump information with EHR data to analyze the most error-prone phases of the medication use process, recognizing that linkage of complex data has its challenges [4]. Smart pump data lack clinical usefulness without appropriate identification of both patient information and medication being used at the time of infusion. One of the main findings was that 29%-34% of the smart pump data were not valid because of missing patient or medication information (Table 1). Most of these missing data were because of invalid patient information, which was caused by mistakes during ID entry on the pumps or unfamiliarity with documentation guidelines during infusion pump programming (Table 2). A large portion of invalid IDs was because of misentries such as missing or adding extra digits, invalid letters, and typographical errors. In addition, a patient ID might be replaced with the patient name or encounter ID, suggesting a mistake in following the documentation guidelines or a workaround. Missing medication information occurred when a basic infusion was selected without specifying the medication being administered. This occurs most commonly as a workaround when the correct medication cannot be found in the smart pump library. When patient or medication identifiers are incorrect or missing, linking smart pump data with order logs or MARs, particularly in real time, becomes vastly more complicated and unreliable. Inference by time of administration is difficult because commonly administered medications (eg, TPN) might have been concurrently ordered for several patients in the same unit.

We identified higher SPR discrepancies than MAR discrepancies (Table 3), suggesting that SPRs could be a more reliable source of error detection than EHR data. This finding also implies that the frequency of medication errors reported in the literature might be underestimated when limited to analysis of EHR data alone [1,9,10]. Both MARs in EHRs and smart pump programming rely on manual data entry and are prone to human error. For example, bar code scanning inputs MAR data into the EHR based on medication label information, but clinical staff must validate the dose, which may change as medications are titrated. Similarly, without a closed-loop system where the pump is automatically programmed by an order, smart pump programming also relies on human data entry. However, compared with MARs, smart pump entries are closest to a patient, representing the truest reflection of what the patient receives. The SPR discrepancies we identified may represent different types of errors. They may be secondary to unintentional misprogramming (ie, the nurse programs an incorrect rate or drug concentration) or misunderstanding (ie, the nurse does not understand an order or misses an order modification), but we are unable to determine the exact causes in this study using only retrospective data. Further studies should investigate the distribution of error types for MARs and SPRs and discuss the effectiveness of corresponding error prevention strategies.

Not all discrepancies have clinical significance, and for most medications, very small discrepancies are not as important as large ones. As observed in our studies, minor discrepancies are typically more numerous (Table 4) [10]. Other studies have also noted that medication errors are numerous but are often small and associated with low rates of harm [7,16-19]. The risk of calling out these frequent, small discrepancies is an increase in workload and decrease in overall attention. It is widely known, for example, that EHR alerts that identify frequent events are perceived as *noisy* (ie, providing erroneous or irrelevant information) and are overridden at high rates [20,21]. As such, we measured the MoDs to assess their severity. Although the analysis demonstrated many minor results, we also detected a notable amount of substantial dosing discrepancies (Table 4). In particular, discrepancies in substantial dosing were dominated by certain medications (eg, epinephrine; Figure 2) and were more commonly detectable from SPR data (Figure 3). These findings suggest the necessity of integrating SPR data into medication error detection, which informs further development of our real-time notification system for medication error events [10]. Imagining a future where multiple data sources are incorporated to detect medication errors in real time, one can see the benefit that the MoD holds. The assessment conveys to a provider not just that an error event has occurred but also the severity of the event to guide his or her clinical response.

Recognizing that SPR and MAR discrepancies may occur together or individually, we developed an assessment to measure their concordance in the same medication cycle (Table 5). Although there were limitations to this methodology, as we had to limit the analysis to only 60.5% (1397/2306) of orders containing both MARs and SPRs, the use of the concordance assessment allowed us to separate documentation issues (MAR-only discrepancies) from true discrepancies (SPR-only discrepancies and matched MAR or SPR discrepancies). Over 85% of the discrepancies were captured by SPRs, implying that the majority were true discrepancies. This trend was consistent across all targeted medications. The kappa statistics suggested that there was little overlap between MAR and SPR discrepancies, and only 21.7% of the discrepancies were captured by both data sources. These novel findings again indicated the necessity of incorporating SPRs into understanding the medication use process. They make smart pump data more clinically and safety relevant, connecting to our ultimate goal of repurposing clinical data to improve the quality of clinical care.

Given that medication discrepancies occur with relative frequency, efforts to improve smart pump use must continue. Several studies have demonstrated the importance of continuous quality improvement with regular assessment of smart pump data [11,12]. Methods that require less device programming, such as the use of barcode scanning pumps, may help reduce pump programming and patient identification errors. Although closed-loop systems have been implemented in several institutions as another method to address these concerns, there have been issues with titration of medications and specific error types remain unmitigated [22-24]. As such, efforts must focus on ways to utilize medication use information from smart pumps to recognize and address errors as quickly as possible. Our ongoing work focuses on incorporating smart pump data into a real-time error notification system and developing new approaches to visualize the medication use process as a means to help frontline clinical providers utilize and learn from the information at hand. By integrating data from multiple sources, we will move medication error detection systems from retrospective and reactive to prospectively preventive and proactive.

Limitations

There are limitations to this study. First, although we were able to utilize approximately 70% of the data, excluding SPRs with missing patient and medication information may have resulted in data bias. Efforts have been initiated to improve the data quality of SPRs via quality improvement. Second, the institution's IT infrastructure does not allow for the delivery of real-time smart pump information, limiting us to medication discrepancy detection and not intervention. Consequently, we must categorize identified events as discrepancies, as we lacked the real-time clinical information to define them as errors. To mitigate this issue, we will increase the frequency of SPR review to a daily basis to capture more real-time information. Third, we chose to focus our work on high-risk continuous medications, and the rates of discrepancy and smart pump issues may differ for intermittent medications. In addition, the study focused on detecting discrepancies in medication doses or rates, which did not capture other error types such as prematurely stopping medications. Future work has been initiated to extend our analysis for intermittent medications and other types of medication use errors. Finally, our data reflected ordering practices and smart pump utilization in a single intensive care unit at a single institution. To assess the generalizability of our findings, project planning and communication are in progress to implement the study in an adult health care institution.

Conclusions

In this study, we integrated smart infusion pump information with EHR data to analyze the most error-prone phases of the medication lifecycle. We identified more discrepancies from SPRs compared with EHR MARs. The MoD assessment also demonstrated that substantial dosing discrepancies were more commonly detectable from SPRs. The concordance analysis showed little overlap between MAR and SPR discrepancies, with most discrepancies captured by the SPR data. The findings suggested that SPRs could be a more reliable data source for medication error detection. Ultimately, it is imperative to integrate SPR information with EHR data to fully detect and mitigate medication administration errors in the clinical setting.

Acknowledgments

The authors wish to acknowledge Teresa Couch for her guidance and domain expertise in infusion pumps and nursing workflows. The authors would also like to thank Eric Hall for his input and assistance in procuring and analyzing the data. This research was supported by the National Institute of Health under award number R01LM012230 (sponsored by the National Library of Medicine).

Authors' Contributions

YN conceptualized the study, coordinated the data extraction, developed the automated algorithms, analyzed the results, created the tables and figures, and wrote the manuscript. TL coordinated the algorithm development, coordinated the result analysis, and contributed to the manuscript. HH coordinated the result analysis and manuscript preparation and contributed to the manuscript. KT reviewed the errors and contributed to the manuscript. KM and EK conceptualized the study, supervised the work, reviewed the errors, provided suggestions in the results analysis, and contributed to the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Li Q, Melton K, Lingren T, Kirkendall ES, Hall E, Zhai H, et al. Phenotyping for patient safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care. J Am Med Inform Assoc 2014;21(5):776-784 [FREE Full text] [doi: 10.1136/amiajnl-2013-001914] [Medline: 24401171]
2. Walsh KE, Adams WG, Bauchner H, Vinci RJ, Chessare JB, Cooper MR, et al. Medication errors related to computerized order entry for children. Pediatrics 2006 Nov;118(5):1872-1879. [doi: 10.1542/peds.2006-0810] [Medline: 17079557]
3. Brown SL, Bogner MS, Parmentier CM, Taylor JB. Human error and patient-controlled analgesia pumps. J Intraven Nurs 1997;20(6):311-316. [Medline: 9423393]
4. Kirkendall ES, Ni Y, Lingren T, Leonard M, Hall ES, Melton K. Data challenges with real-time safety event detection and clinical decision support. J Med Internet Res 2019 May 22;21(5):e13047 [FREE Full text] [doi: 10.2196/13047] [Medline: 31120022]
5. Ohashi K, Dalleur O, Dykes PC, Bates DW. Benefits and risks of using smart pumps to reduce medication error rates: a systematic review. Drug Saf 2014 Dec;37(12):1011-1020. [doi: 10.1007/s40264-014-0232-1] [Medline: 25294653]
6. Rothschild JM, Keohane CA, Cook EF, Orav EJ, Burdick E, Thompson S, et al. A controlled trial of smart infusion pumps to improve medication safety in critically ill patients. Crit Care Med 2005 Mar;33(3):533-540. [doi: 10.1097/01.ccm.0000155912.73313.cd] [Medline: 15753744]
7. Melton KR, Timmons K, Walsh KE, Meinzen-Derr JK, Kirkendall E. Smart pumps improve medication safety but increase alert burden in neonatal care. BMC Med Inform Decis Mak 2019 Nov 7;19(1):213 [FREE Full text] [doi: 10.1186/s12911-019-0945-2] [Medline: 31699078]
8. Fontan J, Maneglier V, Nguyen VX, Loirat C, Brion F. Medication errors in hospitals: computerized unit dose drug dispensing system versus ward stock distribution system. Pharm World Sci 2003 Jun;25(3):112-117. [doi: 10.1023/a:1024053514359] [Medline: 12840964]
9. Li Q, Kirkendall ES, Hall ES, Ni Y, Lingren T, Kaiser M, et al. Automated detection of medication administration errors in neonatal intensive care. J Biomed Inform 2015 Oct;57:124-133 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.012] [Medline: 26190267]
10. Ni Y, Lingren T, Hall ES, Leonard M, Melton K, Kirkendall ES. Designing and evaluating an automated system for real-time medication administration error detection in a neonatal intensive care unit. J Am Med Inform Assoc 2018 May 1;25(5):555-563 [FREE Full text] [doi: 10.1093/jamia/ocx156] [Medline: 29329456]
11. Chaturvedi RR, Etchegaray JM, Raaen L, Jackson J, Friedberg MW. Technology isn't the half of it: integrating electronic health records and infusion pumps in a large hospital. Jt Comm J Qual Patient Saf 2019 Oct;45(10):649-661. [doi: 10.1016/j.jcjq.2019.07.006] [Medline: 31500950]
12. Biltoft J, Finneman L. Clinical and financial effects of smart pump-electronic medical record interoperability at a hospital in a regional health system. Am J Health Syst Pharm 2018 Jul 15;75(14):1064-1068. [doi: 10.2146/ajhp161058] [Medline: 29987060]
13. Maddox PR, Williams CK, Fields M. Intravenous infusion safety initiative: collaboration, evidence-based best practices, and 'smart' technology help avert high-risk adverse drug events and improve patient outcomes. Adv Patient Saf 2008:-. [Medline: 21249948]
14. Westbrook JI, Rob MI, Woods A, Parry D. Errors in the administration of intravenous medications in hospital and the role of correct procedures and nurse experience. BMJ Qual Saf 2011 Dec;20(12):1027-1034 [FREE Full text] [doi: 10.1136/bmjqs-2011-000089] [Medline: 21690248]
15. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
16. Schnock KO, Dykes PC, Albert J, Ariosto D, Call R, Cameron C, et al. The frequency of intravenous medication administration errors related to smart infusion pumps: a multihospital observational study. BMJ Qual Saf 2017 Feb;26(2):131-140. [doi: 10.1136/bmjqs-2015-004465] [Medline: 26908900]

17. Arenas-López S, Stanley IM, Tunstell P, Aguado-Lorenzo V, Philip J, Perkins J, et al. Safe implementation of standard concentration infusions in paediatric intensive care. J Pharm Pharmacol 2017 May;69(5):529-536. [doi: 10.1111/jphp.12580] [Medline: 27334458]
18. Husch M, Sullivan C, Rooney D, Barnard C, Fotis M, Clarke J, et al. Insights from the sharp end of intravenous medication errors: implications for infusion pump technology. Qual Saf Health Care 2005 Apr;14(2):80-86 [FREE Full text] [doi: 10.1136/qshc.2004.011957] [Medline: 15805451]
19. Lyons I, Furniss D, Blandford A, Chumbley G, Iacovides I, Wei L, et al. Errors and discrepancies in the administration of intravenous infusions: a mixed methods multihospital observational study. BMJ Qual Saf 2018 Nov;27(11):892-901 [FREE Full text] [doi: 10.1136/bmjqs-2017-007476] [Medline: 29627799]
20. Kirkendall ES, Kouril M, Minich T, Spooner SA. Analysis of electronic medication orders with large overdoses: opportunities for mitigating dosing errors. Appl Clin Inform 2014;5(1):25-45 [FREE Full text] [doi: 10.4338/ACI-2013-08-RA-0057] [Medline: 24734122]
21. Skledar SJ, Niccolai CS, Schilling D, Costello S, Mininni N, Ervin K, et al. Quality-improvement analytics for intravenous infusion pumps. Am J Health Syst Pharm 2013 Apr 15;70(8):680-686. [doi: 10.2146/ajhp120104] [Medline: 23552046]
22. Shah PK, Irizarry J, O'Neill S. Strategies for managing smart pump alarm and alert fatigue: a narrative review. Pharmacotherapy 2018 Aug;38(8):842-850. [doi: 10.1002/phar.2153] [Medline: 29883535]
23. Tran M, Ciarkowski S, Wagner D, Stevenson JG. A case study on the safety impact of implementing smart patient-controlled analgesic pumps at a tertiary care academic medical center. Jt Comm J Qual Patient Saf 2012 Mar;38(3):112-119. [doi: 10.1016/s1553-7250(12)38015-x] [Medline: 22435228]
24. Trbovich PL, Pinkney S, Cafazzo JA, Easty AC. The impact of traditional and smart pump infusion technology on nurse medication administration performance in a simulated inpatient unit. Qual Saf Health Care 2010 Oct;19(5):430-434 [FREE Full text] [doi: 10.1136/qshc.2009.032839] [Medline: 20427310]

Abbreviations

BCMA: bar code medication administration
CCHMC: Cincinnati Children's Hospital Medical Center
CRC: clinical research coordinator
EHR: electronic health record
MAR: medication administration record
MoD: magnitude of discrepancy
NICU: neonatal intensive care unit
SPR: smart pump record
TPN: total parenteral nutrition

Edited by C Lovis; submitted 01.05.20; peer-reviewed by E Orenstein, J Chaparro; comments to author 24.06.20; revised version received 01.07.20; accepted 03.07.20; published 02.09.20.

Please cite as:
Ni Y, Lingren T, Huth H, Timmons K, Melton K, Kirkendall E
Integrating and Evaluating the Data Quality and Utility of Smart Pump Information in Detecting Medication Administration Errors: Evaluation Study
JMIR Med Inform 2020;8(9):e19774
URL: https://medinform.jmir.org/2020/9/e19774
doi:10.2196/19774
PMID:32876578

Original Paper

Secure Record Linkage of Large Health Data Sets: Evaluation of a Hybrid Cloud Model

Adrian Paul Brown[1], BSc; Sean M Randall[1], DPhil

[1] Centre for Data Linkage, Curtin University, Bentley, Australia

Corresponding Author:
Adrian Paul Brown, BSc
Centre for Data Linkage
Curtin University
Kent Street
Bentley, 6021
Australia
Phone: 61 892669253
Email: adrian.brown@curtin.edu.au

Abstract

Background: The linking of administrative data across agencies provides the capability to investigate many health and social issues with the potential to deliver significant public benefit. Despite its advantages, the use of cloud computing resources for linkage purposes is scarce, with the storage of identifiable information on cloud infrastructure assessed as high risk by data custodians.

Objective: This study aims to present a model for record linkage that utilizes cloud computing capabilities while assuring custodians that identifiable data sets remain secure and local.

Methods: A new hybrid cloud model was developed, including privacy-preserving record linkage techniques and container-based batch processing. An evaluation of this model was conducted with a prototype implementation using large synthetic data sets representative of administrative health data.

Results: The cloud model kept identifiers on premises and uses privacy-preserved identifiers to run all linkage computations on cloud infrastructure. Our prototype used a managed container cluster in Amazon Web Services to distribute the computation using existing linkage software. Although the cost of computation was relatively low, the use of existing software resulted in an overhead of processing of 35.7% (149/417 min execution time).

Conclusions: The result of our experimental evaluation shows the operational feasibility of such a model and the exciting opportunities for advancing the analysis of linkage outputs.

(JMIR Med Inform 2020;8(9):e18920) doi:10.2196/18920

KEYWORDS

cloud computing; medical record linkage; confidentiality; data science

Introduction

Background

In the last 10 years, innovative development of software apps, wearables, and the internet of things has changed the way we live. These technological advances have also changed the way we deliver health services and provide a rapidly expanding information resource with the potential for data-driven breakthroughs in the understanding, treatment, and prevention of disease. Additional information from patient devices, including mobile phone and Google search histories [1], wearable devices [2], and mobile phone apps [3], provides new opportunities for monitoring, managing, and improving health outcomes in new and innovative ways. The key to unlocking these data is in relating details at the individual patient level to provide an understanding of risk factors and appropriate interventions [4]. The linking, integration, and analysis of these data has recently been described as *population data science* [5].

Record linkage is a technique for finding records within and across one or more data sets thought to refer to the same person, family, place, or event [6]. Coined in 1946, the term describes the process of assembling the principal life events of an individual from birth to death [7]. In today’s digital age, the capacity of systems to match records has increased, yet the aim remains the same: linking records to construct individual chronological histories and undertake studies that deliver significant public benefit.

An important step in the evolution of data linkage is to develop infrastructure with the capacity to link data across agencies to create enduring integrated data sets. Such resources provide the capability to investigate many health and social issues. A number of collaborative groups have invested in a large-scale record linkage infrastructure to achieve national linkage objectives [8,9]. The establishment of research centers specializing in the analysis of *big data* also recognizes the issue of increasing data size and complexity [10].

As the demand for data linkage increases, a core challenge will be to ensure that the systems are scalable. Record linkage is computationally expensive, with a potential comparison space equivalent to the Cartesian product of the record sets being linked, making linkage of large data sets (in the tens of millions or greater) a considerable challenge. Optimizing systems, removing manual operations, and increasing the level of autonomy for such processes is essential.

A wide range of software is currently used for record linkage. System deployments range from single desktop machines to multiple servers, with most being hosted internally to organizations. The functional scope of packages varies greatly, with manual processes and scripts required to help manage, clean, link, and extract data. Many packages struggle with large data set sizes.

Many industries have moved toward cloud computing as a solution for high computational workloads, data storage, and analytics [11]. An overview of cloud computing types and service models is shown in Table 1. The business benefits of cloud computing include usage-based costing, minimal upfront infrastructure investment, superior collaboration (both internally and externally), better management of data, and increased business agility [12]. Despite these advantages, uptake by the record linkage industry has been slow. One reason for this is that the storage of identifiable information on cloud infrastructure has been assessed as high risk by data custodians. Although security in cloud computing systems has been shown to be more robust than some in-house systems [13], the media reporting of data breaches has created an impression of insecurity and vulnerability [14]. A culture of risk aversion leaves the record linkage units with expensive, dedicated equipment and computing resources that require managing, maintaining, and upgrading or replacing regularly.

Table 1. Overview of cloud computing types and service models.

Name	Description
Types of cloud computing	
Public	All computing resources are located within a cloud service provider that is generally accessible via the internet.
Private	Computing resources for an organization that are located within the premises of the organization. Access is typically through local network connections.
Hybrid	Cloud services are composed of some combination of public and private cloud services. Public cloud services are typically leveraged in this situation for increasing capacity or capability.
Service models	
IaaS[a]	The provider manages physical hardware, storage, servers, and virtualization, providing virtual machines to the consumer.
PaaS[b]	In addition to the items managed for IaaS, the provider also manages operating systems, middleware, and platform runtimes. The consumer leverages these platform runtimes in their own apps.
SaaS[c]	The provider manages everything, including apps and data, exposing software endpoints (typically as a website) for the consumer.

[a] IaaS: Infrastructure as a Service.

[b] PaaS: Platform as a Service.

[c] SaaS: Software as a Service.

To leverage the advantages of cloud computing, we need to explore operational cloud computing models for record linkage that consider the specific requirements of all stakeholders. In addition, linkage infrastructure requires the development and implementation of robust security and information governance frameworks as part of adopting a cloud *solution*.

Related Work

Some research on algorithms that address the computational burden of the comparison and classification tasks in record linkage has been undertaken. Most work on distributed and parallel algorithms for record linkage is specific to the MapReduce paradigm [15], a programming model for processing large data sets in parallel on a cluster. Few sources detail the comparison and classification tasks themselves, with the focus on load balancing algorithms to address issues associated with data skew. These works attempt to optimize the workload distribution across nodes while removing as many true negatives from the comparison space as possible [16-19]. Load balancing algorithms typically use multiple MapReduce jobs and different indexing methods to tackle the data skew problem. Indexing methods include standard blocking [17,18], density-based blocking [16], and locality sensitive hashing (LSH) [20], with varying success in optimizing the workload distribution.

Pita et al [21] have built on the MapReduce-based work and demonstrated good performance and quality using a Spark-based workflow for probabilistic linkage. Spark was chosen for in-memory processing, ease of programming, scalability, and the new resilient distributed data set model. Like MapReduce, Spark continues to be used to address the issues with linkage and data skew on larger data sets. Spark solutions for full entity resolution are being developed, with different indexing techniques used to address workload distribution. The SparkER tool by Gagliardelli et al [22] uses LSH, meta-blocking, and a block purging process to remove high-frequency blocking keys. Mestre et al [23] presented a sorted neighborhood implementation with an adaptive window size, which uses three Spark transformation steps to distribute the data and minimize data skew.

Outside of the Hadoop ecosystem, which MapReduce and Spark are a part of, there have been some efforts to address the linkage of larger data sets through other parallel processing techniques. Sehili et al [24] presented a modified version of PPJoin, called P4Join, that can parallelize record matching on graphics processing units (GPUs), claiming an execution time improvement of up to 20 times. Despite its potential for significant improvements in runtime performance, there has not been any further work published on P4Join using larger data sets or on clusters of GPU nodes. More recently, Boratto et al [25] evaluated a hybrid algorithm using both GPUs and central processing units (CPUs) with much larger data sets. Although restricted to single (highly specified) machines, these evaluations show promise provided that the approach can be applied within a compute cluster. Again, there is not yet any further work available.

The blocking techniques used in these studies are based on the same techniques used for traditional probabilistic and deterministic linkages [15]. There are many blocking techniques used in these conventional approaches to record linkages that reduce the comparison space significantly, even when running a linkage on a single machine [26]. However, these approaches become inefficient as data set sizes become larger. They also come with a trade-off; the creation of blocks that reduce the comparisons required for linkage will inevitably reduce the coverage of true matches, resulting in more missed matches.

Much of the work in distributed linkage algorithms is focused on performance, often at the expense of linkage accuracy. Adapting these blocking techniques to distribute workload across many compute nodes has reduced the comparisons efficiently. Unfortunately, this increased efficiency has impacted the accuracy further, reducing comparisons at the expense of missing more true matches. There is still a trade-off between performance and accuracy, and further work is required to address it.

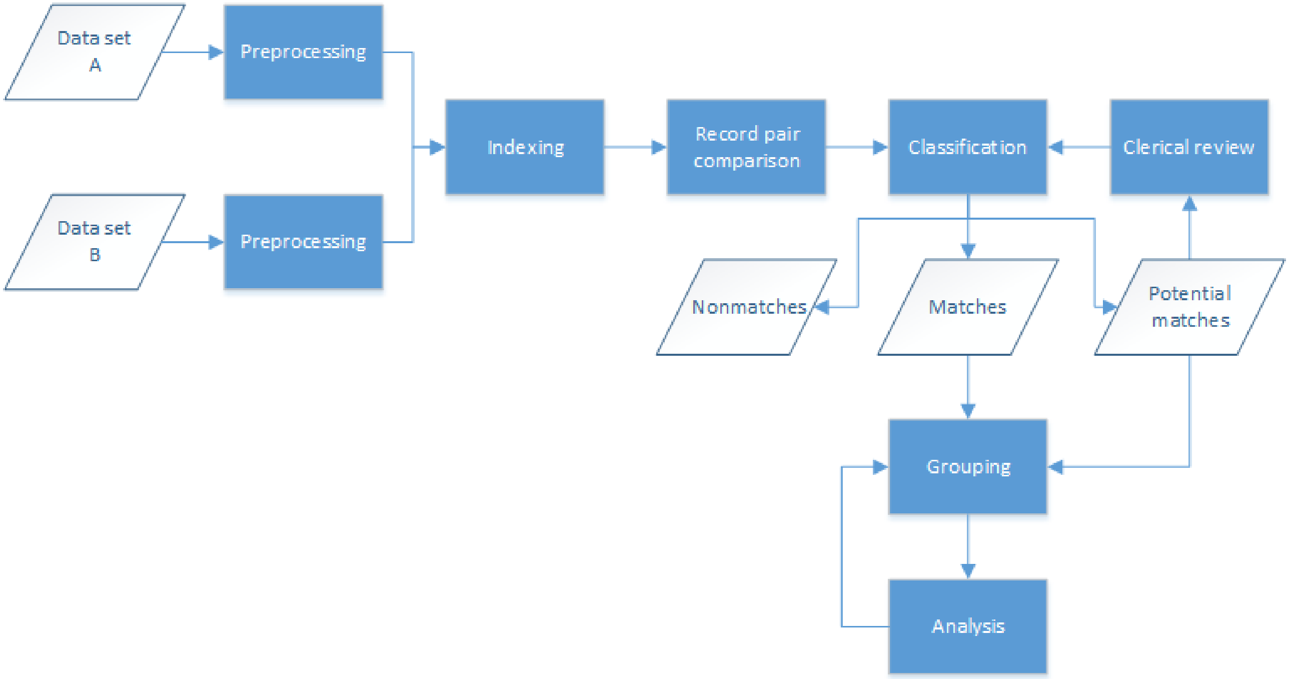
Data Flow and Release for Record Linkage

As data custodians are responsible for the use of their data, researchers must demonstrate to custodians that all aspects of privacy, confidentiality, and security have been addressed. The release of personal identifiers for linkage can be restricted, with privacy regulations such as the Health Insurance Portability and Accountability Act Privacy Rules [27] or EU regulations [28] mandating the use of encrypted identifiers. Standard record linkage methods and software are often unsuitable for linkage based on encrypted identifiers. Privacy-preserving record linkage (PPRL) techniques have been developed to enable linkage on encrypted identifiers [29]. These techniques typically use Bloom filters to store encrypted identifiers, a probabilistic data structure that can be used to approximate the equality of two sets. The emergence of these PPRL methods means that data custodians are not required to release personal identifiers. The use of PPRL methods in operational environments is still in its infancy, with limited tooling available. Available software includes the proprietary LinXmart [30], an R package called PPRL developed by the German Record Linkage Center [31], LSHDB [32], LinkIT [33], and Secure Open Enterprise Master Patient Index [34]. There is little published data on how much these systems are used outside of the organizations that created them. PPRL is a key technology that greatly opens the acceptability of cloud solutions for record linkage.

Record Linkage Process

Record linkage typically follows a standard process for the matching of two or more data sets, as shown in Figure 1. The data sets first undergo some preprocessing, a cleaning and standardization step to ensure consistency with the formatting of fields across data sets. The next step (indexing) attempts to reduce the number of record-level comparisons required (the latter often referred to as the comparison space), removing comparisons that are most likely to be false matches. The indexing step typically groups data sets into overlapping blocks or clusters based on sets of field values and can provide up to 99% reduction in the comparison space. Record pair comparisons occur next, within the blocks or clusters determined during the indexing step; this comparison step is the most computationally expensive and often requires large data sets to be broken down into smaller subsets. Classification of the record pairs into matches, nonmatches, and potential matches results in groups of entities (or individuals) based on the match results. Potential matches can be assessed manually or through special tooling to determine whether they should be classified as matches or nonmatches. A common approach to grouping matches is to merge all records that link together into a single group; however, different approaches can be used to reduce linkage error [35]. Analysis of the entity groups is the last step, where candidate groups are clerically reviewed to determine if and how the records in these groups should be regrouped.

Figure 1. Typical data matching process.



This paper presents 2 contributions to record linkage. First, it offers a model for record linkage that utilizes cloud computing capabilities while providing assurance that data sets remain secure and local. Lessons learned from many real-world record linkage projects, including several PPRL projects, have been instrumental in the design of this cloud model [30,36,37]. Second, the use of containers to distribute linkage workloads across multiple nodes is presented and evaluated within the cloud model.

Methods

Design of a Cloud Model for Record Linkage

The standard record linkage process relies on one party (known as the trusted third party [TTP]) having access to all data sets. Handling records containing identifiable data requires a sound information governance framework with controls in place that manage potential risks. Even with a well-managed information security system in place, access to some data sets may still be restricted. The TTP also requires infrastructure that can help manage data sets, matching processes and linkage key extractions over time. As the number and size of data sets grow, the computational needs and storage capacity must grow with it. However, the computation requirements for data linkage are often sporadic bursts of intense workloads, leaving expensive hardware sitting idle for extended periods.

Dedicated data linkage units in government and academic institutions exist across Australia, Canada, and the United Kingdom, acting as trusted third parties for data custodians. These data linkage units were established from the need to link data for health research at the population level. Some data linkage units are involved in the linkage of other sectors such as justice; however, the primary output of these organizations is linked data for health research. It is essential that a cloud model for record linkage takes into account the linkage practices and processes that have been developed by these organizations.

Our cloud model for record linkage addresses the limitations of data release and the computational needs of the linkage process. Data custodians and linkage units retain control of their identifiable information, while the matching of data sets between custodians occurs within a secure cloud environment.

Tenets of the Record Linkage Cloud Model

The adopted model was founded on 3 overarching design principles:

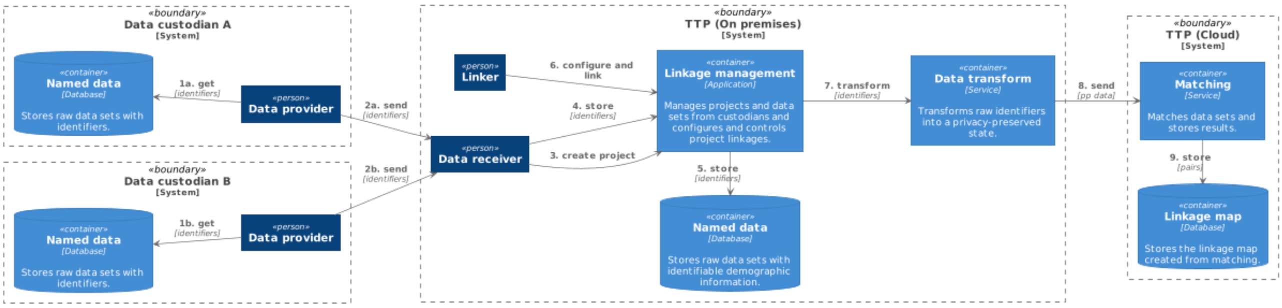
1. *The privacy of individuals in the data is protected.* One of the most important responsibilities for data custodians and linkage units is information security. Data sets contain private, and often sensitive, information on people, and it is vital that appropriate controls are in place to mitigate any potential risks. Some data sets have restrictions on where they can be held, requiring them to be kept local and protected. All computation and storage within the cloud infrastructure must be done on privacy-preserved versions of these data sets.
2. *Computation and storage are outsourced to* the cloud infrastructure. Computation requirements for data linkage are often sporadic bursts of intense workloads, followed by periods of low use or even inactivity. The ability to provision resources for computation as and when required means you only pay for what you use. This computation is generally associated with large sets of input and output data, so it makes sense to keep these data as close to the computation as possible. Storage may not necessarily be cheap, but many cloud computing providers guarantee high levels of durability and availability, with encryption and redundancy capabilities.
3. *Cloud platform services are used over infrastructure services.* Once data are stored within a cloud environment,

additional Platform as a Service offerings for analysis of the data should be leveraged. These are managed services over the top of infrastructure services (such as virtual machines) and can be started and stopped as needed.

High-Level Architectural Model

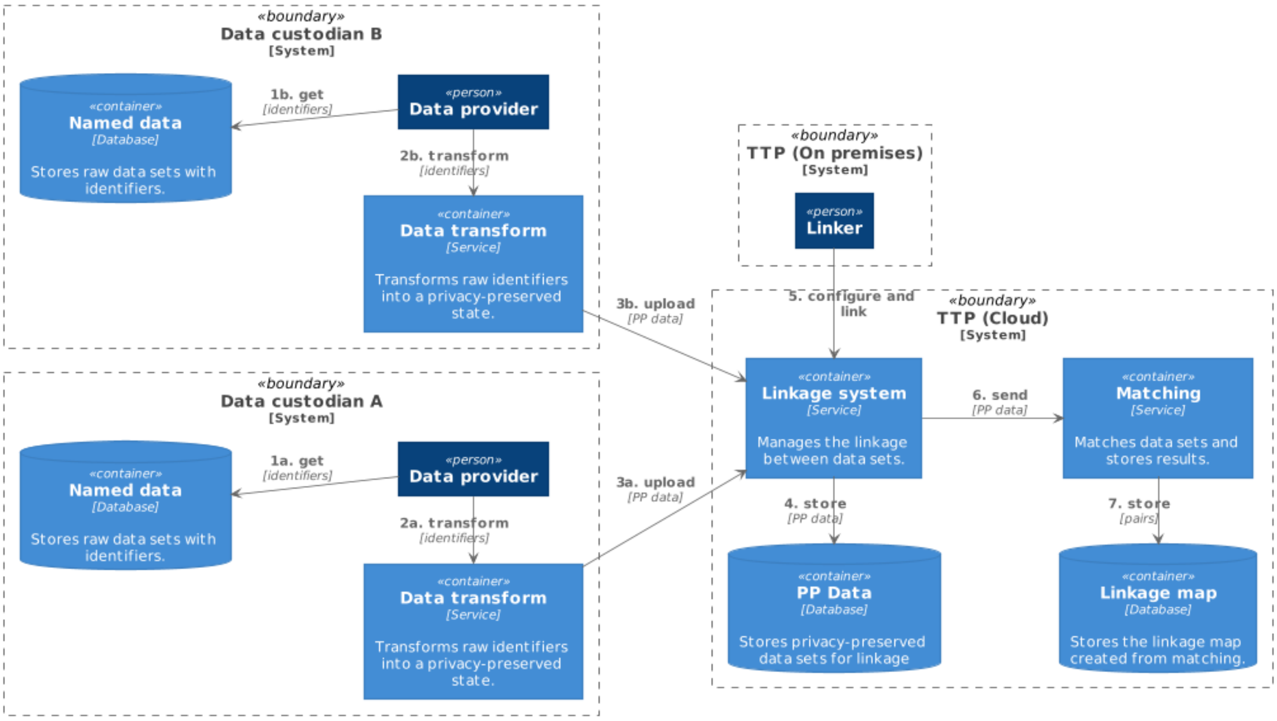
Not all storage and computation can be performed within a cloud environment without impacting privacy; the storage of raw identifiers (such as name, date of birth, and address) must often remain on-premises. The heavy-computational workloads for record linkage, the record pair comparisons and classification, are therefore undertaken on privacy-preserved versions of these data sets. These privacy-preserved data sets must be created on premises and uploaded to cloud storage. The remainder of the linkage process continues within the cloud environment. However, some parts of the classification and analysis steps may be done interactively by the user from an on-premises client app, annotating results from cloud-based analytics with locally stored details (ie, identifiers). An overview of the components and data flows involved in the hybrid TTP model is shown in Figure 2. This model satisfies our cloud model tenets and provides the linkage unit with the ability to scale their infrastructure on-demand. The matching (classification) component can utilize scalable platform services available by the cloud provider to match large privacy-preserved data sets as required. All major cloud providers have platform services that can provide computation on-demand for the processing of big data. The linkage map persists as it contains no identifiable information and can also be analyzed using available cloud platform services.

Figure 2. Hybrid cloud trusted third party model. PP: privacy-preserved; TTP: trusted third party.



Keeping identifiers at the data custodian level (on-premises) while matching on privacy-preserved data within cloud infrastructure enables linkages of data sets *between* data custodians. This model does not require any raw identifiers to be released, and thus, a hybrid model is no longer necessary. The TTP can then be hosted fully in the cloud, as shown in Figure 3. There are 2 immediate ways to achieve this: either one of the custodians manages the cloud infrastructure themselves or an independent third party controls it and provides it as a service to all custodians. A custodian could act as a TTP for all custodians involved in the linkage if this is acceptable to the parties involved. Otherwise, it may be more amenable to go with an independent TTP.

Figure 3. Full cloud trusted third party model. PP: privacy-preserved; TTP: trusted third party.



Although the full cloud TTP model may be useful in some situations, it is unlikely that this would be a desirable model with the dedicated data linkage units. Processes in cleaning, standardization, and quality analysis with personal identifiers have developed and matured over many years. Switching to a model where they no longer have access to personal identifiers would affect the accuracy of the linkage and ultimately the quality of the health research that used the linked data. The hybrid model replaces only the matching component, allowing many existing linkage processes to remain.

Scaling Computation-Heavy Workloads

Record pair comparison and classification tasks are the most computationally intensive tasks in the linkage process, although they are heavily affected by the indexing method used. The single process limitation of most linkage apps makes it difficult to cater to increasingly large data sets, regardless of indexing. Increasing memory and CPU resources for these single-process apps provides some ability to increase capacity, but this may not be sustainable in the longer term.

Although MapReduce appears to be a promising paradigm for addressing large-scale record linkage, 2 issues emerge. First, they consider only the creation of record pairs, whether matches or potential matches, without any thought as to how these record pairs are to come together to form entity groups. The grouping task is also an important part of the data matching process, and the grouping method used can significantly reduce matching errors [35]. Second, MapReduce algorithms do not appear to be readily used, if at all, within an operational linkage environment. Organizational change can be slow, and there is much investment in the existing matching algorithms and apps currently used. It may be operationally more acceptable to continue using these apps where possible.

The comparison and classification tasks of the record linkage process are an embarrassingly parallel problem if the indexing task can produce disjoint sets of record pairs (blocks) for comparison. With the rapid uptake of containerization and the availability of container management and orchestration capability, a viable option for many organizations is to reuse existing apps deployed in containers and run in parallel. Matching tasks on disjoint sets can be run independently and in parallel. The matches and potential matches produced by each matching task can, in turn, be processed independently by grouping tasks. The number of sets that are run in parallel would then only be limited by the number of container instances available.

Indexing solutions are imperfect on real-world data; however, producing disjoint sets for matching is difficult without an unacceptable drop in *pairs completeness* (a measure of the coverage of true positives). There is inevitably some overlap between blocks, as multiple passes with different blocking keys are typically used to ensure accurate results. This overlap prevents independent processing and can be handled in 1 of the 2 ways: (1) the blocks of pairs for classification can be calculated in full before duplicates are removed and the classification task can be run or (2) duplicate matches and potential matches are removed following the classification task. The main disadvantage of option 1 is that this requires a potentially massive set of pairs to be created upfront, as the comparison space is typically orders of magnitude larger than the set of matches and potential matches. Many linkage systems combine their indexing and classification tasks for efficiency, and it is often easier to ignore duplicate matches until completion. The disadvantage of option 2 is that overlapping block sets result in overlapping match sets, preventing the independent grouping of matches from each classification task.

Regardless of the indexing method used to reduce the comparison space for matching, the resulting blocks require grouping into manageable size bins that can be distributed to parallel tasks. A *bin*, therefore, refers to a subset of record pairs grouped together for efficient matching. Block value frequencies are calculated across data sets and used to calculate the size of the total comparison space. Records from these data sets are then copied into separate bins such that each bin has a comparison space of approximately equal size to every other bin.

Using this method, the comparison and classification of each bin are free to be executed on whatever compute capability is available. A managed container cluster is an ideal candidate; however, the container's resources (CPU, memory, and disk) and the bin characteristics (eg, maximum comparison space) need to be carefully chosen to ensure efficient resource use.

Development and Experimental Evaluation of the Prototype

An evaluation of the hybrid cloud linkage model was conducted through the deduplication of different sized data sets on a prototype system. The experiments were designed to evaluate parallel matching using an existing matching app on a cluster of containers; to measure encryption, transfer, and execution times; and to assess the remote analysis of the matching pairs created.

A prototype system was developed with the on-premises component running on Microsoft Windows 10 and the cloud components running on Amazon Web Services (AWS). The prototype focused on the matching part of the linkage model and utilized platform services where available. These services are described in Table 2.

Table 2. Amazon Web Services used.

AWS[a]	Description
S3	Provides an object (file) storage service with security, scalability, and durability.
Glue	A fully managed extract, transform, and load service, providing table definition, schema discovery, and cataloging. Used in conjunction with S3 to expose cataloged files to other AWS services.
Step function	A managed state machine with workflows involving other AWS. The output of a step that uses a particular service can then be used as the input for the next step.
Batch	A fully managed service for running batches of compute jobs. Compute resources are provisioned on-demand.
Athena	An interactive query service for analyzing data in S3 using standard Structured Query Language.

[a] AWS: Amazon Web Services.

Test Data

Three synthetic data sets were generated to simulate population-level data sets: 7 million records, 25 million records, and 50 million records. Although 7 million records may not necessarily represent a large data set, a 50 million record data set is challenging for most linkage units. The data sets were created with a deliberately large number of matches per entity to increase the comparison space and to challenge the matching algorithm.

Data generation was conducted using a modified version of the Febrl data generator [38], an open-source data linkage system written in Python. Frequency distributions of the names and dates of birth of the population of Western Australia were used to generate the synthetic data sets. Randomly selected addresses were sourced from Australia’s National Address File, a publicly available data set [39]. Each data set contained first name, middle name, last name, date of birth, sex, address, and postcode fields. Each field had its own rate of errors and distribution of types of errors. These were based on previously published synthetic data error rates, deliberately set high to challenge matching accuracy [40]. Type of errors included replacement of values, field truncation, misspellings, deletions, insertions, use of alternate names, and values set to missing. Records had anywhere between zero to many thousands of duplicates within the data sets.

All available fields were used for matching in a probabilistic linkage. Two separate blocks were used: first name initial and last name Soundex, and date of birth and sex. Each pair output from the matching process included two record IDs, a score, a block (strategy) name, and the individual field-level comparison weights used to calculate the score.

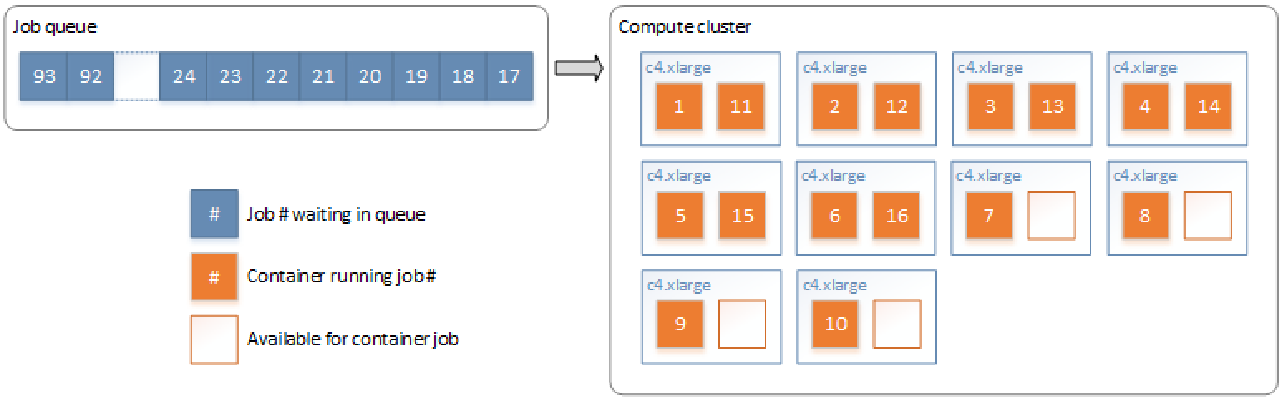
Experiments

The on-premises component first transformed data sets containing named identifiers into a privacy-preserved state using Bloom filters. String fields were split into bigrams that were hashed 30 times into Bloom filters 512 bits in length. Numeric fields (including the specific date of birth elements) were cryptographically hashed using hash-based message authentication code Secure Hash Algorithm 2 (SHA2). These privacy-preserved data sets were compressed (using gzip) before being uploaded to Amazon’s object storage, S3. A configuration file was also uploaded, containing the necessary linkage parameters required for the probabilistic linkage. An AWS step function (a managed state machine) was then triggered to run through a set of tasks to complete the deduplication of the file as defined in the parameter file.

All step function tasks used on-demand resource provisioning for computation. A compute cluster managed by AWS Batch was configured with a maximum CPU count of 40 (10×c4.xlarge instance type). Each container was configured with 3.5 GB RAM and 2 CPUs, allowing up to 20 container instances to run at any one time.

The first task ran as a single job, splitting the file into many bins of approximately equal comparison space, using blocking variables specified in the configuration file. By splitting on the blocking variables, the comparison space for the entire linkage remains unchanged. Each bin was stored in an S3 location with a consistent name suffixed with a sequential identifier. The second task ran a node array batch job, with a job queued for each bin to run on the compute cluster. Docker containers running a command-line version of the LinXmart linkage engine were executed on the compute nodes to deduplicate each bin independently. AWS Batch managed the job queue, assigning jobs to available nodes in the cluster, as shown in Figure 4.

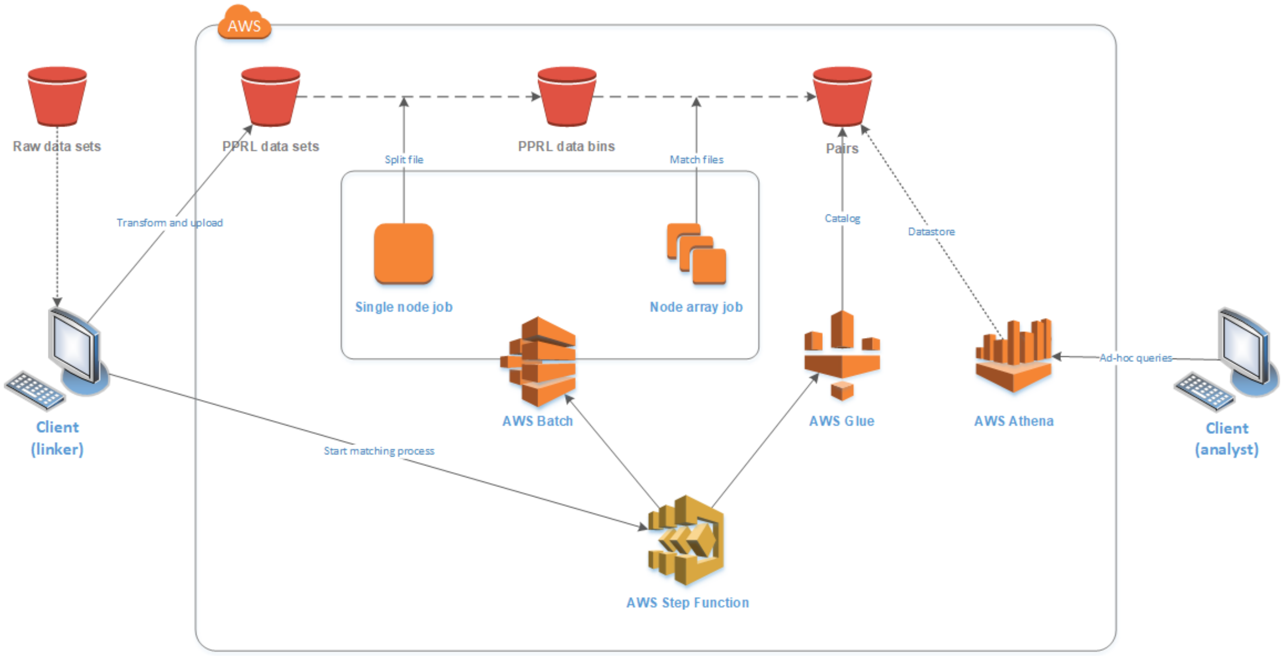
Figure 4. Matching jobs running on compute cluster (one job per bin).



LinXmart is a proprietary data linkage management system, and the LinXmart linkage engine was used because of our familiarity with the program and its ability to run as a Linux command-line tool. It accepts a local source data set and parameter file as inputs and produces a single pairs file as output. There were no licensing issues running LinXmart on AWS in this instance, as our institution has a license allowing unrestricted use. This linkage engine could be substituted, if desired, for others that similarly produce record pair files. The container was bootstrapped with a shell script that downloaded and decompressed the source files from S3 storage, ran the linkage engine program, and then compressed and uploaded the resulting pairs file to S3 storage. Each job execution was passed a sequential identifier by AWS Batch, which was used to identify a source bin datafile to download from S3 and mark the resulting pair file to upload to S3.

The third step function task classified and cataloged all new pairs files, using AWS Glue, making them available for use by other AWS analytical services. The results for each original data set were then able to be presented as a single table, although the data itself were stored as a series of individual text files. The prototype's infrastructure and data flow are shown in Figure 5.

Figure 5. Prototype on Amazon Web Services. AWS: Amazon Web Services; PPRL: privacy-preserved record linkage.



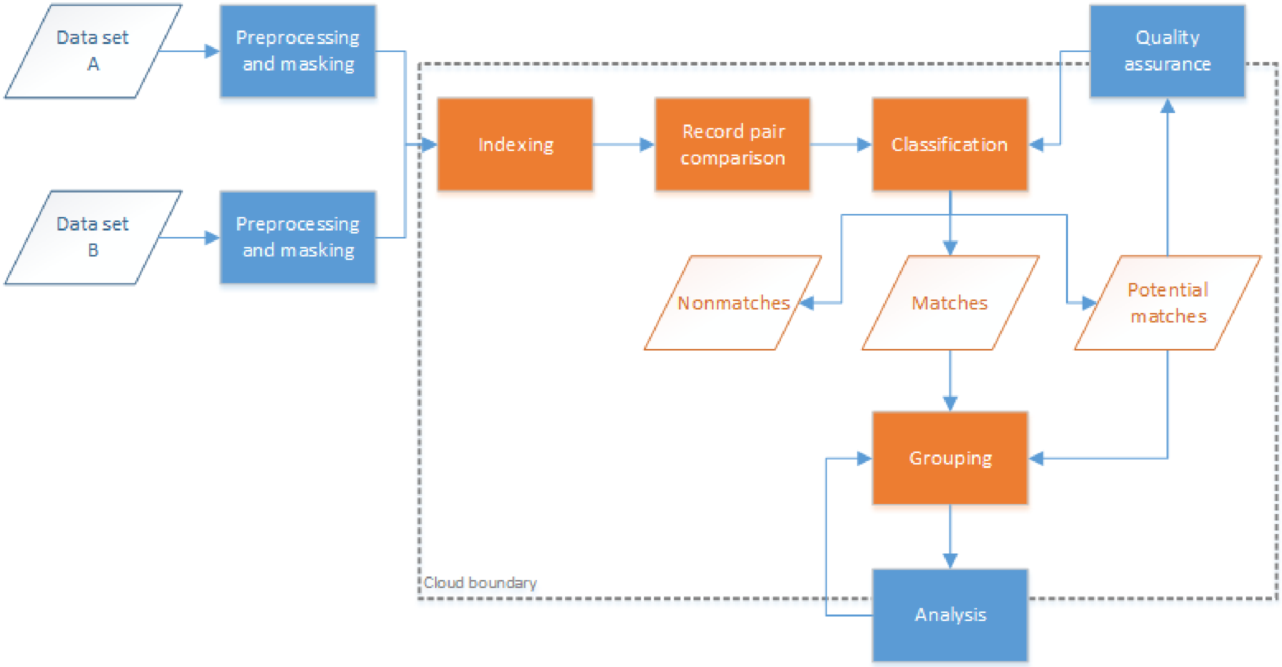
Once the deduplication linkages were complete, the on-premises component of the prototype was employed to query each data set's pair table. The queries were typical of those used following a linkage run: pair count, pair score histogram, and pairs within a pair score range. This query component used the AWS Athena application programming interface (API) to execute the queries, which used Presto (an open-source distributed query engine) to apply the ad hoc structured query language queries to the cataloged pairs tables.

Results

Design of a Cloud Model for Record Linkage

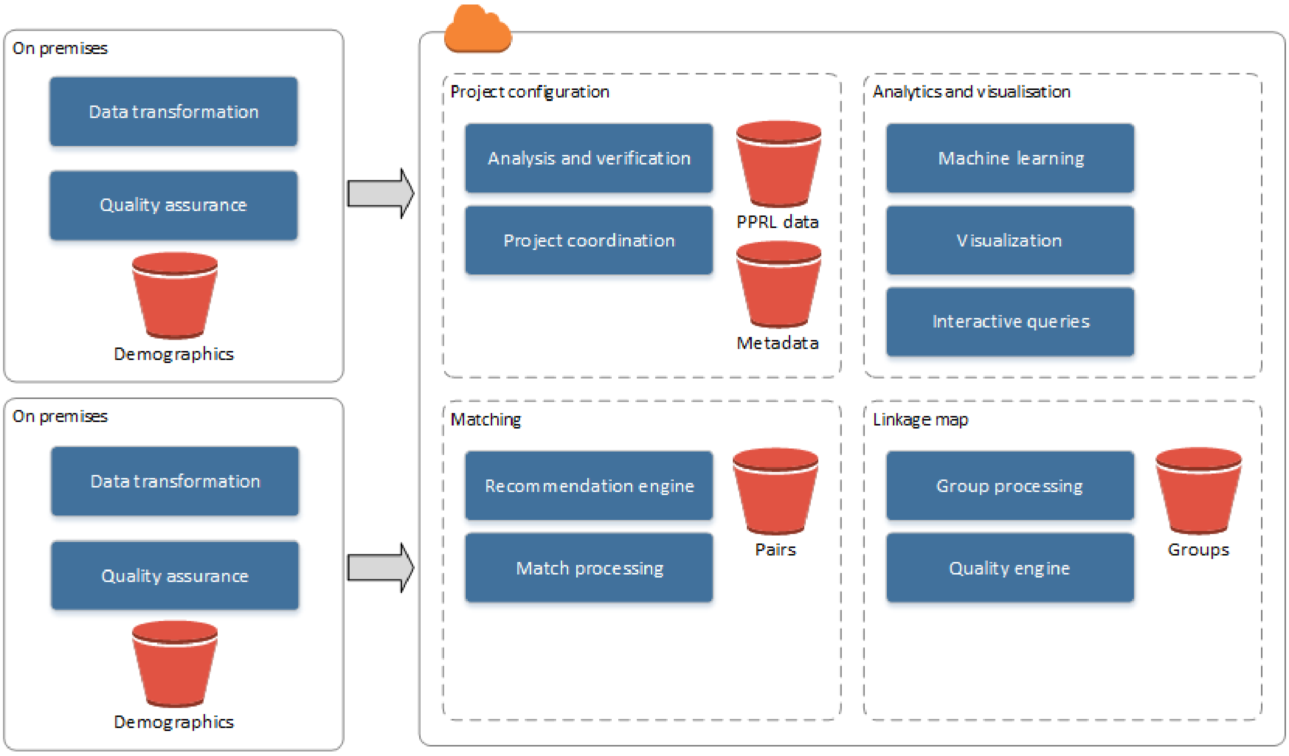
The cloud model data matching process is shown in Figure 6. Essentially, every step in the record linkage process from indexing to group analysis is pushed to cloud infrastructure. Preprocessed data sets are transformed into a privacy-preserved state (masking) and uploaded to the cloud service for linking. The services within the cloud boundary now act as a TTP. The quality assurance and analysis steps sit on the boundary of the cloud as computation and query occur on the hosted cloud infrastructure, but the interactive analysis is performed by the analysis client on premises. If the analysis client has access to one or more of the raw data sets used in the linkage, these data can be annotated onto query results, giving the clerk more informed decisions and an experience to which they are accustomed.

Figure 6. New cloud model data matching process.



As shown in the high-level architectural model in Figure 7, the demographic data (containing personal identifiers) continue to remain on premises with the data custodian. Responsibilities of the data custodians are limited to data transformation and quality assurance management. The responsibilities of the cloud services are covered under 4 main categories: project configuration, matching, linkage map, and analytics and visualization. The project configuration includes the services required for coordinating projects within and across separate data custodians. Privacy-preserved data sets are stored here as well as metadata on the data sets as a result of analysis and verification performed on the uploaded data sets. The matching category includes all match processing (classification) and pairs output as well as services for providing recommendations on linkage parameters (such as *m* and *u* likelihood estimates for probabilistic linkages) for linkages between privacy-preserved data sets [41]. The linkage map category holds the entity group information, the map between individual records, and the group in which they belong. This category also contains services for processing and creating groups from pairs as well as quality estimation and analysis. Analytics and visualization contain all analytical services provided to the on-premises clients.

Figure 7. High-level architecture of record linkage cloud model. PPRL: privacy-preserved record linkage.

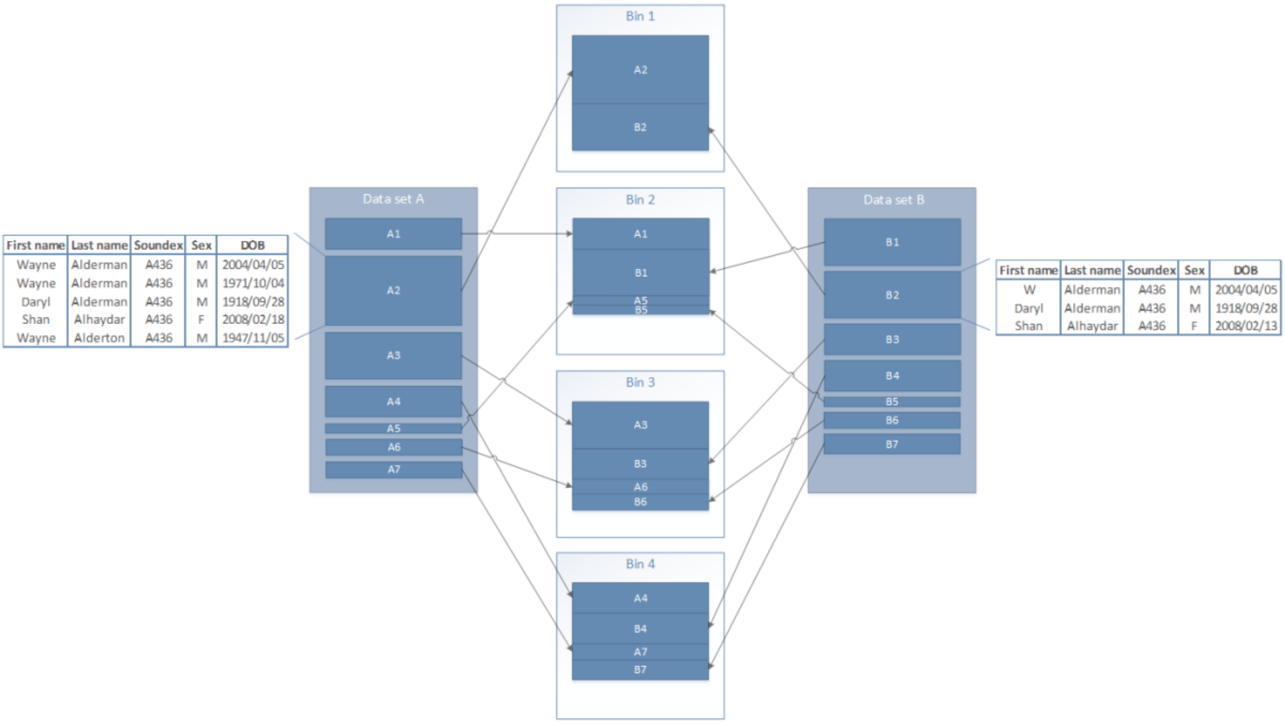


This model also allows computation to be pushed onto inexpensive, on-demand hardware in a privacy-preserving state while retaining the advantage of seeing raw identifiers during other phases of the linkage process (eg, quality assurance and analysis).

Experimental Evaluation of the Prototype

Each deduplication consisted of a single node job to split the data set into multiple bins, followed by a node array job for the matching of records within each bin. The split of data into bins is shown in Figure 8. In this example, all records with the same Soundex value will end up in the same bin.

Figure 8. Datafiles split into independent bins (by Soundex block values) for matching. DOB: date of birth.



First name	Last name	Soundex	Sex	DOB
Wayne	Alderman	A436	M	2004/04/05
Wayne	Alderman	A436	M	1971/10/04
Daryl	Alderman	A436	M	1918/09/28
Shan	Alhaydar	A436	F	2008/02/18
Wayne	Alderton	A436	M	1947/11/05

First name	Last name	Soundex	Sex	DOB
W	Alderman	A436	M	2004/04/05
Daryl	Alderman	A436	M	1918/09/28
Shan	Alhaydar	A436	F	2008/02/13

The total comparison space was calculated using the blocking field frequencies in the data set. These frequencies represent the number of times each blocking field value occurs in the data, providing the ability to calculate the number of comparisons that will be performed for each blocking field value. First, the comparison space for each blocking field was calculated using the frequency of the value within the file. The total comparison space was the sum of each, and the bin count was determined by dividing this by the maximum desired comparison space for a single bin. The blocking field value with the largest comparison space was assigned to the first bin. The blocking field value with the next largest comparison space was assigned to the second bin. This process continued for each blocking field value, returning to the first bin when the end was reached. A file was created for each bin, which was then independently deduplicated. Blocking field values with a very high frequency are undesirable as they are usually less useful for linkage and are costly in terms of computation. Any blocking field value with a frequency higher than the maximum desired comparison space was discarded.

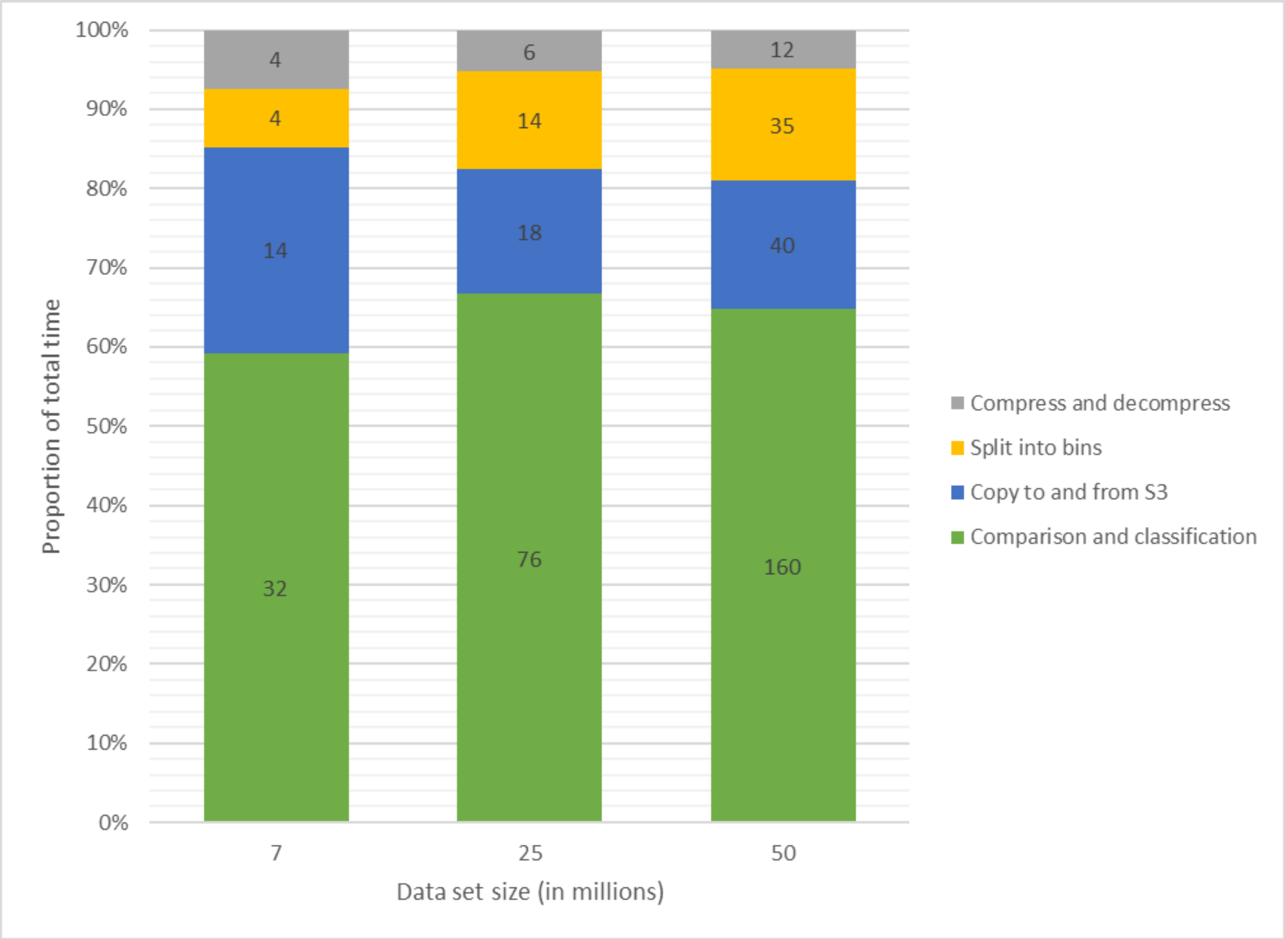
The total comparison space used for each data set, along with the bin count and pair count, is presented in Table 3. The two blocks used for the creation of separate bins for distribution across the processing cluster resulted in some duplication of comparisons and, thus, duplication of pairs.

Table 3. Comparison space and pairs created during classification.

Data set size (millions)	Comparison space, n	Bins, n	Total pairs, n	Unique pairs, n	Pairs files size (GB)
7	2,745,977,009	28	634,544,432	415,444,583	9
25	18,458,616,866	93	2,169,337,646	1,594,343,961	22
50	53,848,633,907	270	4,424,983,776	3,260,509,561	44

Approximately 60% of the time was spent on comparison and classification by each container (Figure 9). Much of the time was spent managing data in and out of the container itself. Splitting a data set into bins for parallel computation took between 7% (4/54 minutes) and 14% (35/247 minutes) of the total task time, a reasonable sacrifice considering the scalability factor this gives for the classification jobs. Provisioning of the compute resources took between 2 and 4 min for each data set.

Figure 9. Task execution time (in minutes) and the proportion of total time.



Running times of ad hoc queries on data sets are shown in Table 4; these were each executed 5 times on the client and run through the AWS Athena API. The mean execution time did not vary greatly across the differently sized data sets. With a simple count query taking around 25 seconds, there appears to be some initial setup time for provisioning the backend Presto cluster. This is expected and should not be considered an issue, particularly with all queries of the largest data set of 4.4 billion pairs taking less than 1 min to execute.

Table 4. Mean execution times for sample queries on full pairs set.

Data set size (millions)	Pairs count (millions)	Sample queries		
		Count (seconds)	Pair score histogram (seconds)	Fetch pairs in score range 15-16 (seconds)
7	635	26	52	51
25	2169	27	56	53
50	4424	24	52	54

In terms of costs associated with the use of AWS cloud services for our evaluation, there were 2 main types. First, the cost of on-demand processing, which is typically charged by the second. This totaled just over US $20 for the linkage processing used for all 3 data sets. The second is the cost of storage, which is charged per month. To retain the pairs files generated for all 3 data sets, it cost only US $2 per month. Querying data via the Athena service is currently charged at US $5 per terabyte scanned.

Discussion

Principal Findings

Our results show that an effective cloud model can be successfully developed, which extends linkage capacity into cloud infrastructure. A prototype was built based on this model. The execution times of the prototype were reasonable and far shorter than one would expect when running the same software on a single hosted machine. Indeed, it is likely that on a single hosted machine, the large data set (50 million) would need to be broken up into smaller chunks and linkages on these chunks run sequentially.

The splitting of data for comparison into separate bins worked well for distributing the work and mapped easily to the AWS Batch mechanism for execution of a cluster of containers. The creation of an AWS step function to manage the process from start to end was relatively straightforward. Step functions provide out-of-the-box support for AWS Batch. However, custom Lambda functions were required to trigger the AWS Glue crawler and retrieve the results from the first data-split task so that the appropriate size batch job could be provisioned.

As the fields used for splitting the data were the same as those used for blocking on each node, the comparison space was not different from running a linkage of the entire data set on a single machine. With the same comparison space and probabilistic parameters, the accuracy of the linkage is also identical. Having a mechanism for distributing linkage processing on multiple nodes with no reduction in accuracy is certainly a massive advantage for data linkage units looking to extend their linkage capacity.

The AWS Batch job definition's retry strategy was configured with five attempts, applying to each job in the batch. This provides some resilience to instance failures, outages, and failures triggered within the container. However, in our evaluation, this feature was never triggered. The timeout setting was set to a value well beyond what was expected as jobs that time out are not retried, and our prototype did not handle this particular scenario. Although our implementation of the step function provided no failure strategies for any task in the workflow, handling error conditions is supported and retry mechanisms within the state machine can be created as desired. An operational linkage system would require these failure scenarios to be handled.

Improvements to the prototype will address some of the other limitations found in the existing implementation. For example, S3 data transfer times could be reduced by using a series of smaller result files for pairs and uploading all of these in parallel. The over-matching and duplication of pairs could be addressed by improving the indexing algorithm used to split data. Although there is inevitably going to be some overlap of blocks, our naïve implementation could be improved. Our algorithm for distributing blocks attempts to distribute workload as evenly as possible based on the estimated comparison space. Discarding overly large blocks helps prevent excessive load on single matching nodes. However, it relies on secondary blocks to match the records within and only partly prevents imbalanced load distribution. The block-based load balancing techniques developed for the MapReduce linkage algorithms can be applied here to mitigate data skew further, where record pairs are distributed for matching instead of blocks.

As improvements to PPRL techniques are developed over time, these changes can be factored in with little change to the model. Future work on the prototype will look to extend the capability of PPRL to use additional security advances such as homomorphic encryption [42] and function-hiding encryption [43].

Conclusions

The model developed and evaluated here successfully extends linkage capability into the cloud. By using PPRL techniques and moving computation into cloud infrastructure, privacy is maintained while taking advantage of the considerable scalability offered by cloud solutions. The adoption of such a model will provide linkage units with the ability to process increasingly larger data sets without impacting data release protocols and individual patient privacy. In addition, the ability to store detailed linkage information provides exciting opportunities for increasing the quality of linkage and advancing the analysis of linkage outputs. Rich analytics, machine learning, automation, and visualization of these additional data will enable the next generation of quality assurance tooling for linkage.

Acknowledgments

AB has been supported by an Australian Government Research Training Program Scholarship.

Authors' Contributions

The core of this project was completed by AB as part of his PhD project. SR provided assistance and support for requirements analysis, testing, and data analysis. Both authors read, edited, and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Abebe R. Using search queries to understand health information needs in Africa. ArXiv 2019 (Forthcoming) [FREE Full text]
2. Radin JM, Wineinger NE, Topol EJ, Steinhubl SR. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. Lancet Digit Health 2020 Feb;2(2):e85-e93. [doi: 10.1016/s2589-7500(19)30222-5]
3. Lai S, Farnham A, Ruktanonchai NW, Tatem AJ. Measuring mobility, disease connectivity and individual risk: a review of using mobile phone data and mHealth for travel medicine. J Travel Med 2019 May 10;26(3) [FREE Full text] [doi: 10.1093/jtm/taz019] [Medline: 30869148]
4. Khoury MJ, Iademarco MF, Riley WT. Precision public health for the era of precision medicine. Am J Prev Med 2016 Mar;50(3):398-401 [FREE Full text] [doi: 10.1016/j.amepre.2015.08.031] [Medline: 26547538]
5. McGrail K, Jones K. Population data science: the science of data about people. Int J Population Data Sci 2018 Sep 6;3(4). [doi: 10.23889/ijpds.v3i4.918]
6. Christen P. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin/Heidelberg: Springer Science & Business Media; 2012.
7. Dunn HL. Record linkage. Am J Public Health Nations Health 1946 Dec;36(12):1412-1416. [Medline: 18016455]
8. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61-81 [FREE Full text] [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]
9. Population Health Research Network: 2013 Independent Panel Review. The Population Health Research Network (PHRN). 2014. URL: https://www.phrn.org.au/media/80607/phrn-2013-independent-review-findings-and-recommendations-v2-_final-report-april-17-2014-2.pdf [accessed 2020-09-15]
10. Annual Report 2015. Centre for Big Data Research in Health (CBDRH). 2015. URL: https://cbdrh.med.unsw.edu.au/sites/default/files/CBDRH_Annual%20Report_2015_160609_Final.pdf [accessed 2020-09-15]
11. Predicts 2019: Cloud Adoption and Increasing Regulation Will Drive Investment in IT Vendor Management. Gartner. 2019. URL: https://www.gartner.com/en/documents/3896211/predicts-2019-cloud-adoption-and-increasing-regulation-w [accessed 2020-09-15]
12. Vasiljeva T, Shaikhulina S, Kreslins K. Cloud computing: business perspectives, benefits and challenges for small and medium enterprises (case of Latvia). Procedia Eng 2017;178:443-451. [doi: 10.1016/j.proeng.2017.01.087]
13. Gunadham T, Kuacharoen P. Security concerns in cloud computing for knowledge management systems. J Appl Stat 2019;1:52-60. [doi: 10.1109/itng.2013.127]
14. John J. Major vulnerabilities and their prevention methods in cloud computing. In: Advances in Big Data and Cloud Computing. New York, USA: Springer; 2019.
15. El-Ghafar R. Record Linkage Approaches in Big Data: a State of Art Study. In: 13th International Computer Engineering Conference. 2017 Presented at: ICENCO'17; December 27-28, 2017; Cairo, Egypt. [doi: 10.1109/icenco.2017.8289792]
16. Dou C. Probabilistic Parallelisation of Blocking Non-matched Records for Big Data. In: IEEE International Conference on Big Data. 2016 Presented at: Big Data'16; December 5-8, 2016; Washington, DC, USA. [doi: 10.1109/bigdata.2016.7841009]
17. Chu X, Ilyas IF, Koutris P. Distributed Data Deduplication. In: Proceedings of the VLDB Endowment. 2016 Presented at: VLDB'16; October 14, 2016; New Delhi, India. [doi: 10.14778/2983200.2983203]
18. Gazzarri L, Herschel M. Towards task-based parallelization for entity resolution. SICS Softw-Inensiv Cyber-Phys Syst 2019 Aug 26;35(1-2):31-38. [doi: 10.1007/s00450-019-00409-6]
19. Papadakis G, Skoutas D, Thanos E, Palpanas T. Blocking and filtering techniques for entity resolution. ACM Comput Surv 2020 Jul;53(2):1-42. [doi: 10.1145/3377455]
20. Karapiperis D, Verykios VS. A fast and efficient Hamming LSH-based scheme for accurate linkage. Knowl Inf Syst 2016 Feb 3;49(3):861-884. [doi: 10.1007/s10115-016-0919-y]
21. Pita R, Pinto C, Melo P, Silva M, Barreto M, Rasella D. A Spark-based Workflow for Probabilistic Record Linkage of Healthcare Data. CEUR-WS. 2015. URL: http://ceur-ws.org/Vol-1330/paper-04.pdf [accessed 2020-09-14]

22. Gagliardelli L, Simonini G, Beneventano D, Bergamaschi S. SparkER: scaling entity resolution in spark. Adv Data Technol 2019;2019:602-605. [doi: 10.5441/002/edbt.2019.66]
23. Mestre DG, Pires CE, Nascimento DC, de Queiroz AR, Santos VB, Araujo TB. An efficient spark-based adaptive windowing for entity matching. J Syst Software 2017 Jun;128:1-10. [doi: 10.1016/j.jss.2017.03.003]
24. Sehili Z, Kolb L, Borgs C, Schnell R, Rahm E. Privacy Preserving Record Linkage with PPJoin. Abteilung Datenbanken Leipzig - Universität Leipzig. 2015. URL: https://dbs.uni-leipzig.de/file/P4Join-BTW2015.pdf [accessed 2020-09-15]
25. Boratto M, Alonso P, Pinto C, Melo P, Barreto M, Denaxas S. Exploring hybrid parallel systems for probabilistic record linkage. J Supercomput 2018 Mar 21;75(3):1137-1149. [doi: 10.1007/s11227-018-2328-3]
26. Christen P. A survey of indexing techniques for scalable record linkage and deduplication. IEEE Trans Knowl Data Eng 2012 Sep;24(9):1537-1555. [doi: 10.1109/tkde.2011.127]
27. Trinckes Jr JJ. The Definitive Guide to Complying With the HIPAA/HITECH Privacy and Security Rules. Boca Raton, Florida, United States: CRC Press; 2012.
28. Regulation (EU) 2016/679 of the European Parliament. Publications Office of the EU - Europa EU. 2016. URL: https://op.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-en [accessed 2020-09-15]
29. Vatsalan D, Sehili Z, Christen P, Rahm E. Privacy-preserving record linkage for big data : current approaches and research challenges. Handbook Big Data Technol 2017:851-895 [FREE Full text] [doi: 10.1007/978-3-319-49340-4_25]
30. Boyd JH, Randall S, Brown AP, Maller M, Botes D, Gillies M, et al. Population data centre profiles: centre for data linkage. Int J Population Data Sci 2020 Mar 11;4(2). [doi: 10.23889/ijpds.v4i2.1139]
31. Schnell R. PPRL: Privacy Preserving Record Linkage. Springer. 2019. URL: https://link.springer.com/content/pdf/10.1007%2F978-3-319-63962-8_17-1.pdf [accessed 2020-09-15]
32. Karapiperis D, Gkoulalas-Divanis A, Verykios VS. LSHDB: A Parallel and Distributed Engine for Record Linkage and Similarity Search. In: 16th International Conference on Data Mining Workshops. 2016 Presented at: ICDMW'16; December 12-15, 2016; Barcelona, Spain. [doi: 10.1109/icdmw.2016.7867099]
33. Bonomi L, Xiong L, Lu JJ. LinkIT: Privacy Preserving Record Linkage and Integration via Transformations. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. 2013 Presented at: SIGMOD'13; June 22-27, 2013; New York, USA URL: http://dl.acm.org/citation.cfm?doid=2463676.2465259 [doi: 10.1145/2463676.2465259]
34. Toth C, Durham E, Kantarcioglu M, Xue Y, Malin B. SOEMPI: a secure open enterprise master patient index software toolkit for private record linkage. AMIA Annu Symp Proc 2014;2014:1105-1114 [FREE Full text] [Medline: 25954421]
35. Randall S, Boyd J, Ferrante A, Brown A, Semmens J. Grouping Methods for Ongoing Record Linkage. In: Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference. 2015 Presented at: ACM-SIGKDD PopInfo'15; April 11, 2015; Sydney, Australia.
36. Irvine K, Smith M, de Vos R, Brown A, Ferrante A, Boyd J, et al. Real world performance of privacy preserving record linkage. Int J Population Data Sci 2018 Sep 10;3(4). [doi: 10.23889/ijpds.v3i4.990]
37. Lee YA. Medicineinsight: Scalable and Linkable General Practice Data Set. In: Health Data Analytics. 2019 Presented at: HDA'19; October 16, 2019; Sydney, Australia.
38. Christen P, Pudjijono A. Accurate synthetic generation of realistic personal information. Adv Know Discovery Data Mining 2009;5476:507-514 [FREE Full text] [doi: 10.1007/978-3-642-01307-2_47]
39. PSMA Geocoded National Address File (G-NAF). Australian Government. 2016. URL: https://data.gov.au/data/dataset/19432f89-dc3a-4ef3-b943-5326ef1dbecc [accessed 2020-09-15]
40. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating data Linkage software. J Biomed Inform 2012 Feb;45(1):165-172 [FREE Full text] [doi: 10.1016/j.jbi.2011.10.006] [Medline: 22061295]
41. Brown AP, Randall SM, Ferrante AM, Semmens JB, Boyd JH. Estimating parameters for probabilistic linkage of privacy-preserved datasets. BMC Med Res Methodol 2017 Jul 10;17(1):95 [FREE Full text] [doi: 10.1186/s12874-017-0370-0] [Medline: 28693507]
42. Randall S. Privacy Preserving Record Linkage Using Homomorphic Encryption. In: Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference. 2015 Presented at: ACM-SIGKDD PopInfo'15; April 11, 2015; Sydney, Australia.
43. Lee J, Kim D, Song Y, Shin J, Cheon J. Instant Privacy-Preserving Biometric Authentication for Hamming Distance. Semantic Scholar. 2018. URL: https://api.semanticscholar.org/CorpusID:57760611 [accessed 2020-09-15]

Abbreviations

API: application programming interface
AWS: Amazon Web Services
CPU: central processing unit
GPU: graphics processing unit
LSH: locality sensitive hashing
PPRL: privacy-preserving record linkage
TTP: trusted third party

Edited by G Eysenbach; submitted 27.03.20; peer-reviewed by S Albakri, A Al-Muqarm, T Enamorado; comments to author 23.06.20; revised version received 02.08.20; accepted 23.08.20; published 23.09.20.

Please cite as:
Brown AP, Randall SM
Secure Record Linkage of Large Health Data Sets: Evaluation of a Hybrid Cloud Model
JMIR Med Inform 2020;8(9):e18920
URL: http://medinform.jmir.org/2020/9/e18920/
doi:10.2196/18920
PMID:32965236

Viewpoint

Including Social and Behavioral Determinants in Predictive Models: Trends, Challenges, and Opportunities

Marissa Tan[1], MPH, MD; Elham Hatef[1,2,3], MD, MPH, FACPM; Delaram Taghipour[1], MPH, MBA, MD; Kinjel Vyas[4], MS; Hadi Kharrazi[2,4], PhD, MD; Laura Gottlieb[5], MPH, MD; Jonathan Weiner[2], DrPH

[1]General Preventive Medicine Residency Program, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[2]Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Center for Population Health Information Technology, Baltimore, MD, United States

[3]Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[4]Division of Health Sciences Informatics, Johns Hopkins School of Medicine, Baltimore, MD, United States

[5]Social Interventions Research and Evaluation Network, Center for Health & Community, University of California, San Francisco, CA, United States

Corresponding Author:
Elham Hatef, MD, MPH, FACPM
Department of Health Policy and Management
Johns Hopkins Bloomberg School of Public Health
Center for Population Health Information Technology
624 N. Broadway, Room 502
Baltimore, MD, 21205
United States
Phone: 1 4109788006
Email: ehatef1@jhu.edu

Abstract

In an era of accelerated health information technology capability, health care organizations increasingly use digital data to predict outcomes such as emergency department use, hospitalizations, and health care costs. This trend occurs alongside a growing recognition that social and behavioral determinants of health (SBDH) influence health and medical care use. Consequently, health providers and insurers are starting to incorporate new SBDH data sources into a wide range of health care prediction models, although existing models that use SBDH variables have not been shown to improve health care predictions more than models that use exclusively clinical variables. In this viewpoint, we review the rationale behind the push to integrate SBDH data into health care predictive models and explore the technical, strategic, and ethical challenges faced as this process unfolds across the United States. We also offer several recommendations to overcome these challenges to reach the promise of SBDH predictive analytics to improve health and reduce health care disparities.

(JMIR Med Inform 2020;8(9):e18084) doi:10.2196/18084

KEYWORDS

social determinants of health; information technology; health care disparities; population health

Social and Behavioral Determinants of Health and Predictive Analytics

Since the Health Information Technology for Economic and Clinical Health act of 2009, the majority of US health care systems have adopted electronic health records (EHRs) for patient care [1]. Faced with increased financial incentives to improve population health, care coordination, and quality of care, health care providers and payers now use EHRs and other digital data sources to understand how past associations and trends in their patient populations can be used to forecast health care–related outcomes, a component of the widely known strategy of *predictive analytics* [1,2].

Predictive analytics uses extensive data, modeling, and algorithms to predict individual and population events and has a long history in commercial industries [3]. For better or worse, commercial industries have developed innovative techniques to *mine* demographic, socioeconomic, and consumer behavior data as part of the forecasting and analytics process. For example, web-based sellers and banks collect personal information on purchase histories, credit data, consumer behaviors, and life events that are available in various digital databases. These institutions use such data to make predictions for various goals, such as determining ideal customers for specific products or services and how much institutions should offer to whom [4].

There are 2 broad approaches to predictive analytics. The modeling and simulation approach is used to test hypotheses or assess the consequences of scenarios where the rules of the models are developed from theories. Such models also employ data to initialize variables, to calibrate free parameters, or for validation. Alternatively, predictive analytics may also use machine learning in which models are exclusively built from data via algorithms and tested on data that mirror the calibration and validation steps of modeling and simulation, respectively. These approaches can be combined in complex systems [5]. This paper focuses on machine learning and provides several observations that apply to modeling and simulation. Generally, the modeling and simulation approach is useful in systems where the dynamics are well known, whereas machine learning is useful when accurate simulations cannot be performed and there are enough data to determine a model [5]. On the basis of the specific prediction goal, different types of data and methods are required and thus have different associated limitations and challenges.

In health care, the same techniques are used with different goals. Over the last decade, health insurance plans have ramped up the use of predictive analytics, employing patient demographics, insurance claims data, and clinical characteristics derived from EHRs to create statistical models of future health care risks and resource utilization [6]. Analysts have also developed predictive models for health and health care. These data science techniques generally involve larger and more complex databases but represent an application of traditional statistical forecasting methods using a wide range of techniques such as deep neural networks, natural language processing (NLP), random forest, and decision tree algorithms [7,8].

The growing awareness of associations between social and behavioral factors and health has led predictive modeling to explore the incorporation of social and behavioral determinants of health (SBDH) into forecasting [9,10]. For example, on an individual level, diet and physical activity affect health care use and costs [11,12]. At the community level, characteristics of neighborhoods, such as food access and transportation, play significant roles in health outcomes, morbidity, and mortality [13-15].

Although SBDH factors have been incorporated in the predictive modeling process to forecast health care–related outcomes, there are limitations related to the use of such factors. For instance, machine learning methods are not generally developed to capture changing SBDH factors. They mainly address the stationary distributions of the SBDH factors. A change in the data requires providing longitudinal data to the model to perform time series modeling and to capture these changes. If a change in the distribution of data is necessary (eg, to reflect potential trends in SBDH over time), then the approach of modeling and simulation may be used to explore various scenarios. An example is the common use of event-driven simulations in health care research [16].

A growing crop of initiatives uses SBDH to predict health care use in the United States [17]. Although the methods and evidence underlying these new models that incorporate SBDH are nascent and have not shown improved predictions over traditional clinical measures, the medical community's interest in SBDH needs in conjunction with predictive analytics continues to increase [18,19].

The Rationale for Including Social and Behavioral Determinants in Predictive Models

Studies in the United States and worldwide have suggested that SBDH, such as educational attainment, have a greater impact on premature mortality than clinical care access and quality [10,20]. A meta-analysis in the United States found that income inequality, social support, segregation, individual and neighborhood poverty, and education level were responsible for 50% of deaths [9]. Some literature on mortality estimates the lack of quality medical care to encompass 10% to 20% of deaths [10,21,22]. Entities such as the World Health Organization have recognized the role of SBDH factors in health equity and committed to action on these determinants [23].

Several national agencies have recognized and advocated for the incorporation of SBDH into health care practices and the standard use of health data. The National Academies of Science, Engineering, and Medicine have identified 5 complementary activities that can facilitate the integration of social care into health care. These activities include the following: "(1) identify the social risks and assets of defined patients and populations; (2) focus on altering clinical care to accommodate identified social barriers; (3) reduce social risk by assisting in connecting patients with relevant social care resources; (4) understand existing social care assets in the community, organize them to facilitate synergies, and invest in and deploy them to positively affect health outcomes; and (5) work with partner social care organizations to promote policies that facilitate the creation and redeployment of assets or resources to address health and social needs." [24] Moreover, the eHealth initiative, a national coalition focused on health data interoperability in the United States, advocates the use of SBDH data to coordinate care, evaluate interventions that address social needs, identify gaps in community resources, predict health risk, and develop SBDH-sensitive interventions to improve health [25].

Potential Benefits of Including Social and Behavioral Determinants in Predictive Models

Bolstered by the initiatives of the national organizations, incorporation of SBDH into predictive models could help to (1) identify patients and populations who need more resources, (2) improve health care reimbursement for providers who serve patients with social needs, (3) reduce health and health care disparities, and (4) improve the quality of health care.

Predictive analytics and SBDH risk segmentation could facilitate efforts to identify patients who would benefit from more resources and targeted services. This may lessen the resource burden of universal social risk screening or social care delivery

[26]. For example, a systematic risk analysis could help identify patients with modifiable social risks at a higher risk of poor medical outcomes. This type of segmentation could help health systems target appropriate resources, for example, referrals to case management, social service agencies, or government support programs such as the Supplemental Nutrition Assistance Program [27-29]. In addition to using SBDH-sensitive analytics to identify vulnerable individuals, this approach could also help health care organizations or partner agencies identify disadvantaged communities, such as neighborhoods with food deserts [26,30]. A health care system truly desiring to maximize its impact on the health of a community could more effectively increase food access at the neighborhood level by working with farmers' markets and grocery stores in addition to individual-level interventions.

Under the present federal regulations for Medicaid-managed care, social and behavioral services such as care coordination are reimbursed through capitation. Predictive analytics and SBDH risk segmentation could support new payment models to adequately reflect the medical and social complexity of patients [31]. Beyond capitated or global payments, contextualizing patients with their SBDH needs enables health care payers to more accurately assess providers' care for vulnerable populations who require more health care resources, thus impacting their fee-for-service payments [27]. Present Medicaid-managed care regulations could support value-added services that would not be reimbursed under capitation alone but would address the health needs of members, such as interventions that assess environmental triggers of asthma [31]. Several states (eg, Rhode Island, Minnesota, and Oregon) have adopted the Accountable Care Organization models that reward health care providers for addressing their Medicaid populations' SBDH with adjusted payment structures [32,33]. Patient protection laws in the United States regarding insurance denials and premium payments should be upheld to ensure that SBDH risk segmentation does not increase the burden of health care costs to disadvantaged populations [34].

Identifying and accounting for the increased risk of poor health outcomes and associated health care utilization is critical to the elimination of disparities in care for vulnerable populations. The spread of COVID-19 across the United States and worldwide is a great example of how predictive modeling could help health care systems and public health officials address health disparities and potentially change the course of the pandemic. The COVID-19 pandemic has highlighted long-standing health disparities [35,36]; neighborhoods with the highest proportion of racial and ethnic minorities and people living in poverty are experiencing higher rates of hospitalization and death [37-40]. In response, several research teams have started to include information on SBDH in predictive modeling and assessment of COVID-19–related risk and outcomes [39,41].

Exclusion of SBDH-related variables in risk-adjusted reimbursement models would result in lower reimbursement for patients with greater social needs, which dissuades providers from caring for these patients in capitated systems [42]. Employing SBDH in risk-adjusted capitated payment models could translate into improved health care policy by supporting organizations to more effectively meet the needs of individuals and communities with greater social needs.

Beyond payment adjustment, stratifying patients by their SBDH risk levels could reveal health disparities as well as promote health care quality by establishing a mechanism to fairly evaluate providers' care of patients with social disadvantages [42]. Health systems and payers could further evaluate the quality of health care by developing specific SBDH-dependent quality indicators that bolster equity in health care across the range of patients served [42].

Present State of Including Social and Behavioral Determinants in Predictive Analytics

Although there is a strong and compelling body of literature on the observed associations between SBDH and health, to date, diagnosis-based forecasting models used to predict cost and utilization have not yet shown the incremental value of adding SBDH risk factors to predictions. Some published reports using community-level SBDH data contribute only slightly to the predictive model performance beyond individual patient characteristics extracted from EHR data [43,44].

Similarly, SBDH-oriented predictive models using newer applications of machine learning techniques have shown varying levels of performance in predictions. A neural network predictive model that incorporates SBDH was found to identify, with 78% accuracy, over two-third of the Medicare patients in their sample who would not respond to automated medication refill requests and may benefit from targeted outreach [45]. Seligman et al [46] applied linear regression and different machine learning techniques to predict systolic blood pressure, BMI, waist circumference, and telomere length using SBDH variables of gender, income, wealth, education, public benefits, family structure, and health behaviors. Although neural networks outperformed other machine learning techniques as fit for their sample, most of their tested machine learning models performed similar to the simpler regression models, and all models had poor out-of-sample prediction [46]. Applying random survival forest methods to develop a predictive model using the poverty status and EHR data, Bhavsar et al [44] did not find that risk prediction for health care services and hospitalization outcomes improved beyond models using traditional EHR data. Similarly, a machine learning model using random forest decision methods on structured and unstructured SBDH only improved sensitivity (67.6%) by 0.1% and showed decreased specificity (69.6%) by 1.9% compared with their tested non-SBDH models in predicting referrals for social needs [47].

Given the evidence-based expectation that SBDH should improve predictive models, why have published predictive models not shown enhanced predictions? Although insufficient data and suboptimal methods are potential explanations common to all research, triple challenges unique to the SBDH context include the diversity of data sources and health outcomes used in existing models as well as the lack of transparency, which together pose an important question about model accuracy.

Diversity of Data Sources

A wide range of SBDH variables and data sources are used in predictive models and no guidelines exist to distinguish which variables and data sources would best improve the performance of the predictive model. A rapid review of social, behavioral, and environmental determinants of health used with clinical data identified 744 variables among 178 articles, in which the majority of articles included socioeconomic and material conditions [48]. Data sources vary from individual-level EHR data and insurance claims to community-level data from the United States census and similar sources as well as commercial data such as information from credit reporting agencies.

Health plans have historically used insurance claims, which include diagnostic and prior utilization information of varying completeness across health care settings, for predictive modeling to forecast utilization and cost [27,49]. More recently, health payers and other private health care companies have obtained consumer and financial data, such as information on household size, income, and wealth measures, from credit reporting agencies to better assess their members' needs [27,29,50]. For instance, one company mines public data on education, law enforcement records, birth records, voter registration, and derogatory records such as a history of evictions and liens [27].

Rather than commercial data, academic centers and government organizations have primarily relied on individual-level clinical information derived from structured and unstructured EHRs [51] and relevant risk factors on a community level extracted from public surveys [52], such as the United States Census Bureau American Community Survey, which includes multiple indicators of neighborhood deprivation [43,53]; the Food Access Research Atlas, which describes food deserts [54,55]; and the American Housing Survey, which contains information on housing characteristics [56,57]. In one systematic review of predictive models using EHR data, 36 of the 106 unique studies included SBDH data in one of their final predictive models [58]. However, the social determinants included were limited to race or ethnicity alone in 19 of the 36 EHR-based studies [58]. The same systematic review included behavioral determinants in 30 of these EHR-based predictive models. However, 12 of these studies' behavioral variables were limited to tobacco use or smoking alone [58]. As another case example, a Kaiser predictive model that uses race and ethnicity as one variable to develop a hypoglycemia risk model omitted race in their final, simpler model on finding that race was not one of the strongest predictors of hypoglycemia compared with clinical factors [59].

In addition to survey-collected data aggregated at the geographic level, academic centers are expanding this community-level framework to include geocentric data such as transit data, which contains data on access to transportation [60], the Environmental Protection Agency's Air Quality Index data [61], and food desert data from the United States Department of Agriculture's Food Access Research Atlas [62].

As expected with predictive models, the performance of a model varies depending on the selected SBDH variables and data sources [43,44]. When analyzing SBDH variables, the diversity of data sources has implications for a model's ability to address challenges associated with SBDH, such as accurately assessing the temporal duration of SBDH and determining the spatial-level effects of population-level SBDH data. Researchers need to critically analyze SBDH variables and data sources to ensure the selection of variables and high-quality data sources that accurately and authentically capture SBDH factors to be tested.

Diversity of Health Outcomes

Health care–based predictive models that integrate SBDH risk factors have been used to forecast a wide range of health care–relevant endpoints. Although, most often, the predicted outcomes include health care costs and utilization, such as emergency department visits, hospitalizations, and readmissions [27,44,63,64]. There is no consensus on which health outcomes are the most appropriate to predict with specific SBDH factors. Within the public health, academic, and health policy sectors, models have expanded their focus outside the realm of medical care. For example, the Centers for Disease Control and Prevention (CDC), the CDC Foundation, and the Robert Wood Johnson Foundation collaboratively created 500 Cities, a tool that uses community-level socioeconomic characteristics to predict city-level health behaviors, mortality, and morbidity [65,66].

Similar to challenges related to data sources, the diversity of health outcomes as the endpoint for the predictive models will impact assessing the performance of their methods and determining the best methods to address specific SBDH variables or to set the stage for standardized guidelines for specific SBDH variables and outcomes.

Lack of Transparency

Many predictive models that incorporate SBDH data have been developed and are used in the private sector and are therefore not only proprietary but also unavailable for public review and scrutiny. Consequently, other researchers cannot replicate the methods used in these predictive models. Several predictive modeling companies that have made use of only clinical risk factors now extensively market the inclusion of SBDH data in their predictive risk models [27,29,50]. One company relies exclusively on consumer data, rather than medical data, to develop as many as 70 different models to predict patients at risk for general poor health and high health care costs [67]. For example, one commercial model developer described a case study using its socioeconomic score model to predict the risk of common chronic diseases, highlighting the score's successful prediction in the top 10% and bottom 10% of the score risk data, although it did not describe how the model performed in the remaining 80% of the population covered [68].

However, the lack of transparency also extends to the academic sector. When data used for a data-driven model, source code, and the model itself are not made open source, the derived models cannot be replicated, a problem known as the *reproducibility crisis* in machine learning [69]. When available, analysts would ideally search out the code and data for models in code repositories to learn how models are organized [70]. However, in a survey of 400 artificial intelligence conference papers with algorithms, only 6% shared the code and about one-third shared their data [69]. Reasons for avoiding sharing range from dependence on another unpublished code and desire to maintain a competitive advantage to its proprietary nature or institutional review board restrictions [69]. Without the training data and code, the reproducibility of machine learning is dismal.

Given the relative novelty of SBDH in predictive analytics and the lack of standardization around data sources and outcomes assessed as well as challenges related to transparency of models in the private sector, models that incorporate SBDH factors are fraught with questions about accuracy. The lack of transparency makes it very difficult to assure model accuracy, precludes replicability, and portends clinicians' mistrust of these models. Such challenges highlight the need for greater transparency in model development and sharing across institutions.

Recommendations to Address Challenges and Improve SBDH Predictive Models

Advancing SBDH predictive analytics will require overcoming several challenges. As the field of health care predictive modeling grows, the incorporation of SBDH factors into predictions will face challenges similar to those of traditional models. Predictive models should follow guidelines in the *Transparent Reporting of a multivariate prediction model for Individual Prognosis or Diagnosis* (TRIPOD) initiative [71]. The TRIPOD guidelines are concerned with how general health care predictive models are reported and serve as the framework for predictive model development, validation, and modification in health care contexts [71]. This initiative was developed in response to the growing field of health-related predictive analytics and concerns about the lack of transparency, standardization, and oversight [72]. As the field of health care predictive analytics matures, it is time to apply the TRIPOD initiative's guidelines to this rapidly evolving area of health services analytics regarding SBDH factors. Consequently, we offer several recommendations to advance the use of SBDH in health and health care predictive analytics (Textbox 1).

Textbox 1. Recommendations to advance the use of social and behavioral determinants of health in health care predictive analytics.

Privacy standards, patient consent, and ethical use of social and behavioral determinants of health (SBDH) data

- Develop consensus on transparency, privacy protections, and ethical uses of SBDH data in predictive models
- Create guidelines to reduce inherent bias in predictive models

Technical challenges associated with SBDH data sources and analytics

- Determine best practice guidelines for SBDH data sources and predictive model design as well as open-source access
- Expand standardized coding and taxonomies of SBDH risk factors that enhance interoperability

Expanding the knowledge base to inform best practice guidelines for SBDH analytics

- Support national shared research and development to advance the SBDH predictive model development and application
- Establish a national agenda to create a shared evidence base regarding the importance of SBDH factors and the best approach for including SBDH in analytics

Privacy Standards, Patient Consent, and Ethical Use of Social and Behavioral Determinants Data

Develop Consensus on Transparency, Privacy Protection, and Ethical Uses of SBDH Data in Predictive Models

As expected, many consumers are unsettled by the unregulated use of personal and commercial information to predict sensitive behaviors or health outcomes [4]. An example of such unregulated use of personal information is Google's acquisition of large amounts of personal health data, from hospitals and clinics across 21 US states, used to predict health and health care use, undisclosed to patients and other parties [73,74]. Social determinants cover sensitive topics, such as poverty, substance misuse, food insecurity, and homelessness. Individuals may fear stigmatization from health providers in revealing their SBDH information [75]. Similarly, individuals may be concerned about the social, employment, and legal effects of revealing SBDH when their data are not protected [75].

To address such concerns, there needs to be an established discourse leading to a national consensus and clear guidelines regarding the ethical use of patients' SBDH data in the context of a health care predictive model [76]. Lack of transparency in methods, applications, and data protection results in little accountability to ensure that SBDH risk predictions are not used to achieve profits at the expense of health care quality or access, such as using SBDH data to exclude vulnerable patients from a health intervention to ensure greater health care profits [76,77]. Establishing robust and meaningful national guidelines for using SBDH data will require insights from a variety of clinical, social science, and technical perspectives as well as views of patients, community members, policy makers, and ethicists. In particular, patients should participate and be involved in the research that is developing models to safeguard the ethical and transparent use of patient data [78]. Without the perspectives of patients and community members at the forefront of these discussions, rather than moving to a new level of health care equity and access, SBDH predictive analytics could easily slide into domains that many would consider inappropriate use, especially given a special concern and focus on the highest risk members of our communities [76].

Create Guidelines to Reduce Inherent Bias in Predictive Models

One important ethical and technical challenge of SBDH analytics, mostly in the application of statistical modeling, is ingrained model bias. For instance, vulnerable patients, such as those with more social and behavioral risk factors, may not be adequately represented in the data sources used to build the predictive model, leading to the model's inaccurate predictions for these individuals. Machine learning models on the other hand can address this issue through over- or undersampling. Therefore, being at risk for bias from the original sample is normally corrected in a standard process [79].

The data sources might also lack information on the key SBDH variables that affect the desired outcomes. An example of this challenge might be a predictive model that focuses on health care utilization as the desired outcome and lacks data on health care access for vulnerable populations. Such a model may indicate that individuals with poor access to health care have a low likelihood of future utilization. A model with such ingrained bias would thus underestimate the actual requirement for the greater amount of health care resources necessary to achieve the same health outcomes once these individuals have access to health care [42]. Recently, this situation was observed in a study by Obermeyer et al [80] who assessed a large, commercial health plan's predictive algorithm. The model systematically underestimated the health needs of African American patients by assuming that health care costs served as an adequate proxy for health needs. The bias arose because the unequal access to care among African American patients resulted in less money spent caring for those patients compared with White patients.

Although many researchers use health care utilization and costs as outcomes for SBDH research, models with these outcomes, proxied for health needs, are biased in that the data underrepresents those with lower access to health care. In recognition of the ingrained model bias, one approach might be to develop guidelines that recommend stratifying the population for key SBDH risk factors. Therefore, separate models would assess health care utilization for each stratum, taking into account unmeasured SBDH risk factors impacting health care utilization (eg, socioeconomic status, which defines insurance type and access to health care).

Technical Challenges Associated With Data Sources and Analytics

Determine Best Practice Guidelines for SBDH Data Sources and Predictive Model Design As Well As Open-Source Access

The future of SBDH-centric predictive modeling faces several challenges related to data sources and model design. One *big data*–related challenge is that most social and behavioral data found within providers' EHRs are unstructured, free-text clinical notes and are not standardly interoperable. Although ubiquitous, this information is captured inconsistently and depends on the use of NLP to render the data useful in analytics [81,82]. When NLP is utilized, the SBDH language in the health record may not describe the level of SBDH precisely enough to accurately determine social risk as social determinants such as neighborhood disadvantage may need to reach a threshold to have a significant impact on health-related outcomes [83].

Another important challenge is related to the use of population-level SBDH variables and whether such variables are interpreted as proxies for individual-level factors that cannot be measured, such as low household income, or represent population-level spatial elements, such as a high concentration of low household income in a neighborhood [84]. Proxies are based on assumptions to confer population-level characteristics to an individual. In contrast, geospatial models investigate population-level elements based on the principle of spatial autocorrelation, meaning that data located close together are interrelated by nature [85]. Addressing this challenge is critical to the interpretation of models and requires sufficiently transparent models that allow the proper distinction between the two implications of the population-level SBDH variables.

There are also several technical challenges related to the analytic approach, spanning the choice of analytic model, data sources, discriminatory power, and SBDH temporality. Statistical models, spatial analysis, and machine learning have all been used alone and in combination with various SBDH predictive models. Most often, health care predictive analytics uses regression models for their simplicity and acceptability [86]. However, machine learning models may be useful for finding new dimensions that can accurately classify outcomes according to their predictive characteristics in nonlinear data [86]. However, not all machine learning techniques, which range from transparent decision tree algorithms to unsupervised neural networks, are appropriate for use with SBDH predictive models. Highly autonomous machine learning models may select characteristics that are not clinically relevant for the outcome (eg, family meetings as a predictive characteristic for hospital mortality) when researchers do not remove these characteristics [86]. Models should instead reflect appropriate domain expertise as well as appropriate machine learning techniques. Moreover, for techniques that depend on unsupervised neural networks, there are long-standing controversies regarding the disadvantages of nontransparent, one-of-a-kind models versus more readily explainable logistic regression models [7,86].

There are also challenges related to using SBDH data at the geographic level in predictive modeling, which are often needed to identify SBDH on a population level and for community-level interventions [26]. Geospatial analysts need to choose the appropriate granularity for a model, which may be associated with a model's discriminatory power to help distinguish those at high- versus low-risk levels [87]. Furthermore, analyzing SBDH data at different geographic levels (eg, census block group, census tract, county, and state) is methodologically complex.

The discriminatory power to distinguish patients with and without social needs also poses a challenge in nongeospatial modeling with the potential to introduce higher-than-desirable false positives and/or negatives [74]. For instance, a study of food security among Medicare patients using clinical data and a needs assessment survey could not accurately predict which patients would benefit from a referral to community resources [88]. Similarly, a predictive model that uses random forest decision methods applied to socioeconomic data did not improve referral rates to community services once at-risk patients were identified [28]. When SBDH data are operationalized in a poorly functioning algorithm, these false positives and negatives indicate that a health system spends unnecessary resources evaluating several patients not at high risk, whereas groups of patients needing social services remain unidentified [74,89]. To address this phenomenon, algorithms may need to be tested with new data as predictive analytics methods that use SBDH risk data have evidenced limited generalizability outside of the original sample data where the model was developed [26,46].

Within a model's discriminatory power is the challenge of temporality in analytic models. Specifically, further research and development are necessary to determine how to capture changing social risk factors related to changing life circumstances throughout a person's life or epoch [90]. For example, by structural design, a model may overlook an individual's loss of income through unemployment or community changes not reflected in neighborhood data [74]. Thus, time-oriented models will be better able to elucidate the persistence or amelioration of disparities.

Further guidance on analytic challenges, such as optimizing the appropriate separation of high- and low-risk cases, will be crucial as part of future, wide-scale dissemination of SBDH-focused predictive modeling tools. To advance predictive analytics and increase generalizability across the United States, there should also be open-source SBDH resources for methods and databases that leverage previous SBDH research and development [91,92]. Globally, the Research Data Alliance could create a working group to spearhead the creation of open-source SBDH data sources and facilitate work toward interoperability [93].

Expand Standardized Coding and Taxonomies of SBDH Risk Factors That Enhance Interoperability

Once a single health care system renders SBDH data useful through advanced data science, they must find ways to disseminate these advances. The lack of standardization of SBDH data and collection processes prevents the interoperability and integration of modeling into diverse platforms [91,92] and impacts the creation of SBDH products for EHRs [94]. For greater interoperability, we need a standard, practical coding system for SBDH factors that goes beyond vendor-specific coding [91,92]. Such an endeavor is presently being pioneered by the Social Interventions Research and Evaluation Network through the HL7 *Gravity Project* [95].

Expanding the Knowledge Base to Inform Best Practice Guidelines

Support National Shared Research and Development to Advance SBDH Predictive Model Development and Application

In recognition of the emerging field of SBDH predictive analytics, steps toward developing consensus and further evaluative work are needed to produce best practice guidelines for the use of SBDH data in predictive modeling [91]. There is wide variability in the choice of data sources, risk factors, targeted outcomes, geographic levels, and analytic approaches in the SBDH predictive models. Each of these model components can impact a tool's accuracy and appropriateness for use in a particular setting or context. At present, there is a very limited understanding of the impact of these parameters on the effectiveness of the SBDH predictive model. Although endpoints such as health care cost and utilization may seem similar, the choice of health outcome in a model can obscure the path from social risk to health. Best practice guidelines should include transparency of model validation methods for various outcomes to ensure that modeling methods can be replicated in other populations [91]. The use of SBDH variables in predictive modeling is relatively new. Developing consensus might be premature in such circumstances and evaluative work must occur beforehand. However, to form guidelines, it is critical to consider standardization in SBDH predictive analytics and to organize the discourse early on. Such discourse would facilitate data sharing, create open-source tools and algorithms, and set expectations.

Establish a National Agenda to Create a Shared Evidence Base Regarding the Importance of SBDH Factors and the Best Approach for Including SBDH in Analytics

Although the methods and analyses addressing SBDH have matured substantially over the past decades, an expanded data infrastructure and more research are necessary to gain a full understanding of how SBDH manifests throughout a person's life [96]. Present health analytics platforms are generally not built to advance our knowledge base in this area. Rather, they are often intended to give health systems or insurers a leg-up over their competition in achieving financial or pay-for-performance targets. There should be a national agenda to develop and share technology and human resources and strategies to support efficient data extraction, evidence-based development, and effective analytics and reporting within and across institutions in the United States [92]. For-profit entities also have a vested interest to create better predictive models. Such shared desire would be an incentive for them to participate in the development of a shared evidence base, resulting in the creation of better predictive models.

Conclusions

In the face of great challenges and perhaps even greater benefits, we have identified a series of potential approaches for advancing the present state of predictive analytics within the SBDH context. The future of predictive modeling involving SBDH will require key stakeholders—including policy makers, payers, providers, researchers and analysts, patients, and their advocates—to reach a consensus regarding ethical frameworks, data sharing, technical parameters, and model transparency. Such a consensus will help ensure that the ultimate promise of SBDH analytics, improving health and reducing health disparities, is achieved in health care systems and communities across the United States.

Acknowledgments

The work by LB on this project was supported by a grant from the Robert Wood Johnson Foundation.

Authors' Contributions

All the authors contributed significantly to the project and writing of the manuscript. All the authors reviewed the final paper and provided comments as deemed necessary. MT drafted the manuscript and revised it using input from other authors. EH supervised the literature review and development of the overall manuscript. MT, DT, and KV performed the literature review and provided a summary of available studies that address SBDH in predictive modeling. HK and LG provided insight into the application of SBDH in predictive analytics. JW was the principal investigator of the project, who designed the overall scope and goals of the study and supervised the day-to-day operations of the project.

Conflicts of Interest

LG reports receiving funding from the Commonwealth Fund, Episcopal Health Foundation, Kaiser Permanente, NIMHD, and AHRQ for work unrelated to this manuscript. She received support from the Robert Wood Johnson Foundation for her work on this manuscript. The remaining authors declare no conflicts of interest.

References

1. Digital Data Improvement Priorities for Continuous Learning in Health and Health Care: Workshop Summary. Washington, DC: National Academies Press; 2013.
2. Public Health and Promoting Interoperability Programs. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/ehrmeaningfuluse/introduction.html [accessed 2019-10-26]
3. Engelgau MM, Khoury MJ, Roper RA, Curry JS, Mensah GA. Predictive analytics: helping guide the implementation research agenda at the national heart, lung, and blood institute. Glob Heart 2019 Mar;14(1):75-79 [FREE Full text] [doi: 10.1016/j.gheart.2019.02.003] [Medline: 31036305]
4. Duhigg C. How Companies Learn Your Secrets. The New York Times. 2012. URL: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html [accessed 2019-09-08]
5. Deist T, Patti A, Wang Z, Krane D, Sorenson T, Craft D. Simulation-assisted machine learning. Bioinformatics 2019 Oct 15;35(20):4072-4080 [FREE Full text] [doi: 10.1093/bioinformatics/btz199] [Medline: 30903692]
6. Gurley V. Using Predictive Analytics to Address Social Determinants of Health. Population Health Learning Network. 2018. URL: https://www.managedhealthcareconnect.com/article/using-predictive-analytics-address-social-determinants-health [accessed 2019-09-01]
7. Beam AL, Kohane IS. Big data and machine learning in health care. J Am Med Assoc 2018 Apr 3;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
8. Lovis C. Unlocking the power of artificial intelligence and big data in medicine. J Med Internet Res 2019 Nov 8;21(11):e16607 [FREE Full text] [doi: 10.2196/16607] [Medline: 31702565]
9. Galea S, Tracy M, Hoggatt KJ, Dimaggio C, Karpati A. Estimated deaths attributable to social factors in the United States. Am J Public Health 2011 Aug;101(8):1456-1465. [doi: 10.2105/AJPH.2010.300086] [Medline: 21680937]
10. 2019 County Health Rankings Key Findings Report. County Health Rankings & Roadmaps. 2019. URL: https://www.countyhealthrankings.org/reports/2019-county-health-rankings-key-findings-report [accessed 2020-08-24]
11. Xu X, Bishop EE, Kennedy SM, Simpson SA, Pechacek TF. Annual healthcare spending attributable to cigarette smoking: an update. Am J Prev Med 2015 Mar;48(3):326-333 [FREE Full text] [doi: 10.1016/j.amepre.2014.10.012] [Medline: 25498551]
12. Ley SH, Ardisson Korat AV, Sun Q, Tobias DK, Zhang C, Qi L, et al. Contribution of the nurses' health studies to uncovering risk factors for type 2 diabetes: diet, lifestyle, biomarkers, and genetics. Am J Public Health 2016 Sep;106(9):1624-1630. [doi: 10.2105/AJPH.2016.303314] [Medline: 27459454]
13. Walker RE, Keane CR, Burke JG. Disparities and access to healthy food in the United States: a review of food deserts literature. Health Place 2010 Sep;16(5):876-884. [doi: 10.1016/j.healthplace.2010.04.013] [Medline: 20462784]
14. Leonardi C, Simonsen NR, Yu Q, Park C, Scribner RA. Street connectivity and obesity risk: evidence from electronic health records. Am J Prev Med 2017 Jan;52(1S1):S40-S47. [doi: 10.1016/j.amepre.2016.09.029] [Medline: 27989291]
15. Nelson K, Schwartz G, Hernandez S, Simonetti J, Curtis I, Fihn SD. The association between neighborhood environment and mortality: results from a national study of veterans. J Gen Intern Med 2017 Apr;32(4):416-422 [FREE Full text] [doi: 10.1007/s11606-016-3905-x] [Medline: 27815763]
16. Katsaliaki K, Mustafee N. Applications of simulation within the healthcare context. J Oper Res Soc 2011;62(8):1431-1451 [FREE Full text] [doi: 10.1057/jors.2010.20] [Medline: 32226177]
17. Gusoff G. Professional medical association policy statements on social health assessments and interventions. Perm J 2018;22:18-92. [doi: 10.7812/tpp/18-092]
18. Friedman NL, Banegas MP. Toward addressing social determinants of health: a health care system strategy. Perm J 2018(22):18-95. [doi: 10.7812/TPP/18-095]

19. Kankanhalli A, Hahn J, Tan S, Gao G. Big data and analytics in healthcare: introduction to the special section. Inf Syst Front 2016 Mar 9;18(2):233-235. [doi: 10.1007/s10796-016-9641-2]
20. World Health Organization. The Economics of Social Determinants of Health and Health Inequalities: A Resource Book. Geneva, Switzerland: World Health Organization; 2013.
21. Center for Prevention Services. Ten Leading Causes of Death in the United States. Atlanta, GA: Centers for Disease Control and Prevention; 1977.
22. McGinnis JM, Williams-Russo P, Knickman JR. The case for more active policy attention to health promotion. Health Aff (Millwood) 2002;21(2):78-93. [doi: 10.1377/hlthaff.21.2.78] [Medline: 11900188]
23. World Conference on Social Determinants of Health. World Health Organization. 2020. URL: http://www.who.int/social_determinants/sdhconference/background/en/ [accessed 2020-08-24]
24. Integrating Social Care Into the Delivery of Health Care: Moving Upstream to Improve the Nation's Health. Washington, DC: National Academies Press; 2019.
25. Guiding Principles for Ethical Use of Social Determinants of Health Data. EHealth Initiative. 2019. URL: https://www.ehidc.org/resources/guiding-principles-ethical-use-social-determinants-health-data [accessed 2019-10-27]
26. Nau C, Adams JL, Roblin D, Schmittdiel J, Schroeder E, Steiner JF. Considerations for identifying social needs in health care systems. Med Care 2019;57(9):661-666. [doi: 10.1097/mlr.0000000000001173]
27. Socioeconomic Health Scores. LexisNexis Risk Solutions. URL: https://risk.lexisnexis.com/products/socioeconomic-health-score [accessed 2019-09-01]
28. Vest JR, Menachemi N, Grannis SJ, Ferrell JL, Kasthurirathne SN, Zhang Y, et al. Impact of risk stratification on referrals and uptake of wraparound services that address social determinants: a stepped wedged trial. Am J Prev Med 2019 Apr;56(4):e125-e133. [doi: 10.1016/j.amepre.2018.11.009] [Medline: 30772150]
29. Simpson M, Genovese A. Carrot Health - Leveraging Consumer Data to Grow Medicare Market Share. The Carrot MarketView. 2016. URL: https://info.carrothealth.com/hubfs/Brochures%20and%20Whitepapers/Carrot%20Health%20-%20Leveraging%20Consumer%20Data%20to%20Grow%20Medicare%20Market%20Share.pdf?__hstc=122733652.515b5d9ff7a33417378b5a218fdca83f.1567383407805.1567386846670.1567398132639.3&__hssc=122733652.1.1567398132639 [accessed 2019-09-02]
30. Predmore Z, Hatef E, Weiner JP. Integrating social and behavioral determinants of health into population health analytics: a conceptual framework and suggested road map. Popul Health Manag 2019 Dec;22(6):488-494. [doi: 10.1089/pop.2018.0151] [Medline: 30864884]
31. McGinnis T, Crumley D, Chang D. Implementing Social Determinants of Health Interventions in Medicaid Managed Care: How to Leverage Existing Authorities and Shift to Value-Based Purchasing. AcademyHealth. 2018. URL: https://www.academyhealth.org/sites/default/files/implementing_sdoh_medicaid_managed_care_may2018.pdf [accessed 2019-09-24]
32. Matulis R, Lloyd J. The History, Evolution, and Future of Medicaid Accountable Care Organizations. Center for Health Care Strategies. 2020. URL: https://www.chcs.org/resource/history-evolution-future-medicaid-accountable-care-organizations/ [accessed 2020-08-24]
33. Artiga S, Hinton E. Beyond Health Care: The Role of Social Determinants in Promoting Health and Health Equity. Kaiser Family Foundation. 2018. URL: https://www.kff.org/disparities-policy/issue-brief/beyond-health-care-the-role-of-social-determinants-in-promoting-health-and-health-equity/ [accessed 2019-09-08]
34. How Insurance Companies Set Health Premiums. HealthCare. URL: https://www.healthcare.gov/how-plans-set-your-premiums/ [accessed 2020-05-12]
35. Chin T, Kahn R, Li R, Chen J, Krieger N, Buckee C, et al. US county-level characteristics to inform equitable COVID-19 response. medRxiv 2020 Apr 11:11 epub ahead of print [FREE Full text] [doi: 10.1101/2020.04.08.20058248] [Medline: 32511610]
36. Owen WF, Carmona R, Pomeroy C. Failing another national stress test on health disparities. J Am Med Assoc 2020 Apr 15;323(19):1905-1906 epub ahead of print. [doi: 10.1001/jama.2020.6547] [Medline: 32293642]
37. Cases in the US. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html [accessed 2020-06-14]
38. Weekly Updates by Select Demographic and Geographic Characteristics: Provisional Death Counts for Coronavirus Disease 2019 (COVID-19). Centers for Disease Control and Prevention. 2020. URL: https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm [accessed 2020-06-14]
39. Wadhera RK, Wadhera P, Gaba P, Figueroa JF, Maddox KE, Yeh RW, et al. Variation in COVID-19 hospitalizations and deaths across New York City boroughs. J Am Med Assoc 2020 Apr 29;323(21):2192 [FREE Full text] [doi: 10.1001/jama.2020.7197] [Medline: 32347898]
40. Braithwaite R, Warren R. The African American petri dish. J Health Care Poor Underserved 2020;31(2):491-502. [doi: 10.1353/hpu.2020.0037]
41. Mapping High Risk Areas for COVID-19. Health Landscape. URL: https://www.healthlandscape.org/coronavirus/ [accessed 2020-06-14]
42. National Academies of Sciences. Accounting for Social Risk Factors in Medicare Payment: Criteria, Factors, and Methods. Washington, DC: National Academies of Sciences Engineering Medicine; 2016.

43. Hatef E, Searle KM, Predmore Z, Lasser EC, Kharrazi H, Nelson K, et al. The impact of social determinants of health on hospitalization in the veterans health administration. Am J Prev Med 2019 Jun;56(6):811-818. [doi: 10.1016/j.amepre.2018.12.012] [Medline: 31003812]
44. Bhavsar NA, Gao A, Phelan M, Pagidipati NJ, Goldstein BA. Value of neighborhood socioeconomic status in predicting risk of outcomes in studies that use electronic health record data. JAMA Netw Open 2018 Sep 7;1(5):e182716 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.2716] [Medline: 30646172]
45. Brar Prayaga R, Agrawal R, Nguyen B, Jeong EW, Noble HK, Paster A, et al. Impact of social determinants of health and demographics on refill requests by medicare patients using a conversational artificial intelligence text messaging solution: cross-sectional study. JMIR Mhealth Uhealth 2019 Nov 18;7(11):e15771 [FREE Full text] [doi: 10.2196/15771] [Medline: 31738170]
46. Seligman B, Tuljapurkar S, Rehkopf D. Machine learning approaches to the social determinants of health in the health and retirement study. SSM Popul Health 2018 Apr;4:95-99 [FREE Full text] [doi: 10.1016/j.ssmph.2017.11.008] [Medline: 29349278]
47. Kasthurirathne S, Vest J, Menachemi N, Halverson P, Grannis S. Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services. J Am Med Inform Assoc 2018 Jan 1;25(1):47-53. [doi: 10.1093/jamia/ocx130] [Medline: 29177457]
48. Golembiewski E, Allen KS, Blackmon AM, Hinrichs RJ, Vest JR. Combining nonclinical determinants of health and clinical data for research and evaluation: rapid review. JMIR Public Health Surveill 2019 Oct 7;5(4):e12846 [FREE Full text] [doi: 10.2196/12846] [Medline: 31593550]
49. All-Payer Claims Databases. The Agency for Healthcare Research and Quality. 2017. URL: https://www.ahrq.gov/data/apcd/index.html [accessed 2020-08-24]
50. Enhance Healthcare Analytics with Consumer Data. SlideShare. 2020. URL: https://www.slideshare.net/RayPun/enhance-healthcare-analytics-with-consumer-data [accessed 2020-08-24]
51. Hatef E, Rouhizadeh M, Tia I, Lasser E, Hill-Briggs F, Marsteller J, et al. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. JMIR Med Inform 2019 Aug 2;7(3):e13802 [FREE Full text] [doi: 10.2196/13802] [Medline: 31376277]
52. Guo Y, Zheng G, Fu T, Hao S, Ye C, Zheng L, et al. Assessing statewide all-cause future one-year mortality: prospective study with implications for quality of life, resource utilization, and medical futility. J Med Internet Res 2018 Jun 4;20(6):e10311 [FREE Full text] [doi: 10.2196/10311] [Medline: 29866643]
53. American Community Survey (ACS). US Census Bureau. URL: https://www.census.gov/programs-surveys/acs/ [accessed 2019-08-10]
54. Food Access Research Atlas. The Economics of Food, Farming, Natural Resources, and Rural America. URL: https://www.ers.usda.gov/data-products/food-access-research-atlas/ [accessed 2019-08-10]
55. Berkowitz SA, Basu S, Venkataramani A, Reznor G, Fleegler EW, Atlas SJ. Association between access to social service resources and cardiometabolic risk factors: a machine learning and multilevel modeling analysis. BMJ Open 2019 Mar 12;9(3):e025281 [FREE Full text] [doi: 10.1136/bmjopen-2018-025281] [Medline: 30862634]
56. American Housing Survey (AHS). US Census Bureau. URL: https://www.census.gov/programs-surveys/ahs.html [accessed 2019-08-11]
57. Hughes HK, Matsui EC, Tschudy MM, Pollack CE, Keet CA. Pediatric asthma health disparities: race, hardship, housing, and asthma in a national survey. Acad Pediatr 2017 Mar;17(2):127-134 [FREE Full text] [doi: 10.1016/j.acap.2016.11.011] [Medline: 27876585]
58. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
59. Schroeder EB, Xu S, Goodrich GK, Nichols GA, O'Connor PJ, Steiner JF. Predicting the 6-month risk of severe hypoglycemia among adults with diabetes: development and external validation of a prediction model. J Diabetes Complications 2017 Jul;31(7):1158-1163 [FREE Full text] [doi: 10.1016/j.jdiacomp.2017.04.004] [Medline: 28462891]
60. FAQ. AllTransit. URL: https://alltransit.cnt.org/faq/ [accessed 2019-08-11]
61. Air Quality Index Report. United States Environmental Protection Agency: US EPA. URL: https://www.epa.gov/outdoor-air-quality-data/air-quality-index-report [accessed 2019-08-11]
62. Food Access Research Atlas. The Economics of Food, Farming, Natural Resources, and Rural America. URL: https://www.ers.usda.gov/data-products/food-access-research-atlas/ [accessed 2019-10-25]
63. Grinspan ZM, Patel AD, Hafeez B, Abramson EL, Kern LM. Predicting frequent emergency department use among children with epilepsy: a retrospective cohort study using electronic health data from 2 centers. Epilepsia 2018 Jan;59(1):155-169 [FREE Full text] [doi: 10.1111/epi.13948] [Medline: 29143960]
64. Chaiyachati KH, Hubbard RA, Yeager A, Mugo B, Lopez S, Asch E, et al. Association of rideshare-based transportation services and missed primary care appointments: a clinical trial. JAMA Intern Med 2018 Mar 1;178(3):383-389. [doi: 10.1001/jamainternmed.2017.8336] [Medline: 29404572]

65. Bresnick J. Social Determinants of Health Dashboard Expands to 500 Cities. HealthITAnalytics. URL: https://healthitanalytics.com/news/social-determinants-of-health-dashboard-expands-to-500-cities [accessed 2019-09-02]
66. 500 Cities: Local Data for Better Health. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/500cities/index.htm [accessed 2019-09-02]
67. Bringing Consumerism to Healthcare: Powered by Social and Behavioral Determinants of Health (SDoH). Carrot Health. URL: https://carrothealth.com/wp-content/uploads/2018/04/MarketView-For-Payers.pdf [accessed 2019-11-09]
68. Benaroya R. Health Plan Member Engagement Strategies That Improve Satisfaction and Outcomes. Cecelia Health. 2020. URL: https://www.ceceliahealth.com/blog/health-plan-member-engagement-strategies-that-improve-satisfaction-and-outcomes [accessed 2020-08-24]
69. Hutson M. Artificial intelligence faces reproducibility crisis. Science 2018 Feb 16;359(6377):725-726. [doi: 10.1126/science.359.6377.725] [Medline: 29449469]
70. Christopher V, Rao D, Giabbanelli P. How Do Modelers Code Artificial Societies? Investigating Practices and Quality of Netlogo Codes from Large Repositories. In: Spring Simulation Conference. 2020 Presented at: SS'20; March 29-April 1, 2020; Fairfax, VA, USA. [doi: 10.22360/springsim.2020.hsaa.007]
71. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). Ann Intern Med 2015 May 19;162(10):735-773. [doi: 10.7326/l15-5093-2]
72. Wharam J, Weiner J. The promise and peril of healthcare forecasting. Am J Manag Care 2012 Mar 1;18(3):e82-e85 [FREE Full text] [Medline: 22435964]
73. Copeland R. Google’s ‘Project Nightingale’ Gathers Personal Health Data on Millions of Americans. The Wall Street Journal. 2019. URL: https://www.wsj.com/articles/google-s-secret-project-nightingale-gathers-personal-health-data-on-millions-of-americans-11573496790 [accessed 2019-11-14]
74. Steiner J, Clift M, Nau C, Schroeder E. Issue Brief: Survey Results Update. SONNET. 2018. URL: https://sonnet.kaiserpermanente.org/products.html [accessed 2020-08-24]
75. McGraw D. Privacy concerns related to inclusion of social and behavioral determinants of health in electronic health records. In: Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2. Washington, DC: The National Academies Press; 2015.
76. Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Aff (Millwood) 2014 Jul;33(7):1139-1147. [doi: 10.1377/hlthaff.2014.0048] [Medline: 25006139]
77. Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. Health Aff (Millwood) 2014 Jul;33(7):1148-1154. [doi: 10.1377/hlthaff.2014.0352] [Medline: 25006140]
78. Beier K, Schweda M, Schicktanz S. Taking patient involvement seriously: a critical ethical analysis of participatory approaches in data-intensive medical research. BMC Med Inform Decis Mak 2019 Apr 25;19(1):90 [FREE Full text] [doi: 10.1186/s12911-019-0799-7] [Medline: 31023321]
79. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal 2002;6(5):429-449. [doi: 10.3233/ida-2002-6504]
80. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
81. White S. A review of big data in health care: challenges and opportunities. Open Access Bioinforma Macclesfield 2014 Oct:13. [doi: 10.2147/oab.s50519]
82. Oreskovic NM, Maniates J, Weilburg J, Choy G. Optimizing the use of electronic health records to identify high-risk psychosocial determinants of health. JMIR Med Inform 2017 Aug 14;5(3):e25 [FREE Full text] [doi: 10.2196/medinform.8240] [Medline: 28807893]
83. Kind AJ, Jencks S, Brock J, Yu M, Bartels C, Ehlenbach W, et al. Neighborhood socioeconomic disadvantage and 30-day rehospitalization: a retrospective cohort study. Ann Intern Med 2014 Dec 2;161(11):765-774 [FREE Full text] [doi: 10.7326/M13-2946] [Medline: 25437404]
84. Bazemore A, Cottrell E, Gold R, Hughes L, Phillips R, Angier H, et al. 'Community vital signs': incorporating geocoded social determinants into electronic records to promote patient and population health. J Am Med Inform Assoc 2016 Mar;23(2):407-412. [doi: 10.1093/jamia/ocv088] [Medline: 26174867]
85. Haining R. Spatial autocorrelation. In: Smelser NJ, Baltes PB, editors. International Encyclopedia of Social & Behavioral Sciences. Oxford, UK: Pergamon; 2001.
86. Chancellor L, Baijal S. Optimizing Healthcare Analytics: How to Choose the Right Predictive Model. EXL Service, Digital Intelligence, Analytics & Operations. URL: https://www.exlservice.com/resources/assets/library/documents/EXL_WP_HC_OptimizingHealthcareAnalytics.pdf [accessed 2020-08-12]
87. Dhar V. Big data and predictive analytics in health care. Big Data 2014 Sep;2(3):113-116. [doi: 10.1089/big.2014.1525] [Medline: 27442491]

88. Steiner JF, Stenmark SH, Sterrett AT, Paolino AR, Stiefel M, Gozansky WS, et al. Food insecurity in older adults in an integrated health care system. J Am Geriatr Soc 2018 May;66(5):1017-1024. [doi: 10.1111/jgs.15285] [Medline: 29492953]
89. Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. J Am Med Assoc 2018 Jul 3;320(1):27-28. [doi: 10.1001/jama.2018.5602] [Medline: 29813156]
90. Phelan J, Link B. Controlling disease and creating disparities: a fundamental cause perspective. J Gerontol B Psychol Sci Soc Sci 2005 Oct;60(Spec No 2):27-33. [doi: 10.1093/geronb/60.special_issue_2.s27] [Medline: 16251587]
91. Amarasingham R, Audet AJ, Bates DW, Cohen IG, Entwistle M, Escobar GJ, et al. Consensus statement on electronic health predictive analytics: a guiding framework to address challenges. EGEMS (Wash DC) 2016;4(1):1163 [FREE Full text] [doi: 10.13063/2327-9214.1163] [Medline: 27141516]
92. Ross TR, Ng D, Brown JS, Pardee R, Hornbrook MC, Hart G, et al. The HMO research network virtual data warehouse: a public data model to support collaboration. EGEMS (Wash DC) 2014;2(1):1049 [FREE Full text] [doi: 10.13063/2327-9214.1049] [Medline: 25848584]
93. Research Data Sharing Without Barriers. Research Data Alliance. URL: https://www.rd-alliance.org/ [accessed 2020-08-12]
94. Freij M, Dullabh P, Lewis S, Smith SR, Hovey L, Dhopeshwarkar R. Incorporating social determinants of health in electronic health records: qualitative study of current practices among top vendors. JMIR Med Inform 2019 Jun 7;7(2):e13849 [FREE Full text] [doi: 10.2196/13849] [Medline: 31199345]
95. The Gravity Project: A National Collaborative to Advance Interoperable Social Determinants of Health Data. UCSF Siren. URL: https://sirenetwork.ucsf.edu/TheGravityProject [accessed 2019-10-13]
96. Glass TA, McAtee MJ. Behavioral science at the crossroads in public health: extending horizons, envisioning the future. Soc Sci Med 2006 Apr;62(7):1650-1671. [doi: 10.1016/j.socscimed.2005.08.044] [Medline: 16198467]

Abbreviations

CDC: Centers for Disease Control and Prevention
EHR: electronic health record
NLP: natural language processing
SBDH: social and behavioral determinants of health
TRIPOD: Transparent Reporting of a multivariate prediction model for Individual Prognosis or Diagnosis

Edited by C Lovis; submitted 02.02.20; peer-reviewed by J Steiner, J op den Buijs, P Giabbanelli; comments to author 06.05.20; revised version received 17.06.20; accepted 20.07.20; published 08.09.20.

Please cite as:
Tan M, Hatef E, Taghipour D, Vyas K, Kharrazi H, Gottlieb L, Weiner J
Including Social and Behavioral Determinants in Predictive Models: Trends, Challenges, and Opportunities
JMIR Med Inform 2020;8(9):e18084
URL: http://medinform.jmir.org/2020/9/e18084/
doi:10.2196/18084
PMID:32897240

Original Paper

Medical Insurance Information Systems in China: Mixed Methods Study

Yazi Li[1], PhD; Chunji Lu[1], MD; Yang Liu[1], MD

The Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China

Corresponding Author:
Yazi Li, PhD
The Institute of Medical Information
Chinese Academy of Medical Sciences
No. 3 Yabao Road, Chaoyang District, Beijing
Beijing
China
Phone: 86 13810084409
Email: li.yazi@imicams.ac.cn

Abstract

Background: Since the People's Republic of China (PRC), or China, established the basic medical insurance system (MIS) in 1998, the medical insurance information systems (MIISs) in China have effectively supported the operation of the MIS through several phases of development; the phases included a stand-alone version, the internet, and big data. In 2018, China's national medical security systems were integrated, while MIISs were facing reconstruction. We summarized China's experience in medical insurance informatization over the past 20 years, aiming to provide a reference for the building of a new basic MIS for China and for developing countries.

Objective: This paper aims to sort out medical insurance informatization policies throughout the years, use questionnaires to determine the status quo of provincial MIIS-building in China and the relevant policies, provide references and suggestions for the top-level design and implementation of the information systems in the transitional period of China's MIS reform, and provide a reference for the building of MIISs in developing countries.

Methods: We conducted policy analysis by collecting the laws, regulations, and policy documents—issued from 1998 to 2020—on China's medical insurance and its informatization; we also analyzed the US Health Insurance Portability and Accountability Act and other relevant policies. We conducted a questionnaire survey by sending out questionnaires to 31 Chinese, provincial, medical security bureaus to collect information about network links, system functions, data exchange, standards and specifications, and building modes, among other items. We conducted a literature review by searching for documents about relevant laws and policies, building methods, application results, and other documents related to MIISs; we conducted searches using PubMed, Elsevier, China National Knowledge Infrastructure, and other major literature databases. We conducted telephone interviews to verify the results of questionnaires and to understand the focus issues concerning the building of China's national MIISs during the period of integration and transition of China's MIS.

Results: In 74% (23/31) of the regions in China, MIISs were networked through dedicated fiber optic lines. In 65% (20/31) of the regions in China, MIISs supported identity recognition based on both ID cards and social security cards. In 55% (17/31) of the regions in China, MIISs at provincial and municipal levels were networked and have gathered basic medical insurance data, whereas MIISs were connected to health insurance companies in 35% (11/31) of the regions in China. China's MIISs are comprised of 11 basic functional modules, among which the modules of business operation, transregional referral, reimbursement, and monitoring systems are widely applied. MIISs in 83% (20/24) of Chinese provinces have stored data on coverage, payment, and settlement compensation of medical insurance. However, in terms of data security and privacy protection, pertinent policies are absent and data utilization is not in-depth enough. Respondents to telephone interviews universally reflected on the following issues and suggestions: in the period of integration and transition of MISs, close attention should be paid to the top-level design, and repeated investment should be avoided for the building of MIISs; MIISs should be adapted to the health care reform, and efforts should be made to strengthen the informatization support for the reform of payment methods; and MIISs should be adapted for the widespread application of mobile phones and should provide insured persons with more self-service functions.

Conclusions: In the future, the building of China's basic MIISs should be deployed at the national, provincial, prefectural, and municipal levels on a unified basis. Efforts should be made to strengthen the development of standard codes, data exchange, and data utilization. Work should be done to formulate the rules and regulations for security and privacy protection and to balance the right to be informed with the mining and utilization of big data. Efforts should be made to intensify the interconnectivity between MISs and other health systems and to strengthen the application of medical insurance information in public health monitoring and early warning systems; this would ultimately improve the degree of trust from stakeholders, including individuals, medical service providers, and public health institutions, in the basic MIISs.

(JMIR Med Inform 2020;8(9):e18780) doi:10.2196/18780

KEYWORDS

medical insurance; medical insurance information system; health information exchange; information infrastructure; big data; policy review; privacy protection

Introduction

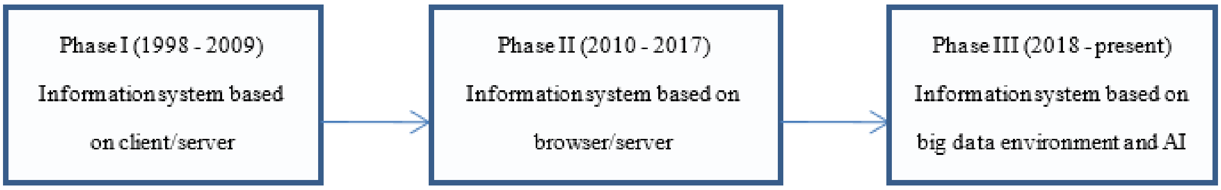
Background

China's medical insurance information systems (MIISs) are connected to over 700,000 medical institutions at various levels; cover more than 1.35 billion people [1]; record medical insurance data regarding coverage, payment, claim, and compensation; offer information management tools for the world's largest medical security network; and provide innovation means for health care reform and change in medical insurance systems (MISs). Many countries in the world have built their own national-level medical security information networks. For example, France built the National Health Insurance Inter-regime Information System [2]; the United States established a medical security network that covers all medical service providers nationwide through the Health Information Technology for Economic and Clinical Health Act of 2009 [3]; and South Korea built the National Health Information Database [4], which provides support for the operation of its MIS.

Over the past 20 years, China's MISs consisted of the following three main parts: (1) medical insurance for urban employees established in 1998, (2) the New Rural Cooperative Medical System (NRCMS) for rural residents established in 2003, and (3) medical insurance for urban residents established in 2008. Medical insurance for urban employees and urban residents was overseen by the Ministry of Human Resources and Social Security of the People's Republic of China (PRC), and the NRCMS for rural residents was overseen by the Ministry of Health of the PRC [5]. In 2018, after the completion of China's system reform and national institutional reform, China's basic MISs were integrated, and the National Healthcare Security Administration (NHSA) of the PRC was established, while two systems were retained; these two systems were the medical insurance for urban employees and the basic medical insurance for urban and rural residents, resulting from the consolidation of medical insurance for urban residents and the NRCMS. The function of assistance for the disadvantaged was transferred from the Ministry of Civil Affairs of the PRC to the NHSA, the functions of pricing of medical service items and bidding procurement of medicines (ie, drugs) were transferred from the National Development and Reform Commission (NDRC) of the PRC to the NHSA, and the function of collection and payment of medical insurance funds would be implemented by the taxation departments [6,7].

Since 1998, the building of China's MIISs has gone through three phases: Phase I, where the stand-alone version realized the handling of medical insurance business; Phase II, where extensive interconnection was based on the internet; and Phase III, where comprehensive decision-making functions based on big data were realized. China's MIISs were established on the basis of the pooling level and have the functions of fundraising, payment, online reimbursement and settlement, and capital settlement, among others. Figure 1 shows the three phases of the development of China's MIISs.

Figure 1. The three phases of the development of China's medical insurance information systems (MIISs).


Phase I (1998 - 2009)
Information system based
on client/server
Phase II (2010 - 2017)
Information system based on
browser/server
Phase III (2018 - present)
Information system based on
big data environment and AI

Phase I took place from 1998 to 2009, when local MIISs were established separately in different pooling-based regions in China according to the needs for business development. Over 2000 NRCMS information systems were established in China at the county level, and more than 300 information systems were established at the prefectural or municipal level in China for urban employees and residents. In terms of functions, these information systems mainly functioned to manage the accounts and insurance participation registration for the insured entities, families, and individuals, and they managed fund collection and expenditure. Before 2009, China's MIISs were based on a stand-alone version [8], which lacked overall planning and design; there existed such serious problems as repeated investment and isolated islands of information systems, while circumstances of poor system interaction and repeated insurance participation occurred from time to time [9].

Phase II took place from 2010 to 2017, when the medical insurance management system tended to be integrated, and MIISs entered the integration phase [10], focusing on network interconnection. The NRCMS information systems at the county level in 24 Chinese provinces were integrated into MIISs, while the situation remained unchanged in the remaining seven Chinese provinces. On the one hand, China continued to expand the coverage of the insured groups, and MIISs continued to expand in coverage; the systems further emphasized the interconnection among the central government, provinces, cities, and counties. As well, data were centralized from the bottom to the top, and data applications were further deepened [11,12]. Improvements included the following: the in-depth application of intelligent audit was realized; the diagnosis, treatment, and service behaviors of medical institutions, physicians and doctors, and pharmacies and drugstores were put under monitoring [13]; settlement and reimbursement for cross-provincial hospitalization were extensively carried out; people participating in the medical insurance scheme could transfer their medical insurance coverage to other provinces; public services, such as payment and direct reimbursement and settlement, became more convenient [14]; and there were more and more big data–based analyses and utilizations, as well as applications and studies on macro decision making [15].

In Phase III, information systems were built in line with the functions of the new NHSA, including the new medical security bureaus at all levels. In the context of big data—in addition to the realization of such functions as management and handling of medical insurance—price administration of medical services, and bidding and procurement of drugs, more data analysis and utilization can be realized; this would provide support for actuarial service for insurance and policy formulation. The new information systems will be built with 14 subfunctions in four categories. First, information systems in the category of internal management include the internal unified portal system and the process control system. Second, information systems in the category of business management include the basic information management system, the credit rating management system, the medical and drug price management system, and the payment method management system. Third, information systems in the category of production handling include the system for basic management of business handling and public services, the bidding and procurement system for drugs and medical supplies, and the cross-provincial transregional medical service management system. Fourth, information systems in the category of data analysis include the operation monitoring system, the intelligent regulation system, and the macro decision-making application system based on big data mining technology.

Objective

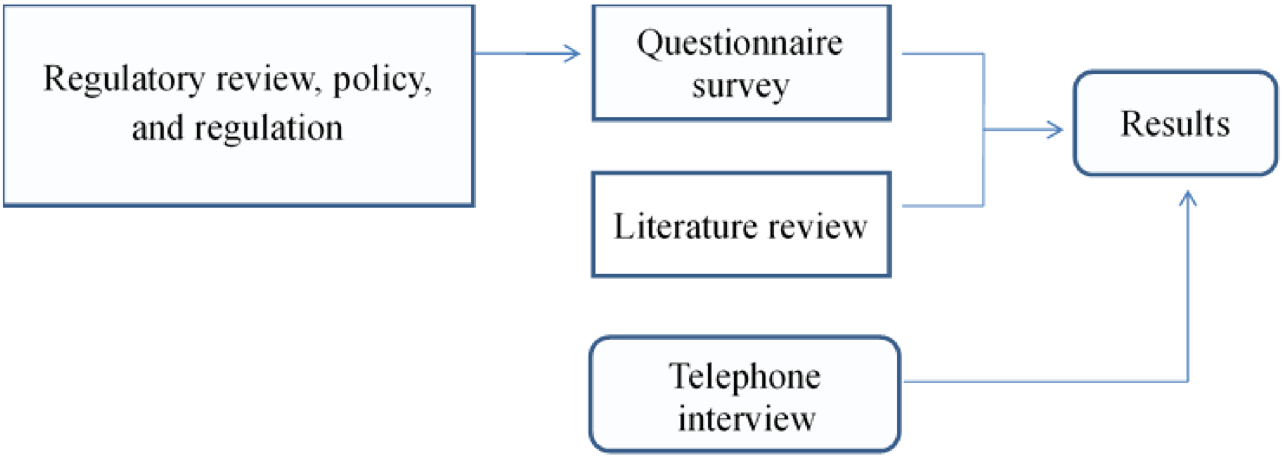
This paper aims to collect and analyze the medical insurance informatization policies throughout the years, survey the status quo of provincial MIIS-building in China by means of questionnaires about the building of MIISs and their relevant policies, provide references and suggestions for the top-level design and implementation of the information systems in the transitional period of China’s MIS reform, and provide a reference for the building of MIISs in developing countries.

Methods

Overview

We conducted a literature review to learn about the relevant policies, regulations, methods, and application results of MIIS-building practices at home and abroad. We conducted a policy analysis by collecting the laws, regulations, and policy documents on medical insurance and medical insurance informatization issued from 1998 to 2020. We conducted a questionnaire survey by designing and distributing questionnaires to the medical security bureaus of 31 Chinese provinces. We conducted telephone interviews to verify the results of the questionnaires and to understand the focus issues concerning the building of China’s national MIISs during the period of integration and transition of China's MIS. Figure 2 shows the design of the methods used for this paper for the policy analysis, literature review, questionnaire survey, and telephone interview.

Figure 2. Flowchart of methods used in the study.



Literature Review

We used the following keywords to search in literature databases, including PubMed and ScienceDirect: “medical insurance information system,” “health insurance information system,” “claim data,” “health information system,” “HIPAA” (Health Insurance Portability and Accountability Act), “privacy policy,” “big data,” “health information exchange,” “health information standard,” and “health insurance database.” We also used the following keywords to search in the China National Knowledge Infrastructure platform, a Chinese literature database: “basic medical insurance information system,” “new rural cooperative medical scheme information system,” “medical insurance for urban employees information system,” “medical insurance information system,” “medical insurance database,” “big data,” and “medical insurance.” The publication dates of literature searched ranged from January 1998 to September 2019.

Policy Review

Policy List

We searched the contents of the policies and technical standards related to China’s medical insurance from the policy and regulation columns on the official websites of the Chinese Government, the NHSA, the National Health Commission of the PRC, the Ministry of Human Resources and Social Security of the PRC, the NDRC of the PRC, and the Ministry of Civil Affairs of the PRC, among others; from these, we collected 124 content entries in total. In addition, we also searched the documents of foreign policies related to medical insurance from the official websites of the US Centers for Medicare & Medicaid Services, the US Department of Health and Human Services, and the French Ministry of Social Affairs and Health, among others; from these, we collected more than 30 content entries in total. Table 1 shows the list of 14 entries of representative policies and regulations [16-32] closely related to medical insurance informatization among the 124 entries of policy documents from China.

Table 1. List of China’s policies and regulations related to medical insurance informatization from 1998 to the present.

Regulation and policy names	Year (document No.)	Summary of informatization document
Decision on Establishing the Basic Medical Insurance System for Urban Employees[a] [16]		
1. Key points of building the planning for the labor and social insurance management information system [17]	1998 (LS-BH[1998]138)	To standardize the labor and social insurance management information system, of which the contents under management involve the microinformation of laborers, enterprises, and other labor organizations; strengthen the management of personnel, wages, job positions, labor relations, and social insurance relations; and solve fundraising, payment, and other handling of businesses
2. Notice of guidance on the building of the basic medical insurance management information system for urban employees [18]	2000 (LSTH[2000]30)	To establish network connections with designated medical institutions, designated retail pharmacies, banks, tax departments, and other relevant departments through establishing a computer management information system; focus on the unification of classification standards, interface standards, and network transmission standards; and promote the application of identity recognition media for ID cards
3. Opinions on comprehensive implementation of the Jinbao Program for unified building of a labor security information system [19]	2003 (LS-BH[2003]174)	To plan to establish a labor security information system with unified standards, which covers business handling of medical insurance, unemployment insurance, and industrial injury insurance; emphasize data centralization at the provincial and national levels; and build data centers at different levels
Opinions of the Ministry of Health, Ministry of Finance, and Ministry of Agriculture on the Establishment of a New Rural Cooperative Medical Insurance System[b] [20]		
4. Basic specifications of the information system for the New Rural Cooperative Medical System (NRCMS) (trial) [21]	2005 (WBN-WF[2005]108)	To build an information system for handling the business of the NRCMS at county level, which comprises six modules, including NRCMS participation management, medical service compensation management, fund collection and expenditure management, accounting, and statistical analysis
Guidelines on Launching a Pilot Program of Basic Medical Insurance for Urban Residents[c] [22]		
5. Notice on carrying out unified implementation of some application software for the Jinbao Program [23]	2008 (RSTH[2008]284)	To build private networks at national, provincial, and municipal levels; upgrade the stand-alone version to a networking operation; and unify the core software functions of the financial exchange library software, fund statement software, fund regulation software, and transregional national service hotline software
6. Basic specifications of the management information system for the NRCMS (2008 revised edition) [24]	2008 (WBN-WF[2008]127)	To upgrade the *Basic specifications of the information system for the NRCMS (trial)* and to put forward the basic architecture for building the information system for the NRCMS China-wide, the building specification of physical environment and infrastructure, the functional specifications of the information system, and the datasets and code specifications for cross-system data exchange, among others
7. Notice on carrying out unified upgrading implementation of some application software for the Jinbao Program [25]	2008 (RSXXH[2008]2)	To revise and upgrade the *Notice on carrying out unified implementation of some application software for the Jinbao Program*, improve the building of a network based on nationwide connectivity, and accelerate data centralization at the national level
8. The scheme for connectivity technology of the national-level information platform for the NRCMS (trial) [26]	2013 (WBN-WH[2013]456)	To establish a national-level information platform for the NRCMS, establish a data exchange network connecting all provincial platforms, collect data from all provinces, and explore the functions of cross-provincial cost verification and reimbursement
9. Notice of the General Office of the Ministry of Human Resources and Social Security on comprehensively promoting the intelligent monitoring of medical services for basic medical insurance [27]	2015 (RSTF[2015]56)	To carry out all-around intelligent monitoring of outpatient service, hospitalization, and drug purchase in pharmacies and drugstores through an information system, identify suspected violations, and then verify and handle such violations

Regulation and policy names	Year (document No.)	Summary of informatization document
10. Notice of the General Office of the Ministry of Human Resources and Social Security on carrying out the building of the information system for registration of universal participation in insurance [28]	2015 (RSTF[2015]86)	To plan to further raise the level of information system building at the provincial level, promote social security cards, advance the building of the information system for registration of universal participation in insurance, and push data at the national level; in addition, establish a list of qualifications for developers undertaking the building of information systems for the medical security industry
11. Implementation plan for networked settlement and reimbursement for transregional hospitalization in the national NRCMS [29]	2016 (GWJCF[2016]23)	To build a national NRCMS network for cross-provincial hospitalization settlement and realize the functions of hospital visits, referral, hospitalization registration, and discharge reimbursement for patients participating in the NRCMS
12. Notice of the General Office of the Ministry of Human Resources and Social Security on accelerating the building of a cross-provincial hospitalization settlement system [30]	2016 (RSTF[2016]185)	To realize functions such as filing of off-site urban employees participating in medical insurance and settlement for discharge reimbursement
China established the National Healthcare Security Administration (NHSA)[d] [31]		
13. Notice of the NHSA on printing and distributing the guidance on medical security informatization [32]	2019 (YBF[2019]39)	After its inception, the new NHSA proposed to establish a new, nationally integrated, standard-unified Medical Security Information System, which covers such functions as management of basic knowledge (eg, dictionary directory), business handling management, public service management, and data analysis management; the system developed 15 content standards, such as diagnosis and surgery standards, among others
14. Notice of the NHSA on carrying out the pilot work of medical security informatization	2019 (YBF[2019]22)	The NHSA selected 16 pilot provinces for implementing information system functions first, carrying out 15 standards, and realizing cross-provincial business linkage and public services through the building of the national platform

[a] In 1998, the State Council of the People's Republic of China (PRC) issued the Decision on Establishing the Basic Medical Insurance System for Urban Employees (No. GF[1998]44), enforcing a basic medical insurance system (MIS) for urban employees throughout China, and exploring the establishment of socialized medical insurance for the population with labor and employment relationships [16].

[b] In 2003, the State Council of the PRC issued the Opinions of the Ministry of Health, Ministry of Finance, and Ministry of Agriculture on the Establishment of a New Rural Cooperative Medical Insurance System (No. ZF[2002]13) [20], aiming to establish an MIS for rural residents on a pilot basis.

[c] In 2007, the State Council of the PRC issued the Guidelines on Launching a Pilot Program of Basic Medical Insurance for Urban Residents [22], aiming to establish a basic MIS for nonemployed urban residents and those with nonfixed-employment relationships.

[d] In 2018, China reformed its state institutions and established the NHSA, which functions to unify the administration of the basic MIS [31].

Policy Analysis

We analyzed the evolution of policies from the five main elements that constitute management information systems (ie, organizational structure, process, data, business rules, and system functions).

Questionnaire Survey

We surveyed the health care security bureaus, or their information centers, of 31 Chinese provinces in terms of network infrastructure, identity recognition media, information system functions, data centers, standards, and specifications. We sent out 31 questionnaires and recovered all 31 of them.

Telephone Interview

We interviewed more than 30 persons by telephone from the departments of health care security administration, medical insurance handling management, health administration, and hospital management, among others. Issues discussed during telephone interviews included the functions urgently needed by MIISs and concerns about information system building in the integration period of MISs.

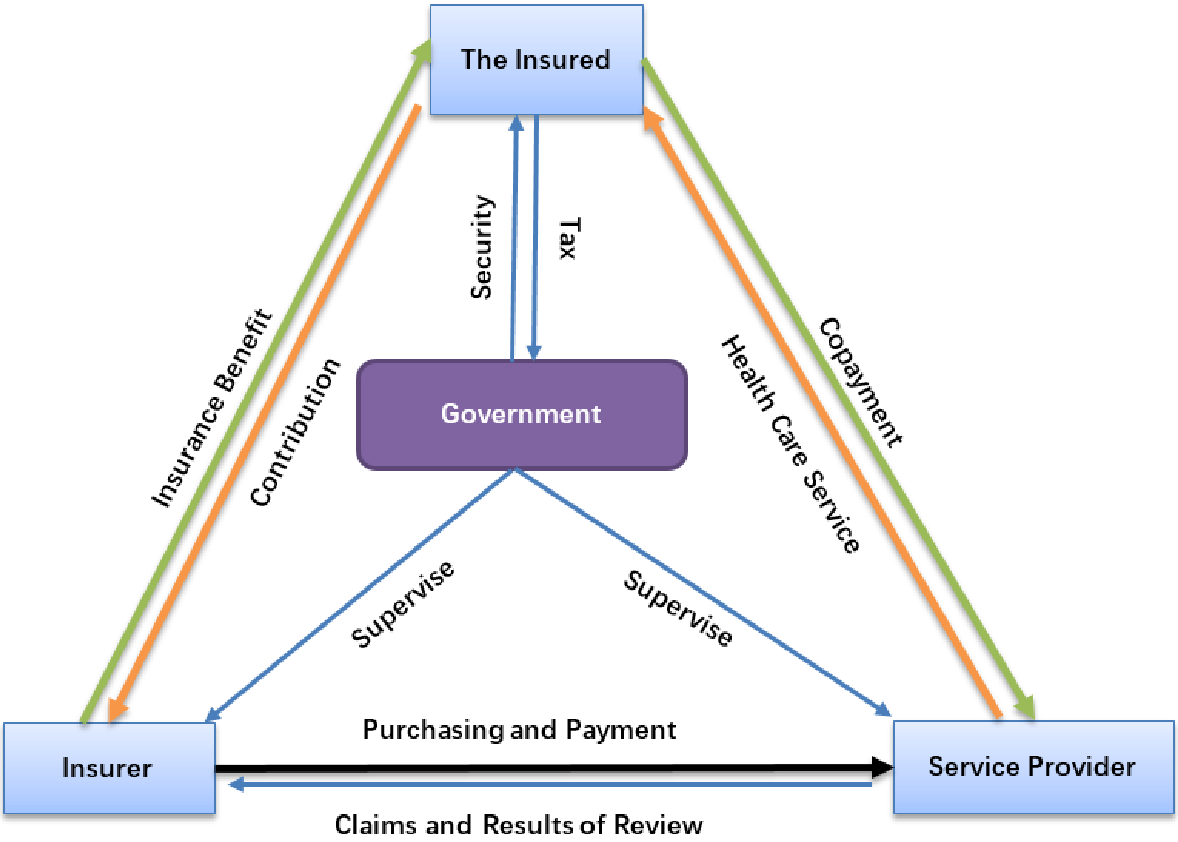
Results

Analysis of Functions Realized by Historical Information Systems

Business Modeling for China's MIISs

China's MIISs support the handling and management of the medical insurance business. The operation of the Chinese MIS is led by the government. Funds can be raised through the taxation of people with fixed employment, such as urban employees. People who have no fixed employment, such as farmers and urban residents, pay cash directly to medical insurance handling institutions. Medical service providers provide medical services to patients, handling institutions provide medical institutions with the settlement of medical insurance funds, and handling institutions manage patient reimbursement and other services. Figure 3 shows the business model, the main elements of China's MIISs, and the interrelationship therein [4]. Medical insurance administrative departments, medical insurance handling institutions, medical service providers, and insured persons play different roles and interact with each other. Medical insurance administrative departments make and regulate medical insurance policies, while medical insurance handling institutions, as the specific implementers of medical insurance policies, provide medical insurance handling and management services for the insured and medical service providers. Affected by the functional adjustment of the national-level administrative department, the administrative level and management authority of China's medical insurance administrative departments as well as medical insurance handling institutions have changed several times, but there has been no significant change in the substance of their respective functions.

Figure 3. Business model chart of China's medical insurance information systems (MIISs), modified from the the National Health Information Database of the National Health Insurance Service in South Korea (Cheol Seong et al, 2017).



Analysis of the Evolution of Information System–Related Policies

Organizational Structure Analysis

Organizational structure includes medical insurance administrative departments, medical insurance handling institutions, medical service providers, and patients, among others. The policy under document No. LSBH[1998]138 shows that social security is administered by two departments—the Ministry of Personnel of the PRC and the Ministry of Labor and Social Security of the PRC—covering enterprises and employees with employment relationships; medical services are provided mainly by medical institutions as well as pharmacies and drugstores subject to designated administration. The policies under document Nos. ZF[2002]13 and GF[2007]20 include rural and urban residents without employment relationships in the coverage for basic medical insurance. The policy under document No. RSTH[2008]284 shows the merger of the Ministry of Personnel of the PRC, the Ministry of Labor of the PRC, and the social security bureaus of China into one ministry, namely, the Ministry of Human Resources and Social Security of the PRC, which is in charge of medical insurance for urban employees and urban residents. The policy under document No. RSTF[2015]56 emphasizes universal coverage of the MIS, aiming for covering all those who should be covered. The policy under document No. EBF[2019]39 shows that the administrative department for basic medical insurance was merged into the NHSA of the PRC, which covers the administration of drug suppliers and manufacturers, who need to participate in bidding for drug procurement.

Business Process Analysis

Since the issuance of the policy under document No. RSXXH[2008]2, as shown in the business model chart of China's MIISs (see Figure 3), the medical insurance business process focuses on payment by insured persons as well as hospitalization and reimbursement at the place of insurance participation. Medical institutions provide medical services and settle accounts with medical insurance handling institutions. Medical insurance handling institutions are mainly responsible for raising funds and clearing up of medical expenses with medical institutions as well as pharmacies and drugstores. The Chinese government formulates policies and supervises all stakeholders. Since the issuance of the policy under document No. RSXXH[2008]2, the Chinese government began to pay attention to transregional transfer and continuation of insurance participation relationships as well as hospitalization and reimbursement; especially since the issuance of the 2016 policy, China has launched a large-scale promotion of cross-provincial hospitalization reimbursement, in order to meet the needs of population mobility and employment all over the country. Since 2018, drug procurement through bidding has been included in the scope of medical insurance administration.

Data Analysis

Analysis by data type covers administrative divisions, enterprise entities, hospitals, pharmacies and drugstores, fundraising, hospitalization behaviors, fund reimbursement, and directory data, such as disease diagnosis, drugs, and medical devices. Analysis by data level shows the following: before 2010, data flowed, were stored, and were utilized mainly at and below the prefecture level; since the issuance of the 2018 policies under document Nos. RSXXH[2008]2 and WBNWF[2008]127, data gradually flow toward the provincial and national levels; the 2016 policies under document Nos. GWJCF[2016]23 and RSTF[2016]185 aim to realize cross-provincial data flow through a national-level platform and support collaborative businesses, such as transregional transfer and continuation of insurance participation relationships and hospitalization reimbursement.

Business Rules Analysis

Business rules have adapted to the medical insurance administration functions and covered the insurance participation and hospitalization reimbursement for urban employees in 1998, for rural residents in 2003, and for urban residents in 2007. Meanwhile, the policies of insurance participation and medical insurance reimbursement were embedded into information systems, in order to standardize the behaviors of medical insurance participation, medical insurance handling, and hospitalization reimbursement. The policy under document No. RSTF[2015]56 emphasizes the use of big data analysis technology to carry out intelligent monitoring of patients' hospitalization behaviors, doctors' diagnosis and treatment behaviors, and handling institutions' handling behaviors. The policy under document No. YBF[2019]22 incorporates drug procurement bidding rules into information system administration; incorporates such payment methods as diagnosis-related groups into the process of patient hospitalization, reimbursement, and fund clearing, step by step; carries out extensive interconnection with such departments as health, taxation, and public security, as well as with such entities as banks and insurance companies; and gradually establishes a nationwide medical insurance credit system.

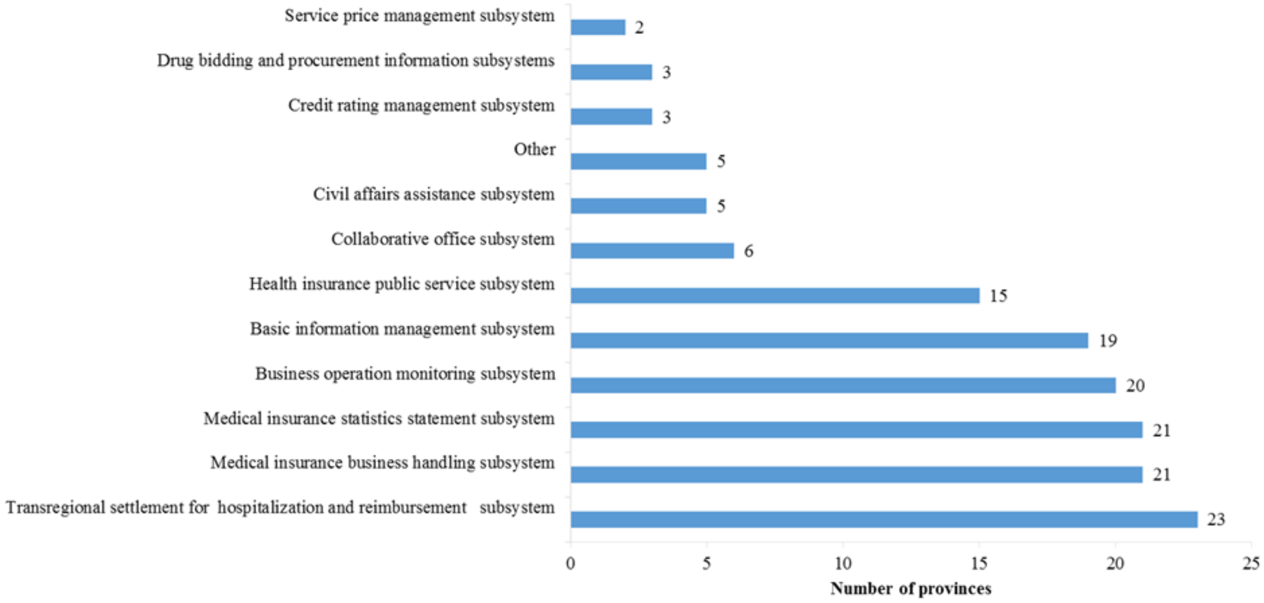
Analysis of Information System Functions

Before the issuance of the 2018 policies under document Nos. RSXXH[2008]2 and WBNWF[2008]127, China's MIISs were mainly responsible for managing basic functions for patients, medical institutions, and medical insurance handling institutions in terms of insurance participation, payment, reimbursement, and fund clearing, as well as managing such standards as disease diagnosis standards and drug lists. Since the issuance of the 2015 policy, China's MIISs gradually strengthened the collection and utilization of data and realized business supervision. The policy under document No. YBF[2019]39 shows that big data gradually plays a role in promoting the fine management of medical insurance and providing evidence-based support for formulation and evaluation of policies.

Main Functions of China's MIISs

The existing MIISs mainly function to describe the status quo before the institutional integration (ie, before 2018). The related functions of the MIISs are scattered throughout multiple ministries and commissions. The information systems related to these functions include the medical insurance handling subsystem, the civil affairs assistance subsystem, and the drug bidding and procurement subsystem, among others. Figure 4 shows the distribution of business function modules used in various provinces in China according to questionnaire feedback. Details of each module are shown below:

Figure 4. Distribution of business function modules used in various provinces in China according to questionnaire feedback.



1. The medical insurance statistics statement subsystem can be used to understand the data regarding insured persons, insured entities, insured rate, fund collection and expenditure, number of hospital visits, average reimbursement amount, and compensation ratio in each region.
2. The medical insurance business handling subsystem functions to manage the insured entities, insured persons, employment relationships, payments by entities and individuals, medical insurance card issuance, personal identity recognition, hospital visits by insured persons, handling of reimbursement and compensation procedures, and compensation for serious illness insurance—if the individual out-of-pocket payment exceeds a certain amount, the serious illness insurance will compensate them again—among others.
3. The civil affairs assistance subsystem functions to manage the groups who receive civil affairs assistance, the distribution of medical assistance funds, and the compensation for medical services—compensation will be provided again on the basis of the basic medical insurance compensation—among others.
4. The drug bidding and procurement information subsystems are established by the responsible administrative departments: development and reform commission, health commission, health care security bureau, etc. The responsible provincial administrative departments, on behalf of hospitals, negotiate with drug manufacturers and suppliers and complete the tender process through this information system. Some provincial administrative departments directly pay funds to drug manufacturers and suppliers, while funds for drug supply are paid by hospitals to drug manufacturers and suppliers in some other provinces.
5. The transregional hospitalization and settlement subsystem supports transregional or transprovincial reimbursement of insured persons and covers the following functions: hierarchical referral, identify recognition and identification of insured status, discharge settlement, and window-based reimbursement, as well as fund clearing among transprovincial handling institutions.
6. The service price management subsystem functions to monitor the sales prices of medical institutions and pharmacies and drugstores, as well as to analyze their changing trends, so as to provide a reference basis for price formulation and payment standards of medical services, drugs, and medical devices.
7. The credit rating management subsystem functions to manage credit rating for the insured entities, insured persons, medical institutions, pharmaceutical production and circulation enterprises, and medical workers, and to establish a credit management system for medical services and medical insurance handling.
8. The public service subsystem for medical insurance caters to insured persons; provides self-services, such as insurance participation, payment, and referral filing; supports mobile online payment; and allows the inquiry of personal historical behaviors, such as insurance participation and hospital visits, so as to assist health management.
9. The basic information management subsystem functions to manage the qualification information for designated medical institutions and designated pharmacies and drugstores so they may join the MIS. It also functions to manage the information of medical workers and maintain the dictionary codes for disease diagnosis, diagnosis and treatment services, drugs, and consumables.
10. The business operation monitoring subsystem functions to monitor the compliance of medical services and the balance of revenue and expenditure of medical insurance funds, among others, by collecting data on insurance participation, medical treatment, treatment behavior, reimbursement and compensation, among others. It also functions to make forecasts on partial trends of medical insurance participation, fund collection, and expenditure.
11. The collaborative office subsystem integrates all office automation work within a jurisdiction into the information system for unified management, such as routine management of mail delivery as well as drafting, approval, and receipt of official documents, among others.

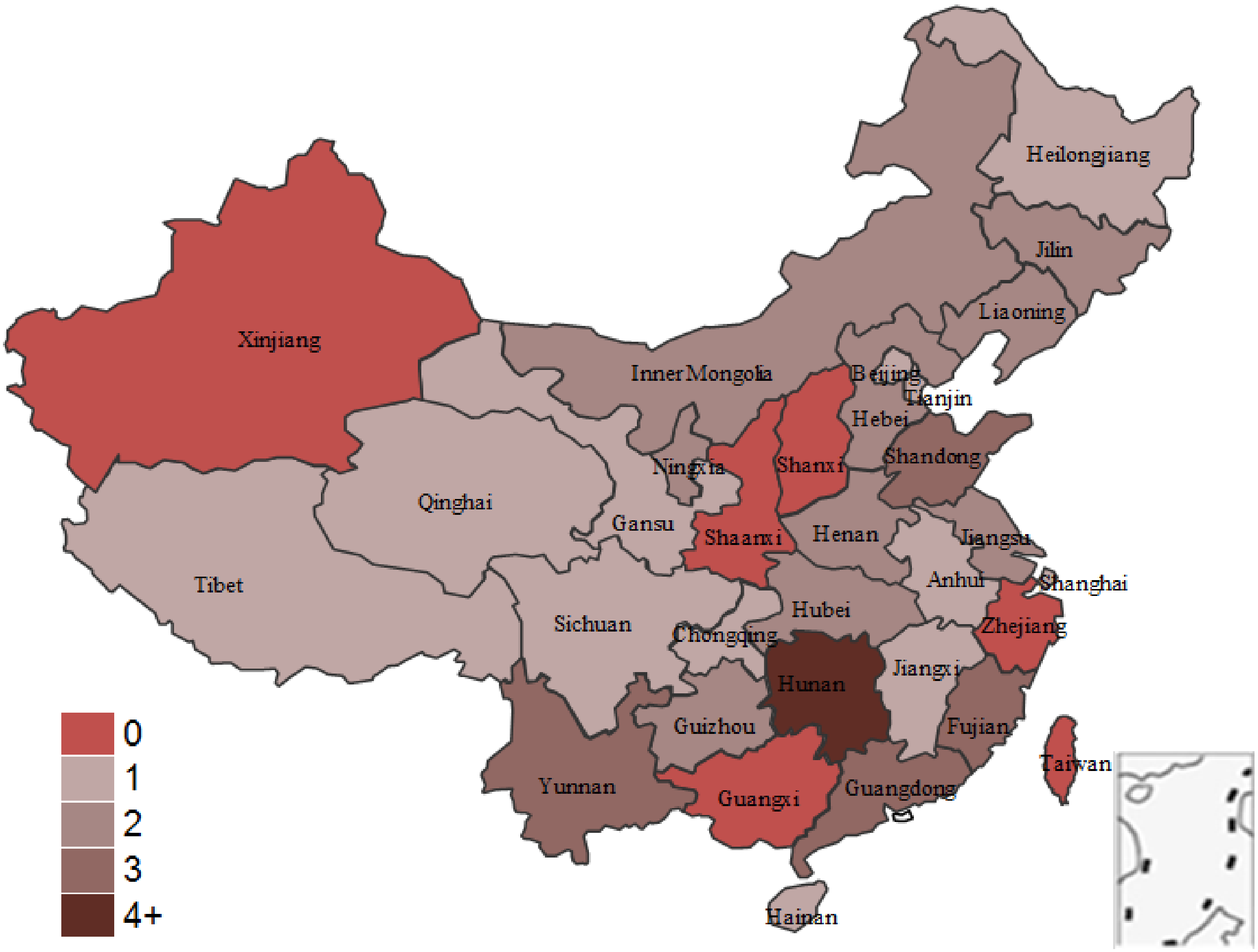
The 11 business function modules above, just like components of a jigsaw puzzle, build up the framework of MIISs in each province. Some functions, such as electronic medical record management, budget management, financial management, and account book management, have also been mentioned in some regions.

Analysis of Infrastructure and Identity Recognition Media

Networking Mode

Four main networking modes include the following: fiber optic private network (FOPN), e-government extranet (EGE), virtual private network (VPN), and the internet. FOPN features high-security performance, good transmission performance, and high cost. EGE connects medical insurance handling institutions at all levels and features lower cost and better security performance, but it is only connected to government agencies and is not yet connected to hospitals. VPN establishes a virtual safe channel based on the internet and features low cost and better security performance. The internet is characterized by good transmission performance and low cost, but its security performance is low. Figure 5 shows the distribution of the four networking modes above.

Figure 5. Analysis of networking modes adopted by medical insurance information systems (MIISs) in each province in China.



A total of 26 provinces of China, including provinces, autonomous regions, and municipalities directly under the central government, have responded with their networking modes; only Hunan Province has used all four modes. Four provinces—Fujian, Shandong, Guangdong, and Yunnan—use three networking modes. Nine provinces at all levels—Hebei, Inner Mongolia, Liaoning, Jilin, Jiangsu, Henan, Hubei, Guizhou, and Ningxia—use two networking modes. A total of 12 provinces—Beijing, Tianjin, Heilongjiang, Shanghai, Anhui, Jiangxi, Hainan, Chongqing, Sichuan, Tibet, Gansu, and Qinghai—use one networking mode.

Regarding the main networking modes used, 23 out of 31 (74%) provinces use FOPN; Anhui, Tibet, and Hainan are the only provinces that do not use FOPN for networking. Six provinces—Fujian, Henan, Hunan, Guangdong, Yunnan, and Ningxia—also use EGE for networking.

Identity Recognition Media

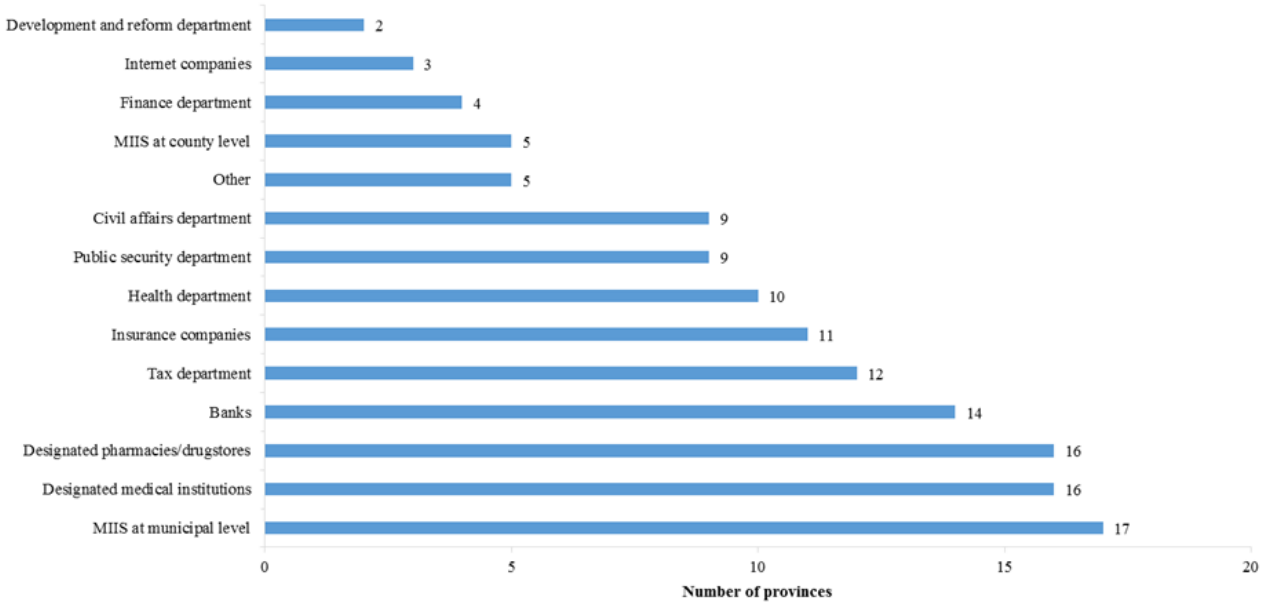
There are mainly two kinds of identity recognition media: ID cards and social security cards. A total of 20 provinces out of 31 (65%) support both ID cards and social security cards for identity recognition, while some provinces, such as Inner Mongolia and Chongqing, support only one kind of media for identity recognition.

Analysis of Data Storage and Data Utilization

Interconnectivity of Information Systems

A provincial MIIS vertically connects municipal MIISs and regional pharmacy information systems; it exchanges referral and transregional hospitalization settlement data with municipal MIISs and exchanges drug purchase information via personal accounts with designated pharmacies and drugstores. A provincial MIIS horizontally connects the information systems of such departments and entities as tax, finance, development and reform, health, civil affairs, public security, banks, insurance companies, and internet companies. This MIIS exchanges information with different departments and entities as follows: (1) it exchanges information on collection and payment of insured expenses with the tax department, (2) it exchanges electronic medical record information with the health department, (3) it exchanges personal identity and credit information with the public security department, (4) it exchanges medical insurance fund transfer information with banks, (5) it exchanges poverty alleviation and social assistance information with the civil affairs department, (6) it exchanges serious illness insurance information with insurance companies, and (7) it exchanges online payment information with internet companies, among others. MIISs in some provinces also exchange information with departments of audit, industry, and commerce, among others. Figure 6 shows the connection between provincial MIISs and peripheral information systems.

Figure 6. The interconnection between the medical insurance information systems (MIISs) and the surrounding information systems in each province.



In 55% (17/31) of the regions in China, MIISs at provincial and municipal levels were networked and have gathered basic medical insurance data. Provincial MIISs in 11 provinces out of 31 (35%) exchange data with insurance companies; the main data item exchanged includes the secondary compensation information for serious illness insurance following basic medical insurance compensation. Provincial MIISs in 12 provinces out of 31 (39%) exchange data with tax departments. In adaption to the adjustment to their functions in the medical insurance reform, the tax departments are responsible for the collection and payment of insurance premiums.

Data Storage Types at Provincial Information Centers

Feedback from 24 provinces in China shows that some types of medical insurance data are stored in these provinces by establishing provincial-level medical insurance data centers; such data centers are under construction in two provinces: Shandong and Henan. Seven provinces did not fill out the questionnaire. Many data storage types included insurance participation, payment, settlement, and expenses. Among these types, information about insured individuals and insured entities accounted for 92% (22/24), the highest share; information related to payment, settlement compensation, details of outpatient and inpatient expenses, and cross-provincial hospitalization accounted for over 83% (20/24), the second-highest share. No provincial-level medical insurance data center has stored the procurement, sales, and inventory data from designated pharmacies and drugstores. The storage is also very limited for procurement, sales, and inventory data from hospitals (1/24, 4%) and data from bidding and procurement of drugs (3/24, 13%). Efforts should be made to further expand the scope of data exchange and sharing, so as to match the functions of MIISs in terms of bidding and procurement of drugs as well as use of drugs and consumables. Feedback from all provinces that were surveyed shows that Heilongjiang Province has stored the highest number of data types (15/17, 88%); this province has covered all data types listed in Multimedia Appendix 1 except the procurement, sales, and inventory data from hospitals as well as pharmacies and drugstores. A total of 76% or more of relevant information has been stored in such provinces and municipalities as Tianjin (13/17, 76%), Shanghai (13/17, 76%), Guangdong (14/17, 82%), Hainan (13/17, 76%), and Guizhou (13/17, 76%). According to the feedback from the questionnaires, the provincial-level medical insurance data center in Hubei Province has collected the payment information on medical insurance participation of urban and rural residents in the whole province, as well as information on medical insurance treatment for urban and rural residents in individual prefectures and cities, but it did not fill out the information about other types of data storage. Henan Province did not fill out information about other data storage types except for data related to insurance participation and payment. As a result, these two provinces—Hubei and Henan—have a relatively lower percentage of data storage types. In those provinces with high economic development levels, solid informatization foundations, and lower pooling regions, their provincial-level medical insurance data centers have stored relatively more types of data; however, as a whole, the degrees of concentricity and comprehensiveness of medical insurance data are still low. Multimedia Appendix 1 shows the corresponding matrix of provincial medical insurance data centers and data storage types.

Analysis of Data Utilization as well as Security and Privacy Protection Policies

China's medical insurance data are mainly used for business handling; however, these data have not been fully exploited. Through the Medicaid Statistical Information System database [33], the United States supports the formulation of Medicaid fundraising and compensation policies. France built its national database EGB (Echantillon Generaliste des Beneficiaires) [34], conducts research on the effects of drugs, and conducts research and development on new drugs through claims data [35]. South Korea built its National Health Insurance Research Database, which functions to make secondary compensation for catastrophic expenditures; it also provides specific datasets to researchers for scientific research, especially for public health governance and infectious disease monitoring through data correlation [36]. The medical insurance database of Taiwan is used for research on cancer treatment [37].

In terms of policies and regulations related to medical insurance information security and privacy protection, the HIPAA was passed by the US Congress in 1996; its applicability was adjusted continuously, especially in the era of big data [38]. In 2016, the European Union issued the General Data Protection Regulation, which stipulates in detail the collection, transmission, processing, and utilization of medical security or health information [39]. China has not yet issued a special law or regulation on security and privacy protection of health information or medical insurance information; some of the existing Chinese policies are scattered among the Social Insurance Law of the PRC, the Law of the PRC on Basic Healthcare and Health Promotion, the Law of the PRC on the Prevention and Treatment of Infectious Diseases, and the Provisions of the PRC on the Disclosure of Government Information. However, the length of such content is extremely limited; there are provisions of a framework nature, but there are no acts or laws with operability.

Urgent Needs in Various Regions During the Period of Integration and Transition of MISs

The information system building model is guided by national top-level planning; the deployment level of information systems should be higher than the unified pooling level of medical insurance funds. In addition, attention should be paid to data coding and standardization and building efforts should be made on the basis of the original information system building, in order to save unnecessary investment.

In terms of functions of information systems, efforts should be made to strengthen the informatization support for the reform of payment mode, in order to adapt it to the promotion of prepayment modes such as diagnosis-related groups. In the era of big data, work should be done to establish an intelligent medical insurance audit system, which will function to monitor behaviors (eg, medical treatment and fund compensation) through algorithms, such as data machine learning, and will identify such events as medical insurance fraud and abuse. Meanwhile, work should also be done to establish a credit database containing information on lawbreakers to be able to take certain punishment measures against such persons.

In the context of the universal application of mobile phones, efforts should be made to provide more self-services for the public, such as online hospitalization appointment, mobile payment, self-service insurance participation and payment, policy access, and information inquiry.

Discussion

Building Model of China's Basic MIISs in the Future

In the future, we should overcome the disadvantage of the lack of overall planning for the building of China's MIISs in the first and second phases and arrange for the building of MIISs and hospital information systems at the national, provincial, and municipal levels on a unified basis. We should establish a deployment mode higher than the fund pooling level that is at least not lower than the municipal deployment level. In addition, we should try our best to realize provincial deployment and to deploy prefectural and municipal information systems by utilizing the cloud computing model. There are more than four networking modes, which need further planning to establish a safer and faster private network mode. In terms of identity recognition media, this should expand to identity recognition based on electronic ID cards; moreover, by adopting the QR (Quick Response) code, we can use mobile phones to recognize identities. Information system building standards should be formulated and issued first, covering the planning of medical insurance business, specification of information system modules, and standard codes for data exchange, among others.

Information Exchange and Data Utilization

Information exchange involves the integration of internal information systems in the MIS, the collection of data at different levels, and the exchange of data within fields or between different fields. The MIS internal information systems receiving close attention exceed the 11 subsystems listed above in the *Main Functions of China's MIISs* section. These systems are designed and developed by many developers; as a result, they have diversified data structures and codes. First, internal integration of these information systems shall avoid repeated investment. Second, we should carry out the mode of deployment of MIISs at the national, provincial, prefecture, and municipal levels; unify the standards for data exchange; and realize bottom-up collection and gathering of data. Third, we should realize data exchange with the information systems in other fields, clarify the operation specifications for business links, and realize business collaboration through data exchange; for instance, exchanging data with the tax department to confirm the qualification of patients to participate in insurance and medical insurance payment, and exchanging medical record data with the health department. Through such practice as standardization of health information exchange data [40] and Federal Enterprise Architecture Framework business [41], we will realize the business interoperation of different systems, such as fundraising and payment to the tax department, reimbursement for inpatients and outpatients at health departments, designated institution certification at industry and commerce departments, and claim settlement at commercial insurance companies.

Balancing the Stake Between Informed Consent for Privacy Protection and Data Mining and Utilization in a Context of Big Data

MIISs cover a massive amount of heterogeneous data, which have the typical 4-V characteristics of big data: volume, variety, velocity, and veracity. For example, residents' participation in medical insurance and hospital visits involve detailed identity information of individuals, health information, economic status, invoice images, and other relevant data. China needs to formulate special-purpose laws for the security and privacy protection of medical insurance information. China also needs to clarify the connotation of medical security information, the rights and interests of the public, the scope of security and privacy protection, the operation requirements for the right to be informed and information disclosure, the contents of exceptional protection of safe harbor, infringements, and punishments, among others, by referring to the HIPAA of the United States and the Personal Information Protection and Electronic Documents Act (PIPEDA) of Canada.

We should balance the stake between *security and privacy protection* and *data analysis and utilization*. New models for assisting in disease diagnosis and treatment have been identified from the big data of health records through utilization of artificial intelligence technology. These new models have been widely used in medical innovation, which involves patients' health histories, treatment methods, and treatment results, and are even associated with such information as genetics; in particular, the association with multisource data makes privacy protection more difficult. We should ensure users' rights of informed consent and enable patients to feel comfortable in providing data for scientific research without degrading safety protection measures. This represents a direction of collection and utilization of medical insurance information, rather than transitional privacy protection [42].

Lessons From COVID-19 for Building of MIISs

China has recently developed hospital information systems. China's MIISs have established a mechanism of data exchange and sharing with each hospital, in order to meet patients' health needs and facilitate settlements and expense compensations. However, China's regional health information platforms of health departments and authorities are relatively isolated, and a normalized mechanism of data exchange with hospitals has not yet been established. Although China has established the world's most extensive surveillance system for infectious diseases, this system is mainly based on a bottom-up reporting approach by manual entry and form filling; as a result, China's system fails to exchange data with hospital information systems in real time [43]. Reporting time is delayed; the system needs a process that begins with identifying a suspected case of an infectious disease and leads to case confirmation, thereof. The source of data generation has no authority to publish information, which has resulted in delays in reporting cases of coronavirus disease 2019 (COVID-19) after case identification as well as delays in information publication to the public after reporting to the central government. During the whole process, the early warning mechanism of the Infectious Diseases Information Network failed to work effectively, which resulted in serious decreases in disease prevention and early warning. This suggests the following: China's new MIISs should be closely combined with its medical and health information systems; efforts should be made to exert the roles of numerous medical institutions as parts of a network foundation, in order to gather data from hospitals to be transferred to the national-level data center in a timely manner; and work should be done to establish a computer-based early warning model, in order to detect the sudden states and development trends of infectious diseases and public health events in various regions [44]. Through information connection between MIISs and health systems, we will be able to capture information about patients' hospitalizations at medical institutions in a timelier manner and more efficiently. Then, through dynamic analysis and summarized reports of data regarding disease types and expenses, etc, we will be able to identify risk factors in a prospective manner, so as to maintain the safety of medical insurance funds.

Conclusions

China's MIISs are the most extensive information systems that could allow network foundations to connect medical institutions. Over the past 20 years, after the three phases of development, China's MIISs have played an important role in medical insurance business management and reimbursement, and have provided strong support for the operation of the world's largest medical security system. Particularly in terms of settlements for transregional hospitalization and reimbursements, China's MIISs have enabled extensive data exchange among the central government, provinces, prefectures, municipalities, cities, and medical institutions, and have realized transprovincial business collaboration. In many developing countries, information system building is an indispensable element to realize universal health coverage and to continuously improve their respective medical security systems. The analysis on the functions, advantages, and disadvantages of China's MIISs at different phases has a sound significance of reference. Currently, China's MIISs are in a period of transformation and transition. In terms of the top-level design and planning of China's national medical insurance informatization, as well as the redeployment and reimplementation of information systems, it is necessary to further consider such focal issues as normalization of business, standardization of data, and interoperation of information systems. In 2019, the outbreak of COVID-19 revealed a poor interoperability between the MIISs and the health information systems. Due to privacy protection and other reasons, data sharing with the public health information network was insufficient, and big data technology was not fully utilized to analyze medical insurance data and provide early warning services for public health. In the future, more detailed laws, regulations, and policies should clearly set forth the contents and ways of exchanging and sharing medical insurance data. The implementation of security and privacy protection policies of MIISs will further improve the degree of trust from individuals, medical service providers, and public health institutions in the information systems.

Conflicts of Interest

None declared.

Multimedia Appendix 1

List of data storage types used by the medical insurance data centers from each province in China.

[DOCX File , 29 KB - medinform_v8i9e18780_app1.docx]

References

1. The National Health Commission of the People's Republic of China. China Health Statistics Yearbook 2019. Beijing, China: Peking Union Medical College Press; 2019.
2. Tuppin P, de Roquefeuil L, Weill A, Ricordeau P, Merlière Y. French national health insurance information system and the permanent beneficiaries sample. Rev Epidemiol Sante Publique 2010 Aug;58(4):286-290. [doi: 10.1016/j.respe.2010.04.005] [Medline: 20598822]
3. Ng-Mak D, Ruetsch C. Association between meaningful use of electronic health records and patient health outcomes in schizophrenia: A retrospective database analysis. Am J Manag Care 2019 Jul;25(9 Suppl):S159-S165 [FREE Full text] [Medline: 31318518]
4. Cheol Seong S, Kim Y, Khang Y, Heon Park J, Kang H, Lee H, et al. Data resource profile: The National Health Information Database of the National Health Insurance Service in South Korea. Int J Epidemiol 2017 Jun 01;46(3):799-800 [FREE Full text] [doi: 10.1093/ije/dyw253] [Medline: 27794523]
5. Meng Q, Xu L, Zhang Y, Qian J, Cai M, Xin Y, et al. Trends in access to health services and financial protection in China between 2003 and 2011: A cross-sectional study. Lancet 2012 Mar;379(9818):805-814. [doi: 10.1016/s0140-6736(12)60278-5]
6. Yip W, Fu H, Chen AT, Zhai T, Jian W, Xu R, et al. 10 years of health-care reform in China: Progress and gaps in universal health coverage. Lancet 2019 Sep;394(10204):1192-1204. [doi: 10.1016/s0140-6736(19)32136-1]
7. Xinhua News Agency. Institutional Reform Plan of the State Council. The State Council of the People's Republic of China. 2018 Mar 17. URL: http://www.gov.cn/guowuyuan/2018-03/17/content_5275116.htm [accessed 2020-02-15]
8. Lu L, Guan H, Jiang H. Current situation of and suggestions for China's medical insurance management information systems [article in Chinese]. Soft Sci Health 2000;14(6):269-272 [FREE Full text]
9. Huang Z. A study on fragmentation issues in medical insurance information systems [article in Chinese]. China Health Insur 2017(5):34-37 [FREE Full text]
10. Ling C, Zhai T, Wang R, Du X, Zhang X. An analysis of the integration of urban and rural medical insurance policies: Based on the analysis perspective of policy tools. Chin Health Econ 2018;37(12):12-17.
11. Li Y, Yu C, Wu C, Wang H, Ling D. A study on the technical route of information system integration in the integration of new rural cooperative medical system and basic medical insurance system for urban residents. Chin Health Econ 2017;1:34-36 [FREE Full text]
12. Huang H. Progress in and thinking on the intelligent monitoring of medical insurance. China Health Insur 2015;12:30-32.
13. Jian W. A study on and application of data integration and data mining technologies in medical insurance information system. Mod Comput (Professional Edition) 2010;11:47-51.
14. Wang S, Liu B. Thoughts on settlement and management of transregional hospitalization for basic medical insurance. Chin Health Resour 2018;21(4):346-350 [FREE Full text] [doi: 10.13688/j.cnki.chr.2018.18067]
15. Shi C, Yang H. Design of a medical insurance decision support system based on a multilevel model. Comput Eng Des 2009;30(5):1252-1254.
16. The National People's Congress of the People's Republic of China. Decision of the State Council of the PRC on Establishing a Basic Medical Insurance System for Urban Employees. The Central People's Government of the People's Republic of China. 2005 Aug 04. URL: http://www.gov.cn/banshi/2005-08/04/content_20256.htm [accessed 2020-02-15]
17. The General Office of the Ministry of Labor and Social Security of the People's Republic of China. Notice on Printing and Distributing the Implementation Outline of Urban Basic Endowment Insurance Management Information System Building (1999-2001) (Document No. LSTH[1999]67). The Ministry of Human Resources and Social Security of the People's Republic of China. 1999. URL: http://www.mohrss.gov.cn/SYrlzyhshbzb/zhuanti/jinbaogongcheng/jbgcshehuibaozhang/jbgcshbzzcwj/200602/t20060228_47488.htm [accessed 2020-02-15]
18. The General Office of the Ministry of Labor and Social Security of the People's Republic of China. Notice on Issuing Guidelines on the Building of the Management Information System for Basic Medical Insurance for Urban Employees (Document No. [2000]30). The Ministry of Human Resources and Social Security of the People's Republic of China. 2005 Dec 14. URL: http://www.mohrss.gov.cn/SYrlzyhshbzb/zhuanti/jinbaogongcheng/jbgczhengcewenjian/200512/t20051214_90313.html [accessed 2020-02-15]
19. The Ministry of Labor and Social Security of the People's Republic of China. Notice on Printing and Distributing the Opinions on Comprehensive Implementation of the Jinbao Program for Unified Building of a Labor Security Information System (Document No. LSBH[2003]174). The Ministry of Human Resources and Social Security of the People's Republic of China. 2003 Dec 23. URL: http://www.mohrss.gov.cn/SYrlzyhshbzb/zhuanti/jinbaogongcheng/jbgczhengcewenjian/200811/t20081126_90341.htm [accessed 2020-02-15]
20. The General Office of the State Council of the People's Republic of China. Notice of the General Office of the State Council to Transmit the Opinions of the Ministry of Health and Other Departments on Establishing the New Rural Cooperative Medical System (Document No. GBF[2003]3). The Central People's Government of the People's Republic of China. 2003. URL: http://www.gov.cn/zwgk/2005-08/12/content_21850.htm [accessed 2020-02-15]

21. The General Office of the Ministry of Health of the People's Republic of China. Notice of the General Office of the Ministry of Health on Printing and Distributing the Basic Specifications of the New Rural Cooperative Medical Information System. The National Health Commission of the People's Republic of China. 2005. URL: http://www.nhc.gov.cn/jws/s3581sg/200804/eaee82fb62df40cb849b70968239b38e.shtml [accessed 2020-02-15]
22. The General Office of the State Council of the People's Republic of China. Guidelines of the State Council of the PRC on Launching a Pilot Program of Basic Medical Insurance for Urban Residents (Document No. GF [2007]20). The Central People's Government of the People's Republic of China. 2007. URL: http://www.gov.cn/zwgk/2007-07/24/content_695118.htm [accessed 2020-02-15]
23. The Ministry of Human Resources and Social Security of the People's Republic of China. Notice on Carrying out Unified Implementation of Some Application Software for the Jinbao Program (Document No. RSTH[2008]284). The Ministry of Human Resources and Social Security of the People's Republic of China. 2008. URL: http://www.mohrss.gov.cn/SYrlzyhshbzb/zhuanti/jinbaogongcheng/jbgczhengcewenjian/200811/t20081126_52104.html [accessed 2020-02-15]
24. The Ministry of Health of the People's Republic of China. Basic Specification of New Rural Cooperative Medical Management Information System (2008 Revision) (Document No. WBNWF (2008) 127). The Ministry of Health of the People's Republic of China. 2008. URL: http://www.doc88.com/p-9465205319558.html [accessed 2020-02-15]
25. The Ministry of Human Resources and Social Security of the People's Republic of China. Notice on Carrying out Unified Upgrading Implementation of Some Application Software for the Jinbao Program (Document No. RSXXH[2010]20). The Ministry of Human Resources and Social Security of the People's Republic of China. 2010. URL: http://www.mohrss.gov.cn/SYrlzyhshbzb/zhuanti/jinbaogongcheng/jbgczhengcewenjian/201112/t20111221_47507.html [accessed 2020-02-15]
26. The General Office of the National Health and Family Planning Commission of the People's Republic of China. Notice on Printing and Distributing the Scheme for Connectivity Technology of the National-Level Information Platform for the New Rural Cooperative Medical System (trial) (Document No. WBNWH[2013]456). Anhui Province. 2013. URL: http://www.ahhzyl.com/NewsInfo.aspx?gID=2c6eea68-9f49-4f86-9c8c-b34aad3a4231 [accessed 2020-02-15]
27. The General Office of the Ministry of Human Resources and Social Security of the People's Republic of China. Notice of General Office of the Ministry of Human Resources and Social Security on Comprehensively Promoting the Intelligent Monitoring of Medical Services for Basic Medical Insurance (Document No. RSTH [2015]56). The Ministry of Human Resources and Social Security of the People's Republic of China. 2015. URL: http://www.mohrss.gov.cn/SYrlzyhshbzb/shehuibaozhang/zcwj/yiliao/201506/t20150630_213006.html [accessed 2020-02-15]
28. The General Office of the Ministry of Human Resources and Social Security of the People's Republic of China. Notice of the General Office of the Ministry of Human Resources and Social Security on the Building of the National Insurance Registration Information System (Document No. RSTF [2015]86). The Ministry of Human Resources and Social Security of the People's Republic of China. 2015. URL: http://www.mohrss.gov.cn/SYrlzyhshbzb/zhuanti/jinbaogongcheng/jbgczhengcewenjian/201506/t20150630_212985.htm [accessed 2020-02-15]
29. The National Health and Family Planning Commission of the People's Republic of China. Implementation Plan for Networked Settlement and Reimbursement for Trans-Regional Hospitalization in the National New Rural Cooperative Medical System (Document No. GWJCF[2016]23). The National Health Commission of the People's Republic of China. 2016 May 26. URL: http://www.nhfpc.gov.cn/jws/s3581sg/201606/e28c759f9a28451fa6ae26d5782a177b.shtml [accessed 2020-02-15]
30. The General Office of the Ministry of Human Resources and Social Security of the People's Republic of China. Notice of the General Office of the Ministry of Human Resources and Social Security of the PRC on Accelerating the Building of Cross-provincial Trans-regional Medical Service Settlement System (Document No. RETF[2016]185). The Ministry of Human Resources and Social Security of the People's Republic of China. 2016 Dec 16. URL: http://www.mohrss.gov.cn/SYrlzyhshbzb/shehuibaozhang/zcwj/201612/t20161222_262613.html [accessed 2020-02-15]
31. Xinhua News Agency. Institutional Reform Plan of the State Council. The Central People's Government of the People's Republic of China. 2018 Mar 17. URL: http://www.gov.cn/xinwen/2018-03/17/content_5275116.htm [accessed 2020-02-15]
32. National Healthcare Security Administration. Response of the National Medical Security Administration to Recommendation No. 8396 of the Second Session of the 13th National People's Congress (Document No. YBH[2019]20). The Central People's Government of the People's Republic of China. 2019 Jul 25. URL: http://www.nhsa.gov.cn/art/2019/7/25/art_26_1569.html [accessed 2020-02-15]
33. MacTaggart P, Foster A, Markus A. Medicaid Statistical Information System (MSIS): A data source for quality reporting for Medicaid and the Children's Health Insurance Program (CHIP). Perspect Health Inf Manag 2011 Apr 01;8(Spring):1d [FREE Full text] [Medline: 21464861]
34. Lin L, Warren-Gash C, Smeeth L, Chen P. Data resource profile: The National Health Insurance Research Database (NHIRD). Epidemiol Health 2018;40:e2018062 [FREE Full text] [doi: 10.4178/epih.e2018062] [Medline: 30727703]
35. Moulis G, Lapeyre-Mestre M, Palmaro A, Pugnet G, Montastruc J, Sailler L. French health insurance databases: What interest for medical research? Rev Med Interne 2015 Jun;36(6):411-417 [FREE Full text] [doi: 10.1016/j.revmed.2014.11.009] [Medline: 25547954]
36. Kim TJ, Lee JS, Kim JW, Oh MS, Mo H, Lee CH, et al. Building linked big data for stroke in Korea: Linkage of stroke registry and national health insurance claims data. J Korean Med Sci 2018 Dec 31;33(53):e343 [FREE Full text] [doi: 10.3346/jkms.2018.33.e343] [Medline: 30595684]

37. Chiang J, Lin C, Wang C, Koo M, Kao Y. Cancer studies based on secondary data analysis of the Taiwan s National Health Insurance Research Database. Medicine 2017;96(17):e6704. [doi: 10.1097/md.0000000000006704]
38. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019 Jan;25(1):37-43 [FREE Full text] [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]
39. Tovino S. The HIPAA Privacy Rule and the EU GDPR: Illustrative comparisons. Seton Hall Law Rev 2017;47(4):973-993 [FREE Full text] [doi: 10.1201/b17548-8]
40. Mihoko O, Haku I, Sunao W, Yoshimune S. Framework of performance measures for health information exchange (HIE). Stud Health Technol Inform 2017;245:1103-1107. [Medline: 29295273]
41. Federal enterprise architecture framework. Centers for Medicare & Medicaid Services. 2016 Nov 17. URL: https://www.cms.gov/Research-Statistics-Data-and-Systems/CMS-Information-Technology/EnterpriseArchitecture/FEAF [accessed 2020-02-15]
42. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019 Jan;25(1):37-43 [FREE Full text] [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]
43. Vlieg WL, Fanoy EB, van Asten L, Liu X, Yang J, Pilot E, et al. Comparing national infectious disease surveillance systems: China and the Netherlands. BMC Public Health 2017 May 08;17(1):415 [FREE Full text] [doi: 10.1186/s12889-017-4319-3] [Medline: 28482830]
44. Bagherian H, Farahbakhsh M, Rabiei R, Moghaddasi H, Asadi F. National communicable disease surveillance system: A review on information and organizational structures in developed countries. Acta Inform Med 2017 Dec;25(4):271-276 [FREE Full text] [doi: 10.5455/aim.2017.25.271-276] [Medline: 29284920]

Abbreviations

COVID-19: coronavirus disease 2019
EGB: Echantillon Generaliste des Beneficiaires
EGE: e-government extranet
FOPN: fiber optic private network
HIPAA: Health Insurance Portability and Accountability Act
MIIS: medical insurance information system
MIS: medical insurance system
NDRC: National Development and Reform Commission
NHSA: National Healthcare Security Administration
NRCMS: New Rural Cooperative Medical System
PIPEDA: Personal Information Protection and Electronic Documents Act
PRC: People's Republic of China
QR: Quick Response
VPN: virtual private network

Edited by G Eysenbach; submitted 18.03.20; peer-reviewed by X Liu, H Qi; comments to author 14.04.20; revised version received 06.05.20; accepted 03.06.20; published 01.09.20.

Please cite as:
Li Y, Lu C, Liu Y
Medical Insurance Information Systems in China: Mixed Methods Study
JMIR Med Inform 2020;8(9):e18780
URL: https://medinform.jmir.org/2020/9/e18780
doi:10.2196/18780
PMID:32673209

Original Paper

Breast Self-Examination System Using Multifaceted Trustworthiness: Observational Study

Rajes Khana[1*], MSc; Manmeet Mahinderjit Singh[1*], PhD; Faten Damanhoori[1*], MSc; Norlia Mustaffa[1*], MSc

[1] Universiti Sains Malaysia, Penang, Malaysia

[*] all authors contributed equally

Corresponding Author:
Manmeet Mahinderjit Singh, PhD
Universiti Sains Malaysia
Jalan Sungai Dua, Gelugor
Penang, 11800
Malaysia
Phone: 60 46533888
Email: manmeet@usm.my

Abstract

Background: Breast cancer is the leading cause of mortality among women worldwide. However, female patients often feel reluctant and embarrassed about meeting physicians in person to discuss their intimate body parts, and prefer to use social media for such interactions. Indeed, the number of patients and physicians interacting and seeking information related to breast cancer on social media has been growing. However, a physician may behave inappropriately on social media by sharing a patient's personal medical data excessively with colleagues or the public. Such an act would reduce the physician's trustworthiness from the patient's perspective. The multifaceted trust model is currently most commonly used for investigating social media interactions, which facilitates its enhanced adoption in the context of breast self-examination. The characteristics of the multifaceted trust model go beyond being personalized, context-dependent, and transitive. This model is more user-centric, which allows any user to evaluate the interaction process. Thus, in this study, we explored and evaluated use of the multifaceted trust model for breast self-examination as a more suitable trust model for patient-physician social media interactions in breast cancer screening.

Objective: The objectives of this study were: (1) to identify the trustworthiness indicators that are suitable for a breast self-examination system, (2) design and propose a breast self-examination system, and (3) evaluate the multifaceted trustworthiness interaction between patients and physicians.

Methods: We used a qualitative study design based on open-ended interviews with 32 participants (16 outpatients and 16 physicians). The interview started with an introduction to the research objective and an explanation of the steps on how to use the proposed breast self-examination system. The breast self-examination system was then evaluated by asking the patient to rate their trustworthiness with the physician after the consultation. The evaluation was also based on monitoring the activity in the chat room (interactions between physicians and patients) during daily meetings, weekly meetings, and the articles posted by the physician in the forum.

Results: Based on the interview sessions with 16 physicians and 16 patients on using the breast self-examination system, honesty had a strong positive correlation (r=0.91) with trustworthiness, followed by credibility (r=0.85), confidence (r=0.79), and faith (r=0.79). In addition, belief (r=0.75), competency (r=0.73), and reliability (r=0.73) were strongly correlated with trustworthiness, with the lowest correlation found for reputation (r=0.72). The correlation among trustworthiness indicators was significant (*P*<.001). Moreover, the trust level of a patient for a particular physician was found to increase after several interactions.

Conclusions: Multifaceted trustworthiness has a significant impact on a breast self-examination system. Evaluation of trustworthiness indicators helps to ensure a trustworthy system and ethical interaction between a patient and physician. A new patient can obtain a consultation by referring to the best physician according to preference of other patients. Patients can also trust a physician based on another patient's recommendation regarding the physician's trust level. The correlation analysis further showed that the most preferred trustworthiness indicator is honesty.

(JMIR Med Inform 2020;8(9):e21584) doi:10.2196/21584

KEYWORDS

trust; trustworthiness; multifaceted trust; breast self-examination; breast cancer; health care system; social media

Introduction

Background

Breast cancer has become the most prevalent type of cancer affecting women in Indonesia and worldwide, and the number of deaths caused by breast cancer is growing every year [1,2]. In 2018, there were an estimated 2,088,849 new cases and 626,679 mortalities related to breast cancer [1]. Breast cancer occurs with an increase in the number of malignant cells originating from the inside layer of the mammary glands [3]. In the United States, breast cancer is detected by mammography (43%, 156/361), breast self-examination (25%, 90/361), clinical screening with breast self-examination (14%, 47/361), and accidents (18%, 64/361) [4]. Breast cancer prevention requires every woman to perform a breast self-examination as an early diagnosis mechanism for all ages after the first menstruation, which is expected to help reduce breast cancer mortality [5-8].

Moreover, curiosity about seeking health care–related information through the internet and social media has been gradually on the rise. Social media users prefer to seek information from social media [9,10], as other users share a substantial amount of information pertaining to breast cancer. Almost 87% of the total posts on Facebook related to breast cancer consist of support groups [11]. Other platforms such as Twitter include surveys on breast cancer education, shared stories about breast cancer survival, treatment plans, and images showing the progress of certain treatments [12]. Consequently, patients prefer to use social media to talk about sensitive body issues (such as breast cancer) as a more convenient venue than face-to-face interaction with a physician [13-16]. At the same time, physicians are actively participating in social media and health care systems related to breast cancer [9,14,17,18], and they tend to use social media for assisting, treating, and consulting on cancer [9,18,19]. Although the physician-patient interaction on social media platforms offers many conveniences, it also has important downsides.

There are reported cases of physicians behaving inappropriately on social media, such as posting incorrect information, misrepresenting their credentials, posting improper content, and false advertising [20-22]. The impact of such unethical behavior [19,23-26] can result in embarrassing patients and losing their trust. Since trustworthiness is an essential factor in any physician-patient relationship, the decrement of trust not only affects the health care business but also causes shame and depression for the patient [20,21,23]. Thus, in this study, we explored a trust model that can support and eliminate this issue. Toward this end, we focused on enhancing the multifaceted trustworthiness model proposed by Quinn et al [27] to be adopted for health care treatment on social media. The current multifaceted trust model calculates a trust score based on social interaction on social media platforms. Thus, the two-way trust evaluation adopted in this trust model is suitable for considering the patient-physician interaction in the health care domain. However, the multifaceted model has its limitations, since there is no credential representation mechanism in building trust context, no informed consent contract between parties trying to build trust, no mechanism to protect confidential data [19], and no preservation of user privacy [14]. Thus, there is a gap to be filled regarding how to best protect confidential information and ensure that each interaction and communication on social media is based on ethics.

The trust in social media is not personalized, specific, and single-faceted, but is rather generalized in a group context [27-29], and trust level cannot be annotated [27,28]. By contrast, the existing multifaceted trust model is personalizable, specializable, and capable of measuring the accuracy of trust recommendation [28]. As a result, Chieng et al [30] presented personalized comments or photos on social media as a user-centric model.

The objectives of this study were to: (1) identify the trustworthiness indicators that are most suitable for a breast self-examination system, (2) design and propose a breast self-examination system, and (3) evaluate the multifaceted trustworthiness interaction between patients and physicians using the breast self-examination system. Implementation of the multifaceted trustworthiness model into the breast self-examination system can identify the most preferred indicator of trustworthiness, offering relationship feedback between the patient and physician based on trust value and trust level. This could ultimately provide a more trustworthy and ethical patient-physician interaction on social media platforms such as Facebook.

Principles of Trust and the Multifaceted Trust Model

The trust theory introduced by Rotter [31] in 1967 is known as interpersonal trust, defined as “an expectancy held by an individual or a group that the word, promise, verbal or written statement of another individual or group can be relied upon.” The trust principle was introduced by Mayer and Davis in 1995 [32], which posits that factors related to the trustor and trustee will lead to trust. The characteristic of the trustor is trust propensity, which is a general willingness to trust others. In other words, trust propensity is a cause of risk behavior. People with different backgrounds, personality types, cultures, and experiences will differ in their propensity to trust.

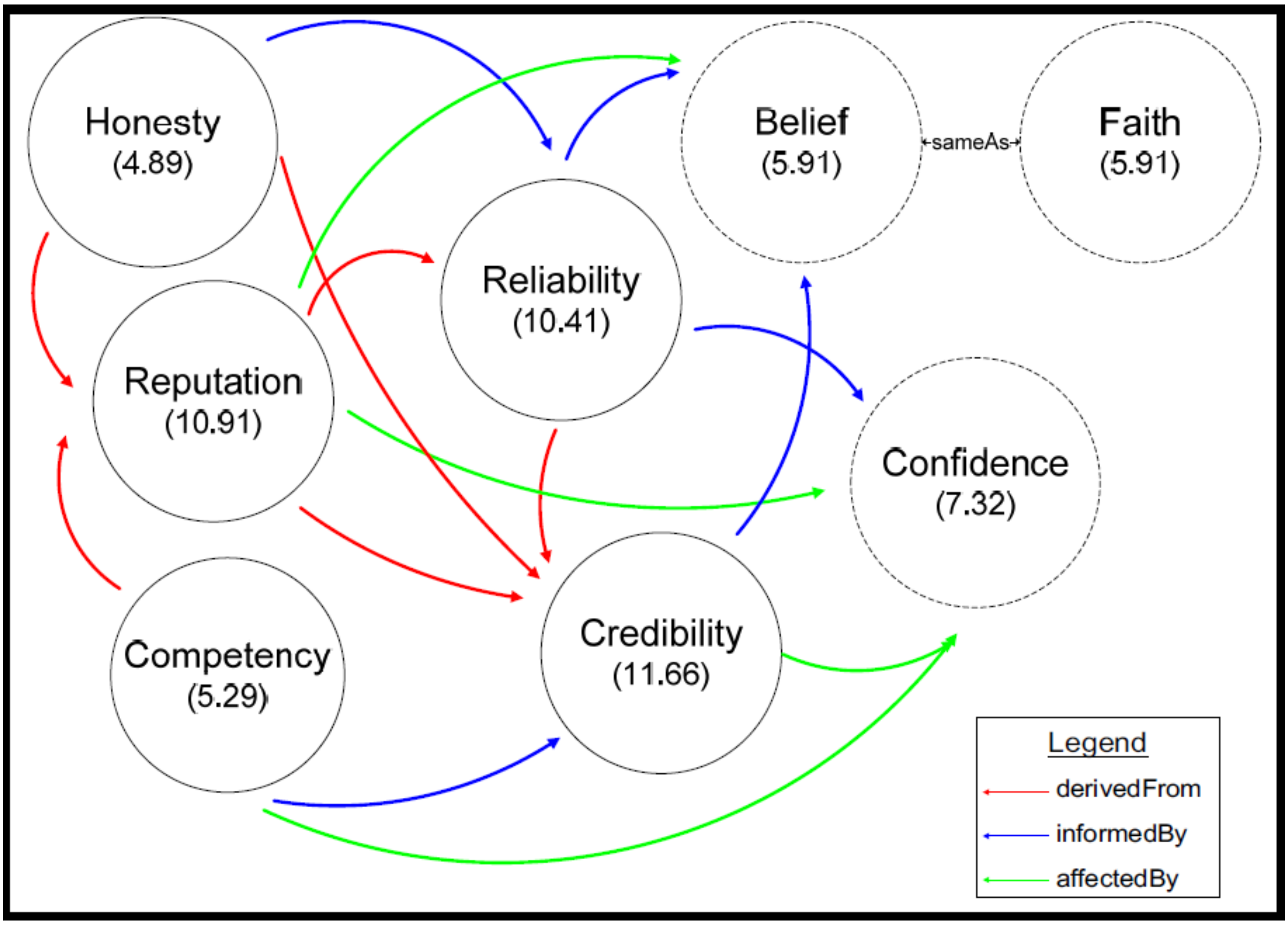
On the other side, the main characteristic of the trustee is trustworthiness, which is measured as the motivation to lie. For example, if a trustee will gain something through lying, they will be seen as less trustworthy [32].

Trust can therefore be defined as the confidence in somebody or a belief that somebody is good and honest [33]. Trustworthiness has been mentioned in the context of character honesty and integrity in health care [34], and it is a context-dependent and personalized characteristic [30]. According to Quinn [28], the multifaceted trustworthiness model is personalizable and specializable [28]. The trust characteristic in social media has been defined according to the following four traits [30,35-37]. The first is asymmetry, as trust between two users is not identical. That is, individual A could trust individual B, whereas individual B might not essentially trust individual A and vice versa. Second, trust is transitive: longer links create less trust. For example, consider that A and B are friends, and they trust each other. B is friends with C, but A does not know C. Since A knows B and trusts B’s friends, A might trust C to a certain extent. However, C is friends with D, whom neither A nor B knows, and thus A finds it difficult to trust D because of the distant connection and the fact that they do not know each other. Third, trust is context-dependent, according to time, situation, and experience. People tend to exhibit differences in trust based on the context. Fourth, trust is personalized as a subjective view. That is, the trustworthiness of a particular person might be viewed differently by two different people.

The key indicators of the multifaceted trust model (Figure 1) are honesty, reputation, competency, reliability, credibility, belief, confidence, and faith [27]. Honesty means the person makes good-faith agreements, tells the truth, and fulfills any promises made. Competency is the ability of one person to fulfill another person's needs. Confidence is "a feeling of certainty or easiness regarding a belief one holds" [38]. Reputation is part of the social notion of trust [39]. Belief is justified and should be accepted (ie, acceptable without argumentative support) [34].

Figure 1 shows how the concept of personalization allows the user to declare the idea that competency is influenced by reputation (ie, competency derived from reputation), credibility is influenced by belief (ie, credibility informed by belief), and so on. This concept can be repeated to construct a trust model to suit the user's needs and to reflect the user's subjective view of trust.

Figure 1. Personalized model of trustworthiness [27].



Trust Measurement Models

The trust measurement model in an open social network can be classified into five main models [36]: online reputation models, Marsh trust management, multicontext trust, trust inference for social networks (TISoN), and action-based trust.

Online Reputation Model

The online reputation model is based on online marketplaces such as Amazon and eBay. These models focus on user performance ratings provided after every part of the transaction is completed. The reputation value is then derived from the total sum of scores on eBay and the mean value from all ratings on Amazon. There would be no mechanism on correction if the user provided false information, as this model runs only based on the increment number of opinions that can create the reliability of reputation value [40].

Marsh Trust Model

Marsh [41] proposed a trust model based only on direct interactions, which can be broken down into basic trust, general trust, and situational trust.

For basic trust, the agent has an independent trusting disposition, which is calculated based on the accumulation of agent experiences. The best experiences bring an excellent disposition to trust, and the minimum experiences bring a bad disposition to trust. Marsh presented the notation T_x^t to identify the trust disposition of agent *x* at time *t*.

For general trust, the trust of the agent does not consider factors of the specific situation. Marsh used the notation $T_x(y)^t$ for expressing general trust between agent *x* and agent *y* at time *t*.

For situational trust, the trust of agents takes into account the specific situation. The following formula is used to calculate situational trust based on the utility of a situation:

in which *x* is the evaluator, *y* is the target agent, and α is the situation. $U_x(\alpha)^t$ *t* represents the utility *x* taken from situation α, $I_x(\alpha)^t$ is the important aspect in the situation α to agent *x*, and is the general trust estimation when identifying all possible data into $T_x(y,\alpha)$.

Multicontext Trust

Based on the Marsh trust model described above, a model in which context trust represent the fields of trust capability was proposed. For this purpose, trust is broken down into different contexts and each context is normalized in the range of 0 to 1 to fulfill future aggregation. The following seven trust functionalities on Facebook are considered [42]: (1) interaction time span (*S*), (2) number of interactions (*N*), (3) number of characters (*C*), (4) interaction regularity (*F*), (5) photo tagging (*P*), (6) group membership (*G*), and (7) common interests (*L*).

These seven contexts are summed to establish the formula of trust aggregation. Marsh [42] multiplied these contexts and used the final values, which identified a vector with several numbers where T_x is the priority given context $P=(T_S, T_N, T_C, T_F, T_P, T_G, T_L)$, and the final value of trust is formulated as:

The method of aggregation is important for attributing a value for each context. As an example, the context-free contribution to the overall trust simply represents decreasing the level of importance to the priority vector [42].

TISoN

The computational model for TISoN was introduced as a hybrid model based on a mathematical model and algorithm. Hamdi et al [43] also generated and evaluated trust values for relative rating. The authors designed a novel trust path–searching algorithm to ensure reliability of the trust path in a wider social network and used the trust inference measure to measure the degree of user trust in others.

Action-Based Trust

Gambhir et al [44] introduced action-based trust as a new model of trust based on user content disclosures such as comments, “likes,” post sharing, image tagging, and video posting. The algorithm of action-based trust involves calculation of trust values for user actions performed that focus on sharing sensitive content in an online social network. This algorithm has also been used by the multifaceted trust model in the context of online social networks [44].

Breast Self-Examination System in Online Social Networks

Breast self-examination is a method for early detection by which women examine their breasts to facilitate detection and alleviate any fear of cancer [45]. Breast self-examination is a regular monthly breast check using a mirror to observe any abnormal changes on the breasts [3], and is also considered to be the best tool for early breast cancer detection [46].

As patients prefer using social media to make appointments, receive reminders, diagnostic test results, provide information about their health, and as a forum for asking general questions related to health care [14], some specific features have been requested by patients as a reference to develop a breast self-examination system. Based on existing breast self-examination systems (Table 1), nine standard features should be embedded in any online breast self-examination system, including user account management, calendar, self-exam wizard, history, chat room, location, knowledge, video tutorial, and forum.

Table 1. Comparison of currently available breast self-examination (BSE) systems.

Existing BSE System	Features								
	User Account	Calendar (cycle period)	Self-Exam Wizard	History	Chat Room	Location	Knowledge	Video Tutorial	Forum
BSE Apps [47]	Absent	Present	Present	Present	Absent	Absent	Absent	Absent	Absent
Keep A Breast App by Luis M [48]	Absent	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent
Beyond The Shock App by NBCF [49]	Present	Absent	Absent	Absent	Absent	Absent	Present	Present	Present
Dr K's Breast Checker App [50]	Absent	Absent	Present	Present	Absent	Present	Present	Absent	Absent
Daisy Wheel App [51]	Absent	Absent	Present	Absent	Absent	Absent	Absent	Present	Absent
Breast Control App [52]	Absent	Present	Present	Present	Absent	Absent	Present	Absent	Absent
Makna (LUDIc) [53]	Absent	Present	Present	Absent	Absent	Present	Present	Present	Absent

Theory of Physician-Patient Interaction

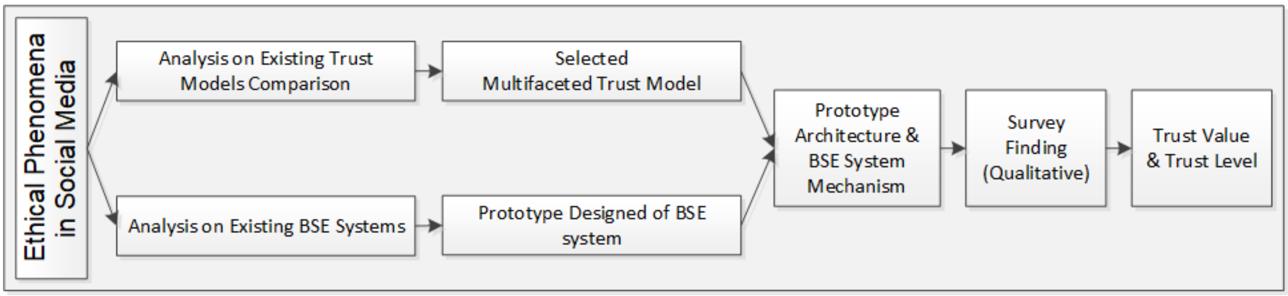
The physician-patient interaction is a communication process that describes the shared nature of the problem, treatment aims, and psychosocial care [54]. The physician-patient interaction emphasizes the behaviors of physicians toward patients. Behavior consent is the content given by the physician on health solutions (instrumental behavior) and the capability of physicians to treat patients (affective behavior) [55].

Methods

Research Flow

The research flow is outlined in Figure 2, following ethical phenomena in social media. The multifaceted trust model was selected for this study, as it is able to provide a subjective view of individual trust for each user. The prototype of the breast self-examination system was based on existing breast self-examination systems (Table 1). The architecture of the breast self-examination system with trustworthiness indicators was designed as a rating for the trust value for each patient consultation with a physician. Based on a survey with patients and physicians, the relationship between a particular physician trust value and trust level was identified. Thus, any patient who wants to choose a physician for consultation could refer to the physician's trust level.

Figure 2. Research flow. BSE: breast self-examination system.


Ethical Phenomena in Social Media
Analysis on Existing Trust Models Comparison
Selected Multifaceted Trust Model
Analysis on Existing BSE Systems
Prototype Designed of BSE system
Prototype Architecture & BSE System Mechanism
Survey Finding (Qualitative)
Trust Value & Trust Level

Existing Breast Self-Examination Systems

The comparison among the existing breast self-examination systems in Table 1 highlights the user account as an essential feature, which is based on the Beyond the Shock app [49] user account for securing patient and physician data. The remaining six apps do not have user accounts because they do not secure patient data, especially with respect to female breasts. The calendar is the feature that establishes the menstrual cycle period for performing a monthly self-exam. Of the seven available apps, four (BSE, Keep A Breast, Breast Control, and Makna) provide a calendar for setting a monthly self-exam reminder. The self-exam wizard feature explains how to perform a self-exam systematically, which should exist in any breast self-examination system. Only Beyond the Shock does not provide this feature. The history feature helps to record the patient's activity and the interaction process with their physician, which is present in several apps (BSE, Dr. K's Breast Checker, and Breast Control). The history feature records data by keeping a medical history for each patient, which will help physicians to trace each patient's performance. It is vital that this feature is secure due to the private nature of the information. The chat room allows for direct interaction between the patient and physician, which is essential for allowing patients to have direct interaction and communication with a physician without a face-to-face meeting. None of the apps currently supports the chat room feature, as they focus only on self-exam without connecting to a physician.

The next feature is location, which identifies available physicians nearest to the patient. This feature helps patients find a physician for further consultation based on their self-exam results. This feature is found in Dr. K's Breast Checker and Makna, which provide an address of a hospital or clinic for consultation with a physician. The knowledge feature provides scientific information related to breast cancer prevention, which is also an essential feature for obtaining breast cancer–related information for patient education on breast cancer. This feature is offered by several apps (Beyond the Shock, Dr. K's Breast Checker, Breast Control, and Makna). The video tutorial feature refers to any related breast cancer information provided via video as guidance. Video tutorials are provided in Beyond the Shock, Daisy Wheel, and Makna as an essential feature to help patients view information related to living with breast cancer. The forum is an open space to find current news or cases from physicians and patients, which offers a space where physicians provide general information to all patients on breast cancer prevention. Beyond the Shock uses a forum as an essential space for discussion between all physicians and patients in the same area. We included all criteria shown in Table 1 in our proposed breast self-examination system.

Integration of the Multifaceted Model of Trustworthiness in the Breast Self-Examination System

We refer to Quinn et al's [27] multifaceted trust model, which uses the idea of implementing a trust management model to act as the subjective view on trust. The breast self-examination system involves the eight indicators of trustworthiness as a rating system in a chat room setting. These indicators will determine the value of trust based on the user's interaction experience that is entirely personalized, transitive, and context-dependent. The personalized view will allow users to choose their trust value (ie, patients will give a value to each indicator in reference to their physician, and vice versa). The transitive view will offer each physician trust value as a reference when a patient recommends a physician to another patient. The context-dependent view will give patients flexibility in rating a physician; they can edit their trust level regarding physicians from time to time based on several consultations.

The trust level task is designed through the average rating value (ARV), which is used to calculate the average trust value given by the patient to their physician. The ARVs were generated based on the idea of Marsh [41] and Daskivich [56] to identify the trustworthiness level of a physician or patient with independent values. The trustworthiness level and independent value of a physician denoted by ARV are shown in Table 2.

Table 2. Trustworthiness level scale.

Trustworthiness level	Average rating value (ARV)	Independent value
High	9.0≤ARV≤10.0	5
Medium	7.5≤ARV<9.0	4
Low-Medium	5.0≤ARV<7.5	3
Low	2.5≤ARV<5.0	2
Distrust	0≤ARV< 2.5	1

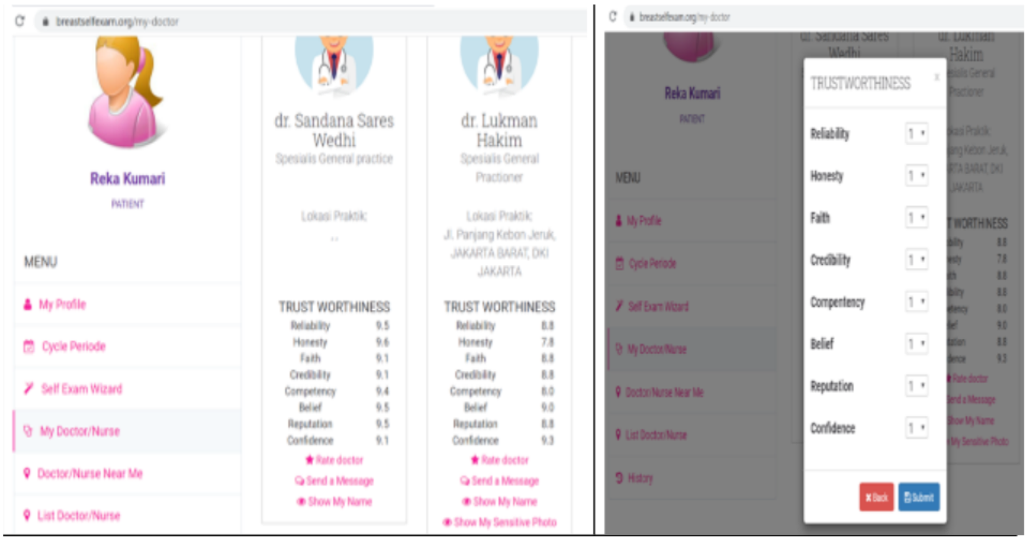
If ARVs are between 9 and 10, the trustworthiness level is considered to be high, and the independent value is 5, whereas lower values indicate higher levels of distrust. Therefore, the correlation analysis will depend on the independent value as the critical element.

Prototype Architecture and Mechanism of a Breast Self-Examination System

The prototype architecture of our breast self-examination system was based on Quinn et al’s [27] trust model, miniOSN [30], and miniOSN2.1 [30]. Patients can personalize accessibility to posted information, comments, and shared history data using the rating feature, as well as limit the physician to view the content of the patient data. The patient allows viewing the trust value and trust level of physicians. A physician’s trust value is based on the average of the trustworthiness indicators [28] and the ARV as the rating trust level [41,56]. The trust level of a physician is then identified by the trust value. To identify the ranking among multifaceted trustworthiness indicators, we evaluated the relationships between each indicator and the trust value. The higher ranking of a trustworthiness indicator is determined by a stronger correlation between the trust value and each indicator. Therefore, the ranking could evaluate the importance of the trustworthiness indicator in the multifaceted model. In the breast self-examination system, patients can personalize accessibility to set the value of posting and comments according to the trustworthiness indicators or trust level.

The trustworthiness mechanism on the breast self-examination system is measured through chat rooms and forums. When a patient request is accepted by a particular physician for consultation, the patient will encourage the trust level option of the physician, which is only seen by the patient. The patient can edit the trust level of a physician by choosing values ranging from 1 to 10 from the trustworthiness indicator’s ARV. The patient can also update the level of trust from time to time. For example, Figure 3 shows that the request of Reka (patient) for a consultation with physicians Sandana and Lukman was accepted by both physicians. After the meeting, she gave the experience a rating of 8 as the value for the physicians.

Figure 3. Left: Patient rated two physician after interaction; Right: Trustworthiness scale.


breastselfexam.org/my-doctor

Reka Kumari
PATIENT

dr. Sandana Sares Wedhi
Spesialis General practice
Lokasi Praktik:
, ,

dr. Lukman Hakim
Spesialis General Practioner
Lokasi Praktik:
Jl. Panjang Kebon Jeruk, JAKARTA BARAT, DKI JAKARTA

TRUST WORTHINESS

	dr. Sandana Sares Wedhi	dr. Lukman Hakim
Reliability	9.5	8.8
Honesty	9.6	7.8
Faith	9.1	8.8
Credibility	9.1	8.8
Competency	9.4	8.0
Belief	9.5	9.0
Reputation	9.5	8.8
Confidence	9.1	9.3

TRUSTWORTHINESS

Reliability	1
Honesty	1
Faith	1
Credibility	1
Compentency	1
Belief	1
Reputation	1
Confidence	1

Action-Based Trust Algorithm for the Breast Self-Examination System

An action-based trust algorithm was further implemented for the computational mechanism on the proposed breast self-examination system. This algorithm can measure the credibility of users on social media. The capabilities for evaluating and calculating the trust factors on user content disclosure include sharing personal records, sharing a post, comment, photo, and posting a message [44]. The computation of trust value is for a physician that acts on user content disclosure, namely as a trust factor. A physician trust factor may decrease or increase based on whether the patient selects a sensitive or not sensitive option. For example, the patient will likely select a sensitive option for a self-exam photo after completing the monthly self-examination. The sensitive option must be accompanied by informed consent from the patient before it is shared with the physician.

The action-based trust algorithm divides the measurement of each user action into weights, including the weight of action (Wa), weight of post (Wp), and weight of category (Wc). At the same time, Wc is a function of the weight for category. These weights, Wa, Wp, and Wc, are identified as the parameters of the trust factor. Table 3 shows the cluster of weights for test cases simulating the algorithm [44].

Table 3. Weighted clusters in the action-based trust algorithm.

Weight type	Weight value
Action	
Share	0.008
Like	0.006
Comment	0.007
Dislike	0.006
Tagging	0.005
Post	0.008
Post	
Photo	0.003
Video	0.002
Link	0.001
Message	0.003
Category	
Sensitive	0.009
Nonsensitive	0.001

Survey

We conducted a survey using open-ended interviews with 32 participants [57,58] and 77 interactions in the breast self-examination system. The survey was conducted from February 3, 2020 to March 30, 2020. The participants were physicians and female outpatients, all of whom had used the breast self-examination system. The 32 participants included 20 females and 12 males, comprising 16 physicians [58] and 16 female outpatients [57]. Of the 16 physicians, 12 were general practitioners and 4 were oncology specialists.

The 16 outpatients were healthy females who were aware of the health care system. Eight of these outpatients were aged 18 to 25 years old and the other eight were aged above 25 years. However, not all 32 participants ultimately completed the interaction task in the chat room and forum due to the consultation period. There were 24 active chat room participants and 22 participants interacted in the forum. The evaluation monitored the activity in the chat room (interaction between physician and patient) and the sharing of information by the physician through the forum. The interview started with an introduction to the research goal and an explanation of the flow on how to use the breast self-examination system.

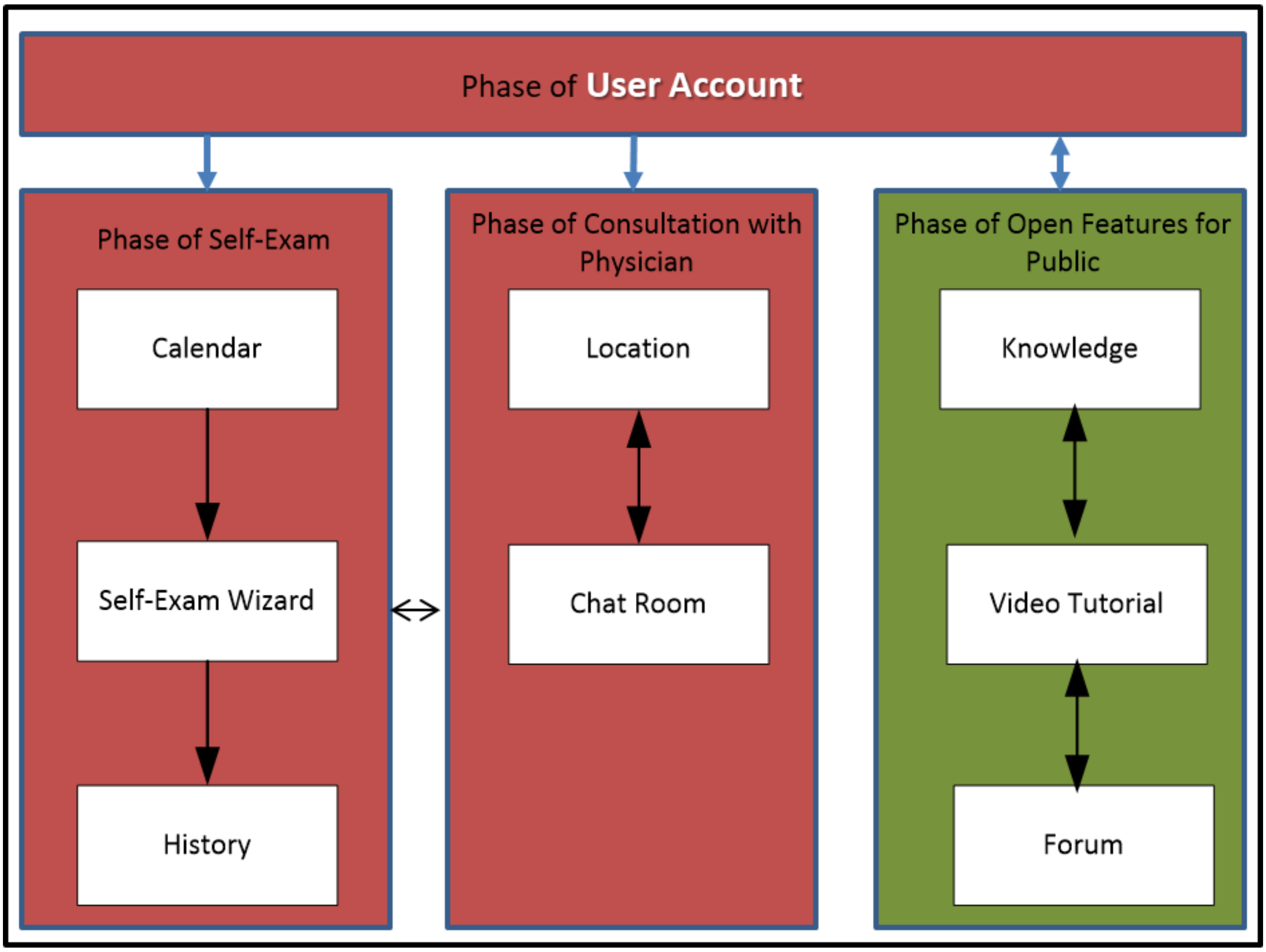
This study was approved by the research ethics board of Esa Unggul University committee (No. 0155-20.133/ DPKE-KEP/ FINAL-EA/ UEU/ V/ 2020).

Results

Design of the Breast Self-Examination Prototype

The prototype follows a module design (Figure 4), which is classified into four phases [59]. The first phase of the user account is registering and logging into the system. The second phase of self-exam is the phase of conducting a personal self-exam on the breast and annotation into the system. The third phase of consultation with a physician is when the user finds a doctor and has a consultation. The fourth phase of open features for the public is the opportunity for the public to access knowledge, video tutorials, and forum features without obtaining a user account. The phrases in each phase are based on user privileges.

Figure 4. Module design of the breast self-examination (BSE) system [59].



Each of the features of the breast self-exam system shown in Figure 4 has its own function. The calendar is a reminder system for performing a breast self-exam. The self-exam wizard is a guide to performing the correct self-exam daily. History serves as a self-exam record and tracks monthly breast self-exams. The location finds the nearest physician for receiving treatment. The chat room is a space for consultations between patients and physicians. Knowledge is a collection of links to news and expert opinions on breast health. Video tutorial is a collection of videos for performing the breast self-examination correctly. The forum is where physicians can share important information for patients related to breast cancer or breast self-exam.

Selection of a Suitable Trust Model

To select the most appropriate trust model for the breast self-examination system, we performed a literature review on papers related to the trust model. Ultimately, 11 articles were selected for comparison analysis among existing trust models. Table 4 shows the analysis of the comparison from several perspectives, including the trust model, related domains, selection of trust factors, methodology, and benefits.

The selection of trust models to suit the breast self-examination system refers to the health care, internet, and social media domains. Among the 11 articles, 7 are related to health care, 2 pertain to the internet, 1 is associated with social media, and the other is related to psychology. The trust model related to social media is the Multifaceted Trust Model for Online Social Network Environment [30]. Initially, the multifaceted trust model was introduced by Quinn [27] for the internet environment. The multifaceted trust factors support a user-centric model that requires users to personalize trust. For instance, Abbas et al [60], Montaque et al [61], and Quinn et al [27] focus on different areas with respect to reliability. Abbas et al [60] focus on the overall reliability of health care software, and Montaque et al [61] focus on the overall reliability of medical technology. In contrast, Quinn et al [27] and Chieng et al [30] introduce reliability as being personalized and specialized to a particular user or things.

From the methodology perspective, 5 of the studies were based on a qualitative approach, 5 were based on a quantitative approach, and the remaining study was based on a structured literature review. The qualitative approaches include evaluation of trust in the relationship between the patient and physician [25,63,64], whereas the quantitative approaches concentrate on a questionnaire to obtain participants' feedback. Therefore, the qualitative approach is more effective and relevant for garnering maximum performance out of the system.

The benefit perspective brings a consistent approach to choosing the trust model for the breast self-examination system. Two of the studies explored the trust model benefit that focused on the personalized and trust recommendation measurement offered by Quinn et al [27] and Chieng et al [30]. In contrast, the remaining trust model benefits focus on the general trust model of theoretical issues such as the instrument created [63], trust theory on the patient-physician relationship [25], patient trust in technology [61], and behavior approach theory [64]. Thus, a trust model related to the user-centric model is relevant to be embedded in the breast self-examination system.

Table 4. Comparison of existing trust models.

Reference	Trust Model	Domain	Trust Factors	Methodology	Benefit
Abbas et al [60]	Trustworthiness health care software model	Health care	Safety, validity, reliability, reusability, scalability, maintainability, performance	Structured literature review	The initial definition of trustworthiness attributes identified.
Velsen et al [63]	A conceptual model of patient trust in telemedicine services	Health care	Trust in the care organization, trust in the care professional, trust in the treatment, trust in the technology	Qualitative method on focus groups with a survey on four factors (trust in care organization, care professional, treatment, and technology)	A valid instrument (PATAT) created to assess patient trust in a telemedicine service and as a benchmark on the same service.
Krot and Rudawska [25]	Model of trust in the doctor-patient relationship	Health care	Macrotrust, microtrust, mesotrust	Qualitative method on the analysis of published comments	Trust in the doctor-patient relationship is a social, complex, and multidimensional phenomenon.
Chieng et al [30]	Multifaceted trust model for online social network environment	Online social network	honesty, reputation, competency, credibility, confidence, reliability, belief, faith	Quantitative method on survey questionnaire data	This model can address trust issues on social networking sites through personalized trust features
Quinn et al [27]	Multifaceted trust model	Internet	honesty, reputation, competency, credibility, confidence, reliability, belief, faith	Quantitative method on survey questionnaire data	A multifaceted model is personalizable and specializable; provides accuracy of trust recommendation
Montaque et al [61]	Model of patient and provider trust in medical technology	Health care	Communication, compassion, privacy, competence, confidentiality, dependability, reliability	Qualitative method with a grounded theory approach	The interaction between the provider and technology influences patient trust in technology
Zahedi and Song [65]	Dynamic model of trust	Health care	ability, benevolence, integrity	Quantitative method on a laboratory experiment	Trust beliefs change depending on web consumers with more experience in health infomediaries.
Corritore et al [66]	Model of online trust	Health care	Credibility, risk, ease of use	Quantitative method on the instrument of a 34-item Likert-scale	To lead the development of the health care website on trust; produce a valid instrument to measure online trust on the health care website; produce a model of online trust for health care websites
Lee and Turban [67]	A trust model for consumer internet shopping	Internet	Ability, benevolence, integrity, trust propensity	Quantitative method on survey questionnaire data	Merchant integrity is a major positive determinant of consumer trust and its effect.
Lewicki et al [64]	Models of interpersonal trust development	Psychology	Ability, benevolence, integrity	Qualitative method on the grounded theory approach	A behavioral and physiological approach theory
Dibben et al [62]	A model of trust development in the patient-physician relationship	Health care	Dispositional trust, learnt trust, situational trust	Qualitative method	"This model is able to identify and map trust levels and thresholds of cooperative behavior and modify the behavior on the interaction between physician and patient."

Based on analysis of the 11 filtered articles, we decided to adopt the multifaceted trust model introduced by Quinn et al [27]. The reason for selecting the multifaceted trust model (Table 4) is that this model can provide a subjective view of individual trust for each user. This model is also a user-centric model that can personalize trust features such as comments and photos on social media [30]. In particular, trust can protect the physician's reputation before the patient makes any decision and allows the user to choose a credible physician [44]. According to Singh and Chin [36], trust is a significant factor to attract a user to use the site for recommendation to others based on a rating feature. That is, patients can consider a physician's credibility for consultations, and physicians can consider the patient's honesty in providing information about their health status [36].

Evaluation of the Multifaceted Trustworthiness of the Breast Self-Examination System

Correlation Analysis Among Trustworthiness Indicators

The Pearson correlation coefficient (r) was used to evaluate the correlations among various trustworthiness indicators in the breast self-examination system based on the following formula:

The Pearson correlation coefficient represents a relationship (*r*) between the independent variable (*x*) and the dependent variable (*y*) based on a numerical variable between –1 and 1, where 0 indicates no correlation, 1 indicates a complete positive correlation, and –1 indicates a complete negative correlation. A correlation coefficient of 0.7 and above indicates a significant and positive relationship between *x* and *y*; that is, when variable *x* increases, variable *y* will also increase. Similarly, if the correlation value is negative, if *x* increases, then *y* also decreases [68]. We conducted a correlation analysis from 77 samples collected from the MySQL database, which included ratings of patients for a doctor in a chat room, ratings of patients for several doctors in the chat room, and ratings of patients for a doctor in the forum. After a participant chatted with a doctor in a chat room, the participant could rate the doctor based on 8 trustworthiness indicators, and then was required to edit the rating after a second consultation. The data were exported from MySQL to a Microsoft Office Excel spreadsheet, and Pearson correlation analysis was performed using SPSS v.24 (IBM) [69-71]. The Cronbach α value was .92, which means that the data are reliable (>.80) [71,72]. The results of the correlation analysis are summarized in Table 5.

Table 5. Correlation coefficients of each trustworthiness indicator and trust value.

Trustworthiness indicators	Correlation coefficient
honesty	0.91
reputation	0.72
competency	0.73
reliability	0.73
credibility	0.85
belief	0.75
confidence	0.79
faith	0.79

A strong positive correlation (>0.70) was found for all trustworthiness indicators and trust value [68], indicating that trustworthiness indicators would predict a tendency to change the trust value; the higher the values of the correlation coefficient, the stronger the relationship between the trustworthiness indicator and trust value. The highest trustworthiness indicator with a strong positive correlation was honesty, followed by credibility, confidence, and faith. Reputation had the lowest correlation value among the trustworthiness indicators (Table 5). The correlations among trustworthiness indicators were significant (P<.001). Overall, these results suggest that honesty has the highest rank with respect to trustworthiness, which means that the most important aspect in the interaction process between a doctor and patient is honest communication. This is followed by credibility and faith as the second most important factors, reflecting the need of patients to have a doctor with good credibility.

Patient-Physician Relationship

The patient-physician interaction was evaluated to measure the trust level of a patient toward a physician's behavior [55]. This analysis was based on different views such as the interaction in different time frames, patient interactions with several physicians, and patient feedback on physicians' articles in the forum.

Patient and Physician Interaction in Different Time Frames

The patients' ratings of a physician in the chat room over multiple interactions are summarized in Table 6. For example, Patient X5 had a consultation with physician Y5 on different occasions. The trust value in the first week was 8.63 and was 10.00 in the second week. Similarly, for patient X13 and physician Y6, the trust value in the first week was 8.13, which increased to 8.50 and 9.25 in the third and fourth week, respectively. This interaction shows that the trust level increased from medium to high, which means that the trust value and trust level of the patient for a particular physician will generally increase after several interactions. Indeed, the time effect in the interaction analysis based on comparing the trust value between the first and second week was statistically significant (P=.03).

Table 6. Interaction between a patient and a doctor over different time frames.

Patient	Doctor	Meeting time	Trustworthiness indicators (ratings)								Trust value	Trust level
			honesty	reputation	competency	reliability	credibility	belief	confidence	faith		
X4	Y3	Week 1	10	10	9	10	10	8	9	10	9.50	High
X4	Y3	Week 2	10	10	10	10	9	9	10	10	9.75	High
X5	Y5	Week 1	10	9	8	8	8	9	9	8	8.63	Medium
X5	Y5	Week 2	10	10	10	10	10	10	10	10	10.00	High
X7	Y10	Week 1	8	7	8	9	9	9	7	8	8.13	Medium
X7	Y10	Week 2	9	8	9	9	9	9	8	9	8.75	Medium
X7	Y10	Week 3	10	9	9	9	9	9	9	8	9.00	High
X11	Y12	Week 1	7	9	9	8	8	9	10	9	8.63	Medium
X11	Y12	Week 2	8	10	7	9	10	9	9	9	8.88	Medium
X12	Y3	Week 1	9	10	10	9	10	10	10	9	9.63	High
X12	Y3	Week 2	10	10	10	10	10	10	10	10	10.00	High
X13	Y6	Week 1	7	9	8	7	9	8	8	9	8.13	Medium
X13	Y6	Week 2	8	9	9	7	9	9	9	8	8.50	Medium
X13	Y6	Week 3	10	10	9	9	10	8	10	8	9.25	High
Total Average			9.00	9.29	8.93	8.86	9.29	9.00	9.14	8.93	9.06	High

As shown in Table 6, the trust value increased when each patient had several consultations with a physician on different occasions. The trust value was taken from each patient's feedback rating, including patients X4, X5, X7, X11, X12, and X13, who provided useful input on trustworthiness regarding their physician after several consultations.

One Patient With Many Physician Interactions

Data interaction in the chat room demonstrated that several patients interacted with more than one physician. Sixteen patients requested communication with several physicians, and only 11 physicians responded to these requests and had an excellent interaction with patients. Patients prefer to chat with physicians who have been rated by another patient. However, not all physicians were receptive to accepting a patient request due to their limited time. Table 7 demonstrates the varying trust values between physicians during interactions with the same patient. The trust value is given by the patients that provided their subjective views when rating a physician after consultation. Based on the total average of trustworthiness indicators, honesty (9.39) emerged as the most important indicator, followed by credibility (9.28) and faith (9.28).

Table 7. Patient ratings of several physicians in a chat room.

Patient	Doctor	Trustworthiness indicators								Trust value	Trust level
		honesty	reputation	competency	reliability	credibility	belief	confidence	faith		
X1	Y1	9	8	9	9	8	9	9	9	8.75	Medium
X1	Y2	9	10	9	10	10	10	10	10	9.75	High
X2	Y3	10	10	10	10	10	10	10	10	10.00	High
X2	Y4	10	8	10	10	10	10	9	10	9.63	High
X3	Y3	10	8	10	10	10	10	10	10	9.75	High
X3	Y6	9	10	9	10	10	10	10	10	9.75	High
X3	Y5	9	10	9	10	10	10	10	10	9.75	High
X4	Y3	10	10	9	10	10	8	9	10	9.50	High
X4	Y5	10	10	10	9	8	9	9	8	9.13	High
X5	Y5	10	9	8	8	8	9	9	8	8.63	Medium
X5	Y6	8	9	8	8	9	8	7	8	8.13	Medium
X6	Y7	9	9	9	9	9	9	9	9	9.00	High
X6	Y8	9	8	9	8	8	9	9	9	8.63	Medium
X7	Y9	9	9	9	8	9	9	10	9	9.00	High
X7	Y2	10	9	10	9	9	9	9	10	9.38	High
X7	Y10	8	7	8	9	9	9	7	8	8.13	Medium
X8	Y11	10	8	9	8	10	8	8	9	8.75	Medium
X8	Y2	10	8	10	8	10	8	9	10	9.13	High
Total Average		9.39	8.89	9.17	9.06	9.28	9.11	9.06	9.28	9.16	High

Rating of Physician-Posted Articles by Patients in the Forum

In the forum, 9 physicians participated in posting information (articles) related to breast cancer, and 13 patients rated the articles. As shown in Table 8, the trustworthiness ratings on the forum revealed the highest trust value for physician Y15, followed by physicians Y7 and Y12. This means that physician Y15 posted the most highly trusted articles even though physician Y12's article was read by more patients.

Table 8. Trust value matrix on patient (X) ratings of articles posted by physicians (Y).

Patient	Y5	Y7	Y12	Y13	Y14	Y15	Y16	Y17	Y18
X1	—[a]	—	—	—	—	—	—	—	8.29
X2	—	—	—	—	—	—	—	8.29	8.71
X3	8.43	—	8.43	—	8.00	—	—	—	—
X4	8.43	—	8.14	—	—	—	—	—	—
X5	—	8.57	—	—	8.71	—	—	—	—
X6	—	—	9.57	—	9.00	9.29	—	—	—
X7	8.57	—	8.57	—	—	—	—	—	—
X9	—	—	—	—	—	8.71	8.57	8.71	—
X10	—	—	—	—	—	9.00	8.86	8.57	—
X11	—	—	—	8.00	—	—	—	—	—
X12	—	—	—	—	—	—	—	—	8.14
X14	—	9.00	—	—	—	—	—	—	—
Total (trust level)	8.48 (medium)	8.79 (medium)	8.68 (medium)	8.00 (medium)	8.57 (medium)	9.00 (high)	8.72 (medium)	8.52 (medium)	8.38 (medium)

[a] —: data not applicable, given no rating for that physician's article.

Table 9 shows the trust values for each physician rated by several patients. Patients rated a physician based on their personal views on the articles posted in the forum by the physicians. For example, physician Y14 was rated by three different patients (X3=8.00, X5=8.71, and X6=9.00), indicating a very high trust level by patient X6. This means that patient X6 considered the article posted by physician Y14 to be of more benefit compared with the views of patients X3 and X5. In this case, honesty (8.86) was the most important indicator, followed by belief (8.81) and confidence (8.81).

Table 9. Patient ratings of articles posted by physicians on the forum.

Doctor	Patient	Trustworthiness indicators								Trust value	Trust level
		honesty	reputation	competency	reliability	credibility	belief	confidence	faith		
Y14	X6	9	9	9	9	9	8	10	8	9.00	High
Y14	X5	9	8	9	9	8	9	9	8	8.71	Medium
Y14	X3	8	9	8	8	8	7	8	9	8.00	Medium
Y15	X6	10	9	9	9	8	10	10	9	9.29	High
Y15	X9	9	8	9	9	8	9	9	8	8.71	Medium
Y15	X10	10	9	9	8	8	10	9	8	9.00	High
Y12	X6	10	9	10	9	9	10	10	9	9.57	High
Y12	X7	9	9	9	9	8	8	8	9	8.57	Medium
Y12	X3	9	9	9	8	7	8	9	8	8.43	Medium
Y12	X4	7	8	8	8	9	8	9	9	8.14	Medium
Y16	X9	8	9	9	9	8	8	9	9	8.57	Medium
Y16	X10	9	8	9	8	8	10	10	7	8.86	Medium
Y17	X9	9	9	8	9	9	8	9	8	8.71	Medium
Y17	X2	8	8	8	8	8	9	9	9	8.29	Medium
Y17	X10	10	9	8	9	7	9	8	8	8.57	Medium
Y5	X7	9	8	8	8	9	9	9	9	8.57	Medium
Y5	X3	8	9	8	9	9	9	7	9	8.43	Medium
Y5	X4	9	9	8	8	9	8	8	7	8.43	Medium
Y18	X1	8	8	9	8	9	8	8	7	8.29	Medium
Y18	X2	9	9	9	8	8	9	9	9	8.71	Medium
Y18	X12	9	8	8	8	9	8	7	7	8.14	Medium
Y7	X5	9	9	8	8	8	9	9	9	8.57	Medium
Y7	X14	10	9	9	9	9	9	8	9	9.00	High
Y13	X11	8	8	8	7	8	10	7	9	8.00	Medium
Total Average		8.88	8.63	8.58	8.42	8.33	8.75	8.67	8.38	8.61	Medium

Discussion

Eight indicators of trustworthiness taken from the multifaceted trust model showed significant positive correlations with trust value, including honesty, credibility, confidence, faith, belief, competency, reliability, and reputation. The following nine features were considered to be important in the design of the breast self-examination system: user account, calendar, self-exam wizard, history, chat room, location, knowledge, video tutorial, and forum. The trust level of a patient for a particular physician was found to increase after several interactions, and the patient can choose the right physician by considering other patients' recommendations based on the physician's trust level.

There are 32 participants registered in the breast self-examination system. Registration is achieved through a user account with approval validation sent by the system to the user's email. The security model of the user account is MD5. Users set their cycle period via the calendar and follow the self-exam wizard by recording their activity in the history feature. If a patient identifies changes on the surface of their breast during a self-exam, they can take a photo of the breast and enter it into the system, which can be annotated as a "sensitive" picture [36,44]. A picture that is annotated as sensitive is then assigned a weight for category (Wc), which means that the picture will require the patient's informed consent before sharing with the physician. The chat room is a convenient space for interaction and communication between a patient and physician. By default, the patients deidentify themselves by showing only their patient ID number to the physician. During the interaction and at the physician's request, the patient shares their history as their medical record. This sharing was identified as a weight for action (Wa). The Wa will lead the patient to share based on the request from the physician. On the other side, physicians are able to post an article to the forum, which is identified as a weight for the post (Wp). The patients looked at several articles posted by the physician in the forum and provided feedback through rating the physician. Each share, post, and category is a confidential activity carried out by the patient and physician [36,44].

The correlation analysis of trustworthiness factors on the breast self-examination system demonstrated that honesty has the highest ranking for trustworthiness overall. This reflects that the interaction process between physicians and patients requires honest communication through honest information from the patient so that the physician can provide the correct treatment. Honest advice from the physician will create trust on the patient's part, and as a result, the patient will follow the physician's advice. This was followed by credibility as the second most important feature due to the patient requiring a credible doctor [27].

The analysis of patient-physician interaction over different time frames revealed that patient trust will grow when several interactions occur between a patient and physician. The patient's understanding of the physician regarding their reputation and credibility is the first preference. Some feedback from the patients included feeling comfortable talking with physicians based on a recommendation by another patient through seeing the physician trust value. This feedback proves that trust is indeed transitive. The interaction of one patient with several physicians reflects the personal views of the patient about a particular physician based on their convenient communication in the chat room [30].

Patient feedback in the forum related to articles posted by a physician was based on the valuable information received by the patient, indicating that patients have their own views for accessing the useful information provided in each article posted by a physician. The most trusted article was measured by the weight of trust value. Overall, we found that patients' subjective views in taking the information from each posted article on breast cancer benefited the patients based on their own experience and situation (ie, context-dependent effect) [30].

Overall, this study reveals the strong ability of the multifaceted trust model to provide a more trustworthy system, ethical interactions between patients and physicians, and patient control of data. This analysis proves the trust characteristic of social media through interactions between patients and physicians in the breast self-examination system [63,73]. Ultimately, the implementation of multifaceted trust enables patients to make the right choice of a physician by considering other patients' recommendations based on the physician's trust level.

In conclusion, multifaceted trustworthiness indicators have a significant impact on the breast self-examination system. These indicators provide a trustworthy system and ethical interaction between a patient and physician as assessed through the trust value and trust level. Based on the trust value and trust level of physicians, a new patient can obtain a consultation by referring to the most highly preferred physician. In addition, the patient's trust level to a particular physician will increase after several interactions. The correlation analysis also showed that the most preferred trustworthiness indicator was honesty. With more interactions that occur based on weekly meetings, more trust will grow between the patient and physician. This trust will automatically increase the reputation and credibility of the physician.

Multifaceted trustworthiness could be explored in more areas of relevance in the health care system. Several actors in various health care systems should consider adding and reviewing the process of interactions, such as those occurring among the health care provider, patient, physician, system provider, and health care supplier.

Acknowledgments

This research was funded by Yayasan Pendidikan Tinggi 17 Agustus 1945 Jakarta, Indonesia.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1676 KB - medinform_v8i9e21584_app1.pdf]

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018 Nov;68(6):394-424 [FREE Full text] [doi: 10.3322/caac.21492] [Medline: 30207593]
2. InfoDatin Stop Kanker. Pusat Data Dan Informasi Kementrian Kesehatan Republik Indonesia. Jakarta, Indonesia: Kemenkes; 2015 Sep 20. URL: http://www.pusdatin.kemkes.go.id/resources/download/pusdatin/infodatin/infodatin-kanker.pdf [accessed 2020-08-20]

3. Longe JL, editor. The Gale encyclopedia of cancer: a guide to cancer and its treatments. China: Thomson Gale; Apr 01, 2005.
4. Roth MY, Elmore JG, Yi-Frazier JP, Reisch LM, Oster NV, Miglioretti DL. Self-detection remains a key method of breast cancer detection for U.S. women. J Womens Health (Larchmt) 2011 Aug;20(8):1135-1139 [FREE Full text] [doi: 10.1089/jwh.2010.2493] [Medline: 21675875]
5. Miller AB, Baines CJ. The role of clinical breast examination and breast self-examination. Prev Med 2011 Sep;53(3):118-120. [doi: 10.1016/j.ypmed.2011.05.001] [Medline: 21596057]
6. Juanita J, Piyanuch J, Umaporn B. BSE practice and BSE self-efficacy among Nursing Students in Aceh, Indonesia. Nurse Media J Nurs 2013 Jan 31;3(1):557-568 [FREE Full text] [doi: 10.14710/nmjn.v3i1.4496]
7. Hassan LM, Mahmoud N, Miller AB, Iraj H, Mohsen M, Majid J, et al. Evaluation of effect of self-examination and physical examination on breast cancer. Breast 2015 Aug;24(4):487-490. [doi: 10.1016/j.breast.2015.04.011] [Medline: 25977176]
8. Sujindra E, Elamurugan T. Knowledge, attitude, and practice of breast self-examination in female nursing students. Int J Educ Psychol Res 2015;1(2):71. [doi: 10.4103/2395-2296.152216]
9. Modahl M, Tompsett L, Moorhead T. Doctors, Patients & Social Media. QuantiaMD.: QuantiaMD; 2011 Sep 05. URL: http://www.quantiamd.com/q-qcp/social_media.pdf [accessed 2020-08-10]
10. Thompson LA, Dawson K, Ferdig R, Black EW, Boyer J, Coutts J, et al. The intersection of online social networking with medical professionalism. J Gen Intern Med 2008 Jul;23(7):954-957 [FREE Full text] [doi: 10.1007/s11606-008-0538-8] [Medline: 18612723]
11. Hale TM, Pathipati AS, Zan S, Jethwani K. Representation of health conditions on Facebook: content analysis and evaluation of user engagement. J Med Internet Res 2014 Aug 04;16(8):e182 [FREE Full text] [doi: 10.2196/jmir.3275] [Medline: 25092386]
12. Attai DJ, Cowher MS, Al-Hamadani M, Schoger JM, Staley AC, Landercasper J. Twitter Social Media is an Effective Tool for Breast Cancer Patient Education and Support: Patient-Reported Outcomes by Survey. J Med Internet Res 2015 Jul 30;17(7):e188 [FREE Full text] [doi: 10.2196/jmir.4721] [Medline: 26228234]
13. Denecke K, Bamidis P, Bond C, Gabarron E, Househ M, Lau AYS, et al. Ethical Issues of Social Media Usage in Healthcare. Yearb Med Inform 2015 Aug 13;10(1):137-147 [FREE Full text] [doi: 10.15265/IY-2015-001] [Medline: 26293861]
14. Fisher J, Clayton M. Who gives a tweet: assessing patients' interest in the use of social media for health care. Worldviews Evid Based Nurs 2012 Apr;9(2):100-108. [doi: 10.1111/j.1741-6787.2012.00243.x] [Medline: 22432730]
15. Fox S. The Social Life of Health Information. Pew Research Center. Washington DC: Pew Research Center; 2011 May 12. URL: http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011/ [accessed 2015-09-11]
16. White M, Dorman S. Receiving social support online: implications for health education. Health Educ Res 2001 Dec;16(6):693-707. [doi: 10.1093/her/16.6.693] [Medline: 11780708]
17. Kemp S. Digital in 2018: World?s internet users pass the 4 billion mark. wearesocial.: Wearesocial URL: https://wearesocial.com/blog/2018/01/global-digital-report-2018 [accessed 2018-02-13]
18. Brown J, Ryan C, Harris A. How doctors view and use social media: a national survey. J Med Internet Res 2014 Dec 02;16(12):e267 [FREE Full text] [doi: 10.2196/jmir.3589] [Medline: 25470407]
19. Chretien KC, Kind T. Social media and clinical care: ethical, professional, and social implications. Circulation 2013 Apr 02;127(13):1413-1421. [doi: 10.1161/CIRCULATIONAHA.112.128017] [Medline: 23547180]
20. Lagu T, Kaufman EJ, Asch DA, Armstrong K. Content of weblogs written by health professionals. J Gen Intern Med 2008 Oct;23(10):1642-1646 [FREE Full text] [doi: 10.1007/s11606-008-0726-6] [Medline: 18649110]
21. Chretien KC, Farnan JM, Greysen SR, Kind T. To friend or not to friend? Social networking and faculty perceptions of online professionalism. Acad Med 2011 Dec;86(12):1545-1550. [doi: 10.1097/ACM.0b013e3182356128] [Medline: 22030752]
22. Greysen SR, Johnson D, Kind T, Chretien KC, Gross CP, Young A, et al. Online professionalism investigations by state medical boards: first, do no harm. Ann Intern Med 2013 Jan 15;158(2):124-130. [doi: 10.7326/0003-4819-158-2-201301150-00008] [Medline: 23318312]
23. Greysen SR, Chretien KC, Kind T, Young A, Gross CP. Physician violations of online professionalism and disciplinary actions: a national survey of state medical boards. JAMA 2012 Mar 21;307(11):1141-1142. [doi: 10.1001/jama.2012.330] [Medline: 22436951]
24. Greysen SR, Kind T, Chretien KC. Online professionalism and the mirror of social media. J Gen Intern Med 2010 Nov;25(11):1227-1229 [FREE Full text] [doi: 10.1007/s11606-010-1447-1] [Medline: 20632121]
25. Krot K, Rudawska I. The role of trust in doctor-patient relationship?: qualitative evaluation of online feedback. BazKon 2016;9(3):76-88 [FREE Full text]
26. Khana R, Mahinderjit SM, Damanhoori F, Mustaffa N. Investigating the Importance of Implementing Ethical Value on a Healthcare System within a Social Media context. Int J Innov Creat Chang 2020;12(5):352-369 [FREE Full text]
27. Quinn K, Lewis D, O'Sullivan D, Wade VP. An analysis of accuracy experiments carried out over of a multi-faceted model of trust. Int J Inf Secur 2009 Jan 13;8(2):103-119. [doi: 10.1007/s10207-008-0069-7]

28. Quinn K. A Multi-Faceted Model of Trust that is Personalisable and Specialisable. Thesis, Doctor of Philosophy. Dublin, Ireland: University of Dublin, Trinity College; 2006 Sep. URL: http://www.tara.tcd.ie/bitstream/handle/2262/77650/Quinn%2c%20Karl_TCD-SCSS-PHD-2006-10.pdf?sequence=1&isAllowed=y [accessed 2020-09-09]
29. Ruohomaa S, Kutvonen L. Trust Management Survey. 2005 Presented at: Proceedings of the Third International Conference on Trust Management; 2005; Berlin p. 77-92 URL: https://link.springer.com/chapter/10.1007/11429760_6
30. Chieng LB, Singh MM, Zaaba ZF, Hassan R. Multi-Facet Trust Model for Online Social Network Environment. Int J Netw Sec Appl 2015 Jan 31;7(1):1-18. [doi: 10.5121/ijnsa.2015.7101]
31. Rotter J. A new scale for the measurement of interpersonal trust. J Pers 1967 Dec;35(4):651-665. [doi: 10.1111/j.1467-6494.1967.tb01454.x] [Medline: 4865583]
32. Mayer RC, Davis JH, Schoorman FD. An Integrative Model Of Organizational Trust. Acad Manag Rev 1995 Jul;20(3):709-734. [doi: 10.5465/amr.1995.9508080335]
33. McKay JM, Hornby AS. Oxford Advanced Learner's Dictionary of Current English. Oxford: Oxford University Press; Mar 1975:77.
34. Beauchamp T, Childress J. Principles of Biomedical Ethics, 6th edition. New York, USA: Oxford University Press; 2009.
35. Chieng L, Mahinderjit SM, Zaaba Z, Hassan R. User-Centric Personalized Multifacet Model Trust in Online Social Network. Comput Sci Inf Technol 2014 Jun:245-259. [doi: 10.5121/csit.2014.41220]
36. Mahinderjit M, Yi T. Hybrid Multi-faceted Computational Trust Model for Online Social Network (OSN). Int J Adv Comput Sci Appl 2016;7(6):1-11. [doi: 10.14569/ijacsa.2016.070601]
37. Johnson H, Lavesson N, Zhao H, Wu S. On the Concept of Trust in Online Social Networks. In: Salgarelli L, Bianchi G, editors. Trustworthy Internet. Milano, Italia: Springer; Jun 2011:143-157.
38. McKnight D, Chervany N. The Meanings of Trust. In: Salgarelli L, Bianchi G, editors. Trustworthy Internet. Milano, Italy: Springer; Jun 15, 2011:143-157.
39. Golbeck J, Hendler J. Accuracy of Metrics for Inferring Trust and Reputation in Semantic Web-Based Social Networks. In: Motta E, Shadbolt R, Stutt A, Gibbins N, editors. Engineering Knowledge in the Age of the Semantic Web. EKAW 2004. Lecture Notes in Computer Science, vol 3257. Berlin: Springer; 2004.
40. Sabater J, Sierra C. Review on Computational Trust and Reputation Models. Artif Intell Rev 2005 Sep;24(1):33-60 [FREE Full text] [doi: 10.1007/s10462-004-0041-5]
41. Marsh S. Trust as a computational concept. In: SpringerBriefs Comput Sci. New York, USA: Springer Link; 1994:5-24.
42. Švec T, Samek J. Trust evaluation on Facebook using multiple contexts. 2013 Jun 14 Presented at: CEUR Workshop; 2013; Rome, Italy p. 1-10 URL: http://ceur-ws.org/Vol-997/trum2013_paper_3.pdf
43. Hamdi S, Bouzeghoub A, Gancarski A, Ben YS. Trust inference computation for online social networks. In: IEEE Xplore. 2013 Dec 12 Presented at: 12th IEEE International Conference on Trust, Security, and Privacy in Computing and Communications; 2013; Melbourne, Australia p. 210-217 URL: https://ieeexplore.ieee.org/document/6680843 [doi: 10.1109/TrustCom.2013.240]
44. Gambhir M, Doja M. Action-based trust computation algorithm for online social network. In: IEEE Xplore. 2014 Apr 07 Presented at: International Conference on Advanced Computing and Communications Technologies, ACCT; February 8-9, 2014; Rohtak, India p. 451-458 URL: https://ieeexplore.ieee.org/abstract/document/6783496 [doi: 10.1109/ACCT.2014.89]
45. Thornton H, Pillarisetti RR. 'Breast awareness' and 'breast self-examination' are not the same. What do these terms mean? Why are they confused? What can we do? Eur J Cancer 2008 Oct;44(15):2118-2121. [doi: 10.1016/j.ejca.2008.08.015] [Medline: 18805689]
46. Hisham AN, Yip C. Overview of breast cancer in Malaysian women: a problem with late diagnosis. Asian J Surg 2004 Apr;27(2):130-133 [FREE Full text] [doi: 10.1016/S1015-9584(09)60326-2] [Medline: 15140665]
47. Webfoot T. Breast Self-Exam Applications. Webfoot Tech. 2016 Aug 25. URL: https://appadvice.com/app/breast-self-exam/498271480 [accessed 2016-08-25]
48. Mendoza L. Keep a Breast Apps. 2016. URL: http://www.keep-a-breast.org/blog/keep-abreast-and-check-your-self-breast-cancer-mob/ [accessed 2016-08-25]
49. National Breast Cancer Foundation Inc. Beyond The Shock Apps. 2016 Aug 25. URL: http://www.nationalbreastcancer.org/nbcf-programs/beyond-the-shock [accessed 2016-08-25]
50. Dr. K's Breast Checker Application. female.com. URL: https://www.female.com.au/dr-ks-breast-checker.htm [accessed 2016-08-25]
51. Daisy Wheel Application. The Get In Touch Foundation. 2016. URL: http://getintouchfoundation.org/girls-program/daisy-wheel/ [accessed 2016-08-25]
52. Breast Control Application. apkpure. URL: https://apkpure.com/breast-control/com.thebreastcontrol.healthyapp [accessed 2016-08-27]
53. Ludic - Interactive App for Breast Examination. Makna. 1995. URL: http://makna.org.my/services/ludic/ [accessed 2016-08-27]
54. Lewicki RJ, Bunker BB. Trust in relationships: A model of development and decline. In: APA PhsycNet. Columbus, Ohio: Max M Fisher College of Business, Ohio State University; Jun 12, 2008:3275-3287.

55. Wu T, Deng Z, Feng Z, Gaskin DJ, Zhang D, Wang R. The Effect of Doctor-Consumer Interaction on Social Media on Consumers' Health Behaviors: Cross-Sectional Study. J Med Internet Res 2018 Feb 28;20(2):e73 [FREE Full text] [doi: 10.2196/jmir.9003] [Medline: 29490892]
56. Daskivich T, Luu M, Noah B, Fuller G, Anger J, Spiegel B. Differences in Online Consumer Ratings of Health Care Providers Across Medical, Surgical, and Allied Health Specialties: Observational Study of 212,933 Providers. J Med Internet Res 2018 May 09;20(5):e176 [FREE Full text] [doi: 10.2196/jmir.9160] [Medline: 29743150]
57. Wong S, Cooper P. Reliability and validity of the explanatory sequential design of mixed methods adopted to explore the influences on online learning in Hong Kong bilingual cyber higher education. Int J Cyber Soc Educ Pages 2016;9(2):45-64 [FREE Full text]
58. Koch T, Kralik D. Participatory Action Research in Health Care. Singapore: Blackwell Publishing Inc; Feb 07, 2009:177-202.
59. Khana R. BSE system. Jakarta, Indonesia URL: https://breastselfexam.org [accessed 2019-11-12]
60. Abbas R, Carroll N, Richardson I, Beecham S. The Need for Trustworthiness Models in Healthcare Software Solutions. 2017 Presented at: Proceedings Of The 10th International Joint Conference On Biomedical Engineering Systems And Technologies; 2017; Porto, Portugal p. 451-456 URL: https://www.scitepress.org/Papers/2017/62499/62499.pdf [doi: 10.5220/0006249904510456]
61. Montague EN, Winchester WW, Kleiner BM. Trust in Medical Technology by Patients and Health Care Providers in Obstetric Work Systems. Behav Inf Technol 2010 Sep;29(5):541-554 [FREE Full text] [doi: 10.1080/01449291003752914] [Medline: 20802836]
62. Dibben M, Morris S, Lean M. Situational trust and co-operative partnerships between physicians and their patients: a theoretical explanation transferable from business practice. QJM 2000 Jan;93(1):55-61. [doi: 10.1093/qjmed/93.1.55] [Medline: 10623783]
63. Velsen LV, Tabak M, Hermens H. Measuring patient trust in telemedicine services: Development of a survey instrument and its validation for an anticoagulation web-service. Int J Med Inform 2017 Jan;97:52-58. [doi: 10.1016/j.ijmedinf.2016.09.009] [Medline: 27919395]
64. Lewicki RJ, Tomlinson EC, Gillespie N. Models of Interpersonal Trust Development: Theoretical Approaches, Empirical Evidence, and Future Directions. J Manage 2016 Jul;32(6):991-1022. [doi: 10.1177/0149206306294405]
65. Zahedi F, Song J. Dynamics of Trust Revision: Using Health Infomediaries. J Manag Inf Syst 2014 Dec 08;24(4):225-248. [doi: 10.2753/MIS0742-1222240409]
66. Corritore C, Wiedenbeck S, Kracher B, Marble R. Online Trust and Health Information Websites. Int J Technol Hum Interact 2007;8(4):92-115 [FREE Full text] [doi: 10.4018/jthi.2012100106]
67. Lee MKO, Turban E. A Trust Model for Consumer Internet Shopping. Int J Electr Comm 2014 Dec 23;6(1):75-91 [FREE Full text] [doi: 10.1080/10864415.2001.11044227]
68. Nettleton D. Selection of Variables and Factor Derivation. In: Commercial Data Mining. Sebastopol, CA: O’Reilly; 2014:79-104.
69. Boslaugh S, Watters PA. Statistics in A Nutshell. Sebastopol, CA: O’Reilly; 2008:77-101.
70. Stangroom J. Social Science Statistics. 2020. URL: https://www.socscistatistics.com/tests/pearson/ [accessed 2020-07-01]
71. Pallant J. SPSS Survival Manual. Maidenhead, UK: Open University Press-McGraw Hill Education; 2010.
72. Goforth C. Using and Interpreting Cronbach’s Alpha. University of Virginia Library URL: http://data.library.virginia.edu/using-and-interpreting-cronbachs-alpha/ [accessed 2017-05-15]
73. Anderson LA, Dedrick RF. Development of the Trust in Physician scale: a measure to assess interpersonal trust in patient-physician relationships. Psychol Rep 1990 Dec;67(3 Pt 2):1091-1100. [doi: 10.2466/pr0.1990.67.3f.1091] [Medline: 2084735]

Abbreviations

ARV: average rating value
TISoN: trust inference for social networks
Wa: weight of action
Wc: weight of category
Wp: weight of post

Edited by C Lovis, G Eysenbach; submitted 19.06.20; peer-reviewed by S Sarbadhikari, JJ Wu; comments to author 19.07.20; revised version received 09.08.20; accepted 17.08.20; published 23.09.20.

Please cite as:
Khana R, Mahinderjit Singh M, Damanhoori F, Mustaffa N
Breast Self-Examination System Using Multifaceted Trustworthiness: Observational Study
JMIR Med Inform 2020;8(9):e21584
URL: https://medinform.jmir.org/2020/9/e21584
doi:10.2196/21584
PMID:32965225

Original Paper

How Specialist Aftercare Impacts Long-Term Readmission Risks in Elderly Patients With Metabolic, Cardiac, and Chronic Obstructive Pulmonary Diseases: Cohort Study Using Administrative Data

Michaela Kaleta[1,2], MSc; Thomas Niederkrotenthaler[3], PhD; Alexandra Kautzky-Willer[4,5], MD; Peter Klimek[1,2], PhD

[1]Section for Science of Complex Systems, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria

[2]Complexity Science Hub Vienna, Vienna, Austria

[3]Department of Social and Preventive Medicine, Center for Public Health, Medical University of Vienna, Vienna, Austria

[4]Department of Internal Medicine III, Clinical Division of Endocrinology and Metabolism, Medical University of Vienna, Vienna, Austria

[5]Gender Institute, Gars am Kamp, Austria

Corresponding Author:
Peter Klimek, PhD
Section for Science of Complex Systems
Center for Medical Statistics, Informatics and Intelligent Systems
Medical University of Vienna
Spitalgasse 23, BT86
Vienna, 1090
Austria
Phone: 43 140160 ext 36255
Email: peter.klimek@meduniwien.ac.at

Abstract

Background: The health state of elderly patients is typically characterized by multiple co-occurring diseases requiring the involvement of several types of health care providers.

Objective: We aimed to quantify the benefit for multimorbid patients from seeking specialist care in terms of long-term readmission risks.

Methods: From an administrative database, we identified 225,238 elderly patients with 97 different diagnosis (ICD-10 codes) from hospital stays and contact with 13 medical specialties. For each diagnosis associated with the first hospital stay, we used multiple logistic regression analysis to quantify the sex-specific and age-adjusted long-term all-cause readmission risk (hospitalizations occurring between 3 months and 3 years after the first admission) and how specialist contact impacts these risks.

Results: Men have a higher readmission risk than women (mean difference over all first diagnoses 1.9%, *P*<.001), but similar reduction in readmission risk after receiving specialist care. Specialist care can reduce readmission risk by almost 50%. We found the greatest reductions in risk when the first hospital stay was associated with diagnoses corresponding to complex chronic diseases such as acute myocardial infarction (57.6% reduction in readmission risk, SE 7.6% for men [m]; 55.9% reduction, SE 9.8% for women [w]), diabetic and other retinopathies (m: 62.3%, SE 8.0; w: 60.1%, SE 8.4%), chronic obstructive pulmonary disease (m: 63.9%, SE 7.8%; w: 58.1%, SE 7.5%), disorders of lipoprotein metabolism (m: 64.7%, SE 3.7%; w: 63.8%, SE 4.0%), and chronic ischemic heart diseases (m: 63.6%, SE 3.1%; w: 65.4%, SE 3.0%).

Conclusions: Specialist care can greatly reduce long-term readmission risk for patients with chronic and multimorbid diseases. Further research is needed to identify the specific reasons for these findings and to understand the detected sex-specific differences.

(JMIR Med Inform 2020;8(9):e18147) doi:10.2196/18147

KEYWORDS

multimorbity; patient-sharing networks; network analysis; gender medicine; chronic disease; morbidity; elderly; older adults; cohort study

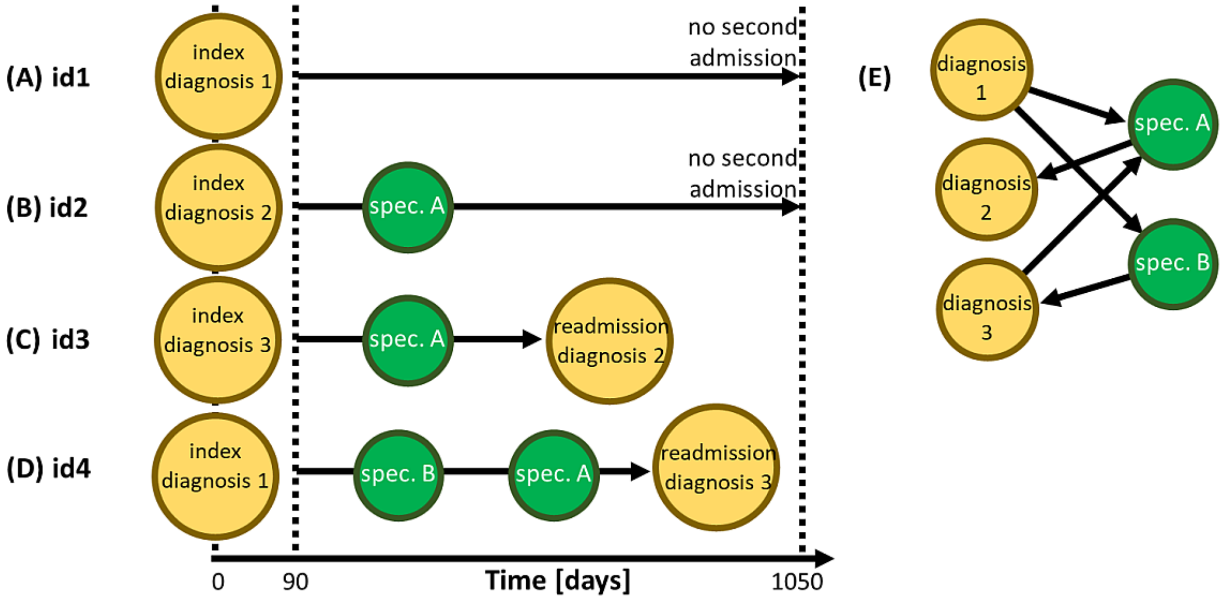
Introduction

The health of elderly patients is typically characterized by more than one disorder [1]. More than 10% of all Austrians aged >50 years accumulate more than 10 different diagnoses over a period of 2 years [2]. Treatment of such highly multimorbid patients often requires the involvement of many different care providers [3,4] taking age-specific and sex-specific differences in physiology and health care–seeking behavior into account [5]. Yet, most health care systems are still configured to treat individual diseases rather than individual multimorbid patients [6]. It is therefore an open challenge to ensure sufficient care coordination among different types of health care providers to adequately treat an aging population [7]. Most findings on long-term readmission risk so far have had an isolated focus on single diseases — for instance, pneumonia [8], colorectal surgery [9], depression [10], or chronic obstructive pulmonary disorder (COPD) [11] — and take only a few predictor variables (eg, medical history) into account [12]. In addition, many studies focus on short-term (30-day or 90-day) readmissions, whereas studies of longer-term risks for patients with chronic complex disorders such as diabetes remain underrepresented in the literature [13].

Digitalization in the health sector has led to increasing availability of observational health care data like electronic health records or medical claims data [14]. The emerging field of network medicine [15,16] strives to leverage such data to improve our understanding of multimorbidity [17] and how care providers coordinate themselves in the treatment of such patients [18]. Complex multimorbid health states of patients can be conceptualized by means of networks (collections of nodes connected by links) in which diseases are nodes that are linked if they tend to co-occur in patients. These comorbidity networks can be used to predict future changes in health as patients are most likely to acquire diseases in close network-proximity to those that they already have [2,19,20]. Networks of care providers have been studied through the analysis of patient-sharing relations [21]. In such patient-sharing networks, providers are represented as nodes connected by links that indicate patient flow between them [18]. It has been shown that the structure of such networks can be related to variations in treatment outcomes [18], cost and intensity of care [22,23], as well as spending for and utilization of health services [24,25].

In this work, we quantified for the first time the long-term readmission risk for 97 frequent diagnoses (ICD-10 3-digit codes) associated with the first hospital stay as a function of age, sex, and the involvement of 13 different types of medical specialists. We propose a novel network statistical modelling approach illustrated in Figure 1. Using an administrative database containing data for almost 2 million patients, we identified all patients aged >50 years with at least one hospital admission (index hospitalization) and followed them for >3 years. There are 4 different types of trajectories that a patient can take after first admission (index hospitalization): (1) no second admission and no specialist contact over the next 3 years (see patient 1 in Figure 1A), (2) one (or several) specialist contact but no second admission (see patient 2 in Figure 1B), (3) second admission without specialist contact (see patient 3 in Figure 1C), or (4) second admission with specialist contact (see patient 4 in Figure 1D). By considering patients with the same diagnosis from the first admission (index diagnosis) and adjusting for age, we can then estimate separately for men and women how contact with a specific type of provider changes the readmission risk for any combination of index diagnosis, readmission diagnosis, and type of specialty.

Figure 1. Illustration of the methodological approach showing examples of timelines for individual patients. Over 3 years following the index hospitalization (yellow circles), patients without a readmission (yellow circles labeled readmission diagnoses) (A) do or (B) do not have specialist contact (green circles), while patients with a readmission (C) do not or do (D) have at least one specialist contact. (E) Trajectories of readmitted patients can be visualized as networks with patient flow between 2 node types (diagnoses and specialists). spec: specialist.



Methods

Study Population

We used a pseudonymized medical claims research dataset from a social insurance carrier in Austria covering the state of Lower Austria [26-28]. The dataset contains 1,861,971 individuals in total who consulted at least one health care provider between January 1, 2006 and March 31, 2012 and were alive during that period. Dead individuals were not included in the data. We extracted the study base, which consisted of all patients with known age and sex that were older than 50 years at the beginning of the observation window and had at least one admission in the time range between January 1, 2006 and December 31, 2008 (n=225,238). For each of these patients, we assessed contact with medical specialists and ICD-10 codes associated with their hospital admissions.

Index Hospitalization

We considered main and secondary diagnoses (3-digit ICD-10 codes from the range A01-N99) from first admissions with at least 1000 occurrences, disregarding codes that are not specific for disorders, such as general examinations, child births, congenital malformations, or unspecific symptoms (Multimedia Appendix 1).

Readmission

For each stay, we identified whether each patient from the study population had a subsequent hospital admission in the time window between 90 and 1050 days after the index hospitalization; if yes, the ICD-10 code of the associated main diagnosis (readmission diagnosis) was noted.

Diagnosis Combinations

The ICD-10 codes of the index diagnosis (d_1) and readmission diagnosis (d_2) form a diagnosis combination: $D=(d_1, d_2)$. All readmission risks were computed with logistic regression models (see the next section) using patients with the same diagnosis combinations (if readmitted) or d_1 as the index diagnosis (no readmission). The models were stratified by sex and considered all diagnosis combinations that occurred in at least 50 cases.

Readmission Risk

The readmission risk measures how likely patients with index diagnosis d_1 were readmitted because of any other diagnosis. The risk was measured separately for men and women. For each patient, we introduced a binary dummy variable for whether the patient was readmitted. We performed logistic regression analysis with this response variable and patient age as the predictor variable. To age-adjust the male (m) and female (f) diagnosis-specific readmission risk (P_{diag}[m/f, d_1]), we evaluated the logistic regression model for the mean population age.

Probabilities of Contact With Medical Specialties

For each patient, we assessed how likely a contact with different types of medical specialists occurred between the index and readmission diagnoses (readmitted) or over a follow-up period of 3 years (controls). Separately for men and women, we performed logistic regression analysis with a dummy variable for contact with a specialty as a response and patient age as the predictor variable. We evaluated the model for the mean population age to obtain the probabilities of contact with a specialist *s* for men and women: $P_{\text{spec}}(\text{m/f},s)$. We included the following specialties: ophthalmology; surgery; dermatovenereal diseases; obstetrics and gynecology; ear, nose, and throat (ENT); pulmonary diseases; neurology; orthopedics; physiotherapy; radiology; accident surgery; urology; labs; psychotherapy and clinical psychology; psychiatry; internal medicine; and outpatient hospital contacts [29].

Health Care Network Construction

A specific subset of patient flow from hospital (re-)admissions to contact with a specialty is summarized graphically in a network representation (see Figure 1E). For each diagnosis combination *D* and specialist contact of type *s* meeting our inclusion criteria, we assumed a direct link in the network from the index diagnosis to the specialty and from the specialty to the readmission diagnosis. For each link, we evaluated the ratio of men to women that followed it. As the full network is too dense to be meaningfully visualized, we applied a standard network filtering method to extract the links that were most significant for each node (type of care provider or diagnosis), the so-called network backbone, by overlapping its maximum spanning tree with the disparity-filtered network [29].

Relative Readmission Risk

Relative readmission risk measures the change in readmission risk associated with contact with a specialty. For each diagnosis combination *D* for men and women separately, we performed logistic regression analysis of whether a readmission because of diagnosis d_2 occurred given that the first diagnosis was d_1. The independent variables were age and a dummy variable for contact with a specialty. This binary dummy variable *s* encoded whether a patient had at least one contact (s=1) between the index and readmission diagnoses (readmitted) or within the 3-year follow-up interval after the index diagnosis (control) or whether no such contact occurred (s=0). For each diagnosis combination *D* and specialty *s*, we obtained the contact-dependent readmission risk $Q(\text{m/f},D,s)$ for men/women by evaluating their models for mean population age. To measure the impact of a contact with specialty *s* on the readmission risk, we evaluated these regression models for patients of mean age that had (s=1) or had not (s=0) such a contact and computed the relative readmission risk, $RR(\text{m/f}, D,s)=Q(\text{m/f},D,s=1)/Q(\text{m/f},D,s=0)$. In terms of the patient timelines in Figure 1, $RR(\text{m/f}, D,s)$ is related to the ratio of frequencies of trajectories (D) to (B), relative to the ratio of trajectories (C) to (A). The *diagnosis-specific* relative readmission risk for men/women, $RR_{\text{diag}}(\text{m/f}, d)$, is given by the medians of $RR(\text{m/f}, D,s)$ over all combinations of readmission diagnoses d_2 and contacts *s*. The *contact-specific* relative readmission risk for specialty *s*, $RR_{\text{spec}}(\text{m/f}, s)$, for men/women is given by the medians of $RR(\text{m/f}, D,s)$ over all diagnosis combinations *D*.

Significance, Multiple Testing, and Robustness Tests

Whether a diagnosis-specific readmission risk is significantly different from 1 was assessed by comparing all related readmission risks that included a specific type of contact (Q[m/f,D,s=1]) with the corresponding risks that did not include such a contact (Q[m/f,D,s=0]). We used a *t* test or sign test depending on whether the individual readmission risks were or were not normally distributed, respectively; normality was assessed by means of a Kolmogorov-Smirnov test. We corrected for multiple testing by controlling the false discovery rate at level α using the Benjamini-Hochberg procedure. To study the robustness of our results, we considered variations of the (1) follow-up interval for the readmission to occur (from 3 years to 1.5 years), (2) minimal number of cases with a diagnoses combination D (from 50 to 25 cases), and (3) inclusion of all patients aged <100 years.

Results

Descriptive statistics of the study population are shown in Table 1. The study population was skewed toward the female sex (130,968/225,238, 58%) with average ages (taken at the beginning of the observation window) of 65 years (men) and 68 years (women). With SDs of 9.7 years (men) and 11 years (women), both sexes had similar and rather broad age distributions. Our inclusion criteria resulted in 97 diagnoses (ICD-10 codes) and 13 different types of specialists. Average numbers of diagnoses and types of specialists involved in the treatment were similar between men and women.

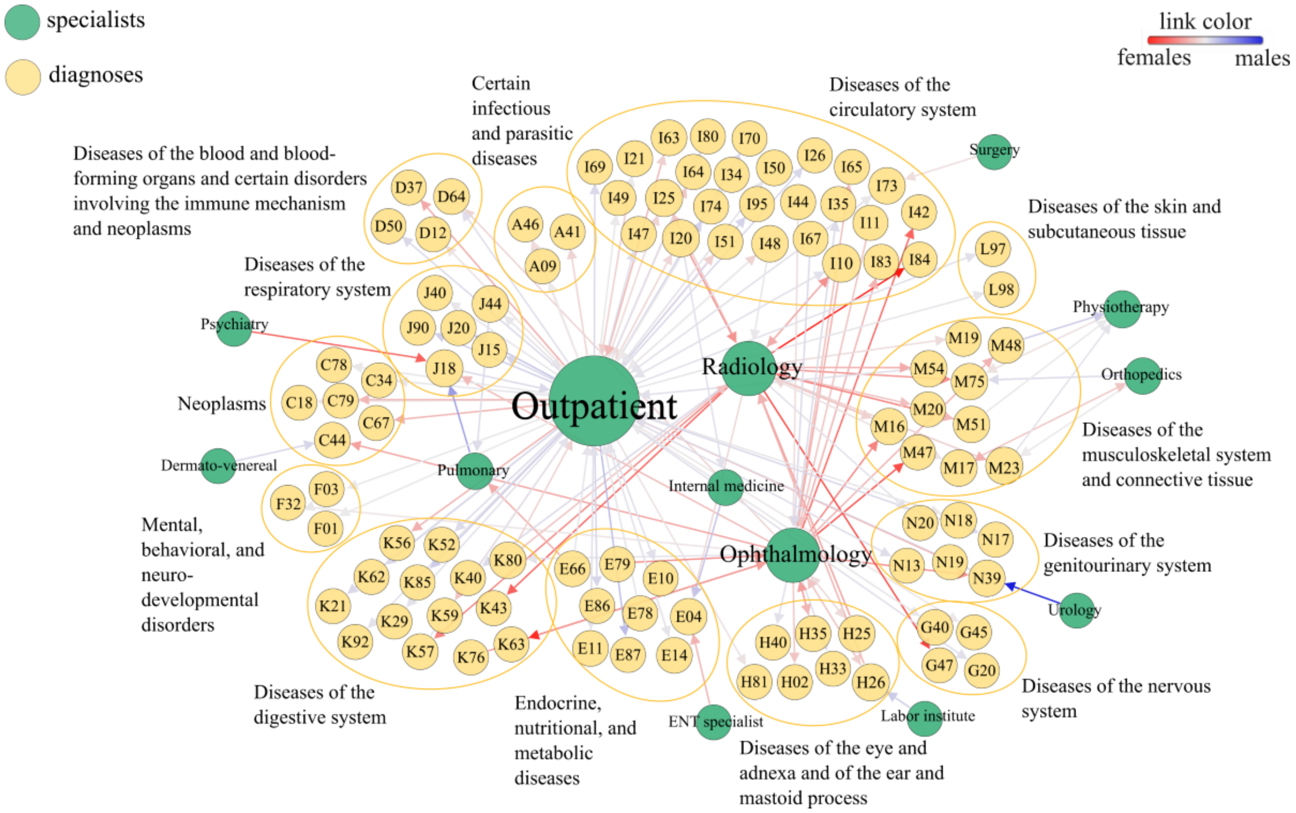
Table 1. Descriptive statistics of the study population.

Variable	Men (n=94,270)	Women (n=130,968)	Entire sample (n=225,238)
Age (years), mean (SD)	65 (9.7)	68 (11.0)	67 (10.0)
Number of diagnoses, mean (SD)	5.1 (4.4)	4.9 (4.4)	5.0 (4.4)
Number of types of specialist, mean (SD)	3.1 (2.8)	3.2 (2.8)	3.1 (2.8)

Network Visualization

A graphical summary of our results is shown in Figure 2 in the form of a network as described in Figure 1E. The node size correlates with the out-degree of the nodes, and the link color shows the ratio of men to women that follow a certain link. Medical specialists with the highest out-degree (connections to different diagnoses) provided outpatient treatments associated with almost the entire spectrum of diagnoses, as well as radiology and ophthalmology with mostly female-dominated links with diseases of the circulatory and musculoskeletal systems. Several specialties were associated with a single diagnosis code, such as dermatovenereal diseases and skin cancer, ENT specialists and nontoxic goiter, urology and urinary tract infections, and psychiatry and pneumonia. To illustrate the results “behind” the network in Figure 2, let us examine the link from psychiatry to pneumonia (J18). Pneumonia was the readmission diagnosis for patient trajectories that included contacts with psychiatry for several index diagnoses. In all but one case, we found reduced relative readmission risks for pneumonia for both men and women, ranging from 46% for women with hypertension to 94% for women with atrial fibrillation; for men with urinary tract infections only, we found a 1% increase in readmission risk. Note that Figure 2 only shows a filtered version of this network. For instance, for type 2 diabetes (E11), there is only a link to outpatient wards. In addition to general practitioners, patients with diabetes also frequently visited internal medicine, radiology, and physiotherapy, which have been filtered out in Figure 2.

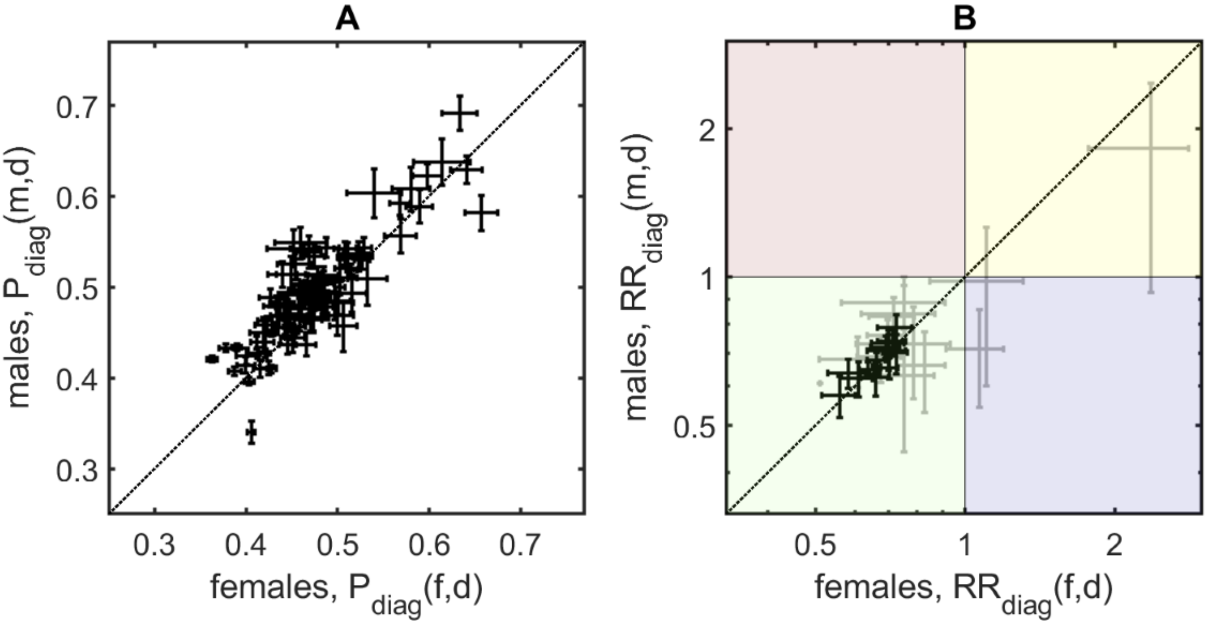
Figure 2. Graphical summary of our results as a network. The network was constructed as described in Figure 1E and filtered for statistical significant links. Node sizes for specialists correlate with the number of outgoing links from the nodes.


specialists
diagnoses
link color
females
males
Certain infectious and parasitic diseases
Diseases of the circulatory system
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism and neoplasms
Diseases of the skin and subcutaneous tissue
Diseases of the respiratory system
Neoplasms
Diseases of the musculoskeletal system and connective tissue
Mental, behavioral, and neuro-developmental disorders
Diseases of the genitourinary system
Diseases of the digestive system
Endocrine, nutritional, and metabolic diseases
Diseases of the eye and adnexa and of the ear and mastoid process
Diseases of the nervous system
Outpatient
Radiology
Ophthalmology
Surgery
Physiotherapy
Orthopedics
Psychiatry
Dermato-venereal
Pulmonary
Internal medicine
Urology
ENT specialist
Labor institute
I63
I80
I70
I69
I21
I64
I34
I50
I26
I65
I49
I25
I74
I95
I44
I35
I73
I47
I20
I51
I48
I67
I11
I42
I10
I83
I84
D37
D64
D50
D12
A46
A41
A09
L97
L98
J40
J44
J90
J20
J15
J18
C78
C34
C18
C79
C67
C44
M19
M48
M54
M75
M20
M51
M16
M47
M17
M23
F03
F32
F01
N20
N18
N17
N13
N19
N39
K56
K52
K80
K62
K85
K40
K21
K29
K59
K43
K92
K57
K76
K63
E66
E79
E10
E86
E78
E04
E11
E87
E14
H40
H35
H25
H33
H81
H02
H26
G40
G45
G47
G20

Results for Diagnosis-Specific Readmission Risks

The long-term all-cause readmission risks for the index diagnoses vary between 30% and 70% (see Figure 3A). We observed the highest readmission risk for secondary neoplasm (C79: 58%, SE 1.8% for men [m]; 66%, SE 2.0% for women [w]; C78: 62%, SE 1.6% [m]; 60%, SE 1.3% [w]) followed by retinopathies including hypertensive retinopathy and macular degeneration (H35: 63%, SE 1.7% [m]; 64%, SE 1.5% [w]) and other retinal disorders such as diabetic retinopathy (H36: 64%, SE 3.1% [m]; 61%, SE 2.5% [w]). Other diagnoses with particularly high readmission risks included colorectal cancer (C18: 69%, SE 1.9% [m]; 63%, SE 1.9% [w]; C20: 61%, SE 2.1% [m]; 58%, SE 2.4% [w]), lung cancer (C34: 59%, SE 1.5% [m]; 59%, SE 2.4% [w]), diabetes (E10: 56%, SE 1.8% [m]; 57%, SE 1.9% [w]), and renal failure (N18: 59%, SE 1.1% [m]; 57%, SE 1.3% [w]; N19: 60.3%, SE 3.0% [m]; 54%, SE 2.7% [w]). In most cases (>70%), the above diagnoses were used as main diagnoses in the index hospitalization except for diabetic retinopathy (H36), which was the main diagnosis in only 33.8% (762/2258) of stays and most frequently occurred as a side diagnosis with type 2 diabetes as the main diagnosis (497/2258, 22.0%). In general, men have higher diagnosis-specific readmission risks than women (mean difference [MD] over all first diagnoses 1.9%, *P*<.001; ie, most points in Figure 3A lie above the diagonal line; see also Multimedia Appendix 1).

Figure 3. Results for diagnosis-specific readmission risks for men and women. Error bars denote SEs; no error bar means only one diagnosis(-specialist) combination contributed to the data point. (A) The diagnosis-specific readmission risks range between 30% and 70%. (B) For several diagnoses, we found significantly decreased (shown in black, as opposed to insignificant results shown in grey) diagnosis-specific relative readmission risks after consulting with medical specialists.



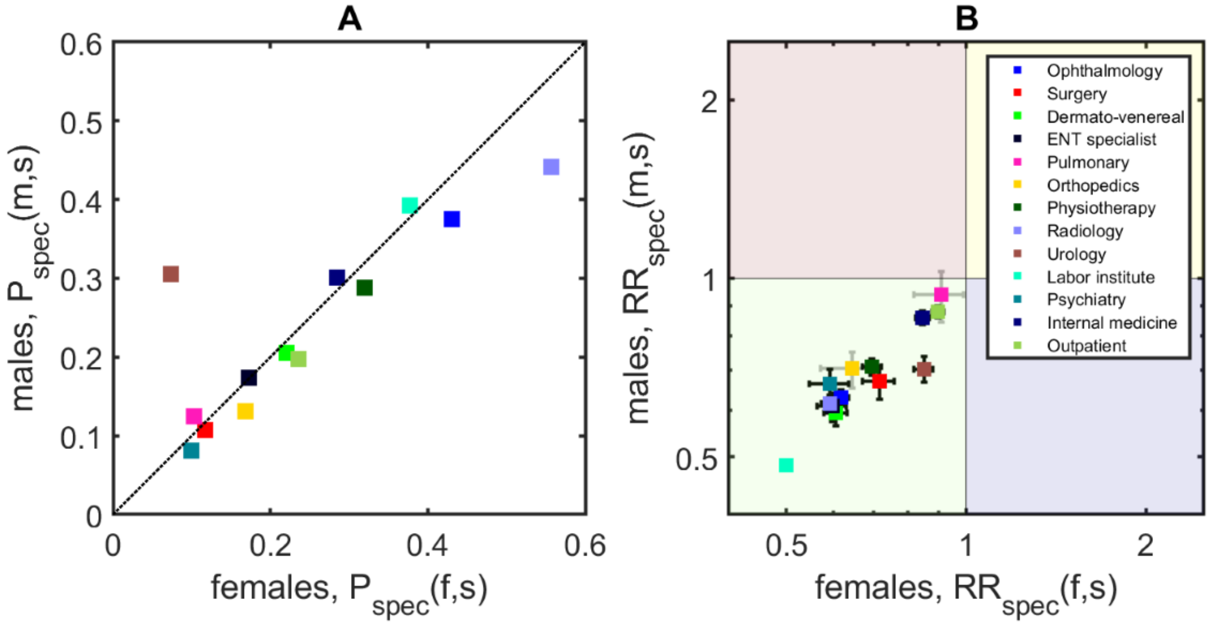
Results for Diagnosis-Specific Relative Readmission Risks

All significant diagnosis-specific readmission risks were <1 for men and women (ie, they lie in the green, bottom left quadrant in Figure 3B). We found the greatest significant reductions in readmission risks upon contact(s) with medical specialists for acute myocardial infarction (I21, male patients with contacts show a reduced readmission risk of 57.6%, SE 7.6% when compared to the risk of patients without such contacts; 55.9%, SE 9.8% [w]), diabetic and other retinal disorders (H35: 62.3%, SE 8.0 [m]; 60.1%, SE 8.4% [w]), COPD (J44: 63.9%, SE 7.8% [m]; 58.1%, SE 7.5% [w]), disorders of lipoprotein metabolism (E78: 64.7%, SE 3.7% [m]; 63.8%, SE 4.0% [w]), and chronic ischemic heart diseases (I25: 63.6%, SE 3.1% [m]; 65.4%, SE 3.0% [w]). There were no significant differences in risk reductions between men and women (MD 1.8%, *P*=.28; see also Multimedia Appendix 2).

Results for Specialist-Specific Readmission Risks

Figure 4A shows the probabilities for men and women in our study population to have contact with a specialty after index admission. Depending on the specialty, these probabilities range from around 8% for psychiatry (10.0%, SE 0.00 % [w]; 8.1%, SE 0.00% [m]) to 56% for radiology (55.7%, SE 0.02% [w]; 44.1%, SE 0.01% [m]). We found no significant differences between men and women in their contact probabilities (MD 0.28%, *P*=.58) with an outlier result for contacts with urology (7.4%, SE 0.00% [w]; 30.1%, SE 0.01% [m]).

Figure 4. Results for specialist-specific readmission risks for men and women. (A) Contact probabilities with certain specialties (colors) range between 8% and 56%. (B) For most specialties, we found significantly reduced readmission risks (black) after contact.



Results for Specialist-Specific Relative Readmission Risks

After contact, all specialties tend to show reduced readmission risks for both men and women; see Figure 4B where all points lie in the green (bottom left) quadrant. There are only 2 specialties for which the readmission risks were not significantly reduced, namely for pulmonary disease specialists and orthopedics. We found the greatest reductions for lab testing (50.1%, SE 1.7% [w]; 48.5%, SE 1.8% [m]), radiology (59.2%, SE 3.2% [w]; 61.6%, SE 3.0% [m]), psychiatry (59.3%, SE 6.7% [w]; 66.3%, SE 6.7% [m]), dermatovenereal diseases (60.5%, SE 4.7% [w]; 59.3%, SE 4.6% [m]), and ENT specialists (59.8%, SE 5.6% [w]; 60.9%, SE 5.6% [m]). Overall, men and women showed similar risk reductions (MD –0.2%, *P*=.96).

Robustness

Our main result of strongly reduced relative readmission risks remained robust under changes of the parameters in the analysis. We show the results for the relative readmission risks for 3 different robustness tests (reducing the minimal number of cases required for a diagnosis combination from 50 to 25, using an observation window of 1.5 years instead of 3 years, and including patients <50 years old) for diagnosis-specific and specialist-specific relative readmission risks in Multimedia Appendix 3 and Multimedia Appendix 4, respectively. In each case, all significant results correspond to strongly reduced relative risks.

Discussion

In this work, we present a comprehensive analysis of sex-specific long-term readmission risks (measured from 90 days to 3 years after the index hospitalization) where we systematically tested how contact with 13 medical specialties impacts readmission risks for 97 diagnoses associated with the first stay. The network visualization reveals that our analysis is indeed based on meaningful flows of patients between different care settings. For instance, we found a dominant flow from lab testing to senile cataract consistent with the fact that such testing is often performed in preoperative screenings to detect risk factors for complications such as diabetes. There are multiple meaningful flows from radiology to musculoskeletal diseases, a link from dermatovenereal diseases to skin cancer, or from ENT specialists to nontoxic goiter. In all cases, our results mean that patients that had contact with a specialist showed a tendency later for reduced readmission risks for the given diagnoses compared to patients without such contact (ie, the links in the network do not just show frequent flows of patient but specialty-diagnosis combinations that contribute to the observed reductions in readmission risks). Other links were less clear. For instance, we found a tendency that contacts with psychiatry reduce readmission risks for pneumonia. Recent epidemiological findings suggest that depression is indeed a risk factor for hospitalization due to pneumonia [30] and that psychological distress is related with a higher risk of pneumonia [31]. Furthermore, lifestyle factors (eg, substance abuse), psychiatric conditions (patients' compromised ability to recognize health problems) as well as side effects of antipsychotics (worsened respiratory muscle functioning) might cause this association [32]. Our results could therefore indicate that contact with psychiatric specialists mitigate these risk factors and thereby reduce pneumonia-related readmissions.

Overall, we found the largest readmission risks after hospital stays associated with chronic complex diseases for which high readmission rates have already been described in the literature. These diseases include various types of cancer including rectal cancer, with a 30-day readmission rate of 10.1% [33], and lung cancer, with a 30-day readmission rate of 13% and 90-day rate of 22% [34]. Diabetic and other retinopathies often occur with type 2 diabetes as the main diagnosis, for which 30-day readmission rates are 8.5%-13.5% [13]. For chronic kidney disease, the 90-day readmission rate has been estimated at 11.7% [35]. These risks cannot be directly compared to the long-term readmission risk (where we exclude readmissions within the first 90 days) considered in our study.

The involvement of medical specialists reduces the need for long-term readmissions by up to 50% depending on the index diagnosis. Chronic complex diseases are among those for which we observe the strongest reductions in readmission risk after contact with medical specialists. Our observation of the greatest reduction for patients with acute myocardial infarction is in line with findings of reduced mortality (up to 19% over an 18-month follow-up) for patients with myocardial infarct who receive follow-up care by cardiologists and internists when compared to patients without such contact [36]. The second greatest reduction was observed for diabetic and other retinopathies (H35), which often occurred with type 2 diabetes as the main diagnosis. These findings are in line with reports that a lack of postdischarge outpatient visits in patients with diabetes is one of the strongest risk factors for short-term (30-day) readmissions [37] and that postdischarge office visits to adjust the diabetes regimen contribute to a decreased risk of short-term readmission [38]. While there is mixed evidence to which extent poor glycemic control is also a risk factor for longer-term readmission risk [13], our findings clearly show that specialist care after discharge is related to a strongly significant reduction in readmission risks of down to 62% (men) and 60% (women) compared to patients without such contacts. Similar diabetes-related observations might be relevant for patients with hypercholesterolemia and hyperlipidemia (E78), which are frequent diabetic comorbidities, who all showed significant reductions in readmission risk. We found that the contact related with reductions in readmission risk for diabetes patients was concentrated on visits at outpatient wards, internal medicine, and radiology, among other specialists. Diabetes is indeed a complex disease requiring the involvement of multiple types of health care providers. Treatment should take place in strict agreement with the corresponding guidelines, including quarterly physician visits and a high continuity of care, to minimize the risk for diabetic complications.

For COPD patients, it has been observed that the involvement of physiotherapists and various pulmonary and respiratory specialists can reduce readmission risks, which is consistent with our finding of a strongly reduced readmission risk for patients with COPD [39]. Finally, in relation to our results for the relative readmission risk for ischemic heart disease, it was reported that patients had significantly lower 60-day readmission rates when they were treated by multiple providers, including surgeons and nonsurgeons [40].

Men and women had comparable probabilities of contacting different types of medical specialists after the index admission; there was no significant difference in how likely men and women seek specialists. A Swedish register study found that most of the sex differences in health care consumption can indeed be explained by an increased level of reproduction-associated care (not considered in our work) and women's higher share in mental and behavioral disorders and diseases of the musculoskeletal system [41]. We found contact probabilities that range from around 8% (psychiatry) to more than 56% (radiology). We did not include primary care providers (eg, general practitioners) in this analysis as almost everyone in the study population had such contacts; therefore, their contact probabilities were close to 100%. After contact with specialists, we observed significantly reduced readmission risks (risk reductions of up to 50%) for almost all specialties, including lab testing, radiology, psychiatry, dermatovenereal diseases, and ENT specialists, whereas pulmonary disease specialists and orthopedics show a rather risk-neutral profile. These findings might reflect that follow-up by specialists generally means more tailored risk detection and improved disease management, but also that patients seeking care from specialists might be more engaged and vigilant compared to those patients that do not seek specialist care. In the present form, our analysis does not allow us to disentangle these effects of targeted prevention and health care–seeking patient behavior.

In terms of sex differences, we found that men overall have higher readmission risks than women. While the raw readmission frequencies were similar for men and women (Table 1), the diagnosis-specific analysis clearly revealed that men have increased readmission risks after adjusting for age and pre-existing condition (index diagnosis). Our findings also showed that the difference between men and women in readmission risks are not due to differences in how likely they are to seek contact with a specialist. In the following, we give two plausible mechanisms that could in principle contribute to the observed sex biases (or lack thereof). First, men might utilize the health care system only at more severe stages of disease compared with women, therefore also showing higher readmission risks. Second, it could be that women are more compliant when consulting specialists and therefore show better outcomes (ie, reduced readmission risks). However, the second explanation is at variance with our result that, after having had a specific type of contact, men and women show similar reductions in readmission risk. The assumption that these sex biases are indeed due to differences in utilization is further corroborated by findings of delayed health-seeking behavior in men compared with women [42].

Our work has several limitations that mostly relate to the administrative dataset used. We have no information on which kind of procedures were performed during the admission and contact with a specialist and no knowledge on results from medical tests other than the diagnosis codes. We cannot guarantee that we indeed observed all admissions of the study population, especially since the data cover only a region of Austria. However, as this bias should have similar effects on index hospitalizations and readmissions, as well as men and women, such coverage issues should only have a limited impact on comparisons of readmission risks. Similar biases might influence the probabilities of contact with specialists. We only considered whether at least one contact took place, but not the specific number of contacts.

To conclude, our results emphasize that specialist aftercare can provide a strong contribution to the reduction in long-term rehospitalization. These effects vary substantially across diagnoses and are most pronounced for outcomes such as myocardial infarction where specialist treatment has already been shown to improve survival. While we find tentative evidence for delayed health-seeking behavior towards medical specialists in men when compared to women, both sexes show similar levels of readmission risk reduction after specialist care. These sex biases require further research into their physiological, biological, social, and psychological causative processes.

Acknowledgments

We acknowledge helpful discussions with Sarah Etter and Stefan Thurner. This study was supported by the WWTF “Mathematics and…” project MA16-045 and FFG project 857136.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Diagnoses-specific results and their SEs for the contact-independent diagnose-specific readmission risks for males/females, Pdiag(m/f,d), and the contact-dependent relative readmission risks RRdiag(m/f, d).

[DOCX File , 28 KB - medinform_v8i9e18147_app1.docx]

Multimedia Appendix 2

Results for the contact probabilities with different types of specialists for females, Pspec(f,s), and males, Pspec(m,s), their SEs, and the contact dependent relative readmission risks RRspec(m/f, s).

[DOCX File , 18 KB - medinform_v8i9e18147_app2.docx]

Multimedia Appendix 3

Results of the robustness tests for diagnoses-specific relative readmission risks.

[DOCX File , 124 KB - medinform_v8i9e18147_app3.docx]

Multimedia Appendix 4

Results of the robustness tests for specialist-specific relative readmission risks.

[DOCX File , 110 KB - medinform_v8i9e18147_app4.docx]

References

1. Marengoni A, Angleman S, Melis R, Mangialasche F, Karp A, Garmen A, et al. Aging with multimorbidity: a systematic review of the literature. Ageing Res Rev 2011 Sep;10(4):430-439 [FREE Full text] [doi: 10.1016/j.arr.2011.03.003] [Medline: 21402176]
2. Chmiel A, Klimek P, Thurner S. Spreading of diseases through comorbidity networks across life and gender. New J. Phys 2014 Nov 14;16(11):115013. [doi: 10.1088/1367-2630/16/11/115013]
3. Calderón-Larrañaga A, Poblador-Plou B, González-Rubio F, Gimeno-Feliu LA, Abad-Díez JM, Prados-Torres A. Multimorbidity, polypharmacy, referrals, and adverse drug events: are we doing things well? Br J Gen Pract 2012 Dec;62(605):e821-e826 [FREE Full text] [doi: 10.3399/bjgp12X659295] [Medline: 23211262]
4. Geroldinger A, Sauter SK, Heinze G, Endel G, Dorda W, Duftschmid G. Mortality and continuity of care - Definitions matter! A cohort study in diabetics. PLoS One 2018 Jan 19;13(1):e0191386 [FREE Full text] [doi: 10.1371/journal.pone.0191386] [Medline: 29351547]
5. Kautzky-Willer A, Harreiter J, Pacini G. Sex and Gender Differences in Risk, Pathophysiology and Complications of Type 2 Diabetes Mellitus. Endocr Rev 2016 Jun;37(3):278-316 [FREE Full text] [doi: 10.1210/er.2015-1137] [Medline: 27159875]
6. Moffat K, Mercer SW. Challenges of managing people with multimorbidity in today's healthcare systems. BMC Fam Pract 2015 Oct 14;16(1):129 [FREE Full text] [doi: 10.1186/s12875-015-0344-4] [Medline: 26462820]
7. Bodenheimer T. Coordinating Care — A Perilous Journey through the Health Care System. N Engl J Med 2008 Mar 06;358(10):1064-1071. [doi: 10.1056/nejmhpr0706165]
8. Weinreich M, Nguyen OK, Wang D, Mayo H, Mortensen EM, Halm EA, et al. Predicting the Risk of Readmission in Pneumonia. A Systematic Review of Model Performance. Annals ATS 2016 Sep;13(9):1607-1614. [doi: 10.1513/annalsats.201602-135sr]
9. Damle RN, Alavi K. Risk factors for 30-d readmission after colorectal surgery: a systematic review. J Surg Res 2016 Jan;200(1):200-207 [FREE Full text] [doi: 10.1016/j.jss.2015.06.052] [Medline: 26216748]

10. Pederson JL, Warkentin LM, Majumdar SR, McAlister FA. Depressive symptoms are associated with higher rates of readmission or mortality after medical hospitalization: A systematic review and meta-analysis. J Hosp Med 2016 May 29;11(5):373-380 [FREE Full text] [doi: 10.1002/jhm.2547] [Medline: 26824220]
11. Lindenauer PK, Dharmarajan K, Qin L, Lin Z, Gershon AS, Krumholz HM. Risk Trajectories of Readmission and Death in the First Year after Hospitalization for Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med 2018 Apr 15;197(8):1009-1017. [doi: 10.1164/rccm.201709-1852oc]
12. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan 17;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
13. Rubin DJ. Hospital readmission of patients with diabetes. Curr Diab Rep 2015 Apr 25;15(4):17 [FREE Full text] [doi: 10.1007/s11892-015-0584-7] [Medline: 25712258]
14. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci U S A 2016 Jul 05;113(27):7329-7336 [FREE Full text] [doi: 10.1073/pnas.1510502113] [Medline: 27274072]
15. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011 Jan 17;12(1):56-68 [FREE Full text] [doi: 10.1038/nrg2918] [Medline: 21164525]
16. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. J Gen Intern Med 2013 Sep 25;28 Suppl 3(S3):S660-S665 [FREE Full text] [doi: 10.1007/s11606-013-2455-8] [Medline: 23797912]
17. Hidalgo CA, Blumm N, Barabási AL, Christakis NA. A dynamic network approach for the study of human phenotypes. PLoS Comput Biol 2009 Apr 10;5(4):e1000353 [FREE Full text] [doi: 10.1371/journal.pcbi.1000353] [Medline: 19360091]
18. Landon BE, Keating NL, Barnett ML, Onnela J, Paul S, O'Malley AJ, et al. Variation in patient-sharing networks of physicians across the United States. JAMA 2012 Jul 18;308(3):265-273 [FREE Full text] [doi: 10.1001/jama.2012.7615] [Medline: 22797644]
19. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun 2014 Jun 24;5(1):4022 [FREE Full text] [doi: 10.1038/ncomms5022] [Medline: 24959948]
20. Kannan V, Swartz F, Kiani NA, Silberberg G, Tsipras G, Gomez-Cabrero D, et al. Conditional Disease Development extracted from Longitudinal Health Care Cohort Data using Layered Network Construction. Sci Rep 2016 May 23;6(1):26170 [FREE Full text] [doi: 10.1038/srep26170] [Medline: 27211115]
21. Lee BY, McGlone SM, Song Y, Avery TR, Eubank S, Chang C, et al. Social Network Analysis of Patient Sharing Among Hospitals in Orange County, California. Am J Public Health 2011 Apr;101(4):707-713. [doi: 10.2105/ajph.2010.202754]
22. Barnett ML, Christakis NA, O'Malley J, Onnela J, Keating NL, Landon BE. Physician Patient-sharing Networks and the Cost and Intensity of Care in US Hospitals. Medical Care 2012;50(2):152-160. [doi: 10.1097/mlr.0b013e31822dcef7]
23. Pollack CE, Weissman GE, Lemke KW, Hussey PS, Weiner JP. Patient sharing among physicians and costs of care: a network analytic approach to care coordination using claims data. J Gen Intern Med 2013 Mar 14;28(3):459-465 [FREE Full text] [doi: 10.1007/s11606-012-2104-7] [Medline: 22696255]
24. Landon BE, Keating NL, Onnela J, Zaslavsky AM, Christakis NA, O'Malley AJ. Patient-Sharing Networks of Physicians and Health Care Utilization and Spending Among Medicare Beneficiaries. JAMA Intern Med 2018 Jan 01;178(1):66-73 [FREE Full text] [doi: 10.1001/jamainternmed.2017.5034] [Medline: 29181504]
25. Robbins R, Seixas A, Schoenthaler A. The nature and scope of patient-sharing network research: a novel, important area for network science. Transl Behav Med 2018 Jul 17;8(4):626-628 [FREE Full text] [doi: 10.1093/tbm/iby052] [Medline: 30016522]
26. Klimek P, Kautzky-Willer A, Chmiel A, Schiller-Frühwirth I, Thurner S. Quantification of diabetes comorbidity risks across life using nation-wide big claims data. PLoS Comput Biol 2015 Apr 9;11(4):e1004125 [FREE Full text] [doi: 10.1371/journal.pcbi.1004125] [Medline: 25855969]
27. Thurner S, Klimek P, Szell M, Duftschmid G, Endel G, Kautzky-Willer A, et al. Quantification of excess risk for diabetes for those born in times of hunger, in an entire population of a nation, across a century. Proc Natl Acad Sci U S A 2013 Mar 19;110(12):4703-4707 [FREE Full text] [doi: 10.1073/pnas.1215626110] [Medline: 23487754]
28. Rinner C, Sauter SK, Endel G, Heinze G, Thurner S, Klimek P, et al. Improving the informational continuity of care in diabetes mellitus treatment with a nationwide Shared EHR system: Estimates from Austrian claims data. Int J Med Inform 2016 Aug;92:44-53 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.05.001] [Medline: 27318070]
29. Serrano MA, Boguñá M, Vespignani A. Extracting the multiscale backbone of complex weighted networks. Proc Natl Acad Sci U S A 2009 Apr 21;106(16):6483-6488 [FREE Full text] [doi: 10.1073/pnas.0808904106] [Medline: 19357301]
30. Davydow DS, Hough CL, Zivin K, Langa KM, Katon WJ. Depression and risk of hospitalization for pneumonia in a cohort study of older Americans. J Psychosom Res 2014 Dec;77(6):528-534 [FREE Full text] [doi: 10.1016/j.jpsychores.2014.08.002] [Medline: 25139125]
31. Hamer M, Kivimaki M, Stamatakis E, Batty GD. Psychological distress and infectious disease mortality in the general population. Brain Behav Immun 2019 Feb;76:280-283 [FREE Full text] [doi: 10.1016/j.bbi.2018.12.011] [Medline: 30579940]

32. Seminog OO, Goldacre MJ. Risk of pneumonia and pneumococcal disease in people with severe mental illness: English record linkage studies. Thorax 2013 Feb 15;68(2):171-176 [FREE Full text] [doi: 10.1136/thoraxjnl-2012-202480] [Medline: 23242947]
33. Doumouras AG, Tsao MW, Saleh F, Hong D. A population-based comparison of 30-day readmission after surgery for colon and rectal cancer: How are they different? J Surg Oncol 2016 Sep 22;114(3):354-360 [FREE Full text] [doi: 10.1002/jso.24334] [Medline: 27334402]
34. Stitzenberg KB, Chang Y, Smith AB, Nielsen ME. Exploring the Burden of Inpatient Readmissions After Major Cancer Surgery. JCO 2015 Feb 10;33(5):455-464. [doi: 10.1200/jco.2014.55.5938]
35. Miric A, Inacio MC, Namba RS. The effect of chronic kidney disease on total hip arthroplasty. J Arthroplasty 2014 Jun;29(6):1225-1230 [FREE Full text] [doi: 10.1016/j.arth.2013.12.031] [Medline: 24556110]
36. Radzimanowski M, Gallowitz C, Müller-Nordhorn J, Rieckmann N, Tenckhoff B. Physician specialty and long-term survival after myocardial infarction - A study including all German statutory health insured patients. Int J Cardiol 2018 Jan 15;251:1-7 [FREE Full text] [doi: 10.1016/j.ijcard.2017.10.048] [Medline: 29092757]
37. Karunakaran A, Zhao H, Rubin DJ. Predischarge and Postdischarge Risk Factors for Hospital Readmission Among Patients With Diabetes. Medical Care 2018;56(7):634-642. [doi: 10.1097/mlr.0000000000000931]
38. Gregory NS, Seley JJ, Dargar SK, Galla N, Gerber LM, Lee JI. Strategies to Prevent Readmission in High-Risk Patients with Diabetes: the Importance of an Interdisciplinary Approach. Curr Diab Rep 2018 Jun 21;18(8):54 [FREE Full text] [doi: 10.1007/s11892-018-1027-z] [Medline: 29931547]
39. Ko FWS, Cheung NK, Rainer TH, Lum C, Wong I, Hui DSC. Comprehensive care programme for patients with chronic obstructive pulmonary disease: a randomised controlled trial. Thorax 2017 Feb 28;72(2):122-128 [FREE Full text] [doi: 10.1136/thoraxjnl-2016-208396] [Medline: 27471050]
40. Hollingsworth JM, Funk RJ, Garrison SA, Owen-Smith J, Kaufman SA, Pagani FD, et al. Association Between Physician Teamwork and Health System Outcomes After Coronary Artery Bypass Grafting. Circ Cardiovasc Qual Outcomes 2016 Nov;9(6):641-648. [doi: 10.1161/circoutcomes.116.002714]
41. Osika Friberg I, Krantz G, Määttä S, Järbrink K. Sex differences in health care consumption in Sweden: A register-based cross-sectional study. Scand J Public Health 2016 May 08;44(3):264-273 [FREE Full text] [doi: 10.1177/1403494815618843] [Medline: 26647097]
42. Galdas PM, Cheater F, Marshall P. Men and health help-seeking behaviour: literature review. J Adv Nurs 2005 Mar;49(6):616-623 [FREE Full text] [doi: 10.1111/j.1365-2648.2004.03331.x] [Medline: 15737222]

Abbreviations

COPD: chronic obstructive pulmonary disorder
ENT: ear, nose, and throat
MD: mean difference

Edited by C Lovis; submitted 06.02.20; peer-reviewed by Q Nguyen, W Zhang; comments to author 14.06.20; revised version received 26.06.20; accepted 28.06.20; published 16.09.20.

Please cite as:
Kaleta M, Niederkrotenthaler T, Kautzky-Willer A, Klimek P
How Specialist Aftercare Impacts Long-Term Readmission Risks in Elderly Patients With Metabolic, Cardiac, and Chronic Obstructive Pulmonary Diseases: Cohort Study Using Administrative Data
JMIR Med Inform 2020;8(9):e18147
URL: http://medinform.jmir.org/2020/9/e18147/
doi:10.2196/18147
PMID:32936077

Original Paper

Human- Versus Machine Learning–Based Triage Using Digitalized Patient Histories in Primary Care: Comparative Study

Artin Entezarjou[1], MD; Anna-Karin Edstedt Bonamy[2,3], PhD, MD; Simon Benjaminsson[4], PhD; Pawel Herman[5*], PhD; Patrik Midlöv[1*], PhD, MD

[1]Center for Primary Health Care Research, Department of Clinical Sciences in Malmö/Family Medicine, Lund University, Malmö, Sweden

[2]Clinical Epidemiology Division, Department of Medicine Solna, Karolinska Institute, Stockholm, Sweden

[3]Doctrin AB, Stockholm, Sweden

[4]Smartera AB, Stockholm, Sweden

[5]Department of Computational Science and Technology, KTH Royal Institute of Technology, Stockholm, Sweden

[*]these authors contributed equally

Corresponding Author:
Artin Entezarjou, MD
Center for Primary Health Care Research
Department of Clinical Sciences in Malmö/Family Medicine
Lund University
Box 50332
Malmö, 202 13
Sweden
Phone: 46 40391400
Email: artin.entezarjou@med.lu.se

Abstract

Background: Smartphones have made it possible for patients to digitally report symptoms before physical primary care visits. Using machine learning (ML), these data offer an opportunity to support decisions about the appropriate level of care (triage).

Objective: The purpose of this study was to explore the interrater reliability between human physicians and an automated ML-based triage method.

Methods: After testing several models, a naïve Bayes triage model was created using data from digital medical histories, capable of classifying digital medical history reports as either in need of urgent physical examination or not in need of urgent physical examination. The model was tested on 300 digital medical history reports and classification was compared with the majority vote of an expert panel of 5 primary care physicians (PCPs). Reliability between raters was measured using both Cohen κ (adjusted for chance agreement) and percentage agreement (not adjusted for chance agreement).

Results: Interrater reliability as measured by Cohen κ was 0.17 when comparing the majority vote of the reference group with the model. Agreement was 74% (138/186) for cases judged not in need of urgent physical examination and 42% (38/90) for cases judged to be in need of urgent physical examination. No specific features linked to the model's triage decision could be identified. Between physicians within the panel, Cohen κ was 0.2. Intrarater reliability when 1 physician retriaged 50 reports resulted in Cohen κ of 0.55.

Conclusions: Low interrater and intrarater agreement in triage decisions among PCPs limits the possibility to use human decisions as a reference for ML to automate triage in primary care.

(JMIR Med Inform 2020;8(9):e18930) doi:10.2196/18930

KEYWORDS

machine learning; artificial intelligence; decision support; primary care; triage

Introduction

Health care digitalization has the potential to mitigate increasing primary care workloads [1,2]. Time-constrained primary care physicians (PCPs) interrupt patient queries within the first 30 seconds of consultations [3], contributing to inadequate gathering of medical histories [4,5]. To reduce PCP workload and to ensure patients are directed to the appropriate level of care, nurse-led telephone triage is commonly used [6,7]. However, nurses face similar time constraints as physicians, which results in incomplete gathering of medical histories [8] and inappropriate levels of care recommended in up to 31% of cases [9,10].

Leveraging the wide use of smartphones, a large portion of patient history can today be acquired before the patient interacts with his/her health care provider. Automated patient interviewing software has been shown to gather reliable and relevant clinical information [11], and may thus save clinicians time and reduce workloads.

Existing "symptom checkers" can provide triage recommendations directly to patients. However, their accuracy is low, ranging from 33% to 78%, with higher accuracy reported only for more acute conditions [12]. Furthermore, patient adherence to symptom checker recommendations seems low at just 65% [13], compared with 81%-100% adherence to advice from triage nurses [7]. Thus, clinician decision-support software may be a better solution for optimizing triage.

With rapid developments in machine learning (ML), labeled automated patient interviewing software data offer a promising opportunity for enhancing triage software accuracy, providing appropriate access to primary care. Recent research shows promising utility of ML to aid in emergency department triage compared with commonly used algorithms [14]. However, the performance of such a system compared with human triage has, to the best of our knowledge, never been evaluated. Furthermore, ML research in the primary care setting is lacking, despite over 60% of health care visits being conducted in primary care [15].

Thus, this study sought to investigate interrater reliability between human physicians and an automated ML-based triage method, as well as evaluating interrater reliability of triage decisions between a panel of physicians assessing the same patient histories from an automated patient interviewing software.

Methods

Context

The automated patient interviewing software technology used in this study (produced by Doctrin AB, Stockholm, Sweden) is being used by several primary care providers in Sweden since 2017. Patients access the platform using their smartphone, tablet, or computer, choosing their chief complaint from a prespecified list. An automated medical history is then taken, allowing patients to briefly formulate ideas, concerns, and expectations in free-form text, and subsequently answer a symptom-specific multiple-choice survey. The software selects suitable subsequent survey questions based on the patient's answers (Table 1).

Table 1. Examples of automated patient interviewing software survey questions. Chosen answers subsequently appear in reports used for triage.

Survey question	Answer format
"How long have you had a cough?"	Short answer: specify number of days, months or years
"How has your cough been since it started"	Multiple choice (one option allowed): "Not changing" "Getting worse" "Improving" "Gone away"
"Do you have any of the following symptoms?"	Multiple choice (multiple options allowed): "Runny nose" "Shortness of breath" "Chest pain" "Sore throat" "Swollen glands" "Fever"
If a patient reports fever: "What was the highest temperature you have had when you measured it?"	Multiple choice: "37°C" […] "Over 40 C"
"How many days in a row have you had fever?"	Short answer: specify number of days

Answers are presented to a PCP as a summarized report for review and further doctor–patient communication may occur asynchronously through a live text chat (eVisit). Physicians can prescribe medications, order laboratory samples, provide patient information, or remain available online for up to 72 hours for conservative management. Anonymized data from the automated patient interviewing software report and subsequent chat are saved in a database used for this study. Clinical decisions regarding triage and treatment are, however, recorded separately in the patient medical record and were not accessible for study.

Data for Classification

Data used in this study were composed of 2 subsets. The first subset consisted of 300 automated patient interviewing software reports labeled by a selected expert PCP with over 10 years of clinical experience and a year of experience with online consultations. The reports represented the 10 most common chief complaints in the platform (common cold, cough, eye redness, genital problems, hay fever, rash, headache, sinus symptoms, sore throat, and urinary tract infections) with an equal marginal distribution between chief complaints. Automated patient interviewing software reports were triaged by the expert PCP to one of 4 levels: (1) Start a digital chat-based consultation; (2) Refer the patient to a primary care center for nonurgent care; (3) Refer the patient to a primary care center for urgent care; or (4) Refer the patient to the emergency department.

The second subset was 300 new automated patient interviewing software reports labeled by a panel of 5 PCPs (1 intern [AE], 2 residents, and 2 specialists). Sample sizes were chosen for feasibility reasons. Each PCP individually triaged automated patient interviewing software reports with an identical distribution of chief complaints as in the first subset. Each automated patient interviewing software report was labeled with a triage level as determined by a majority vote by the panel.

Triage categories in both subsets were then dichotomized into 2 triage levels used for further analyses: (1) No need for urgent physical examination (triage levels 1 and 2) or (2) Need of urgent physical examination (triage levels 3 and 4).

Exclusion Criteria

Because of incorrect formatting of one of the reports in the triage interface used by the panel, 299 automated patient interviewing software reports were triaged instead of 300.

Automated patient interviewing software reports describing cases with an ongoing medical contact or a different chief complaint from the one specified were classified as inappropriate for triage, which occurred in 37 reports classified by at least one panel member. These were manually reviewed by one of the authors (AB) for inclusion or exclusion by expert opinion, resulting in the exclusion of 17 cases from the analysis.

If the panel voting strategy did not result in a majority for 1 triage level, the automated patient interviewing software report was also excluded from the analysis, which occurred in 6 cases.

Initially, 22 automated patient interviewing software reports had missing triage data from some panel members. After applying the exclusion criteria, 16 automated patient interviewing software reports with missing triage data remained for analysis.

Model Analyses

To examine the potential of our ML-based approach for triage, we used the available data and corresponding dichotomized triage categories in a series of classification tests with 3 classifiers: (1) a simple linear naïve Bayes classifier, which assumes statistical independence of input features; (2) logistic regression, commonly used for binary classification problems; and (3) random forest, an ensemble decision tree approach, which is considered particularly suitable for high-dimensional problems.

Because of many questions from the automated patient interviewing software reports only appearing very rarely in the small-sized training data, feature space was reduced by only including those which were used in more than 5% of the training samples. This resulted in 243 features. As a few fields included brief free-form text, the classifiers were trained and tested both with and without information extracted from these text data. Text was handled by first removing common Swedish stop words. The remaining commonly used words appearing in more than 10% of the training samples were included as a bag-of-words model where each word was treated as an input feature to the classifier [16]. This resulted in a total of 53 features.

First, we trained the models on the first subset and tested them in a single pass on the second subset with labels based on the majority vote of the 5 PCPs. We complemented this analysis with a cross-validation approach on the data without text information to better estimate generalization capabilities across the 2 subsets of data. We performed 10-fold cross-validation by dividing the union of the 2 subsets into 10 data clusters, where the mixture of the 2 subsets in 9 out of 10 clusters was used for training and the remaining cluster accounting for 10% (ie, 1/10) served as a test set. By applying this scheme 10 times with different 10% test folds, we could obtain an estimate of the second moment of the generalization classification performance. The cross-validation results were followed up with a nonparametric Friedman test.

We made an attempt at investigating the key input features that had a decisive role in classification. To this end, we ranked the coefficients in the regression models built using naïve Bayes and logistic regression methods as well as variable importance with a random forest approach [17]. We employed the correlation of rank, Kendall τ estimator, to examine the consistency of feature ranking produced by the 3 classifiers:

$$\tau = [(n_c - n_d)]/[n(n - 1)/2]$$

where n is the number of features, n_c is the number of concordant feature pairs, and n_d is the number of discordant feature pairs. The pairwise relation between feature pairs (f_i, g_i) and (f_j, g_j) is considered as concordant if the ranking order between features f is the same as for features g, that is, rank (f_i,) > rank (f_j,) and rank (g_i,) > rank (g_j,), or rank (f_i,) < rank (f_j,), and rank (g_i,) < rank (g_j,). If neither of these relation pairs is preserved, feature pairs are referred to as discordant.

Finally, in order to exploit diagnostic evaluation made by each individual PCP in the second data subset, rather than directly considering the majority vote as the data sample label, we built 5 independent naïve Bayes classifiers. Each one of them was trained on labels from the second subset corresponding to 1 of the 5 panel PCPs. We then evaluated the majority vote of the dichotomized responses of individual classifiers and employed a cross-validation scheme to estimate generalization properties.

Human Versus Model Analysis

To measure the agreement between the PCPs and a classification model, we chose a naïve Bayes approach (referred to as “the model”). Cohen κ [18] was calculated to evaluate interrater reliability of triage level within the panel, as well as interrater reliability between the model results and the panel:

$$\kappa = (p_o - p_e)/(1 - p_e)$$

where p_o is the observed ratio of agreement between 2 raters and p_e is the probability of chance agreement. Cohen κ provides a measure of agreement between raters while accounting for chance agreements. This is in contrast to percentage agreement, which merely quantifies the ratio of cases with the same classification in relation to different classifications made by 2 or more assessors, without accounting for chance agreements. A Cohen $\kappa<0.20$ is generally regarded as low, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1.00 as almost perfect agreement [18].

Additional Analyses

To explore how the brief free-form text influenced the classification, the classifier was retrained without features extracted from the brief free-form text. This analysis was conducted with a linear naïve Bayes approach.

To evaluate intrarater reliability of the training data, 50 of the 300 automated patient interviewing software reports available were chosen for retriage by the same expert PCP. These reports were chosen randomly from the full set but checked to include an even variation of all available symptoms. Cohen κ was used to assess agreement with prior triage.

Furthermore, to evaluate the impact of missing data on our results, we reran the analyses with automated patient interviewing software reports with missing triage data excluded.

Ethical Considerations

The study was approved by the Swedish Ethical Review Authority on April 24, 2019 (reference number 2019-01516).

Data Sharing Statement

Data on triage decisions made by panel members and our expert PCP are available to the Department of Clinical Sciences in Malmö at Lund university, to the Department of Computational Science and Technology at the Royal Institute of Technology, and to Doctrin AB, Stockholm Sweden 10 years following publication. Data can be accessed for a prespecified purpose after approval by all 3 parties above.

Results

Comparisons Between the Three Models

After exclusion, 276 automated patient interviewing software reports were usable as labeled test-set data (Figure 1). The single-pass test results as well as cross-validation outcomes are presented in Table 2. There was no evidence for rejecting the null hypothesis (P>.10), so the performance of all 3 classifiers is considered comparable even though one can observe a trend favorable for random forest.

Figure 1. Flowchart of automated patient interviewing software report exclusion criteria.

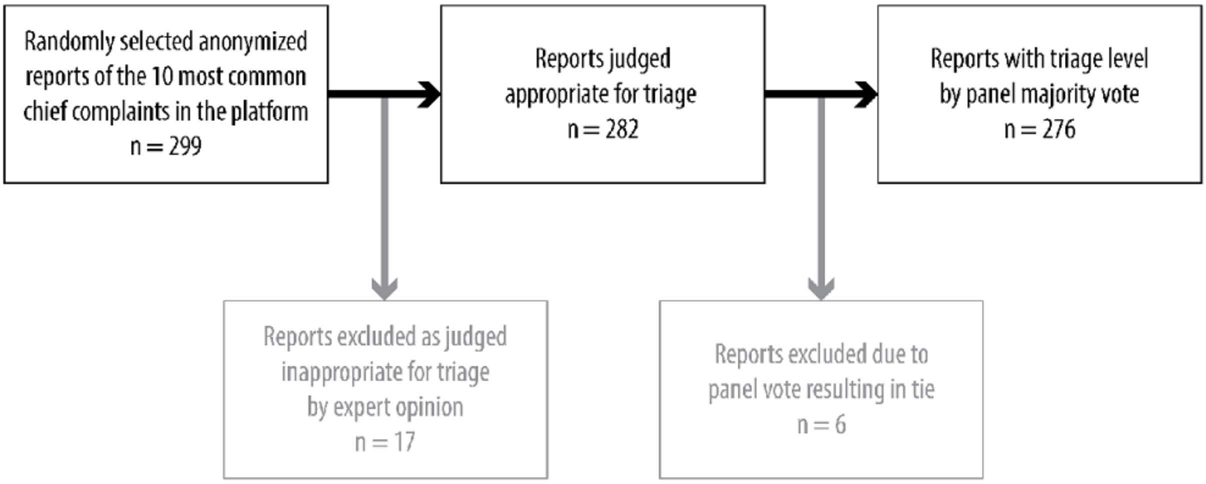

Randomly selected anonymized reports of the 10 most common chief complaints in the platform n = 299
Reports judged appropriate for triage n = 282
Reports with triage level by panel majority vote n = 276
Reports excluded as judged inappropriate for triage by expert opinion n = 17
Reports excluded due to panel vote resulting in tie n = 6

Table 2. Classification results obtained with naïve Bayes, logistic regression, and random forest in a single-pass test as well as in 10-fold cross-validation over the entire combined data set.

Classifier	Test results (training on the first and test on the second data subset), %	10-fold cross-validation (the first and second subsets combined), %[a]
Naïve Bayes	64.1	66.6 (7.6)
Logistic regression	60.1	64.5 (9.0)
Random forest	67.4	69.5 (7.7)

[a] The values for cross-validation are the mean and standard deviation of the classification accuracy obtained over 10 test folds.

Five Classifiers Versus One

Mean cross-validation accuracy calculated using the ensemble performance (majority vote) of the 5 naïve Bayes classifiers, each trained on the labels of one panel member, was 65.3% (SD 8.2%). Comparing this with the model, that is, the single naïve Bayes classifier (mean cross-validation accuracy 66.7% [SD 8.0%]), the null hypothesis could not be rejected (Wilcoxon signed-rank test, n=10, *P*>.24).

Decisive Features for Classification

Because the 3 classification approaches offer insights into feature weighing in the regression function that determines the classification boundary, we investigated more closely the distribution of such feature importance factors (see the "Methods" section). The results are inconclusive as the distribution is rather uniform and the pairwise correlations between feature rankings, Kendell τ (see the "Methods" section), produced by the classifiers are moderate (max 0.32 between naïve Bayes and random forest). This result implies that the given average level of accuracy can be achieved based on different sets of features.

Agreement Between Model and Human Triage

Because there was no statistically significant difference in the performance reported by the 3 classifiers, we decided to rely on the naïve Bayes approach in the next stages of our work due to its intuitive linear formulation. Cohen κ between the naïve Bayes model and the panel majority vote triage was 0.17 (Table 3), with 64% agreement. Excluding the information contained in brief free-form text resulted in the corresponding Cohen κ of 0.15. Within the reference group, average Cohen κ was 0.20, ranging from 0.10 to 0.30.

These results did not differ when analyses were rerun with missing cases excluded. No statistically significant difference in distribution of chief complaint symptoms could be found between reports with and without missing data (chi-square test, *P*>.99).

Using panel majority vote as the gold standard, the model correctly classified 74% (138/186) of nonurgent cases, but only 42% (38/90) of urgent cases. Adding free-form text data had a negligible effect on these numbers (Table 4).

When 50 automated patient interviewing software reports were selected for retriage by our selected expert PCP, Cohen κ was 0.55 with 78% agreement between retriage and previous triage.

Table 3. Assessment of the triage performance: agreement between the naïve Bayes model and each panel member as well as their majority vote, and average interrater agreement among the panel members.[a]

Panel	Panel member versus naïve Bayes model (Cohen κ)	Panel member versus rest panel members (Cohen κ)
PCP1	0.09	0.21
PCP2	0.03	0.21
PCP3	0.24	0.18
PCP4	0.08	0.21
PCP5	0.13	0.17
Majority vote	0.17	N/A

[a] PCP1 had the least amount of clinical experience, whereas PCP4 and PCP5 had the most amount of clinical experience.

Table 4. Contingency table of model triage with panel majority vote as the gold standard.

	Truly urgent	Falsely nonurgent	Truly nonurgent	Falsely urgent
Naïve Bayes model trained on full information including brief free-form text	42% (38 out of 90 cases voted urgent)	58% (52 out of 90 cases voted urgent)	74% (138 out of 186 cases voted nonurgent)	26% (48 out of 186 cases voted nonurgent)
Naïve Bayes model trained with brief free-form text information excluded	42% (38 out of 90 cases voted urgent)	58% (52 out of 90 cases voted urgent)	73% (135 out of 186 cases voted nonurgent)	27% (51 out of 186 cases voted nonurgent)

Discussion

Principal Results

To our knowledge, this is the first study to evaluate human versus ML performance in primary care triage based on a digitalized patient history. The first principal finding of this investigation was that interrater reliability in human triage using automated patient interviewing software reports is low (Cohen κ 0.20). Consequently, our second principal finding was that interrater triage reliability between a statistical model trained on automated patient interviewing software reports and a human panel was low (Cohen κ 0.17).

Findings were robust when cases with missing triage data were excluded from the analysis. The performance of the model was mostly decided by the surveys as removing the free-form text had only marginal impact on Cohen κ (reduced to 0.15). Furthermore, the intrarater reliability was moderate, as seen by retriage of 50 automated patient interviewing software reports by the same PCP (Cohen κ 0.55).

Comparison With Prior Work

While we acknowledge that κ values seldom are comparable across studies [19], previous data have generally found high interrater reliability between triage nurses [20-22]. However, these studies were conducted in high-acuity emergency department settings, where indicators of urgency arguably are more clearly defined [23].

The primary care setting presents a particular challenge in that conditions are of low acuity, making the line between urgent and nonurgent care more difficult to draw. This is supported by the low intrarater agreement for our expert PCP as well as the low agreement between our panel members. Indeed, acquiring a true gold standard for triage is a well-known issue [24]. “Correct” triage is difficult to define, and thus difficult to label and automate using ML. We could not identify any particular features in the data that were linked to the model’s triage decision. As far as the clinicians are concerned, we did not study their clinical reasoning before reaching a triage decision, that is, we do not know on which features their decision was based.

Interpretation

A well-known bottleneck for the creation of reliable ML algorithms is the lack of large enough amounts of labeled training data but this study calls the reliability of labels themselves into question. Labeled data need to be consistent across different raters and over time. Consequently, while adding more automated patient interviewing software data to the training set exploited by the model could improve interrater reliability with humans, the interrater reliability between the humans themselves sets a limit on how useful an algorithm could be if labels are fully decided from human data. While the addition of free-form text did not offer any advantage to the performance of the model, as assessed by our gold standard, it is possible that larger amounts of free-text data would allow the model to leverage these data for improved performance.

Human clinical decision making is likely more prone to be affected by externalities such as stress and mental fatigue [25]. Such externalities may have been present to different extents among our panel, resulting in markedly variable triage decisions compared with each other and the model.

Furthermore, the low agreement between the panel and the model in our study may be due to the fact that variation in human interpretation of text-based cues from automated patient interviewing software data in a primary care setting [26] prevents PCPs from determining urgency as consistently as the model, given access to the same amount of data. It should be noted, however, that in the clinical setting, PCPs would acquire additional data through the eVisit chat before making a triage decision.

The model is trained on triage data from a senior expert PCP, but results show no trend toward higher agreement between more senior PCPs and the model. This suggests that triage decision making depends more on other factors such as PCP temperament and risk aversion than mere experience [27].

Accepting the panel majority vote as the gold standard, nonurgent cases were more often classified correctly compared with urgent cases (74% [138/186] vs 42% [38/90], respectively), even though higher triage accuracy would be expected for urgent conditions where red flags are more well-defined [12]. Selection bias through a disproportionately larger amount of training data on nonurgent automated patient interviewing software reports may explain part of this disparity. On the contrary, this disproportionality may still be representative of a primary care cohort which would utilize such a digital tool for mostly low acuity conditions. However, given the low agreement between panel members, one may also question the suitability of use of the panel majority vote as the gold standard.

Strengths

This study has several strengths. First, it is one of few studies comparing human with ML performance using the same test data set for both groups. It is uniquely conducted in an eVisit primary care setting, where the need for reduced workload is high and where the ML algorithm has access to the same data as the clinician in the eVisit setting would. This contrasts with clinical or electronic health record–based ML tools which may not have access to key clinical data not recorded in the electronic health record [28]. Our data set was largely complete with only 1.4% missing data points. We also used training set data independent of validation test-set data, which is not always the case in other published research in the field [29]. Finally, the findings add nuance to the existing literature of ML versus human physicians [30].

Limitations

The results should be interpreted with consideration to several limitations. Our sample is not representative of a physical primary care population, as reports were acquired from an online consultation service database of self-selected patients being less likely to have life-threatening conditions [31]. Our data did not allow for out-of-sample external validation, as we do not know how these automated patient interviewing software reports ended up being triaged in their clinical setting. Lack of external validation also means that our low interrater reliability was likely overestimated [29]. However, even if externally valid endpoint data could aid in defining a decision as “correct” retrospectively [32], defining “correct” triage prospectively may not be possible as some clinical outcomes cannot be predicted. In addition, the lack of consensus and use of a voting strategy in our panel are unconventional methods of defining a gold standard to compare ML-based performance and make comparison with other studies difficult. Future studies may use consensus techniques such as Delphi [33], incorporating PCP and emergency physician expertise, to mitigate lack of panel triage consensus.

Given the lack of agreement between our panel PCPs, using 1 expert PCP to provide training data may not be optimal. However, we did not observe any significant differences in cross-validation accuracy in this model compared with the ensemble performance of 5 models separately trained by each panel member.

Finally, our data set did not allow us to evaluate how the temporal provision of data affects the triage process in a way that would mimic the iterative clinical decision-making process. Thus, training data sets which make this possible may open up new opportunities for devising ML approaches that better mimic the human decision-making process.

Practical Implications

This study refutes implementation of the current ML model to fully automate binary triage in primary care, despite naïve Bayes being a reasonable ML algorithm to approach this problem. However, in the clinical setting, these reports are used as decision support in the interaction with patients, implying that uncertainties may be addressed by further interaction with the patient. Further development of the model with the suggestions made above may allow for fully automated triage in the future.

Conclusions

While digitalized patient histories have the potential to mitigate primary care workloads, leveraging patient history data to automate triage with ML methods is challenging given the difficulty for human physicians to triage consistently in a primary care setting. Future research should evaluate if external validation and temporal provision of training data may improve automated triage performance, as well as attempt to better identify which features drive triage decisions in a primary care setting.

Acknowledgments

The authors thank Johan Ekegren Gunnarsson and Sonja Petrovic Lundberg for preparing anonymized data sets and creating a graphic user interface for report review and triage. The authors also thank Markus Sandén (Doctrin AB, Stockholm, Sweden) for providing training data as our expert PCP. The authors thank residents Robin Back (Närhälsan Online and Närhälsan Ekmanska vårdcentral, Gothenburg, Sweden) and Lina Nyhlén (Vårdcentralen Lomma, Sweden), and specialists Miriam Pikkemaat (Department of Clinical Sciences in Malmö/Family Medicine, Lund university) and Annika Pahlmblad (Department of Clinical Sciences in Malmö/Family Medicine, Lund university) for being part of our expert panel. This study was partly funded by Vinnova, Sweden's Innovation Agency (Grant No. 2017-02348).

Authors' Contributions

AB, PH, SB, and PM were responsible for study concept and design; AB and AE were responsible for data acquisition; PH and SB performed analysis; AE was responsible for manuscript drafting; all authors were responsible for data interpretation, critical revision of the manuscript for important intellectual content, and final approval of the version to be published.

Conflicts of Interest

AB is the Chief Medical Officer of Doctrin AB, one of the project parties in this Vinnova-financed project. Other authors have no conflicts of interest to declare.

References

1. Colwill JM, Cultice JM, Kruse RL. Will generalist physician supply meet demands of an increasing and aging population? Health Aff (Millwood) 2008 Jan;27(3):w232-w241. [doi: 10.1377/hlthaff.27.3.w232] [Medline: 18445642]
2. van den Berg MJ, van Loenen T, Westert GP. Accessible and continuous primary care may help reduce rates of emergency department use. An international survey in 34 countries. Fam Pract 2016 Feb 28;33(1):42-50. [doi: 10.1093/fampra/cmv082] [Medline: 26511726]
3. Rhoades DR, McFarland KF, Finch WH, Johnson AO. Speaking and interruptions during primary care office visits. Fam Med 2001;33(7):528-532. [Medline: 11456245]
4. Berner ES, Kasiraman RK, Yu F, Ray MN, Houston TK. Data quality in the outpatient setting: impact on clinical decision support systems. AMIA Annu Symp Proc 2005:41-45 [FREE Full text] [Medline: 16778998]
5. Burnett SJ, Deelchand V, Franklin BD, Moorthy K, Vincent C. Missing clinical information in NHS hospital outpatient clinics: prevalence, causes and effects on patient care. BMC Health Serv Res 2011 May 23;11(1):114 [FREE Full text] [doi: 10.1186/1472-6963-11-114] [Medline: 21605359]
6. Campbell JL, Fletcher E, Britten N, Green C, Holt TA, Lattimer V, et al. Telephone triage for management of same-day consultation requests in general practice (the ESTEEM trial): a cluster-randomised controlled trial and cost-consequence analysis. The Lancet 2014 Nov;384(9957):1859-1868. [doi: 10.1016/s0140-6736(14)61058-8] [Medline: 25098487]
7. Marklund B, Ström M, Månsson J, Borgquist L, Baigi A, Fridlund B. Computer-supported telephone nurse triage: an evaluation of medical quality and costs. J Nurs Manag 2007 Mar;15(2):180-187. [doi: 10.1111/j.1365-2834.2007.00659.x] [Medline: 17352701]
8. Shepherd G, Schwartz R. Frequency of incomplete medication histories obtained at triage. Am J Health Syst Pharm 2009 Jan 01;66(1):65-69. [doi: 10.2146/ajhp080171] [Medline: 19106346]
9. Ernesäter A, Engström M, Holmström I, Winblad U. Incident reporting in nurse-led national telephone triage in Sweden: the reported errors reveal a pattern that needs to be broken. J Telemed Telecare 2010 May 10;16(5):243-247. [doi: 10.1258/jtt.2009.090813] [Medline: 20457800]
10. Giesen P, Ferwerda R, Tijssen R, Mokkink H, Drijver R, van den Bosch W, et al. Safety of telephone triage in general practitioner cooperatives: do triage nurses correctly estimate urgency? Qual Saf Health Care 2007 Jun 01;16(3):181-184 [FREE Full text] [doi: 10.1136/qshc.2006.018846] [Medline: 17545343]

11. Zakim D. Development and significance of automated history-taking software for clinical medicine, clinical research and basic medical science. J Intern Med 2016 Sep 12;280(3):287-299 [FREE Full text] [doi: 10.1111/joim.12509] [Medline: 27071980]
12. Semigran H, Linder J, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015 Jul 08;351:h3480 [FREE Full text] [doi: 10.1136/bmj.h3480] [Medline: 26157077]
13. Verzantvoort NCM, Teunis T, Verheij TJM, van der Velden AW. Self-triage for acute primary care via a smartphone application: Practical, safe and efficient? PLoS One 2018 Jun 26;13(6):e0199284 [FREE Full text] [doi: 10.1371/journal.pone.0199284] [Medline: 29944708]
14. Shafaf N, Malek H. Applications of Machine Learning Approaches in Emergency Medicine; a Review Article. Arch Acad Emerg Med 2019;7(1):34 [FREE Full text] [Medline: 31555764]
15. Swedish Association of Local Authorities and Regions (SALAR). Statistics About Health Care and Regional Development 2015: Operations and Economy in Counties and Regions. Stockholm: Swedish Association of Local Authorities and Regions (SALAR); 2016:978-991.
16. Lebanon G, Mao Y, Dillon J. The locally weighted bag of words framework for document representation. J Mach Learn Res 2007;8(12/1/2007):2405-2441. [doi: 10.5555/1314498.1314576]
17. Flaxman AD, Vahdatpour A, Green S, James SL, Murray CJ, Population H. Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. Popul Health Metr 2011 Aug 04;9:29 [FREE Full text] [doi: 10.1186/1478-7954-9-29] [Medline: 21816105]
18. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 2016 Jul 02;20(1):37-46. [doi: 10.1177/001316446002000104]
19. Feinstein AR, Cicchetti DV. High agreement but low Kappa: I. the problems of two paradoxes. Journal of Clinical Epidemiology 1990 Jan;43(6):543-549. [doi: 10.1016/0895-4356(90)90158-l] [Medline: 2348207]
20. Elias P, Damle A, Casale M, Branson K, Churi C, Komatireddy R, et al. A Web-Based Tool for Patient Triage in Emergency Department Settings: Validation Using the Emergency Severity Index. JMIR Med Inform 2015 Jun 10;3(2):e23 [FREE Full text] [doi: 10.2196/medinform.3508] [Medline: 26063343]
21. Grouse AI, Bishop RO, Bannon AM. The Manchester Triage System provides good reliability in an Australian emergency department. Emerg Med J 2009 Jul;26(7):484-486. [doi: 10.1136/emj.2008.065508] [Medline: 19546267]
22. Gerdtz MF, Collins M, Chu M, Grant A, Tchernomoroff R, Pollard C, et al. Optimizing triage consistency in Australian emergency departments: the Emergency Triage Education Kit. Emerg Med Australas 2008 Jun;20(3):250-259. [doi: 10.1111/j.1742-6723.2008.01089.x] [Medline: 18462405]
23. Widgren BR, Jourak M. Medical Emergency Triage and Treatment System (METTS): a new protocol in primary triage and secondary priority decision in emergency medicine. J Emerg Med 2011 Jun;40(6):623-628. [doi: 10.1016/j.jemermed.2008.04.003] [Medline: 18930373]
24. FitzGerald G, Jelinek GA, Scott D, Gerdtz MF. Emergency department triage revisited. Emerg Med J 2010 Feb;27(2):86-92. [doi: 10.1136/emj.2009.077081] [Medline: 20156855]
25. Allan JL, Johnston DW, Powell DJH, Farquharson B, Jones MC, Leckie G, et al. Clinical decisions and time since rest break: An analysis of decision fatigue in nurses. Health Psychol 2019 Apr;38(4):318-324. [doi: 10.1037/hea0000725] [Medline: 30896218]
26. Entezarjou A, Bolmsjö BB, Calling S, Midlöv P, Milos Nymberg V. Experiences of digital communication with automated patient interviews and asynchronous chat in Swedish primary care: a qualitative study. BMJ Open 2020 Jul 23;10(7):e036585 [FREE Full text] [doi: 10.1136/bmjopen-2019-036585] [Medline: 32709650]
27. Considine J, Botti M, Thomas S. Do knowledge and experience have specific roles in triage decision-making? Acad Emerg Med 2007 Aug;14(8):722-726 [FREE Full text] [doi: 10.1197/j.aem.2007.04.015] [Medline: 17656608]
28. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep 2016 Dec 17;6:26094 [FREE Full text] [doi: 10.1038/srep26094] [Medline: 27185194]
29. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/s2589-7500(19)30123-2]
30. Cook TS. Human versus machine in medicine: can scientific literature answer the question? The Lancet Digital Health 2019 Oct;1(6):e246-e247. [doi: 10.1016/s2589-7500(19)30124-4]
31. North F, Crane SJ, Stroebel RJ, Cha SS, Edell ES, Tulledge-Scheitel SM. Patient-generated secure messages and eVisits on a patient portal: are patients at risk? J Am Med Inform Assoc 2013 Nov 01;20(6):1143-1149 [FREE Full text] [doi: 10.1136/amiajnl-2012-001208] [Medline: 23703826]
32. Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Acad Emerg Med 2000 Mar;7(3):236-242 [FREE Full text] [doi: 10.1111/j.1553-2712.2000.tb01066.x] [Medline: 10730830]
33. Fry M, Burr G. Using the Delphi technique to design a self-reporting triage survey tool. Accid Emerg Nurs 2001 Oct;9(4):235-241 [FREE Full text] [doi: 10.1054/aaen.2001.0245] [Medline: 11855763]

Abbreviations

ML: machine learning
PCPs: primary care physicians

Edited by G Eysenbach; submitted 27.03.20; peer-reviewed by R Miotto, A Benis, D Gunasekeran; comments to author 27.04.20; revised version received 22.06.20; accepted 24.06.20; published 03.09.20.

Please cite as:
Entezarjou A, Bonamy AKE, Benjaminsson S, Herman P, Midlöv P
Human- Versus Machine Learning–Based Triage Using Digitalized Patient Histories in Primary Care: Comparative Study
JMIR Med Inform 2020;8(9):e18930
URL: https://medinform.jmir.org/2020/9/e18930
doi:10.2196/18930
PMID:32880578

Original Paper

Identifying Key Predictors of Cognitive Dysfunction in Older People Using Supervised Machine Learning Techniques: Observational Study

Debbie Rankin[1], BSc, PhD; Michaela Black[1], BSc, PhD; Bronac Flanagan[1], BSc, MSc, PhD; Catherine F Hughes[2], BSc, PhD; Adrian Moore[3], BSc, MSc, PhD; Leane Hoey[2], BSc, MSc, PhD; Jonathan Wallace[4], BA, MSc; Chris Gill[2], BSc, PhD; Paul Carlin[5], BSc, MPA; Anne M Molloy[6], PhD; Conal Cunningham[7], MD; Helene McNulty[2], BSc, PhD

[1]School of Computing, Engineering and Intelligent Systems, Ulster University, Derry~Londonderry, United Kingdom

[2]School of Biomedical Sciences, Nutrition Innovation Centre for Food and Health, Ulster University, Coleraine, United Kingdom

[3]School of Geography and Environmental Sciences, Ulster University, Coleraine, United Kingdom

[4]School of Computing, Ulster University, Jordanstown, United Kingdom

[5]School of Health, Wellbeing and Social Care, The Open University, Belfast, United Kingdom

[6]School of Medicine, Trinity College Dublin, Dublin, Ireland

[7]Mercers Institute for Research on Ageing, St James's Hospital, Dublin, Ireland

Corresponding Author:
Debbie Rankin, BSc, PhD
School of Computing, Engineering and Intelligent Systems
Ulster University
Northland Road
Derry~Londonderry, BT48 7JL
United Kingdom
Phone: 44 287167 ext 5841
Email: d.rankin1@ulster.ac.uk

Abstract

Background: Machine learning techniques, specifically classification algorithms, may be effective to help understand key health, nutritional, and environmental factors associated with cognitive function in aging populations.

Objective: This study aims to use classification techniques to identify the key patient predictors that are considered most important in the classification of poorer cognitive performance, which is an early risk factor for dementia.

Methods: Data were used from the Trinity-Ulster and Department of Agriculture study, which included detailed information on sociodemographic, clinical, biochemical, nutritional, and lifestyle factors in 5186 older adults recruited from the Republic of Ireland and Northern Ireland, a proportion of whom (987/5186, 19.03%) were followed up 5-7 years later for reassessment. Cognitive function at both time points was assessed using a battery of tests, including the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS), with a score <70 classed as poorer cognitive performance. This study trained 3 classifiers—decision trees, Naïve Bayes, and random forests—to classify the RBANS score and to identify key health, nutritional, and environmental predictors of cognitive performance and cognitive decline over the follow-up period. It assessed their performance, taking note of the variables that were deemed important for the optimized classifiers for their computational diagnostics.

Results: In the classification of a *low* RBANS score (<70), our models performed well (F_1 score range 0.73-0.93), all highlighting the individual's score from the Timed Up and Go (TUG) test, the age at which the participant stopped education, and whether or not the participant's family reported memory concerns to be of key importance. The classification models performed well in classifying a greater rate of decline in the RBANS score (F_1 score range 0.66-0.85), also indicating the TUG score to be of key importance, followed by blood indicators: plasma homocysteine, vitamin B6 biomarker (plasma pyridoxal-5-phosphate), and glycated hemoglobin.

Conclusions: The results suggest that it may be possible for a health care professional to make an initial evaluation, with a high level of confidence, of the potential for cognitive dysfunction using only a few short, noninvasive questions, thus providing a quick, efficient, and noninvasive way to help them decide whether or not a patient requires a full cognitive evaluation. This approach has the potential benefits of making time and cost savings for health service providers and avoiding stress created through unnecessary cognitive assessments in low-risk patients.

(JMIR Med Inform 2020;8(9):e20995) doi:10.2196/20995

KEYWORDS

classification; supervised machine learning; cognition; diet; aging; geriatric assessment

Introduction

Globally, populations are aging. By 2050, it is estimated that more than 2 billion people will be aged over 60 years [1]. Cognitive function generally declines with age and ranges in severity from mild cognitive impairment (MCI) to dementia. MCI can be defined as cognitive decline greater than that expected for an individual's age and education level, but it does not interfere with activities of daily living, whereas dementia profoundly impacts normal functioning [2,3]. Dementia currently affects 50 million people worldwide, and it is estimated that this will increase to 152 million by 2050. The annual cost of dementia is estimated at US $1 trillion and is expected to more than double by 2030 [4]. Therefore, strategies that promote better brain health and well-being in older age are an urgent public health priority.

Alzheimer disease is the most common form of dementia, with other forms including vascular dementia, dementia with Lewy bodies, frontotemporal dementia, and mixed dementia. Risk factors for dementia are disease dependent but commonly include age, genetics and medical conditions including cardiovascular disease and diabetes, diet, lifestyle, and environmental factors [5]. An important recent report highlighted the complexity of dementia and the potential to prevent or delay the onset of the disease through interventions targeted at modifiable risk factors [6]. In particular, nutrition has been identified as a key area of interest, and emerging evidence links lower levels of certain vitamins with cognitive dysfunction in older adults, whereas certain dietary patterns and components appear to have protective roles in maintaining cognitive health [7].

The application of data mining within health care has become increasingly popular, driven particularly by the large amount of complex data available that test the capabilities of traditional statistical approaches [8]. In health care, as in other areas, data mining has provided a means of accessing and analyzing large volumes of data to better inform and drive change. Classification models, in particular, have been utilized extensively in the understanding of MCI. These models can help us to understand patterns in the behavior of data in terms of diagnosing MCI, specifically in the consideration of key features pertaining to a diagnosis of impairment [9,10] or predicting the progression of the impairment [11]. Furthermore, models have been developed to apply a more objective approach to the MCI diagnosis [12], not to undermine but rather to support a clinician's analysis [13]. Na c [14] investigated the use of noninvasive, easy-to-collect variables that are commonly collected in community health care settings such as sociodemographic, health, functional, and interpersonal variables, for the prediction of cognitive impairment among community-dwelling older adults, using the Korean Longitudinal Study of Aging (KLoSA) data set [15] and a gradient boosting machine classifier.

Many studies apply machine learning approaches to the popular Open Access Series of Imaging Studies [16], Alzheimer Disease Neuroimaging Initiative (ADNI) [17], and Australian Imaging Biomarkers and Lifestyle Flagship Study of Aging (AIBL) [18] data sets consisting of neuroimaging data (eg, magnetic resonance imaging [MRI] and positron emission tomography scan data) from participants ranging from no cognitive impairment to MCI to Alzheimer disease [19]. These data sets also include a range of demographic, biomarker, clinical, and cognitive assessment data. Ding et al [20] used a Bayesian network approach for the classification of Alzheimer disease with heterogeneous features from the AIBL data set and demonstrated that machine learning could be used to select features and their appropriate combinations that are relevant for Alzheimer disease severity classification with high accuracy. Korolev et al [21] used a kernel-based classifier and the ADNI data set to develop a prognostic model for predicting MCI-to-dementia progression over a 3-year period.

The aim of our study is to compare the selection of data analytics techniques to identify determinants of cognitive health in community-dwelling older adults using existing data from the Trinity-Ulster and Department of Agriculture (TUDA) study (ClinicalTrials.gov identifier: NCT02664584). The TUDA study was designed to investigate nutritional, health, and lifestyle factors in the development of diseases related to aging, including dementia. A range of analytical models on the data were developed to determine factors that may predict poorer cognitive performance and cognitive decline over time, assessed using an in-depth neuropsychiatric test.

Methods

Cross-Industry Process for Data Mining Methodology

In this study, the widely used cross-industry process for data mining (CRISP-DM) research methodology was adopted [22]. CRISP-DM has 6 main steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. In the business understanding phase, the objective of this study was to use classification techniques to identify the key patient predictors considered most important in the classification of cognitive dysfunction, which itself is a predictor of dementia. In the data understanding phase, the data quality was examined to understand data collection methods and the features contained within the TUDA data set, as described in the next section (The Data). In the data preparation phase, the TUDA data set was preprocessed to cleanse the data set and select features relevant to the modeling phase. Feature selection methods and the results of feature selection are described in the subsequent sections (*The Data* and *Feature Selection* sections in *Methods* and the *Feature Selection* section in *Results*). In the modeling phase, a number of machine learning modeling techniques were selected and applied to the prepared data and their parameters were calibrated to optimal values to increase the knowledge extracted from the data (described in the Machine Learning Techniques section in Methods and the RBANS Classification and Classifying Cognitive Decline Using the Rate of Change in the RBANS Score sections in Results). Upon building the models that produced the highest quality knowledge from the data analysis perspective, the models were thoroughly evaluated to ensure robustness and achievement of the business objectives. The knowledge gained from the models was then presented to clinical experts in a way that could be used and understood.

The Data

The TUDA cohort provides detailed nutrition and health data, along with related lifestyle, clinical, and biochemical details, on a total of 5186 community-dwelling older adults aged 60 to 102 years, making this cohort one of the most comprehensively characterized cohorts of its kind for aging research internationally. With an overall goal to address the prevention of age-related diseases, the TUDA study is aimed at investigating nutrition and related factors in the development of common diseases of aging. TUDA study participants were recruited between 2008 and 2012 from hospital outpatient or general practice clinics in the Republic of Ireland or Northern Ireland via standardized protocols for participant sampling, assessment, and data recording and with a centralized laboratory analysis. In brief, the inclusion criteria for the TUDA study were being born on the island of Ireland, aged >60 years, and not having an existing diagnosis of dementia. Nonfasting blood samples were collected from all participants, and a wide range of parameters including routine biochemistry and hematological profiles, along with biomarkers of micronutrient status, were measured. A comprehensive health and lifestyle questionnaire was administered as part of the 90-min interview to capture medical and demographic details, along with comprehensive information on medication and vitamin supplement usage. Physiological function tests, blood pressure, bone health (dual-energy x-ray absorptiometry scans), and cognitive function tests were also performed. A subset of approximately 19.03% (987/5186) of participants were reassessed 5 to 7 years after their initial assessment to investigate the progression of risk factors and disease over time.

A summary of the characteristics of the subset of the TUDA cohort (n=2869) analyzed in this study is shown in Table 1. Preprocessing and feature selection performed on the original data set to reach this subset of data are described in the *Feature Selection* sections of the Methods and Results sections.

Cognitive function was assessed at both time points using 3 assessment tools, the Mini-Mental State Examination (MMSE), the Frontal Assessment Battery (FAB), and RBANS, and the rate of cognitive decline was calculated over the 5- to 7-year follow-up period. For the purposes of this study, the cognitive function outcome indicator is categorized based on RBANS. RBANS is an age-adjusted and sensitive neuropsychiatric battery for assessing global cognitive function [23]. This tool has also been validated to assess specific cognitive domains within the brain, including immediate and delayed memory, visual-spatial, language, and attention, which are combined to provide a total score, with lower scores generally indicative of poorer cognitive performance.

The rate of RBANS change over the 5- to 7-year period between the initial assessment and the follow-up assessment was computed as the difference between a participant’s RBANS score at each sampling point, normalized to account for the time between each assessment, where this can differ by up to 2 years across participants (Figure 1).

Table 1. General characteristics of the Trinity-Ulster and Department of Agriculture study participants.

Characteristics	Males (n=1191)	Females (n=1678)
Age (years), mean (SD)	72.1 (7.8)	72.2 (7.8)
Education (years)[a], mean (SD)	16.3 (3.3)	16.1 (2.8)
Health and lifestyle		
BMI (kg/m^2), mean (SD)	28.9 (4.3)	28.7 (5.7)
Waist-to-hip ratio, mean (SD)	0.97 (0.07)	0.88 (0.07)
Instrumental activities of daily living, mean (SD)	25.0 (4.1)	24.9 (3.5)
Physical self-maintenance scale score, mean (SD)	23.3 (1.6)	23.1 (1.7)
Timed Up and Go (seconds), mean (SD)	12.9 (9.1)	13.0 (8.0)
Living alone, n (%)	260 (21.8)	632 (37.7)
Current smoker, n (%)	122 (10.2)	194 (11.6)
Alcohol (units/week), mean (SD)	8.8 (14.6)	2.9 (6.7)
Socioeconomically most deprived, n (%)	291 (24.4)	426 (25.4)
Neuropsychiatric assessment		
MMSE[b] score, mean (SD)	27.8 (1.4)	27.9 (1.4)
RBANS[c] score, mean (SD)	87.3 (14.5)	88.9 (15.2)
RBANS class="low" (target), n (%)[d]	133 (11.2)	168 (10.0)
RBANS class="high" (target), n (%)[d]	1058 (88.8)	1510 (90.0)
FAB[e] score, mean (SD)	15.7 (2.2)	15.9 (2.1)
Depression CES-D[f] score, mean (SD)	4.8 (6.2)	6.1 (7.7)
Anxiety (HADS[g] score), mean (SD)	2.6 (3.2)	3.5 (3.8)
Clinical measures		
White cell count (10^9/L), mean (SD)	7.1 (3.6)	6.9 (3.3)
Hemoglobin (g/DL), mean (SD)	14.2 (1.5)	13.0 (1.3)
Mean corpuscular volume (FL[h]), mean (SD)	90.7 (5.5)	90.6 (5.1)
Platelet count (10^9/L), mean (SD)	229 (59.0)	265 (66.9)
Urea (mmol/L), mean (SD)	7.2 (2.9)	6.7 (2.3)
Creatinine (µmol/L), mean (SD)	98 (31.0)	79 (22.4)
Albumin (g/L), mean (SD)	42 (3.7)	42 (3.4)
Gamma GT (U/L), mean (SD)	43 (47.5)	34 (36.0)
Sodium (mmol/L), mean (SD)	140 (5.1)	139 (3.2)
Potassium (mmol/L), mean (SD)	4.3 (0.5)	4.2 (0.4)
Calcium (mmol/L), mean (SD)	2.3 (0.1)	2.3 (0.1)
Phosphate (mmol/L), mean (SD)	1.0 (0.2)	1.1 (0.2)
Alkaline phosphatase (U/L), mean (SD)	82 (34.2)	82 (25.7)
Low-density lipoprotein (mmol/L), mean (SD)	2.23 (0.8)	2.58 (0.9)
High-density lipoprotein (mmol/L), mean (SD)	1.23 (0.4)	1.55 (0.4)
Triglycerides (mmol/L), mean (SD)	1.78 (1.0)	1.62 (1.0)
C-reactive protein (mg/L), mean (SD)	6.1 (11.1)	5.5 (11.9)
Glycated hemoglobin (%), mean (SD)	6.0 (1.0)	5.9 (0.7)
Parathyroid hormone (pg/mL), mean (SD)	45.2 (30.8)	47.2 (31.9)

Characteristics	Males (n=1191)	Females (n=1678)
Glomerular filtration rate (mL/min), mean (SD)	77.2 (25.3)	67.8 (22.6)
Nutritional biomarkers		
Red blood cell folate (nmol/L), mean (SD)	1053 (591.1)	1100 (582.7)
Serum vitamin B12 (pmol/L), mean (SD)	267 (191.0)	296 (277.3)
Plasma vitamin B6 (nmol/L), mean (SD)	74.1 (53.2)	81.5 (69.7)
Riboflavin (EGRac[i]), mean (SD)	1.35 (0.2)	1.34 (0.2)
Total plasma homocysteine (µmol/L), mean (SD)	15.1 (5.9)	14.1 (5.1)
Total vitamin D (nmol/L), mean (SD)	51.6 (25.9)	56.0 (30.1)

[a] Education refers to the age of stopping formal education.

[b] MMSE: Mini-Mental State Examination.

[c] RBANS: Repeatable Battery for the Assessment of Neuropsychological Assessment.

[d] RBANS score <70 is assigned class *low* and an RBANS score ≥70 is assigned class *high*.

[e] FAB: Frontal Assessment Battery.

[f] CES-D: Centre for Epidemiological Studies Depression.

[g] HADS: Hospital Anxiety and Depression Scale.

[h] FL: femtolitre.

[i] EGRac: erythrocyte glutathione reductase activation coefficient, with a higher EGRac value indicating poorer riboflavin status.

Figure 1. Calculating Repeatable Battery for the Assessment of Neuropsychological Status rate of change over a 5- to 7-year period between initial assessment and follow-up assessment, normalized to account for the time between each assessment.

$$RBANS_{rate\ of\ change} = \frac{RBANS_{assessment\ 2} - RBANS_{assessment\ 1}}{Time\ between\ assessments\ 1\ and\ 2}$$

The data set initially contained 525 variables. During preprocessing, the data were cleansed to detect and correct inaccurate values, identify missing values and ensure consistent coding of these, ensure consistent coding of categorical variables, identify spelling and coding inconsistencies and correct these, transform text variables into categorical variables where possible, ensure numeric values fell within an appropriate and accurate range, check for consistency among dependent variables and correct any errors, and finally check for duplicate data and remove any redundancy. Normalization was carried out on the data table, including nonloss decomposition to decompose the large data table into smaller tables, transforming composite attributes into separate attributes, transforming multivalued attributes, repeating columns into separate tables, and recoding text attributes to categorical attributes where possible. This process reduced the number of variables to 345 within the data set. These variables were a combination of text, categorical, and numerical variables.

Feature Selection

Dimension reduction is an important stage for understanding information in a data set. Typical dimension reduction techniques, such as principal component analysis (PCA) [24], describe all the numerical variables contained within a data set in terms of a number of linear combinations (fewer than the original number of features) of these features. Although a widely used and appreciated method for reducing the number of dimensions within a data set, PCA is only valid for numerical features. In addition, a more transparent feature selection method is often required to remove redundant features of various types to reduce the size of the data set without losing potentially valuable information. Although a range of feature selection techniques exist because of the nature of the features in the TUDA data set and the prior knowledge that a large number of variables were likely to be highly correlated, a correlation analysis and clustering were used in this study to allow highly correlated features to be determined and redundant features to be removed. These methods also helped us to discuss, evaluate, and agree on the features to be retained in collaboration with the data gatekeepers and expert clinicians who had in-depth knowledge of the data. Further feature selection was not carried out as we elected to retain as many features as possible for use in training the classifiers. This section describes the feature selection techniques performed, and the results of feature selection are described in the Results section.

Manual Feature Selection

Manual feature selection was performed to remove features containing large amounts of missing data and, therefore, considered not useful for the analysis. Free-text variables that could not be encoded were also removed. On the basis of expert clinical knowledge, features deemed irrelevant to the study were removed, as well as a number of subjective features where a comparable, objective laboratory-obtained feature existed in the data set.

Correlation and Association

A correlation analysis is necessary before the development of classification models for 2 primary reasons: "Algorithms might 'overfit' predictions to spurious correlations in the data; multicollinear, correlated predictors could produce unstable estimates" [25] and "Perfectly correlated variables are truly redundant in the sense that no additional information is gained by adding them" [26]. In other words, as many machine learning algorithms rely on linearly independent variables, strongly correlated variables must be evaluated and removed to avoid unreliable results. Moreover, 2 variables that follow the same behavior add little to the information gained by the data set and thus are considered redundant. The correlation analysis allows the determination of highly correlated variables, which may undermine the consequential data analysis results. Owing to the difference in categorization of the variables within the data set, correlation coefficients were calculated for numerical-numerical pairs, whereas the strength of association was necessary for categorical-categorical variables and categorical-numerical variables. Correlations between numerical variables were calculated using the Spearman nonparametric correlation coefficient [27], the strength of association between categorical variables was calculated using the Cramér V statistic [28], and the coefficient of determination (R2) was calculated between categorical and numerical variables [29].

Clustering

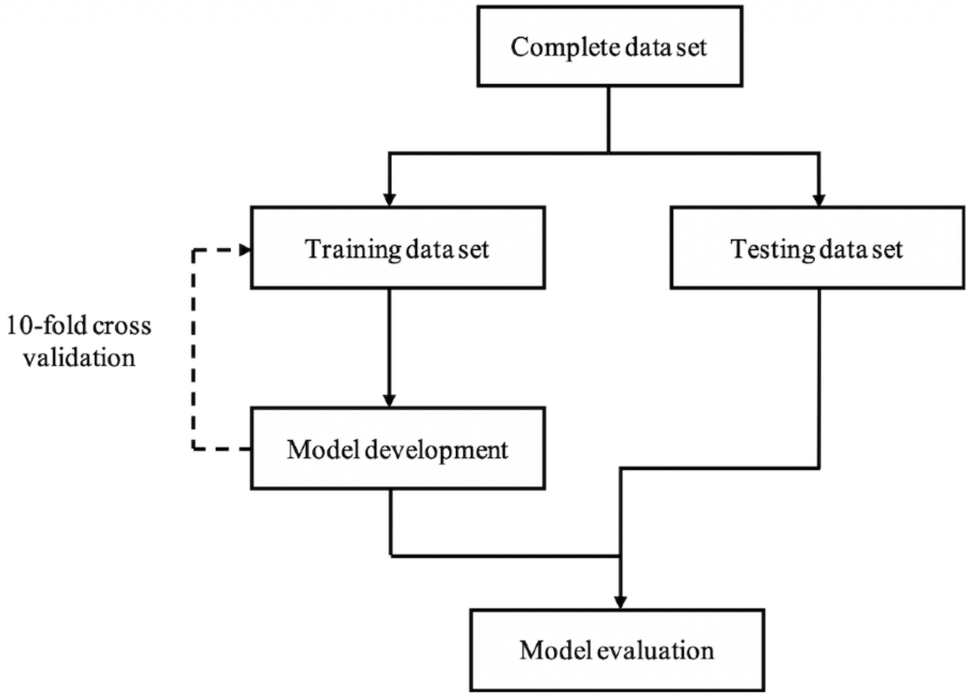
Clustering is useful in feature selection [26] to analyze the data to find structural patterns. Clustering can be used together with correlation analysis to identify those variables that behave in a similar manner; thus, the information offered by the variables may prove redundant. Clustering of variables can take 1 of 2 forms: hierarchical, which outputs an informative hierarchy, and nonhierarchical, which divides the data into clusters, within which the variables may behave similarly. Owing to the nature of the information this study seeks to derive, the focus was placed on hierarchical clustering, illustrated specifically in the form of tree structures or dendrograms.

Ascendant hierarchical clustering can use a mixture of both numerical and categorical variables to arrange variables into homogenous clusters, that is, variables that are strongly related to each other [30]. The algorithm for finding these related clusters follows the concepts of PCA and multiple correspondence analysis (MCA). In PCA and MCA, the data set is analyzed to find new linearly independent variables to describe the same set of data. In this hierarchical clustering, these new synthetic variables are used as the center points of the clusters, and each original variable is then grouped according to its similarity to the cluster center, either using the sum of the correlation ratio, for numeric variables, or the squared correlation, for categorical variables.

Machine Learning Techniques

Machine learning techniques are regularly employed for detecting patterns and dependencies within data, such as within health care data. Specifically, machine learning algorithms can be used to look for combinations of variables and generate rules within data that can be used to reliably predict outcomes [25]. This style of problem relies on classification algorithms, where predictor variables are used to predict an outcome or a class variable. These predictions are based on a training sample of the data, usually consisting of a random sample of about 70% to 80% of the available data. The developed model comprises rules based on these training data and then tested against the remaining data (Figure 2). The training procedure is repeated on a number of different subsets of the data to reduce the likelihood of overfitting the model. In this study, 10-fold cross-validation was used to measure the performance of classifiers. Initially, the data were split into a training set (75%) and an evaluation set (25%). The models were trained using the training set with 10-fold cross-validation applied (with a 90%/10% train/test split at each fold). The modeling techniques of decision trees, random forests, and Naive Bayes were selected for their ease of interpretability. It is crucial that the results of modeling in this study can be explained to clinical experts. The individual algorithms were developed using the R caret package, specifically using the train and predict functions. The evaluation data set was used to evaluate the performance of the model found to be optimal during training for each of the 3 respective techniques considered.

Figure 2. Model development and testing protocol.



Decision Tree

Decision trees are one of the most common machine learning algorithms when using a combination of continuous and categorical variables, chosen for their computational efficiency and readability. The Classification and Regression Tree (CART) [31] algorithm, in particular, lends itself well to explanatory knowledge discovery [32] due to its transparency. CART decision trees are developed using a top-down recursive algorithm, where the data set is split into increasingly smaller subsets according to some predetermined metric, most commonly using either the Gini impurity index or a permutation importance measure. The measures used are described below. The rpart implementation of the CART decision tree algorithm in the R caret package was used in this study. This implementation automatically applies pruning, choosing a range of complexity parameters and automatically selecting the optimal model using the complexity parameter that provides the highest accuracy.

The resulting decision tree easily translates itself to a series of rules that can be used to classify the test data. The advantages of using a decision tree classifier lie in its ease of application, particularly as both numerical and categorical input variables require little to no preprocessing; its transparency for interpretation, as the resulting tree can be explained using Boolean logic; and its computational efficiency, particularly with large data sets. In addition, decision tree classification does not require domain knowledge or parameter setting [32]. However, traditional decision trees are also the least robust of the machine learning classification methods, as they are prone to overfitting and therefore rely substantially on the training data. Often, a small change in the training data can result in large changes in the developed tree. These shortcomings can be addressed using the random forest algorithm.

Random Forest

The random forest algorithm [33] works in a similar manner to decision trees, but where the CART algorithm results in a single tree, the random forest algorithm results in a *forest* of trees. Each of the maximal trees within the random forest will have been developed using a random subset of the predictor variables [34]. Each split within the tree is then calculated according to a given performance metric from only within this subset of variables. Typically, many trees are considered, thus reducing the prediction error, as the model prediction will reflect the average prediction across all trees. As a result, the random forest algorithm is considered robust, flexible, and highly suited to large data sets [35]. The random forest algorithm in the R caret package was used in this study. This implementation chooses a range of mtry parameters, where mtry is the number of variables available for splitting at each tree node, which have a strong influence on predictor variable importance estimates [36]. The mtry parameter providing the highest accuracy was used to select the optimal model.

Naïve Bayes

The Naïve Bayes algorithm for classification is based on Bayes' theorem, which describes the most likely outcome (Y) based on k number of observations ($X=\{x_1,x_2,\ldots,x_k\}$). This can be written as P(Y|X) and, as the algorithm is *naïve* and all variables are considered independent, is calculated using the equation in Figure 3.

Figure 3. Naïve Bayes algorithm.

$$P(Y|\mathbf{X}) = \frac{P(\mathbf{X}|Y)P(\mathbf{X})}{P(Y)} = \frac{1}{P(Y)}\prod_{i}^{k} P(x_i|Y)P(x_i)$$

The probability of an outcome P(Y); the probability of an observation being described by X, P(X); and the probability of an observation being described by X, given that they can be classed by Y, P(X|Y), can all be estimated using the given data set. For its use as a classifier, an observation is classified according to the most likely class based on the random variables the observation describes. A benefit of the Naïve Bayes classifier is its theoretical low error rate; however, based on the underlying independence of the variables, in practice, this may not be the case. The Naïve Bayes algorithm in the R caret package was used in this study.

Importance and Accuracy Measures

Gini Impurity Index

The Gini impurity index describes the likelihood of an incorrect classification using a random variable (var) and is described mathematically as shown in Figure 4.

Figure 4. Gini impurity index.

$$\mathrm{Gini(var)} = 1 - \sum_{\mathrm{i}}^{\mathrm{m}} \mathrm{p_i}$$

Here p_{i} is the probability of a correct classification according to m classes. By considering the variables resulting in a minimal Gini impurity index, this metric will therefore determine the best (most pure) variables to use to split the training data until a convergence criterion is met.

Permutation Importance

Permutation variable importance [33] is calculated by using the effect the variable has on the overall prediction performance. This performance can be predicted using the out-of-bag prediction error, calculated by taking the mean prediction error rate of those trees that did not include the specific variable [35].

Performance Evaluation

To compare the performance of each classification model, a variety of evaluation metrics were used. The accuracy, precision, recall, and F_1 scores were computed. Precision, recall, and F_1 scores take account of true and false positives and negatives, whereas accuracy considers only true-positives and true-negatives [37].

Results

Feature Selection

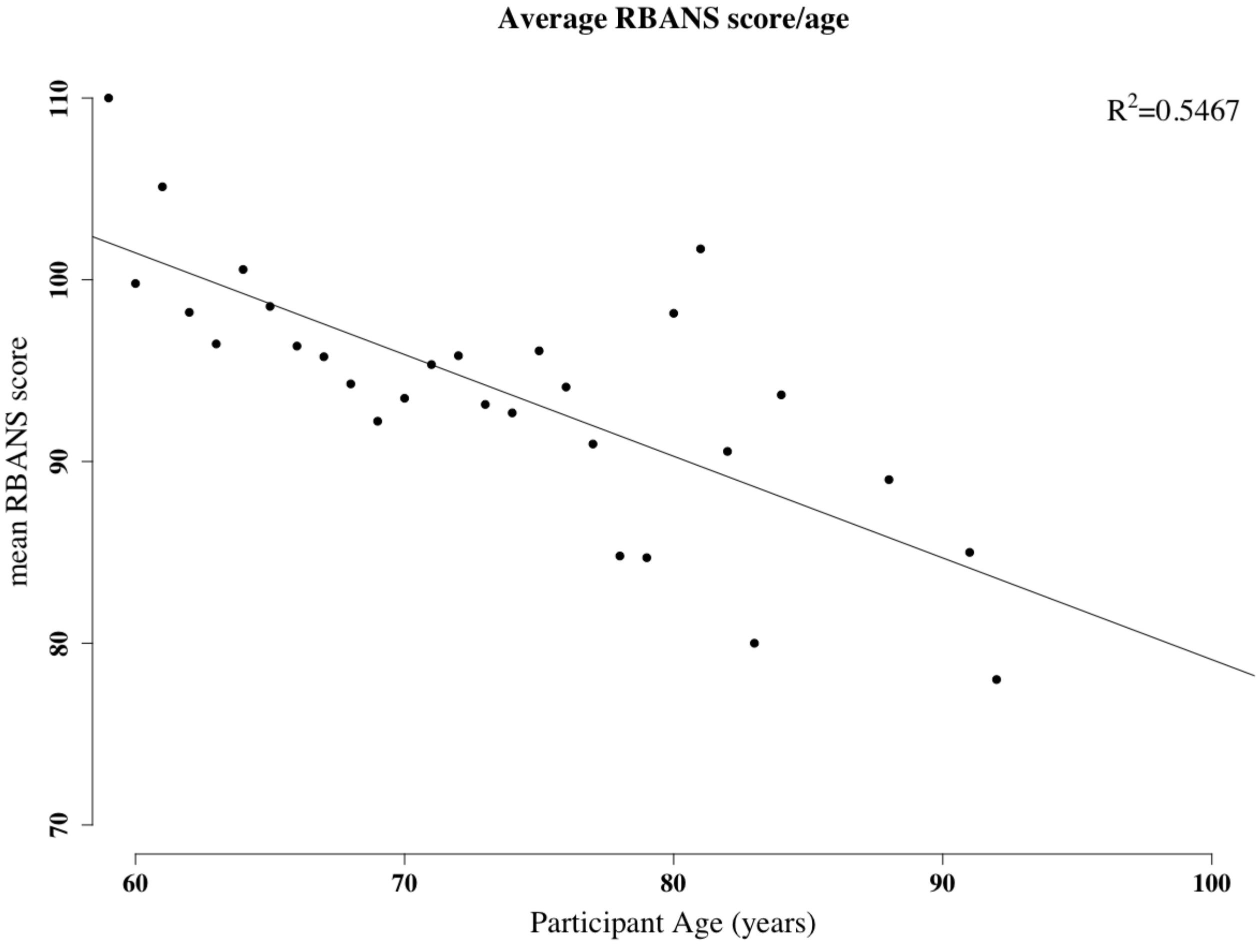
Manual Selection

Initially, 6 features deemed irrelevant for analysis were removed, including participant identification numbers and cohort category (which described the clinic from which the participants were selected). A total of 9 free-text variables and 9 variables with inconsistent questioning were removed. In addition, 94 subjective features were removed in favor of more objective laboratory-obtained results. Several of the removed subjective features had high numbers of missing values; therefore, removal of these in favor of subjective features assisted in handling missing data while ensuring that there was no information loss within the data set and data duplication was also minimized. For example, nutritional status based on blood analysis (eg, measurement of key vitamin biomarkers) was retained over self-reported dietary intake (eg, supplement and fortified food use).

Correlation and Association

Initial investigation into cognitive function with the TUDA data set, as measured using the RBANS score, highlights that as expected RBANS decreases with age (Figure 5).

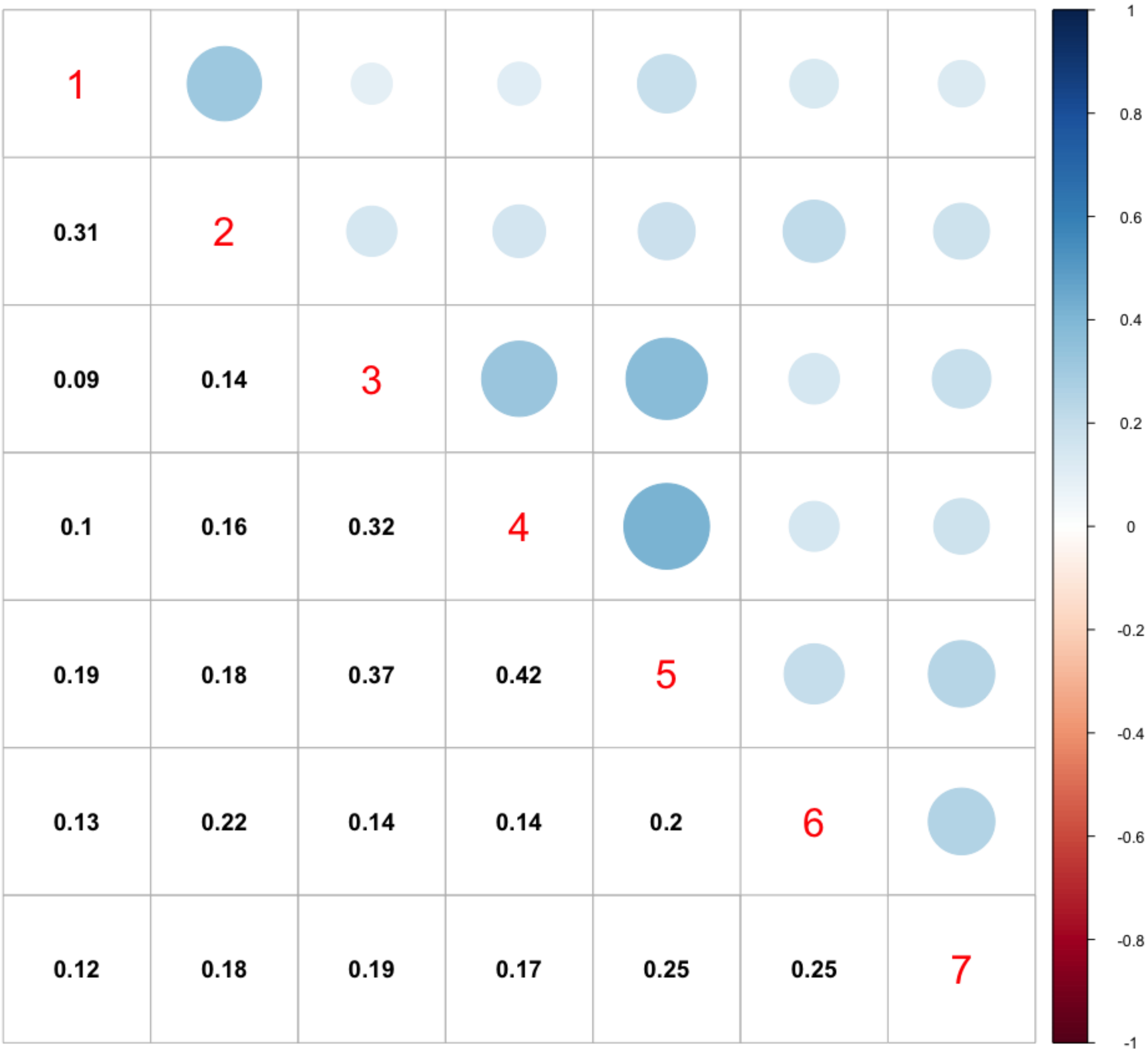
Figure 5. Mean Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) score as a function of participant's age. The graph shows a general decrease in the RBANS score as age increases. RBANS scores have been averaged by age; thus, each point represents the average score for any particular age. One outlier existed for age=86. This was removed and the R value recalculated accordingly.



Correlation and association analyses were carried out. The key results of this analysis are shown in (Multimedia Appendix 1). We observed a relationship between variables concerning follow-up questions within the questionnaire (eg, medication use and duration of use). On the basis of this, 41 features related to follow-up questions were removed. We also observed a high correlation between the use of specific medications (eg, bisphosphonate medications: Risedronate, Ibandronic acid, and Etidronate). These medications could be grouped into bone- and hormone-related categories, and therefore, we amalgamated each subset into a new variable. Specifically, 2 new variables were added for bone- and hormone-related medication, encompassing many types of bone medications, including bisphosphonates and hormone-related medications, from the original data set. This resulted in the removal of 30 features and the addition of 2 new features. Furthermore, scores for each assessment element of RBANS were removed and only the total score was retained. The total RBANS score was later used as the target variable in classification.

We also removed the other neuropsychiatric test results (MMSE, FAB, Hospital Anxiety and Depression Scale, Centre for Epidemiological Studies Depression Scale) and functional test results (instrumental activities of daily living [IADL] and the physical self-maintenance scale [PSMS]) from the data set, as they are clinical assessment tools as opposed to individual predictor variables. This resulted in the removal of 72 additional features. The correlation matrix between these scores is shown in Figure 6.

Figure 6. Correlation matrix using the Spearman (nonparametric) coefficient between participant test scores, ignoring observations with missing data. Variable descriptors are as follows: 1=Hospital Anxiety and Depression Scale total score; 2=depression questionnaire total score; 3=Mini-Mental State Examination total score; 4=Frontal Assessment Battery total score; 5=Repeatable Battery for the Assessment of Neuropsychological Status total score; 6=Physical Maintenance Scale total score; 7=instrumental activities of daily living total score.



The resulting subset of features following this stage of selection reduced the data set from 345 variables to 69 plus the class variable (RBANS score; Multimedia Appendix 2).

Clustering

A cluster analysis was carried out using the ClustOfVar package within R Studio [30] to determine variable clusters and the strengths of their relationships. As expected, the scores from the clinical assessments, RBANS and its subcomponent tests, FAB and MMSE, are closely related (Figure 7). The participant's age was closely related to kidney function, as indicated by the glomerular filtration rate (GFR), and together these form a variable cluster with the scores from the physical diagnostic tests of IADL, TUG, and PSMS indicating a relationship between these variables (Figure 8).

Figure 7. Hierarchical clustering of variables depicted as a dendrogram showing strong relationships between clinical assessment scores from the RBANS, FAB, and MMSE assessments. The variable descriptors are as follows: MMSE_score, Mini-Mental State Examination total score; FAB_score, Frontal Assessment Battery total score; RBANS_index_score_I, Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) immediate memory score; RBANS_index_score_II, RBANS visuospatial constructional score; RBANS_index_score_III, RBANS language score; RBANS_index_score_IV, RBANS attention score; RBANS_index_score_V, RBANS delayed memory score; RBANS_total_score, RBANS total score.

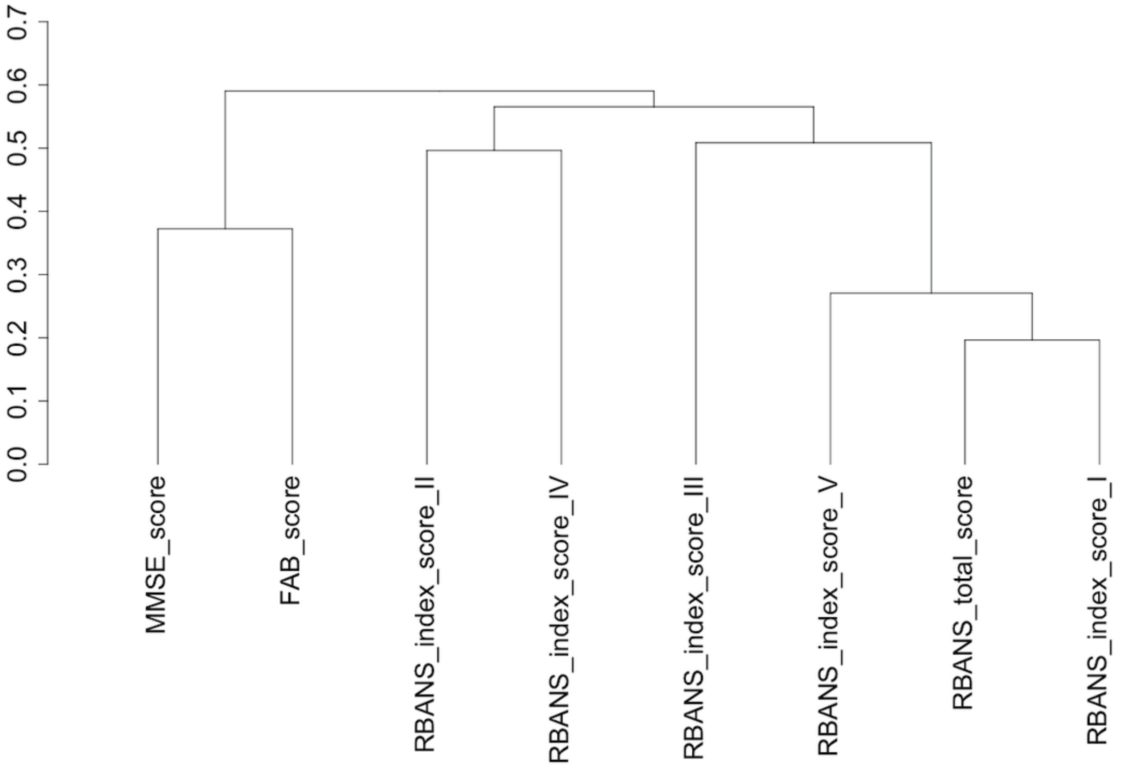
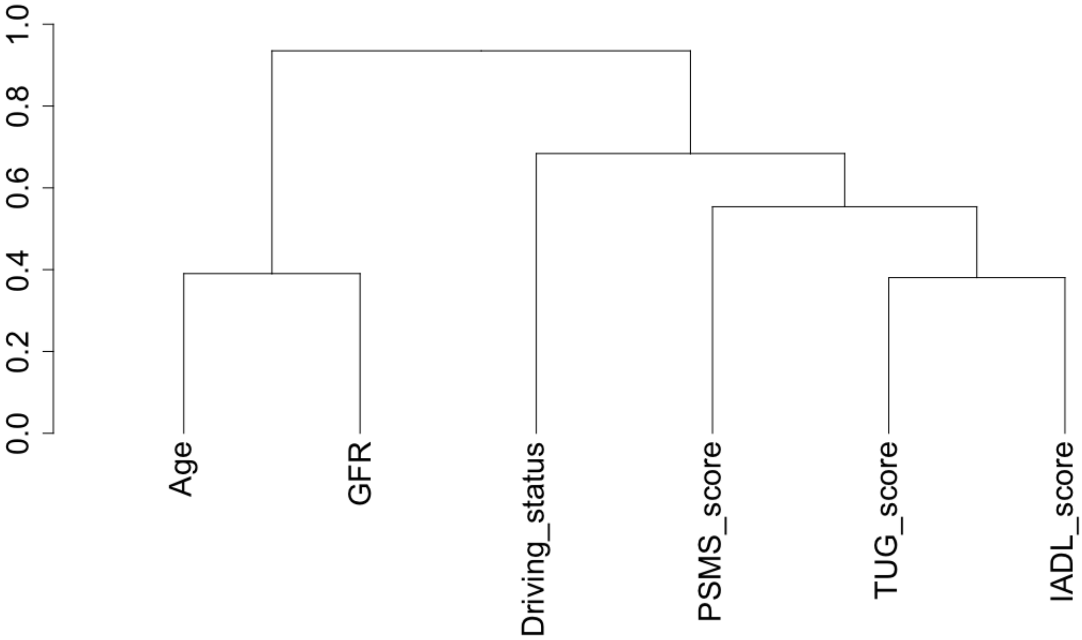


Figure 8. Hierarchical clustering of variables depicted as a dendrogram showing the close relation between a participant’s age and kidney function (glomerular filtration rate [GFR]), which together form a cluster with the physical diagnostic tests of IADL, TUG, and PSMS. The variable descriptors are as follows: age, participant’s age; GFR, kidney function; Driving_status, driving status; PSMS_score, Physical Maintenance Scale total score; TUG score, Timed Up and Go score; IADL_score, Instrumental Activities of Daily Living total score.



Following feature selection, the data set contained 69 features and 5186 observations; however, missing data still remained. To retain as much data as possible while minimizing the chance of statistical bias, participant records were imputed by replacing missing values with the average or expected value, in this case, according to the participant's age and gender. As in other studies on the RBANS score [38], participants with visual (224 participants) or arthritic problems (1445 participants) were omitted as they would have been hindered from carrying out certain tasks within the test, and thus, their results may be unreliable, as were those displaying an MMSE score of <24 (647 participants). Upon removing the relevant records, 2869 observations remained.

RBANS Classification

Classification models were utilized for 2 purposes: to discover if a model could be developed to predict a low RBANS score, representing poorer cognitive function, from the TUDA data set and to determine if the developed model could be used to identify key health, nutritional, and environmental predictors of these low scores.

The target variable in this analysis was the RBANS total score. For this analysis, the RBANS score was categorized using a data-driven clustering approach to find 2 natural groupings within the data identifying those with poorer cognitive performance as having an RBANS score <70 (assigned class *low*) and an RBANS score ≥70 was indicative of normal cognitive performance (assigned class *high*).

Class imbalance [39] within the data set was resolved using oversampling, in which a random sample of the smaller class was replicated until the class sizes were equal.

The supervised modeling techniques of decision trees, random forest, and Naïve Bayes were applied with 69 predictor variables (listed in Multimedia Appendix 2). The data set (n=2869) was split into a training set (2152/2869, 75%) and an evaluation set (717/2869, 25%). The models were trained using the training set with 10-fold cross-validation applied, and the results are shown in Table 2. For the decision tree model, the complexity parameter value of 0.020 for pruning was found to produce the highest accuracy. For the random forest model, the mtry value of 58 was found to produce the highest accuracy.

Table 2. Classification of the Repeatable Battery for the Assessment of Neuropsychological Status score performance measures when models were trained with 10-fold cross-validation (training set size=2152).

Classification technique	Accuracy, mean (SD)	Precision, mean (SD)	Recall, mean (SD)	F_1, mean (SD)
Decision tree	0.737 (0.020)	0.795 (0.037)	0.643 (0.051)	0.709 (0.028)
Naïve Bayes	0.500 (0.000)	0.500 (0.000)	1.000 (0.000)	0.667 (0.000)
Random forest	0.990 (0.006)	1.000 (0.000)	0.981 (0.011)	0.990 (0.006)

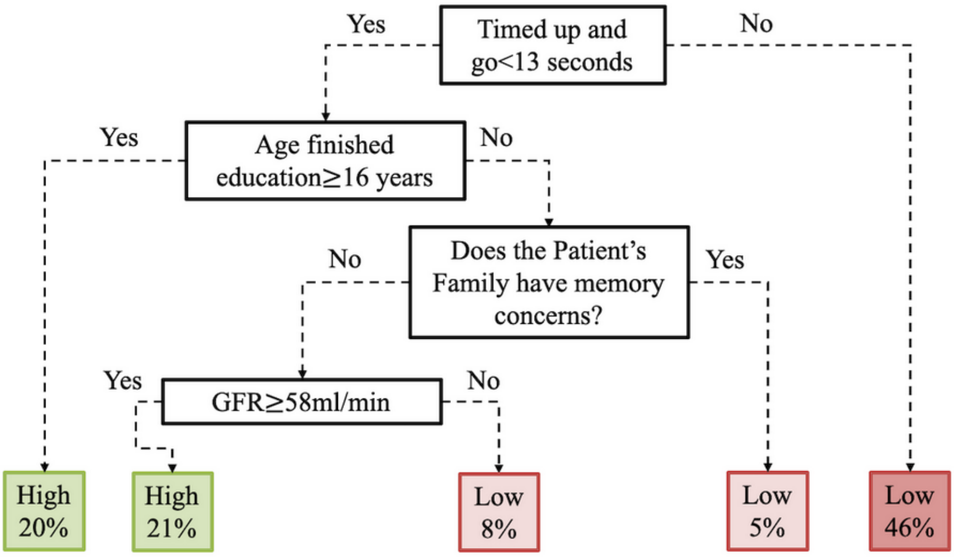
The models were then evaluated using the held out 25% evaluation data set, and the accuracy of these models ranged from 60.4% using the decision tree to 87.7% using the random forest algorithm (Table 3). The random forest algorithm performed best in this comparison in terms of both accuracy and F_1 score, with the decision tree algorithm performing the worst. This is expected in terms of robustness, specifically pertaining to problems with overfitting by the decision tree algorithm, which has been rectified somewhat using multiple trees within the random forest.

Table 3. Classification of the Repeatable Battery for the Assessment of Neuropsychological Status score performance measures when applied to the evaluation data set (training set size=2152; evaluation set size=717).

Classification technique	Overall accuracy	Precision	Recall	F_1 score
Decision tree	0.604	0.926	0.596	0.725
Naïve Bayes	0.876	0.876	0.100	0.934
Random forest	0.877	0.882	0.992	0.934

The key predictors of the RBANS total score in the decision tree were as follows: participants' scores from the TUG functional mobility test, representing the time a participant takes to get out of a chair, walk 3 m, turn around, and walk back to return to his or her original seated position; the age at which the participant stopped education; whether any family members were concerned about the participant's memory; and the participant's GFR, as shown in Figure 9. This decision tree predicted that a person who took under 13 seconds to perform the TUG test and stopped education after 16 years of age was classified as a *high* RBANS scorer (ie, indicative of normal cognitive performance). The decision tree classification model also highlights the importance of the TUG test alone; if a participant took longer than 13 seconds to perform the test, he or she was most likely to be a low scorer, indicative of poorer cognitive performance.

Figure 9. Decision tree classifier of the Repeatable Battery for the Assessment of Neuropsychological Status score. GFR: glomerular filtration rate.


Timed up and go<13 seconds
Yes
No
Age finished education≥16 years
Yes
No
Does the Patient's Family have memory concerns?
No
Yes
GFR≥58ml/min
Yes
No
High 20%
High 21%
Low 8%
Low 5%
Low 46%

Similarly, the Naïve Bayes and random forest algorithms also detect the TUG score, the age at which the participant stopped education, and the participant's age as being highly informative features as shown in Figures 10 and 11 (see Multimedia Appendix 2 for feature descriptions) for Naïve Bayes and random forest models, respectively, with the Naïve Bayes algorithm adding a participant's driving status and the random forest algorithm adding GFR to form the top 4 informative variables within these respective algorithms.

Figure 10. The 20 most important features for classification of the Repeatable Battery for the Assessment of Neuropsychological Status score as detected using feature permutation using a Naïve Bayes classifier. GFR: glomerular filtration rate; LDL: low-density lipoprotein; TUG: Timed Up and Go.

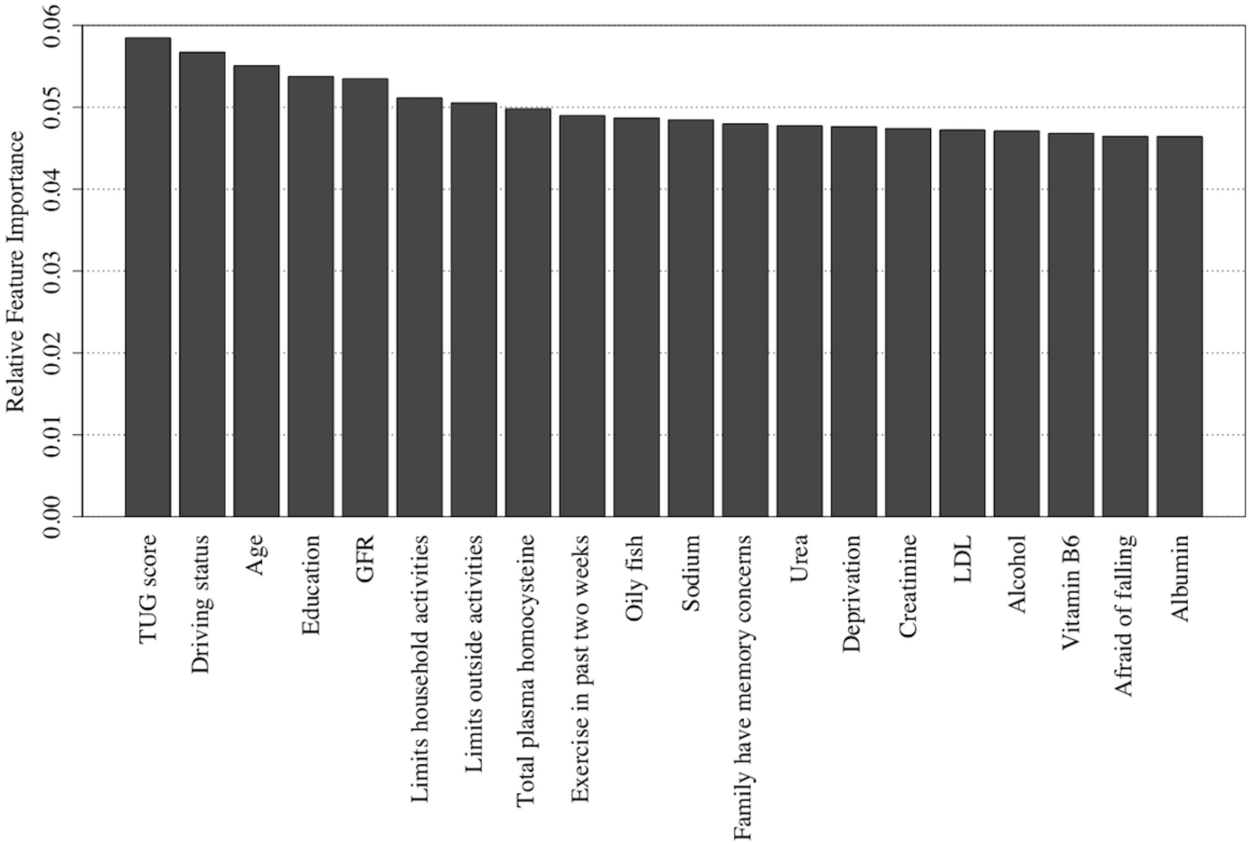
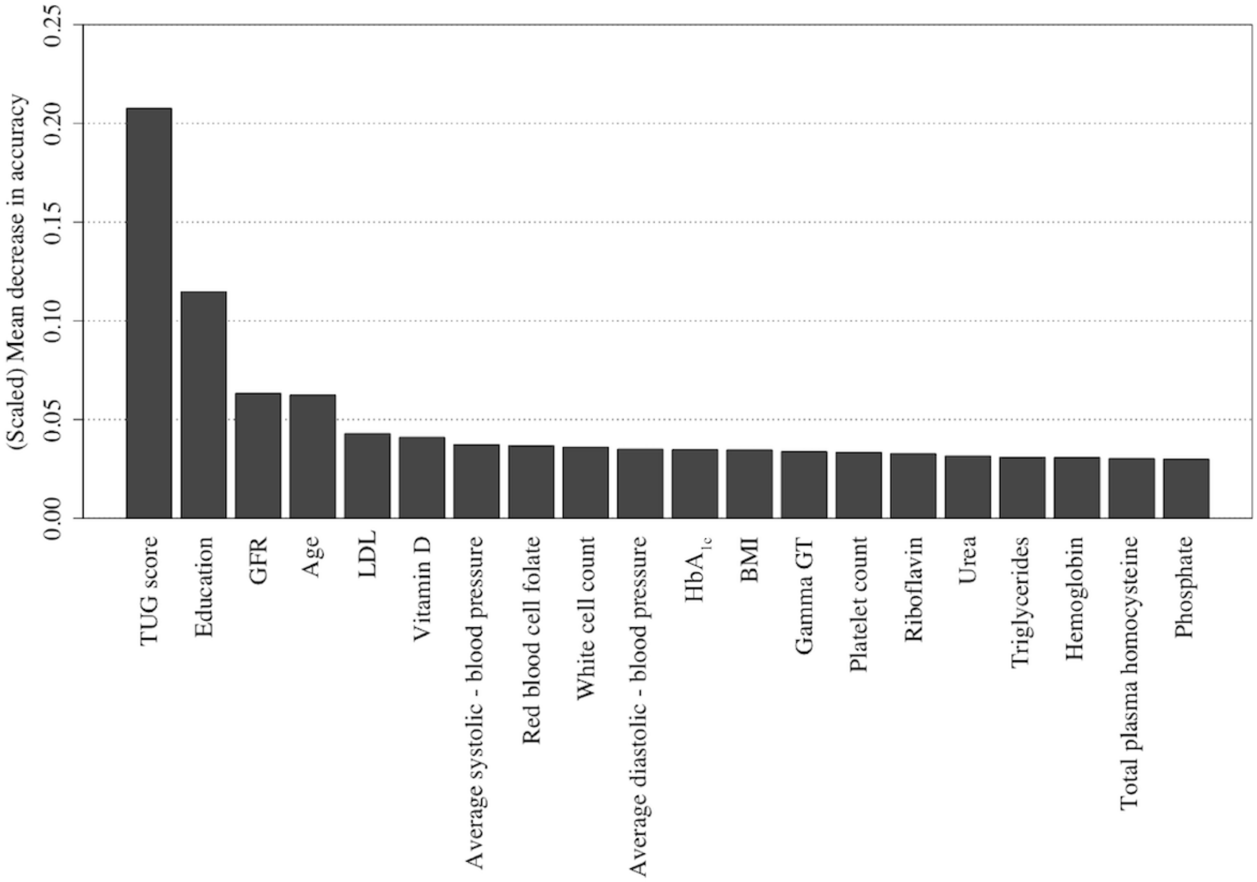


Figure 11. The 20 most important features for classification of the Repeatable Battery for the Assessment of Neuropsychological Status score as detected using feature permutation using a random forest classifier. GFR: glomerular filtration rate; HbA1c: glycated hemoglobin; LDL: low-density lipoprotein; TUG: Timed Up and Go.



The informative nature of the 4 most important features determined by the most accurate classifier (random forest), as shown in Figure 11, was confirmed when these algorithms were rerun using only this subset of 4 features. In addition, 10-fold cross-validation was applied to train the model on the training data set (n=2152), with the results shown in Table 4. For the decision tree model, the complexity parameter value of 0.010 for pruning was found to produce the highest accuracy. For the random forest model, the mtry value of 2 was found to produce the highest accuracy. The models were then evaluated using the held out 25% evaluation data set. Training on the 4 most important features as determined by the random forest model resulted in a decrease in accuracy for the random forest model from 87.7% to 80.1% (Table 5). A larger reduction in accuracy was observed for the Naïve Bayes model, decreasing from 87.6% to 69.3%, whereas the decision tree model increased in accuracy from 60.4% to 72.5% when trained on this reduced data set compared with training on the original data set containing 69 variables.

Table 4. Classification of the Repeatable Battery for the Assessment of Neuropsychological Status score performance measures when models trained with 10-fold cross-validation (training set size=2152) and the 4 key variables: (1) age at which the participant stopped education, (2) the Timed Up and Go score, (3) the glomerular filtration rate measure, and (4) the participant's age.

Classification technique	Accuracy, mean (SD)	Precision, mean (SD)	Recall, mean (SD)	F_1, mean (SD)
Decision tree	0.688 (0.020)	0.702 (0.026)	0.655 (0.045)	0.677 (0.020)
Naïve Bayes	0.693 (0.012)	0.775 (0.021)	0.545 (0.026)	0.640 (0.018)
Random forest	0.929 (0.013)	1.000 (0.000)	0.857 (0.026)	0.923 (0.015)

Table 5. Classification of the Repeatable Battery for the Assessment of Neuropsychological Status score performance measures when models trained using the 4 key variables: (1) age at which the participant stopped education, (2) the Timed Up and Go score, (3) the glomerular filtration rate measure, and (4) the participant's age when applied to the evaluation data set (training set size=2152; evaluation set size=717).

Classification technique	Overall accuracy	Precision	Recall	F_1 score
Decision tree	0.725	0.928	0.732	0.819
Naïve Bayes	0.598	0.946	0.557	0.701
Random forest	0.801	0.878	0.889	0.883

Classifying Cognitive Decline Using the Rate of Change in the RBANS Score

A subset (n=987) of TUDA study participants was reassessed using an identical protocol 5 to 7 years after the initial assessment. The result of this follow-up assessment enabled the creation of a new variable to add to the original TUDA data set for these 987 participants; the rate of change of the RBANS score (calculated using the equation in Figure 1). This variable would act as a measure of predicted cognitive decline (or improvement) over the 5- to 7-year follow-up period. The same classification models of decision tree, Naïve Bayes, and random forest were applied to the TUDA data (n=987), using the new *rate of RBANS change* as the classification variable. If the rate of change of a participant's RBANS score was calculated as more than one half standard deviation below the mean rate of change of the RBANS score across the sample of participants, the participant was considered to have shown *acute decline* over time, otherwise the change in RBANS was considered *normal* or expected. The variable was normalized to adjust for differing periods of time between the first and second RBANS assessments (between 5 and 7 years) among participants. The data set (n=987) was split into a training set (740/987, 75%) and an evaluation set (247/987, 25%). The models were trained using the training set with 10-fold cross-validation applied, and the results are shown in Table 6. For the decision tree model, the complexity parameter value of 0.035 for pruning was found to produce the highest accuracy. For the random forest model, the mtry value of 2 was found to produce the highest accuracy.

Table 6. Classification of the Repeatable Battery for the Assessment of Neuropsychological Status score performance measures when models trained with 10-fold cross-validation (training set size=740).

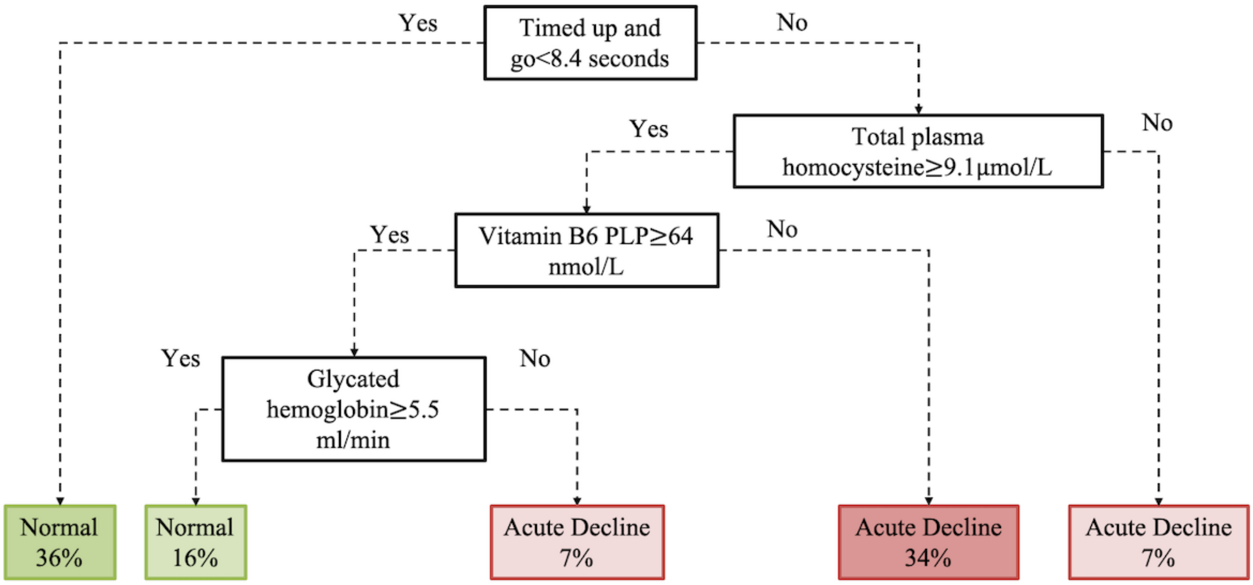
Classification technique	Accuracy, mean (SD)	Precision, mean (SD)	Recall, mean (SD)	F_1, mean (SD)
Decision tree	0.603 (0.045)	0.613 (0.053)	0.571 (0.151)	0.582 (0.083)
Naïve Bayes	0.499 (0.008)	0.499 (0.008)	0.997 (0.009)	0.665 (0.007)
Random forest	0.962 (0.026)	0.978 (0.035)	0.946 (0.031)	0.962 (0.028)

The models were then evaluated using the held out 25% evaluation data set, and the results are shown in Table 7. Although the accuracy of these classification models is lower than that reported for the classification of the RBANS score, approximately 70% versus 90% for random forest classifiers, it nevertheless indicates the possibility of using our existing variables for predicting a perhaps pathological rate of cognitive decline to a reasonable level of accuracy. The decision tree performed the poorest; however, the information it provides (Figure 12) indicates that the TUG test score is again the most informative attribute, followed by the participant's blood measures of total plasma homocysteine, vitamin B6 biomarker pyridoxal-5-phosphate (PLP), and glycated hemoglobin.

Table 7. Classification performance for rate of change of the Repeatable Battery for the Assessment of Neuropsychological Status score when applied to the evaluation data set (training set size=740; evaluation set size=287).

Classification technique	Overall accuracy	Precision	Recall	F_1 score
Decision tree	0.547	0.735	0.605	0.664
Naïve Bayes	0.739	0.739	1.000	0.850
Random forest	0.702	0.735	0.933	0.822

Figure 12. Decision tree classifier of rate of change of the Repeatable Battery for the Assessment of Neuropsychological Status score. PLP: vitamin B6 marker pyridoxal-5-phosphate.



Furthermore, using permutation importance measures (Figures 13 and 14, see Multimedia Appendix 2 for feature descriptions), it has been indicated that the same key variables for the classification of RBANS scores are no longer of such importance for the classification of rate of RBANS score change. Instead, the blood measures of PLP (vitamin B6 biomarker) and urea, coupled with the results of the TUG test and the participant's age, are likely key predictors, particularly using the (best performing) Naïve Bayes algorithm (Figure 13).

Figure 13. The 20 most important features for predicting rate of the Repeatable Battery for the Assessment of Neuropsychological Status change as detected using feature permutation using a Naïve Bayes classifier. Gamma GT: Gamma-glutamyl transferase; GFR: glomerular filtration rate; TUG: Timed Up and Go.

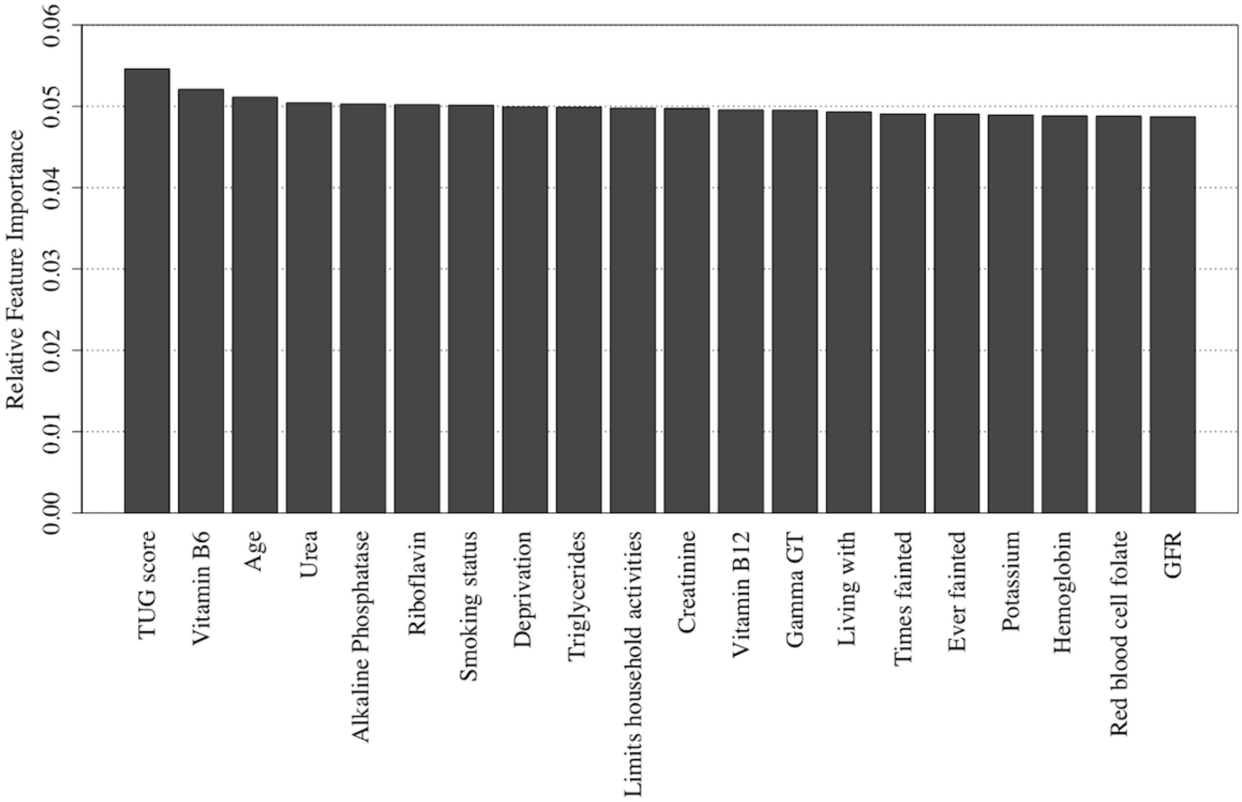
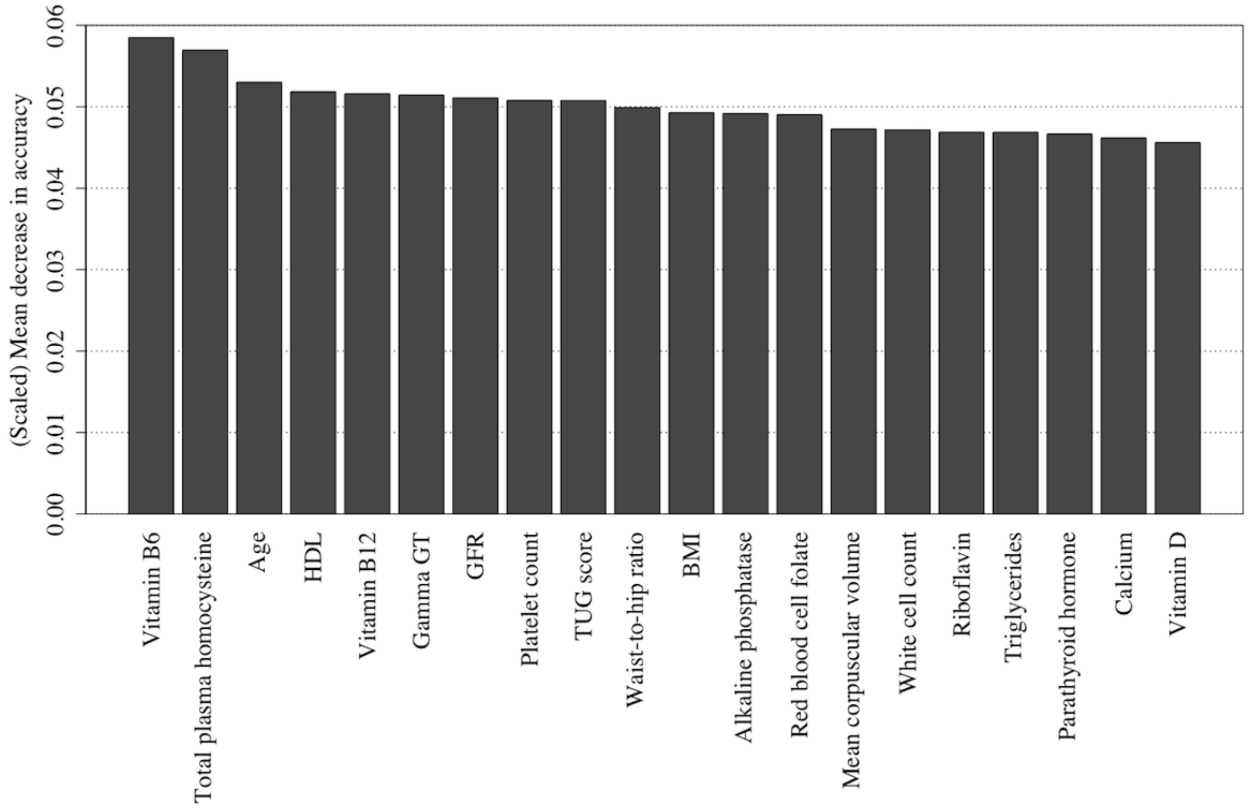


Figure 14. The 20 most important features for predicting rate of the Repeatable Battery for the Assessment of Neuropsychological Status change as detected using feature permutation using a random forest classifier. Gamma GT: Gamma-glutamyl transferase; GFR: glomerular filtration rate; HDL: high-density lipoprotein; TUG: Timed Up and Go.



Discussion

Principal Findings

The results of this study indicate that modeling of a variety of clinical, lifestyle, and sociodemographic factors using machine learning techniques may help predict poorer cognitive function in older people with a high level of accuracy (approximately 90%) and using a small number of noninvasive indicators. The approach is also useful, although slightly less accurate (approximately 70%), in predicting the rate of cognitive decline over a 5- to 7-year period with a small number of measures being the most influential health, nutritional, and environmental predictors. The results are important for clinicians and health service providers, especially at the early stages of engagement and diagnosis of cognitive dysfunction in older patients, by identifying those patients most in need of more intensive investigation. Furthermore, these findings may be useful for informing nutritional and lifestyle interventions aimed at maintaining brain health in the adult population.

The results presented here suggest that it may be possible for a health care professional to make an initial prediction (with a high level of confidence) of cognitive dysfunction using only a few short, noninvasive questions. Although the approach is not a diagnostic instrument for detecting the presence or absence of dementia, it has particular merit in that it could provide a very quick, efficient, and noninvasive screening method to help clinicians decide, at an early consultation stage, whether or not a patient should be investigated further using more in-depth cognitive assessment tools. Similarly, a recent study [14] used a machine learning approach to develop a gradient boosting machine classifier with the KLoSA data set [15], also identified sociodemographic, functional, and health-related factors, among others, as the most important predictors of cognitive impairment. The authors concluded that the model could be used to screen for cognitive impairment in a community health care setting. Using such an approach may offer potential benefits to both health service providers and older patients. It may provide time and cost savings for health service providers reducing the need for cognitive tests that are often laborious to administer (eg, it takes approximately 30 min to complete the RBANS assessment used in this study), and could potentially avoid testing of low-risk patients. As a result, any unnecessary stress associated with cognitive testing may be reduced or avoided in older adults. This study's results also suggest that some additional invasive clinical measures may be required to identify those individuals at greatest risk of future cognitive decline, providing valuable information that could help clinicians design the most appropriate intervention and treatment strategies for patients on a case-by-case basis.

In the prediction of poorer cognitive performance, it is interesting to note that, in addition to participants' age, the models identified noninvasive physical, behavioral, and socioeconomic variables over invasive clinical measures as the most influential predictors (with the exception of GFR), whereas the opposite was true for predicting the rate of change (with TUG being the exception). This suggests that nonclinical factors are much better in predicting poorer cognitive performance in older people, while clinical measures are needed to predict cognitive decline.

Machine learning methods produce the best classification models and predictive outcomes based on the quality and quantity (comprehensiveness) of the input variables. The potential for bias still remains, for example, when a key variable is missing from the data. Consequently, the results from the models need to be evaluated for theoretical and, in health outcome studies, clinical plausibility to determine their value and potential for real-world application [40].

In this study, all 3 models identified TUG and the age at which a participant stopped education as the most important predictive variables. In terms of plausibility, this is encouraging, as both these factors have been frequently identified and cited in the literature in large cohort studies as being important risk factors of cognitive dysfunction [6,41]. In support of these findings, we previously reported using a geodemographic analysis of this cohort that socioeconomic status, namely, area-based deprivation, was an important determinant of cognitive dysfunction alongside age, years of education, depression, and TUG test [42]. The emergence of the age a participant stopped education as the dominant variable from the socioeconomic cluster is particularly interesting as it has consistently been found to be the most important individual socioeconomic factor related to cognitive function across the life cycle [43]. Furthermore, 2 recent population-based longitudinal studies in the United States and the United Kingdom have indicated that higher educational attainment, particularly in early life, could help protect against a decline in cognitive function as people age [44,45]. Reduced physical function, measured using tools such as TUG, has also been associated with lower socioeconomic status [46] and cognitive dysfunction [47]. The TUG test reflects an individual’s strength and mobility, inherently assessing gait, balance, and, to a lesser degree, cognition and vision. It is a screening tool routinely used to assist clinicians in identifying patients at risk of falling [48]. A cutoff of ≥12 seconds is commonly applied to identify individuals at high risk of falls, but these cutoff levels are applied differently across various studies [49]. Within this study, a TUG score of >13 seconds was associated with poor cognitive performance, and a score of >8 seconds predicted future risk of cognitive decline. These selected predictors, and their associated split points, from the machine learning analytics, are consistent with other studies, where poor functional performance was correlated with lower executive function in patients with MCI and Alzheimer disease [50,51], and is associated with future dementia occurrence [52]. Moreover, the TUG test can be considered, in a sense, a global measure of body function. Poor performance has been associated with increased cardiovascular disease and mortality as well as all-cause mortality in older adults [53-55] and in patients with chronic kidney disease [56]. Additional predictors beyond the TUG score selected in the decision trees as informative are also linked with poor cognitive performance, including a measure of kidney function, GFR. Low GFR is associated with poorer cognitive performance [57], with a recent study reporting that individuals with impaired kidney function had lower cognitive performance compared with individuals with normal kidney function. Furthermore, in frail older adults with poor TUG scores, the severity of renal dysfunction is independently correlated with cognitive impairment [58]. Consequently, it is clear that the various machine learning approaches investigated in this study are identifying appropriate factors with known links to cognitive performance.

When the machine learning approaches were applied to identify the predictors of the rate of cognitive decline in TUDA participants over a 5- to 7-year follow-up period, vitamin B6 status (as measured by blood concentrations of the active form of the vitamin, PLP) at baseline emerged, after the TUG test, as one of the key predictors. High proportions of older adults in population-based surveys from the United States and Europe, including the United Kingdom, are reported to have deficient or low B6 status [59]. Vitamin B6 has a number of important biological roles, including immunomodulating effects. In clinical and population-based studies, blood B6 concentrations are found to be inversely associated with inflammatory conditions, neurodegenerative diseases, and depression and to predict the risk of cardiovascular disease and certain cancers [60]. Of note, vitamin B6 and related B vitamins (namely, folate, vitamin B12, and riboflavin) are required as cofactors in one-carbon metabolism, a series of essential reactions involving the transfer of one-carbon units for DNA synthesis and repair and homocysteine metabolism and in the methylation of phospholipids, proteins, DNA, and neurotransmitters [61]. There is a growing body of evidence indicating that one-carbon metabolism and related B vitamins may be important for maintaining cognitive health during aging. The majority of research to date has focused on folate and vitamin B12. Although vitamin B6 has been less extensively investigated, the findings of this study are in agreement with other observational studies. A low vitamin B6 status has been associated with cognitive dysfunction [62,63] and cognitive decline [64,65] in older people. A low vitamin B6 status was associated with cognitive decline in the Veterans Affairs Normative Aging Study [65]. More recently, a low baseline status of vitamin B6 was also associated with a greater-than-expected rate of cognitive decline in a cohort of community-dwelling older adults in Northern Ireland [64]. Of greater importance, a number of randomized controlled trials demonstrated that vitamin B6 supplementation in combination with other B vitamins reduces the rate of cognitive decline in older people [66,67] and a reduced rate of brain atrophy as measured using MRI [68]. Furthermore, other evidence from the TUDA study indicates that vitamin B6, along with folate and riboflavin, is associated with an increased risk of depression [7]. This machine learning approach has identified vitamin B6 as an important determinant of cognitive health in the TUDA study and, whilst biologically plausible and supported by other scientific evidence, the possible beneficial effects of vitamin B6 on cognitive health would need to be confirmed in randomized controlled trials.

What is very interesting from a clinical setting are the changes in the selected predictors within machine learning models when comparing the RBANS total score model versus the rate of change of the RBANS score model. The age at which a participant stopped education is a dominant predictor from the socioeconomic cluster in the RBANS total score model; however, it becomes an uninformative predictor of the rate of change of the RBANS score model and actually disappears from the models. This implies that while this socioeconomic factor is an important predictor of cognitive dysfunction (diagnosis), it is not important when predicting the rate of cognitive decline. Thus, while patients may start off on a different baseline due to socioeconomic predictors, their rate of cognitive decline is not influenced by these socioeconomic predictors.

Although this paper focuses on key health, nutritional, and environmental predictors of cognitive dysfunction and rate of change of cognitive function using machine learning techniques, as part of the project, the research team also sought input from personal and public involvement (PPI): patients, carers, and clinicians. This engagement focused on causation of cognitive dysfunction, particularly in relation to age, activity, and genetics, considered as measures of risk. This aspect of the work in terms of engagement with PPI, their expectations, and how these align with the findings of this work will be the focus of future research publications.

Limitations

This study had several strengths and limitations. The main limitation is that the TUDA study is observational in design and thus residual confounding and reverse causality cannot be ruled out in this analysis. In addition, owing to the low instances of participants with poorer cognitive performance as indicated by an RBANS score below 70 (target class=*low*), this class was underrepresented within the training data set, and therefore, oversampling had to be performed to allow for more balanced classifier training. This artificial approach of boosting the number of samples was necessary for the classifier, but, coupled with the imputation of missing data, no new information would have been attained. This led to an imbalance between the precision and recall accuracy metrics, although this was remedied with the use of the F_1 score. Generally, the algorithms performed well in the classification of the RBANS score. The decision trees performed the poorest, but as explained in the *Results* section, they were still capable of drawing out key and transparent information. Although an extensive comparison of classification approaches was not the focus of this study, we recognize that alternative variations of the algorithms used in this study exist, for example, C4.5 and C5.0 for decision trees as well as other learning algorithms such as neural networks and boosting algorithms. These alternative approaches may yield better results, and we intend to investigate these in the future while ensuring that the interpretability of results remains to be a key objective. In addition, the performance of the classifiers could have been improved using a dimension reduction technique such as PCA; however, this would have impacted the interpretability of the classifier, as was the objective of the study.

The main strength of this study is the utilization of data from the TUDA study, a large and comprehensively characterized cohort of community-dwelling older adults. Furthermore, a subset of the TUDA study cohort was reexamined 5 to 7 years later using standardized protocols at both time points. This enabled changes in cognition to be tracked over time and the rate of cognitive decline to be calculated compared with most observational studies that measure cognition at one time point only. The primary outcome of this study was based on the RBANS test, a sensitive neuropsychiatric battery for global cognitive assessment. As comprehensive data were available, this permitted objective laboratory measures over subjective measures of nutritional status to be included in the analytical models, thus providing more robust data on predictors of cognitive function.

Conclusions

In conclusion, the derived classification models were able to identify a small number of key noninvasive predictors that are able to predict cognitive dysfunction and the rate of change of cognitive function with a high level of accuracy in the TUDA study. The TUG score, the age at which the participant stopped education, and whether or not the participant's family reported memory concerns emerged as key predictors that could potentially be incorporated into a screening tool for cognitive dysfunction for health care professionals to identify individuals in need of further in-depth cognitive evaluation. Given the burden on health care resources, this could result in improvements in the efficiency of dementia screening and present cost and time savings for the relevant health professions. Furthermore, the results provide evidence to identify key targets that could be included in public health strategies aimed at prevention of dementia. Further investigation is necessary to test the accuracy of the identified predictors in other large cohorts and using other cognitive assessment tools. The TUDA data enable extensive opportunities for future investigations of the aging population.

Acknowledgments

The authors would like to acknowledge the support of the eHealth and Data Analytics Dementia Pathfinder Programme and the HSCB eHealth Directorate in this work, under Award 15F-1701. This work was supported by the Irish Department of Agriculture, Food and the Marine and Health Research Board (under the Food Institutional Research Measure) and by the Northern Ireland Department for Employment and Learning (under its Strengthening the All-Island Research Base initiative). The funders of this research had no role in the design, methods, data collection and analysis, or preparation of the paper.

Authors' Contributions

DR, BF, and MB contributed to the design, model development, analysis, and interpretation of the study. CH, LH, AM, CG, HN, PC, and JW contributed to the design and interpretation of data and models. BF and DR drafted the manuscript. CH, LH, AM, CG, and HN contributed to the clinical aspects of the manuscript. All authors reviewed the manuscript critically for scientific and technical content, and all authors gave final approval of this version for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Correlation and association matrix of Trinity-Ulster and Department of Agriculture dataset variables.
[DOC File , 596 KB - medinform_v8i9e20995_app1.doc]

Multimedia Appendix 2
List of remaining quantitative and qualitative variables from the Trinity-Ulster and Department of Agriculture dataset and their descriptions following feature selection stages as determined by manual selection, correlation analysis and clustering.
[DOC File , 77 KB - medinform_v8i9e20995_app2.doc]

References

1. United Nations, Department of Economic and Social Affairs, Population Division. World population prospects: the 2015 revision, key findings and advance tables. Population Dev Rev 2015;41(3):557-561. [doi: 10.1111/j.1728-4457.2015.00082.x]
2. Gauthier S, Reisberg B, Zaudig M, Petersen R, Ritchie K, Broich K, et al. Mild cognitive impairment. Lancet 2006 Apr;367(9518):1262-1270 [FREE Full text] [doi: 10.1016/s0140-6736(06)68542-5] [Medline: 16631882]
3. International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) Mental and Behavioural Disorders. World Health Organisation. 2015. URL: https://icd.who.int/browse10/2016/en#/ [accessed 2020-04-01] [WebCite Cache ID https://icd.who.int/browse10/2016/en#/]
4. World Alzheimer Report 2019: Attitudes to Dementia. Alzheimer's Disease International. 2019. URL: https://www.alz.co.uk/research/WorldAlzheimerReport2019.pdf [accessed 2020-04-01]
5. World Alzheimer Report 2014: Dementia and Risk Reduction. An Analysis of Protective and Modifiable Factors. Alzheimer's Disease International. 2014. URL: https://www.alz.co.uk/research/WorldAlzheimerReport2014.pdf [accessed 2020-04-01]
6. Livingston G, Sommerlad A, Orgeta V, Costafreda S, Huntley J, Ames D, et al. Dementia prevention, intervention, and care. Lancet 2017 Dec 16;390(10113):2673-2734. [doi: 10.1016/S0140-6736(17)31363-6] [Medline: 28735855]
7. Moore K, Hughes CF, Ward M, Hoey L, McNulty H. Diet, nutrition and the ageing brain: current evidence and new directions. Proc Nutr Soc 2018 May;77(2):152-163. [doi: 10.1017/S0029665117004177] [Medline: 29316987]
8. Koh HC, Tan G. Data mining applications in healthcare. J Healthc Inf Manag 2005;19(2):64-72. [Medline: 15869215]
9. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007 Oct 1;23(19):2507-2517. [doi: 10.1093/bioinformatics/btm344] [Medline: 17720704]
10. Bedner P, Steinhäuser C. Crucial role for astrocytes in epilepsy. In: Pathological Potential of Neuroglia: Possible New Targets for Medical Intervention. New York, USA: Springer; 2014.
11. Albert M, Zhu Y, Moghekar A, Mori S, Miller S, Soldan A, et al. Predicting progression from normal cognition to mild cognitive impairment for individuals at 5 years. Brain 2018 Mar 1;141(3):877-887 [FREE Full text] [doi: 10.1093/brain/awx365] [Medline: 29365053]
12. Linz N, Tröger J, Alexandersson J, Wolters M, König A, Robert P. Predicting Dementia Screening and Staging Scores from Semantic Verbal Fluency Performance. In: International Conference on Data Mining Workshops. 2017 Presented at: ICDMW'17; November 18-21, 2017; New Orleans, LA, USA URL: https://doi.org/10.1109/ICDMW.2017.100 [doi: 10.1109/icdmw.2017.100]
13. Petersen R. Mild cognitive impairment as a diagnostic entity. J Intern Med 2004 Sep;256(3):183-194 [FREE Full text] [doi: 10.1111/j.1365-2796.2004.01388.x] [Medline: 15324362]
14. Na K. Prediction of future cognitive impairment among the community elderly: a machine-learning based approach. Sci Rep 2019 Mar 4;9(1):3335 [FREE Full text] [doi: 10.1038/s41598-019-39478-7] [Medline: 30833698]
15. Korean Longitudinal Study of Ageing (KLoSA). Korea Employment Information Service. 2015. URL: https://survey.keis.or.kr/eng/klosa/klosa01.jsp [accessed 2020-06-26]
16. Marcus D, Fotenos A, Csernansky J, Morris J, Buckner R. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. J Cogn Neurosci 2010 Dec;22(12):2677-2684 [FREE Full text] [doi: 10.1162/jocn.2009.21407] [Medline: 19929323]
17. Petersen R, Aisen P, Beckett L, Donohue M, Gamst A, Harvey D, et al. Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. Neurology 2010 Jan 19;74(3):201-209 [FREE Full text] [doi: 10.1212/WNL.0b013e3181cb3e25] [Medline: 20042704]
18. Ellis K, Bush A, Darby D, de Fazio D, Foster J, Hudson P, AIBL Research Group. The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal

study of Alzheimer's disease. Int Psychogeriatr 2009 Aug;21(4):672-687. [doi: 10.1017/S1041610209009405] [Medline: 19470201]

19. Pellegrini E, Ballerini L, Hernandez MD, Chappell F, González-Castro V, Anblagan D, et al. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. Alzheimers Dement (Amst) 2018;10:519-535 [FREE Full text] [doi: 10.1016/j.dadm.2018.07.004] [Medline: 30364671]
20. Ding X, Bucholc M, Wang H, Glass D, Wang H, Clarke D, et al. A hybrid computational approach for efficient Alzheimer's disease classification based on heterogeneous data. Sci Rep 2018 Jun 27;8(1):9774 [FREE Full text] [doi: 10.1038/s41598-018-27997-8] [Medline: 29950585]
21. Korolev I, Symonds L, Bozoki A, Alzheimer's Disease Neuroimaging Initiative. Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification. PLoS One 2016;11(2):e0138866 [FREE Full text] [doi: 10.1371/journal.pone.0138866] [Medline: 26901338]
22. Wirth R, Hipp J. CRISP-DM: Towards a Standard Process Model for Data Mining. In: Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining. 2000 Presented at: ICPAKDDM'00; April 11-13, 2000; Manchester, UK URL: https://pdfs.semanticscholar.org/48b9/293cfd4297f855867ca278f7069abc6a9c24.pdf
23. Randolph C, Tierney M, Mohr E, Chase T. The repeatable battery for the assessment of neuropsychological status (RBANS): preliminary clinical validity. J Clin Exp Neuropsychol 1998 Jun;20(3):310-319. [doi: 10.1076/jcen.20.3.310.823] [Medline: 9845158]
24. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometr Intell Lab 1987 Aug;2(1-3):37-52 [FREE Full text] [doi: 10.1016/0169-7439(87)80084-9]
25. Obermeyer Z, Emanuel E. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219 [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
26. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1107-1135 [FREE Full text] [doi: 10.5555/944919.944968]
27. Campbell M, Swinscow T. Statistics at Square One. Eleventh Edition. Chichester, UK: John Wiley & Sons; 2009.
28. Cramér H. Mathematical Methods of Statistics (PMS-9). Princeton, UK: Princeton University Press; 1999.
29. Glantz S, Slinker B, Neilands T. Primer of Applied Regression & Analysis of Variance. Third Edition. New York, USA: McGraw-Hill Education; 2016.
30. Chavent M, Kuentz-Simonet V, Liquet B, Saracco J. ClustOfVar: an R package for the clustering of variables. J Stat Soft 2012;50(13):1-16 [FREE Full text] [doi: 10.18637/jss.v050.i13]
31. Gordon AD, Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Biometrics 1984 Sep;40(3):874. [doi: 10.2307/2530946]
32. Han J, Kamber M, Pei J. Classification: basic concepts. In: Data Mining: Concepts and Techniques. Third edition. New York, USA: Morgan Kaufmann Publishers; 2011.
33. Breiman L. Random Forests. Mach Learn 2001;45:5-32 [FREE Full text]
34. Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. Pattern Recognit Lett 2010 Oct;31(14):2225-2236 [FREE Full text] [doi: 10.1016/j.patrec.2010.03.014]
35. Nembrini S, König IR, Wright MN. The revival of the Gini importance? Bioinformatics 2018 Nov 1;34(21):3711-3718 [FREE Full text] [doi: 10.1093/bioinformatics/bty373] [Medline: 29757357]
36. Strobl C, Boulesteix A, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics 2008 Jul 11;9:307 [FREE Full text] [doi: 10.1186/1471-2105-9-307] [Medline: 18620558]
37. Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. In: Proceedings of the 23rd international conference on Machine learning. 2006 Presented at: ICML'06; June 25-29, 2006; Pittsburgh, PA URL: https://doi.org/10.1145/1143844.1143874 [doi: 10.1145/1143844.1143874]
38. Hook J, Marquine M, Hoelzle J. Repeatable battery for the assessment of neuropsychological status effort index performance in a medically ill geriatric sample. Arch Clin Neuropsychol 2009 May;24(3):231-235. [doi: 10.1093/arclin/acp026] [Medline: 19549722]
39. Longadge R, Dongre S. Class Imbalance Problem in Data Mining: Review. Int J Comput Sci Netw 2013;2(1):- [FREE Full text]
40. Crown W. Potential application of machine learning in health outcomes research and some statistical cautions. Value Health 2015 Mar;18(2):137-140 [FREE Full text] [doi: 10.1016/j.jval.2014.12.005] [Medline: 25773546]
41. Kivimäki M, Batty G, Pentti J, Shipley M, Sipilä P, Nyberg S, et al. Association between socioeconomic status and the development of mental and physical health conditions in adulthood: a multi-cohort study. Lancet Public Health 2020 Mar;5(3):e140-e149 [FREE Full text] [doi: 10.1016/s2468-2667(19)30248-8] [Medline: 32007134]
42. McCann A, McNulty H, Rigby J, Hughes C, Hoey L, Molloy A, et al. Effect of area-level socioeconomic deprivation on risk of cognitive dysfunction in older adults. J Am Geriatr Soc 2018 Jul;66(7):1269-1275. [doi: 10.1111/jgs.15258] [Medline: 29430638]
43. Norton S, Matthews F, Barnes D, Yaffe K, Brayne C. Potential for primary prevention of Alzheimer's disease: an analysis of population-based data. Lancet Neurol 2014 Aug;13(8):788-794. [doi: 10.1016/S1474-4422(14)70136-X] [Medline: 25030513]

44. Langa K, Larson E, Crimmins E, Faul J, Levine D, Kabeto M, et al. A comparison of the prevalence of dementia in the United States in 2000 and 2012. JAMA Intern Med 2017 Jan 1;177(1):51-58 [FREE Full text] [doi: 10.1001/jamainternmed.2016.6807] [Medline: 27893041]
45. Wu C, Odden M, Fisher G, Stawski R. Association of retirement age with mortality: a population-based longitudinal study among older adults in the USA. J Epidemiol Community Health 2016 Sep;70(9):917-923 [FREE Full text] [doi: 10.1136/jech-2015-207097] [Medline: 27001669]
46. Stringhini S, Carmeli C, Jokela M, Avendaño M, McCrory C, d'Errico S, LIFEPATH Consortium. Socioeconomic status, non-communicable disease risk factors, and walking speed in older adults: multi-cohort population based study. Br Med J 2018 Mar 23;360:k1046 [FREE Full text] [doi: 10.1136/bmj.k1046] [Medline: 29572376]
47. Donoghue O, Horgan NF, Savva G, Cronin H, O'Regan C, Kenny R. Association between timed up-and-go and memory, executive function, and processing speed. J Am Geriatr Soc 2012 Sep;60(9):1681-1686. [doi: 10.1111/j.1532-5415.2012.04120.x] [Medline: 22985141]
48. Barry E, Galvin R, Keogh C, Horgan F, Fahey T. Is the Timed Up and Go test a useful predictor of risk of falls in community dwelling older adults: a systematic review and meta-analysis. BMC Geriatr 2014 Feb 1;14:14 [FREE Full text] [doi: 10.1186/1471-2318-14-14] [Medline: 24484314]
49. Lusardi M, Fritz S, Middleton A, Allison L, Wingood M, Phillips E, et al. Determining risk of falls in community dwelling older adults: a systematic review and meta-analysis using posttest probability. J Geriatr Phys Ther 2017;40(1):1-36 [FREE Full text] [doi: 10.1519/JPT.0000000000000099] [Medline: 27537070]
50. Blackwood J, Shubert T, Forgarty K, Chase C. Relationships between performance on assessments of executive function and fall risk screening measures in community-dwelling older adults. J Geriatr Phys Ther 2016;39(2):89-96. [doi: 10.1519/JPT.0000000000000056] [Medline: 26050194]
51. Ansai J, Andrade L, Nakagawa T, Vale F, Caetano M, Lord S, et al. Cognitive correlates of timed up and go subtasks in older people with preserved cognition, mild cognitive impairment, and Alzheimer's disease. Am J Phys Med Rehabil 2017 Oct;96(10):700-705. [doi: 10.1097/PHM.0000000000000722] [Medline: 28177938]
52. Lee J, Shin D, Jeong S, Son K, Cho B, Yoon J, et al. Association between timed up and go test and future dementia onset. J Gerontol A Biol Sci Med Sci 2018 Aug 10;73(9):1238-1243. [doi: 10.1093/gerona/glx261] [Medline: 29346523]
53. Bergland A, Jørgensen L, Emaus N, Strand B. Mobility as a predictor of all-cause mortality in older men and women: 11.8 year follow-up in the Tromsø study. BMC Health Serv Res 2017 Jan 10;17(1):22 [FREE Full text] [doi: 10.1186/s12913-016-1950-0] [Medline: 28068995]
54. Son K, Shin D, Lee J, Kim S, Yun J, Cho B. Association of timed up and go test outcomes with future incidence of cardiovascular disease and mortality in adults aged 66 years: Korean national representative longitudinal study over 5.7 years. BMC Geriatr 2020 Mar 19;20(1):111 [FREE Full text] [doi: 10.1186/s12877-020-01509-8] [Medline: 32192437]
55. Chua K, Lim W, Lin X, Yuan J, Koh W. Handgrip strength and timed up-and-go (TUG) test are predictors of short-term mortality among elderly in a population-based cohort in Singapore. J Nutr Health Aging 2020;24(4):371-378. [doi: 10.1007/s12603-020-1337-0] [Medline: 32242204]
56. Roshanravan B, Robinson-Cohen C, Patel K, Ayers E, Littman A, de Boer IH, et al. Association between physical performance and all-cause mortality in CKD. J Am Soc Nephrol 2013 Apr;24(5):822-830 [FREE Full text] [doi: 10.1681/ASN.2012070702] [Medline: 23599380]
57. Martens R, Kooman J, Stehouwer C, Dagnelie P, van der Kallen CJ, Koster A, et al. Estimated GFR, Albuminuria, and cognitive performance: the Maastricht study. Am J Kidney Dis 2017 Feb;69(2):179-191. [doi: 10.1053/j.ajkd.2016.04.017] [Medline: 27291486]
58. Coppolino G, Bolignano D, Gareri P, Ruberto C, Andreucci M, Ruotolo G, et al. Kidney function and cognitive decline in frail elderly: two faces of the same coin? Int Urol Nephrol 2018 Aug;50(8):1505-1510. [doi: 10.1007/s11255-018-1900-3] [Medline: 29868939]
59. Bates C, Pentieva K, Prentice A. An appraisal of vitamin B6 status indices and associated confounders, in young people aged 4-18 years and in people aged 65 years and over, in two national British surveys. Public Health Nutr 1999 Dec;2(4):529-535. [doi: 10.1017/s1368980099000713] [Medline: 10656472]
60. Ueland PM, McCann A, Midttun O, Ulvik A. Inflammation, vitamin B6 and related pathways. Mol Aspects Med 2017 Feb;53:10-27. [doi: 10.1016/j.mam.2016.08.001] [Medline: 27593095]
61. McNulty H, Ward M, Hoey L, Hughes CF, Pentieva K. Addressing optimal folate and related B-vitamin status through the lifecycle: health impacts and challenges. Proc Nutr Soc 2019 Aug;78(3):449-462. [doi: 10.1017/S0029665119000661] [Medline: 31155015]
62. Riggs K, Spiro A, Tucker K, Rush D. Relations of vitamin B-12, vitamin B-6, folate, and homocysteine to cognitive performance in the normative aging study. Am J Clin Nutr 1996 Mar;63(3):306-314. [doi: 10.1093/ajcn/63.3.306] [Medline: 8602585]
63. Kim H, Kim G, Jang W, Kim S, Chang N. Association between intake of B vitamins and cognitive function in elderly Koreans with cognitive impairment. Nutr J 2014 Dec 17;13(1):118 [FREE Full text] [doi: 10.1186/1475-2891-13-118] [Medline: 25516359]

64. Hughes C, Ward M, Tracey F, Hoey L, Molloy A, Pentieva K, et al. B-vitamin intake and biomarker status in relation to cognitive decline in healthy older adults in a 4-year follow-up study. Nutrients 2017 Jan 10;9(1):53 [FREE Full text] [doi: 10.3390/nu9010053] [Medline: 28075382]
65. Tucker K, Qiao N, Scott T, Rosenberg I, Spiro A. High homocysteine and low B vitamins predict cognitive decline in aging men: the veterans affairs normative aging study. Am J Clin Nutr 2005 Sep;82(3):627-635. [doi: 10.1093/ajcn.82.3.627] [Medline: 16155277]
66. de Jager CA, Oulhaj A, Jacoby R, Refsum H, Smith A. Cognitive and clinical outcomes of homocysteine-lowering B-vitamin treatment in mild cognitive impairment: a randomized controlled trial. Int J Geriatr Psychiatry 2012 Jun;27(6):592-600. [doi: 10.1002/gps.2758] [Medline: 21780182]
67. Cheng D, Kong H, Pang W, Yang H, Lu H, Huang C, et al. B vitamin supplementation improves cognitive function in the middle aged and elderly with hyperhomocysteinemia. Nutr Neurosci 2016 Dec;19(10):461-466. [doi: 10.1179/1476830514Y.0000000136] [Medline: 24938711]
68. Smith A, Smith S, de Jager CA, Whitbread P, Johnston C, Agacinski G, et al. Homocysteine-lowering by B vitamins slows the rate of accelerated brain atrophy in mild cognitive impairment: a randomized controlled trial. PLoS One 2010 Sep 8;5(9):e12244 [FREE Full text] [doi: 10.1371/journal.pone.0012244] [Medline: 20838622]

Abbreviations

ADNI: Alzheimer’s Disease Neuroimaging Initiative
AIBL: Australian Imaging Biomarkers and Lifestyle Flagship Study of Aging
CART: classification and regression tree
CRISP-DM: cross-industry process for data mining
FAB: frontal assessment battery
GFR: glomerular filtration rate
IADL: instrumental activities of daily living
KLoSA: Korean Longitudinal Study of Aging
MCA: multiple correspondence analysis
MCI: mild cognitive impairment
MMSE: Mini-Mental State Examination
MRI: magnetic resonance imaging
PCA: principal component analysis
PLP: vitamin B6 marker pyridoxal-5-phosphate
PSMS: physical self-maintenance scale
RBANS: Repeatable Battery for the Assessment of Neuropsychological Status
TUDA: Trinity-Ulster and Department of Agriculture
TUG: Timed Up and Go

Edited by G Eysenbach; submitted 03.06.20; peer-reviewed by C Friedrich; comments to author 24.06.20; revised version received 10.07.20; accepted 23.07.20; published 16.09.20.

Please cite as:
Rankin D, Black M, Flanagan B, Hughes CF, Moore A, Hoey L, Wallace J, Gill C, Carlin P, Molloy AM, Cunningham C, McNulty H
Identifying Key Predictors of Cognitive Dysfunction in Older People Using Supervised Machine Learning Techniques: Observational Study
JMIR Med Inform 2020;8(9):e20995
URL: http://medinform.jmir.org/2020/9/e20995/
doi:10.2196/20995
PMID:32936084

Original Paper

Chinese Clinical Named Entity Recognition in Electronic Medical Records: Development of a Lattice Long Short-Term Memory Model With Contextualized Character Representations

Yongbin Li[1], ME; Xiaohua Wang[1], PhD; Linhu Hui[1], ME; Liping Zou[1], ME; Hongjin Li[1], ME; Luo Xu[1], PhD; Weihai Liu[2], MS

[1]School of Medical Information Engineering, Zunyi Medical University, Zunyi, China

[2]Radiology Department, Beilun District People’s Hospital, Ningbo, China

Corresponding Author:
Yongbin Li, ME
School of Medical Information Engineering
Zunyi Medical University
6 Xuefu Road West, Xinpu New District.
Zunyi, 563000
China
Phone: 86 18311545098
Fax: 86 0851 28642668
Email: bynn456@126.com

Abstract

Background: Clinical named entity recognition (CNER), whose goal is to automatically identify clinical entities in electronic medical records (EMRs), is an important research direction of clinical text data mining and information extraction. The promotion of CNER can provide support for clinical decision making and medical knowledge base construction, which could then improve overall medical quality. Compared with English CNER, and due to the complexity of Chinese word segmentation and grammar, Chinese CNER was implemented later and is more challenging.

Objective: With the development of distributed representation and deep learning, a series of models have been applied in Chinese CNER. Different from the English version, Chinese CNER is mainly divided into character-based and word-based methods that cannot make comprehensive use of EMR information and cannot solve the problem of ambiguity in word representation.

Methods: In this paper, we propose a lattice long short-term memory (LSTM) model combined with a variant contextualized character representation and a conditional random field (CRF) layer for Chinese CNER: the Embeddings from Language Models (ELMo)-lattice-LSTM-CRF model. The lattice LSTM model can effectively utilize the information from characters and words in Chinese EMRs; in addition, the variant ELMo model uses Chinese characters as input instead of the character-encoding layer of the ELMo model, so as to learn domain-specific contextualized character embeddings.

Results: We evaluated our method using two Chinese CNER datasets from the China Conference on Knowledge Graph and Semantic Computing (CCKS): the CCKS-2017 CNER dataset and the CCKS-2019 CNER dataset. We obtained F1 scores of 90.13% and 85.02% on the test sets of these two datasets, respectively.

Conclusions: Our results show that our proposed method is effective in Chinese CNER. In addition, the results of our experiments show that variant contextualized character representations can significantly improve the performance of the model.

(JMIR Med Inform 2020;8(9):e19848) doi:10.2196/19848

KEYWORDS

clinical named entity recognition; ELMo; lattice LSTM; deep learning; neural network; sequence tagging

Introduction

Background

Electronic medical records (EMRs) are an important data resource to describe patients' disease conditions or treatment processes. They are records written by clinicians using unstructured free text to describe medical activities for individual patients. By analyzing EMRs, a large amount of patient-related medical knowledge can be mined [1]. With the generation of a larger number of EMRs and the potential demand for medical information services and medical decision support, they have attracted much attention from researchers.

Clinical named entity recognition (CNER) aims to automatically identify clinical entities in EMRs and classify them into predefined categories, such as disease, image review, laboratory examination, operation, drug, and anatomy [2]. CNER is the key component of clinical text mining and EMR information extraction research and is used for clinical decision support in medical informatics [3]. At the same time, CNER can also provide support for disease diagnosis and medical knowledge base construction, so as to improve overall medical quality [4]. Compared with English CNER and due to the complexity of Chinese word segmentation and grammar, Chinese CNER was implemented later and is more challenging. As a public task, Chinese CNER has been introduced three times at the China Conference on Knowledge Graph and Semantic Computing (CCKS), from 2017 to 2019, in order to promote the information extraction of Chinese EMRs. In this paper, we conducted research and experiments with our Chinese CNER approach, based on the CCKS-2017 (Task 2) CNER dataset and the CCKS-2019 (Task 1) CNER dataset.

CNER is generally performed as a sequence tagging problem to identify and extract entity references related to clinical medicine. For the English CNER task, several neural network architectures have been proposed and achieved excellent performance; among them, the most widely used system is a combination of bidirectional long short-term memory (BiLSTM) and conditional random fields (CRFs) [5-7]. Ma and Hovy [8] presented the BiLSTM-convolutional neural network (CNN)-CRF model with CNN and achieved an approximately equal performance. Compared to named entity recognition (NER) in other fields, Chinese CNER is more challenging. Medical texts often use nonstandard abbreviations, or the same entity has multiple forms; for example, "奥沙利铂" (oxaliplatin) is the same as "奥沙利柏" (oxaliplatin) [9]. The more critical problem is that the Chinese grammatical structure is more complex than the English structure, and there is no natural word-segmentation boundary in Chinese, which may lead to word-segmentation error propagation in CNER [10]. In view of the dependence of Chinese word segmentation, Zhang and Yang [11] put forward an innovative lattice long short-term memory (LSTM) model for Chinese NER. Lattice LSTM is character based and effectively utilizes the corresponding potential word information, which is superior to character-based and word-based models in many Chinese general datasets.

Compared with statistical learning methods, which need to design or extract hand-crafted features based on domain-specific knowledge, deep learning methods usually use distributed representation as the input feature. Traditional pretrained character-embedding models, such as word2vec [12] and Global Vectors for Word Representation (GloVe) [13], train embedding based on their syntactic and semantic similarity in sentence-level contexts, but the training result is a context-independent character vector. In fact, a character may have completely different meanings in different contexts. For instance, in the sentence "考虑为腺癌，于5月30日给予TP方案化疗（紫杉醇240MG静脉滴注，顺铂90MG腹腔灌注），过程顺利，无明显副作用," the meanings of both characters "顺" are different depending on their context. Reasonably, the two characters "顺" should have different vector representations. The Embeddings from Language Models (ELMo) [14] model, which provides deep contextualized word representations, allows the same word to have different vector representations in different sentences. The ELMo model was originally proposed for English text and generates specific English word vectors for each sentence, not character vectors. However, the lattice LSTM model is essentially based on Chinese characters; therefore, we modified the ELMo model to replace the character-encoding layer with domain-specific Chinese characters as input, so that the domain-specific ELMo embedding of Chinese characters was obtained.

In this paper, we propose a lattice LSTM model combined with a variant contextualized character representation and CRF layer for Chinese CNER. By taking advantage of the lattice LSTM structure, our approach can control the long-term state with the combination of word information to make full use of EMR information. Moreover, a variant ELMo model is projected into the lattice LSTM model to help it obtain contextual semantic information. Finally, a CRF layer is used to capture the dependencies between adjacent labels. We can summarize the main contributions of our work as follows:

1. We used the medical field texts to train domain-specific character embedding and word embedding; since traditional word embedding is difficult to use for capturing contextual semantics, the addition of the variant ELMo model can help the model combine the contextualized character representations on the basis of character information and potential word information.
2. This is the first time the variant ELMo embedding has been integrated into the lattice LSTM model and applied to Chinese CNER research. Compared with other prevalent models, it has achieved relatively competitive results with F1 scores of 90.13% and 85.02% on two Chinese CNER datasets, respectively.

Prior Work

CNER

In the first research studies on CNER, rule-based methods [15] and dictionary-based methods [16] were the most common methods. For instance, Savova et al [17] and Zeng at al [18] combined manual rules and heuristic rules to identify medical entities with good results. Because of the grammatical complexity of Chinese clinical texts, rule-based methods need a lot of hand-crafted rules, which cannot identify enough entities and are difficult to transfer to other fields. Statistical learning algorithms are mainly based on single-word classification or sequence tagging, which can consider the tagging results of adjacent words jointly [19,20]; these algorithms include support vector machines (SVMs) [21], CRFs [22], and structured SVMs. Finkel et al [23] used CRF to establish an automatic annotation model for NER, which mainly considered the characteristics of words, prefixes, parts of speech sequences, and word morphologies. However, statistical learning methods rely heavily on complex feature engineering and resources for specific tasks. Collobert et al [24] took the lead in solving the NER problem with a neural model, and used the word embedding as the input feature. With the extensive application of deep learning in the field of natural language processing (NLP), various neural networks have been applied to sequence tagging tasks [25].

Systematic research on EMR entity recognition was initiated by i2b2 (Informatics for Integrating Biology and the Bedside) as a public evaluation task in 2010 [26]. This evaluation first classified EMR entities [27], mainly identifying three types of entities: medical problems, treatment, and examination. For Chinese CNER, Feng et al [28] first carried out CNER research on Chinese EMRs, using the CRF model and manually compiled dictionaries. In the Chinese CNER, the open dataset is extremely lacking, and only the CCKS evaluation tasks published the datasets; they were published three times, between 2017 and 2019. The BiLSTM-CRF model, with self-taught and active learning proposed by Xia and Wang [29], reached an F1 score of 88.98% on the CCKS-2017 CNER dataset. Since there is no clear word-boundary information in Chinese text, Chinese CNER systems can be generally divided into character-based and word-based methods. However, the character-based method may lose word-level information, while the word-based method suffers from word-segmentation error propagation.

Word Embedding

In general, the deep learning method uses word embedding trained from a large-scale unlabeled corpus as a model input instead of feature engineering. The most representative, pretrained word vectors—word2vec [12], GloVe [13], and a semisupervised learning method [30]—can capture fine-grained semantic and syntactic information from unlabeled text. Most of the pretrained word-embedding models are trained on the general corpus, and the semantic similarity measurement built for a general purpose is not effective in a specific field. In specific fields such as clinical text mining, there are many clinical entities and syntactic blocks that contain rich domain information, and the semantics of words are closely related to them; therefore, we need to use a specific corpus to train domain-specific embedding [31].

Most of the embedding models only produce context-independent representation for each word, so it is difficult to obtain contextual semantic information. Current research focuses on contextual vector representation; for example, context2vec [32] uses the LSTM model to encode context around a center word or some unsupervised language model [33]. Devlin et al [34] proposed a pretrained language model, Bidirectional Encoder Representations from Transformers (BERT), which achieved state-of-the-art results in many NLP tasks. This paper adopts the contextualized word-embedding (ie, ELMo) model introduced by Peters et al [14] and modifies it to adapt to Chinese characters.

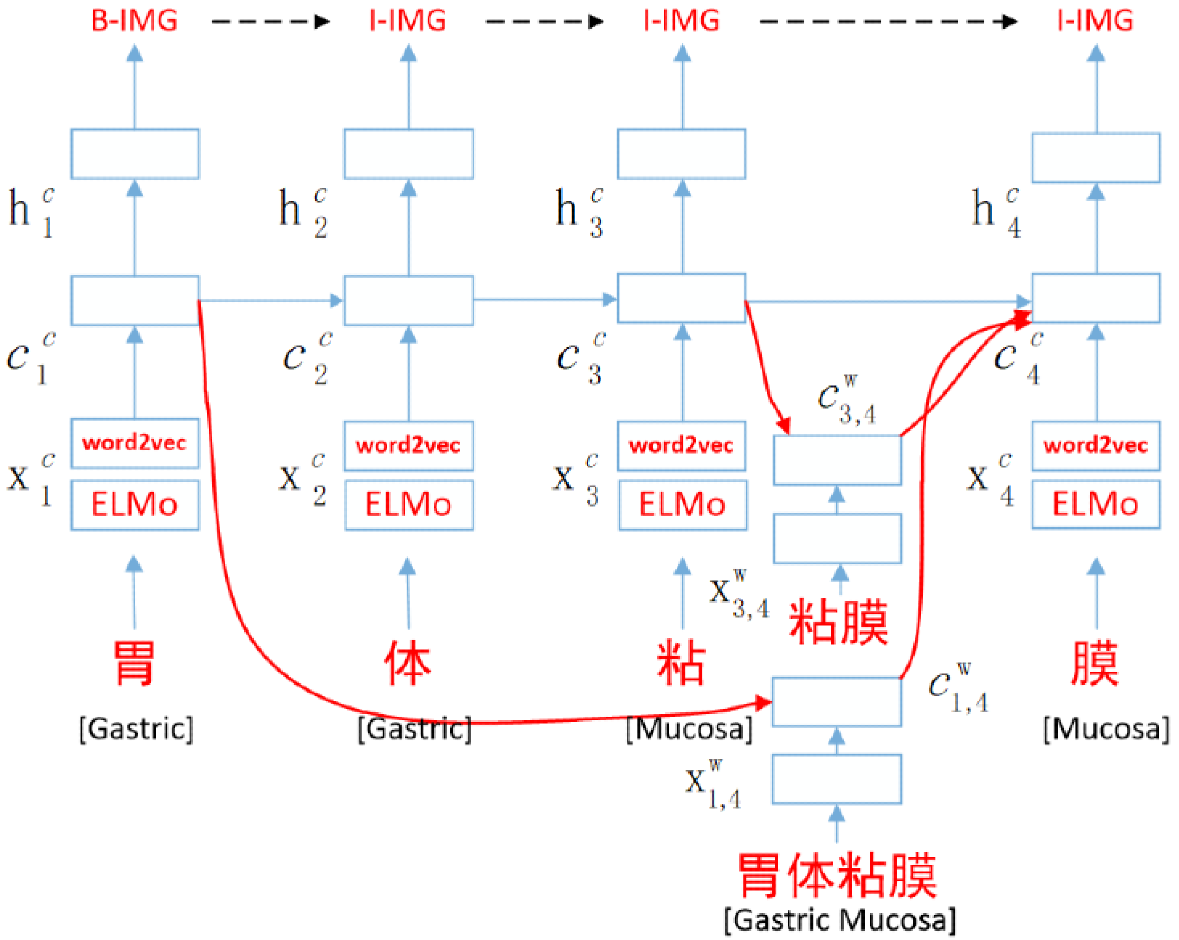
Methods

Model

Overview

In this section, we propose the ELMo-lattice-LSTM-CRF model in detail; its architecture is shown in Figure 1. First, we concatenated the ELMo embedding and the word2vec embedding as the input of the character-embedding part of the lattice LSTM model. Second, embedding of the subsequence from lexicon *D* was used as the input of the word-embedding part. Finally, a CRF layer was used to predict the label probability. We illustrate these three parts of the ELMo-lattice-LSTM-CRF model with real clinical text (ie, “胃体粘膜” [gastric mucosa]) as an example.

Figure 1. Architecture of the ELMo-lattice-LSTM-CRF model. B-IMG: beginning of image entity; c: cell memory; CRF: conditional random field; ELMo: Embeddings from Language Models; h: hidden state; I-IMG: inside of image entity; LSTM: long short-term memory; superscript c: character sign; superscript w: word sign; x: embedding of a character or word.



Lattice LSTM

The lattice LSTM model can be regarded as an extension of the character-based method, which takes the addition of character embedding and weighted-word embedding as the input of the model. The input is a sequence of m characters as ($c_1, c_2,..., c_m$), together with words that are obtained by matching the clinical text in lexicon D. We used the Gensim word2vec tool to train the unlabeled clinical corpus to obtain domain-specific character embedding and word embedding. This clinical corpus includes the CCKS-2017 CNER dataset, the CCKS-2019 CNER dataset, the unlabeled corpora provided by these two tasks, a health care and learning community [35], and the China National Knowledge Infrastructure (CNKI) medical abstracts [36], with a total of 526,631 sentences. In the known literature, there is no publicly available medical domain lexicon D, so we use the annotated entities in the Chinese CNER datasets provided by the CCKS-2017 and CCKS-2019 datasets and the dictionaries captured through open sources; finally, we built a medical terminology dictionary at a scale of about 23 kB. The term $w^d_{b,e}$ denotes the subsequence of matching lexicon D in clinical text, beginning with character index b and ending at index e, as an example in Figure 1; the subsequence $w^d_{1,2}$ is “胃体” (gastric), and $w^d_{1,4}$ is “胃体黏膜” (gastric mucosa). The term $x^w_{b,e}$ is the embedding of subsequence $w^d_{b,e}$. The character-level recurrent LSTM functions are shown below:

$$f^c_t = \sigma (W^{ct}_f [h^c_{t-1}, x^c_t]) + b_f \ \mathbf{(1)}$$

$$o^c_t = \sigma (W^{ct}_o [h^c_{t-1}, x^c_t]) + b_o \ \mathbf{(2)}$$

$$i^c_t = \sigma (W^{ct}_i [h^c_{t-1}, x^c_t]) + b_i \ \mathbf{(3)}$$

$$(c^c_t)\sim = tanh (W^{ct}_c[h^c_{t-1}, x^c_t]) + b_c \mathit{(4)}$$

$$c^c_t = f^c_t \times c^c_{t-1} + i^c_t \times (c^c_t)\sim \ \mathbf{(5)}$$

$$h^c_t = o^c_t \times tanh (c^c_t) \ \mathbf{(6)}$$

where i^c_t, o^c_t, f^c_t, and c^c_t represent input, output, forget gates, and the cell memory, respectively. W and b are model parameters and σ ( ) denotes the sigmoid function.

A word cell $c^w_{b,e}$, which is calculated by the following formula, is used to represent the recurrent state of $x^w_{b,e}$:

$$f^w_{b,e} = \sigma (W^{wt}_f [x^w_{b,e}, h^c_b]) + b_f \ \mathbf{(7)}$$

$$i^w_{b,e} = \sigma (W^{wt}_i [x^w_{b,e}, h^c_b]) + b_i \ \mathbf{(8)}$$

$$(c^w_{b,e})\sim = tanh (W^{wt}_c [x^w_{b,e}, h^c_b]) + b_c \ \mathbf{(9)}$$

$$c^{w}_{b,e} = f^{w}_{b,e} \times c^{c}_{b} + i^{w}_{b,e} \times (c^{w}_{b,e})\sim \quad \mathbf{(10)}$$

where $i^{w}_{b,e}$ is the input gate and $f^{w}_{b,e}$ is the forget gate. Compared with the standard LSTM model, there is no output gate for word units, since label prediction is only on the character sequence.

At each time step, multiple information $c^{w}_{b,e}$ flows into c^{c}_{t} through recurrent paths. Take the previous clinical text as an example: the input resources for c^{c}_{4} include x^{c}_{4} (“膜” [mucosa]), $c^{w}_{3,4}$ (“粘膜” [mucosa]), and $c^{w}_{1,4}$ (“胃体黏膜” [gastric mucosa]). We add all $c^{w}_{b,e}$ with weights $b\,\tilde{}\in(b\,\tilde{}\,|i^{w}_{b\,\tilde{},e}\in D)$ to c^{c}_{e}; an additional gate $i^{c}_{b,e}$ controls the contribution of each subsequence into c^{c}_{e}:

$$i^{c}_{b,e} = \sigma\,([x^{c}_{e}, c^{w}_{b,e}]) + b^{l} \quad \mathbf{(11)}$$

The function for calculating cell values c^{c}_{t} becomes equation 12. Among them, the gate values $i^{c}_{b,t}$ and i^{c}_{j} are normalized (sum to 1) to $\alpha^{c}_{b,t}$ and α^{c}_{t}:

$$c^{c}_{t} = \Sigma_{weights}\alpha^{c}_{b,t} \times c^{w}_{b,t} + \alpha^{c}_{t} \times (c^{c}_{t})\sim \quad \mathbf{(12)}$$

The final hidden vectors h^{c}_{t} are still calculated according to equation 6. According to the above deduction, we find that the lattice LSTM model can focus on relevant words dynamically during NER labeling and can make comprehensive use of the character information and word information of clinical text.

ELMo

Unlike most widely used, pretrained word-embedding models, ELMo [14] word representations are calculated by the entire input sentence. The sentence first passes through a convolutional character-encoding layer; it is then sent to the two-layer bidirectional language model (BiLM) layer, and the resulting vector is sent to the scalar mixer layer to get the ELMo embedding. Specifically, given a sequence of N tokens (t_1, t_2,..., t_N), a BiLM computes and combines the current tokens’ t_k probabilities in both the forward and backward directions. Its goal is to maximize the following likelihood values:

$$\Sigma^{N}_{k=1}\,(logp\,(t_k|t_1,\ldots,t_{k-1};\,\theta_x,\,\theta_{LSTM}(right),\theta_s) + logp\,(t_k|t_{k+1},\ldots,t_N;\,\theta_x,\,\theta_{LSTM}(left),\,\theta_s)) \quad \mathbf{(13)}$$

Where θ_x, θ_s, $\theta_{LSTM}(right)$, and $\theta_{LSTM}(left)$ are the token representation, the Softmax layer, and the forward- and backward-direction LSTM parameters, respectively.

For each token t_k, an L-layer BiLM calculates a set of 2L+1 representations as follows:

$$R_k = \{X^{LM}_{k},\ h^{LM}_{k,j}(right),\ h^{LM}_{k,j}(left)|j=1,\ldots,L\} = \{h^{LM}_{k,j}|j=0,\ldots,L\} \quad \mathbf{(14)}$$

Where $h^{LM}_{k,0}$ *is* the token layer and $h^{LM}_{k,j} = [h^{LM}_{k,j}(right);\ h^{LM}_{k,j}(left)]$ for each BiLSTM layer.

For these representations, the paper makes a scalar mixer with the following formula:

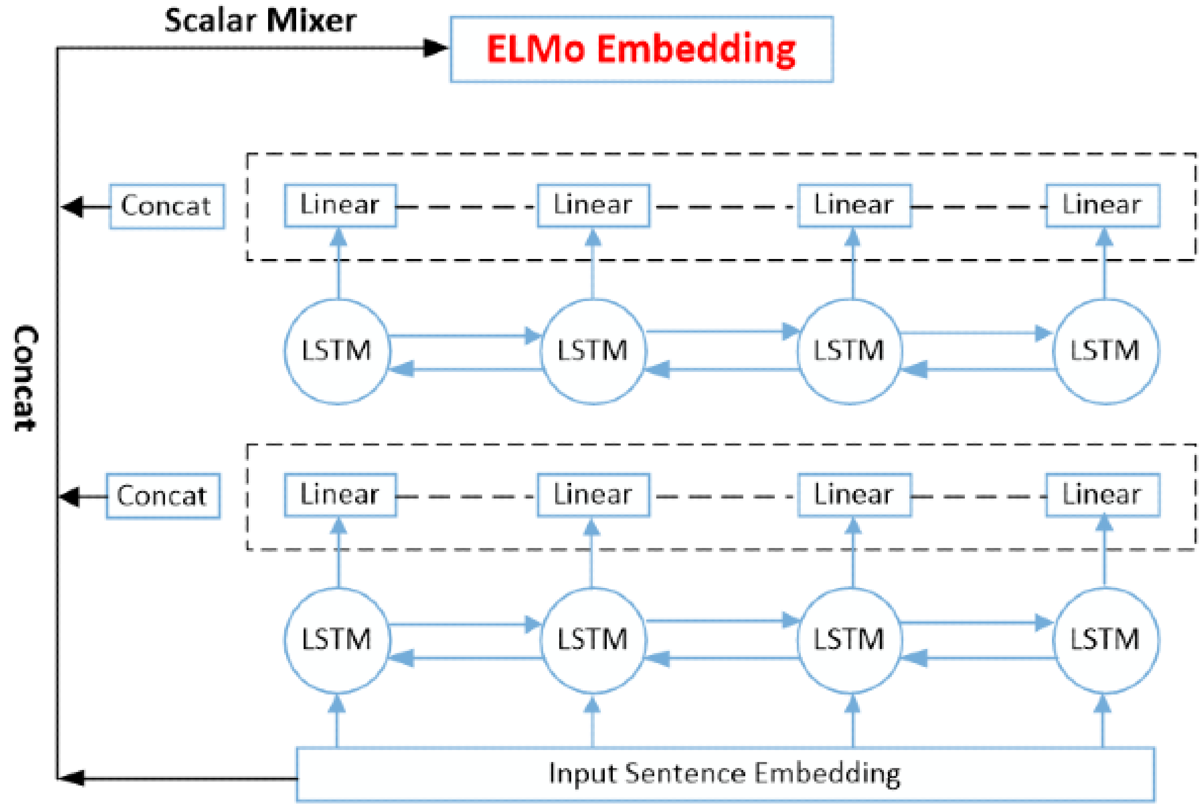
$$ELMo^{task}_{k} = E(R_k;\,\theta^{task}) = \Upsilon^{task}\Sigma^{L}_{j=0}\,s^{task}_{j}\,h^{LM}_{k,j} \quad \mathbf{(15)}$$

Here, s^{task} is the Softmax-normalized weight, and the scalar parameter Υ^{task} is used to scale the whole ELMo vector.

In the specific application, the model is pretrained on a large-scale unlabeled corpus. After the model is trained, a new sentence is input to get the contextualized ELMo embedding of each word in the current context. The original ELMo model was proposed for English text, and English words are divided into English character sequences as input, resulting in ELMo embedding of English words. Che et al [37] applied ELMo to multiple languages, including Chinese. They used the Chinese word-segmentation tools to segment text into words, and then used the ELMo model to obtain the contextualized word embedding.

In the method we proposed, in addition to the standard input of the lattice LSTM model, we integrated the domain-specific, pretrained ELMo embedding of Chinese characters as one of the input features. For obtaining the ELMo embedding of Chinese single characters, we used space to cut the corpus into single-character forms. Then, we modified the ELMo model; the architecture of the variant ELMo model is shown in Figure 2. We removed the convoluted character-encoding layer, and the embedding of Chinese characters was used as the input for training, with the dimension of character embedding set to 100. The input-sentence embedding was sent to the two-layer BiLSTM layer and two-layer representations were obtained. In the original work, the hidden size of the LSTM unit was set to be larger, and the dimension needed to be mapped to 512 through the linear layer, so that the output vector dimension of each character by each BiLSTM layer would be 1024. In our approach, we also modified the linear layer and mapped the hidden size of the LSTM cell to 50 through the linear layer; the output vector dimension of each token by each BiLSTM layer become 100. We then concatenated the input-sentence embedding and two-layer representations of the two-layer BiLSTM; the resulting vector was sent to the scalar mixer layer. Finally, pretrained ELMo embedding of Chinese characters was obtained by equation 15. At the pretrained stage of the ELMo model, we used the same unlabeled clinical corpus as done with the training-character embedding. In the application, a clinical sentence was sent into the pretrained ELMo model, so the ELMo embedding was obtained.

Figure 2. Architecture of the variant Embeddings from Language Models (ELMo) model. concat: concatenate; LSTM: long short-term memory.



CRF

A CRF layer is used on hidden vectors (h^c_1, h^c_2,..., h^c_t). The CRF decodes h^c_t into k-dimensional vectors, which denote label prediction probabilities. The score of the prediction sequence $y = (y_1, y_2, y_3,..., y_n)$ is computed by the following formula:

$$S(X,y) = \Sigma^n_{i=1} p_{i,j} + \Sigma^{n+1}_{i=1} A_{y(i-1),y(i)} \quad \textbf{(16)}$$

where $p_{i,j}$ denotes the probability of label j for word i, A represents the tagging transition matrix, and $A_{i,j}$ represents the score of the transition from label i to j.

Finally, the conditional probability $P(y|X)$ is calculated as follows:

$$P(y|X) = exp(score(X,y)) / \Sigma_{y'} exp(score(X,y')) \quad \textbf{(17)}$$

where $X = (x_1, x_2, x_3,..., x_n)$, which represents the character sequence input.

Model Implementation

In order to evaluate the performance of our approach, we implemented a series of basic models for comparison, as listed below:

1. Char-BiLSTM-CRF. This is a character (char)-based baseline model [29] without word segmentation; domain-specific character embedding was used as input. The pretrained character embedding was trained using the self-constructed clinical corpus mentioned in the Lattice LSTM section, and its dimension is 100.
2. BERT-BiLSTM-CRF. We used the pretrained RoBERTa_middle embedding model [38,39]—an improved version of BERT—as the input into the BiLSTM layer instead of the character embedding.
3. Word-BiLSTM-CRF. This is a word-based baseline model with reference to Wu et al [40]. We used the jieba segmentor [41], which includes the lexicon D, to segment the corpus. The Chinese word embedding in the medical field was trained by the word2vec tool, and the dimension was set to 100.
4. Word-BiLSTM-CRF (char CNN). On the basis of the word-based baseline model, the character-level embedding of words or subsequences was introduced [8]. The Chinese character in a word or subsequence is the smallest semantic unit, which carries certain information. The dimension of character-level embedding was set to 50, and the embedding lookup table was randomly initialized. The final state of character-level embedding was obtained by a CNN model; it was then concatenated with the word embedding to obtain the distributed representation of the word subsequence.
5. Word-BiLSTM-CRF (char LSTM). Similar to the above structure, the difference is that the LSTM model was used to encode character-level embedding [42].
6. ELMo-lattice-LSTM-CRF. This structure was our proposed method. The pretrained word2vec character embedding was combined with the medical field, pretrained, ELMo character embedding as the character part input of the model. The word subsequence was obtained by matching sentences in lexicon D, and its embedding was the same as that of the word-based baseline model.

Parameter Settings

In this study, we cut sentences into character sequences and limited the length to no more than 200. The BIO (beginning, inside, outside) schema was taken to annotate the entity. As mentioned earlier, the pretrained character embedding, word embedding from lexicon *D*, and ELMo embedding were all 100-dimensional vectors. The number of layers of LSTM was 1 and the hidden size was 200. We set the epoch to 10, the batch size to 1, and the dropout rate was 0.5. We adopted categorical cross-entropy to compute the loss function. A stochastic gradient descent optimizer, with a learning of 0.015 and decay rate of 0.05, was used to update parameters. The detailed settings of hyperparameters are shown in Table 1; similar parameters were used in other baseline models. On two Chinese CNER datasets, we used the same parameters, embedding, and lexicon to evaluate our method. Finally, we used the deep learning framework pytorch [43] to implement our model.

Table 1. Hyperparameter settings of the proposed approach.

Parameter	Value
Character-embedding size	100
Embeddings from Language Models (ELMo) embedding size	100
Word-embedding size	100
Dropout rate	0.5
Long short-term memory (LSTM) hidden size	200
LSTM layer	1
Learning rate	0.015
Learning rate decay	0.05
Epoch	10
Batch size	1

Results

Dataset and Evaluation Metrics

We conducted experiments based on two datasets, both of which were processed to delete privacy in the annotation phase. The first dataset was the CCKS-2017 CNER dataset, which contains 1596 labeled EMRs with five categories of clinical entities, including diseases, symptoms, exams, treatments, and body parts. We divided the dataset into two parts: 1198 EMRs were taken as a training set and 398 EMRs were taken as test set. Sequences that are too long will lead to the deterioration of model performance, so punctuation was used to split EMRs into sentences [11]. Therefore, the training set contained 7906 sentences and the test set contained 2118 sentences. The detailed distribution of the count of different types of entities is shown in Table 2.

The second dataset was the CCKS-2019 CNER dataset, which contains 1000 labeled EMRs. We divided the dataset into 900 training EMRs (5872 sentences) and 100 test EMRs (612 sentences). There were six categories of clinical entities in the dataset: disease, image, laboratory, operation, drug, and anatomy. The detailed distribution of the count of different types of entities is shown in Table 3.

In this paper, we used standard evaluation metrics, such as precision, recall, and F1 scores, to evaluate model performance. Meanwhile, the evaluation metrics were strict, which requires that the true label and prediction label have exactly the same entity name, same boundary, and same entity type.

Table 2. The distribution of entities in the China Conference on Knowledge Graph and Semantic Computing (CCKS)-2017 clinical named entity recognition (CNER) dataset.

Dataset	Number of entities in each category					
	Sentence	Disease	Symptom	Exam	Treatment	Body part
Training set	7906	722	7831	9546	1048	10,719
Test set	2118	553	2311	3143	465	3021

Table 3. The distribution of entities in the China Conference on Knowledge Graph and Semantic Computing (CCKS)-2019 clinical named entity recognition (CNER) dataset.

Dataset	Number of entities in each category						
	Sentence	Disease	Image	Laboratory	Operation	Drug	Anatomy
Training set	5872	3755	940	1167	932	1586	7524
Test set	612	362	34	37	116	242	898

Experiments Results

In order to get convincing experimental results, we ran each model five times and calculated the average precision, recall, and F1 scores as the final results. Table 4 shows the results of various models with different architectures on the test set of two Chinese CNER datasets.

Table 4. Results of various models with different architectures on two datasets.

Model	CCKS[a]-2017 CNER[b] dataset			CCKS-2019 CNER dataset		
	Precision, %	Recall, %	F1 score, %	Precision, %	Recall, %	F1 score, %
Char[c]-BiLSTM[d]-CRF[e] (baseline)	88.86	86.78	87.81	81.67	80.01	80.83
BERT[f]-BiLSTM-CRF	87.42	86.37	86.89	79.58	80.67	80.12
Word-BiLSTM-CRF (baseline)	85.87	86.33	86.10	79.63	80.07	79.85
Word-BiLSTM-CRF (char CNN[g])	88.23	86.90	87.56	82.69	81.72	82.20
Word-BiLSTM-CRF (char LSTM[h])	89.86	87.34	88.58	83.58	82.21	82.89
ELMo[i]-lattice-LSTM-CRF	*90.20* [j]	*90.06*	*90.13*	*84.69*	*85.35*	*85.02*

[a] CCKS: China Conference on Knowledge Graph and Semantic Computing.

[b] CNER: clinical named entity recognition.

[c] char: character.

[d] BiLSTM: bidirectional long short-term memory.

[e] CRF: conditional random field.

[f] BERT: Bidirectional Encoder Representations from Transformers.

[g] CNN: convolutional neural network.

[h] LSTM: long short-term memory.

[i] ELMo: Embeddings from Language Models.

[j] The best experimental results are italicized.

We observed that the character-based baseline model was better than the BERT-BiLSTM-CRF model, which is also character based and used the state-of-the-art pretrained BERT embedding. The main reason for this result is that BERT embedding was trained on the general field corpus rather than on the domain-specific corpus, which reflects the complexity of Chinese clinical texts. The character-based baseline model was better than the word-based baseline model as a whole, which shows that the character-based method can make better use of medical text information in Chinese CNER tasks.

It can be seen from the table that the word-BiLSTM-CRF (char LSTM) model outperformed the character-based and word-based baseline models and obtained competitive F1 scores of 88.58% and 82.89% on two datasets, respectively. This shows that the introduction of character-level embedding in the word-based method can make relatively full use of character and word information and can effectively improve the performance of the model. In addition, we also observed that the LSTM model captured the character-level semantic information of words better than did the CNN model.

From the results, we observed that the ELMo-lattice-LSTM-CRF model we proposed, which integrates lattice LSTM structure and variant pretrained ELMo embedding, achieved excellent results compared with the other models on both Chinese CNER datasets. This was seen with the F1 scores that reached 90.13% on the CCKS-2017 CNER dataset and 85.02% on the CCKS-2019 CNER dataset. Compared with the word-BiLSTM-CRF (char LSTM) model, the F1 scores of our method on both datasets were significantly improved by 1.55% and 2.57%, respectively. Table 5 shows the results of our method compared with previous representative systems on these two datasets [42,44,45].

The system in the first line [42] also used both Chinese character embedding and word embedding as feature representations, and an external health domain lexicon was adopted, which achieved an F1 score of 87.95% on the CCKS-2017 CNER dataset. The system in the second line [44] was similar to that in this paper. It adopted a lattice LSTM structure and used an adversarial training approach to improve the performance of the model; it achieved a good result, with an F1 score of 89.64%. The results show that our method surpassed these two systems by 2.18% and 0.49%, respectively. For the CCKS-2019 CNER dataset, Li et al [45] achieved the top performance by adopting the method of transfer learning and ensemble; our method obtained a similar score. By comparing our method with the previous models, the effectiveness of our method is evident.

Table 5. Comparative results between our approach and previous systems on two datasets.

Model	CCKS[a]-2017 CNER[b] dataset			CCKS-2019 CNER dataset		
	Precision, %	Recall, %	F1 score, %	Precision, %	Recall, %	F1 score, %
Recurrent neural network (char[c]-word) [42]	—[d]	—	87.95	—	—	—
AT[e]-lattice-LSTM[f]-CRF[g] [44]	88.98	90.28	89.64	—	—	—
FS[h]-TL[i] (ensemble) [45]	—	—	—	—	—	*85.16* [j]
Our approach	*90.20*	90.06	*90.13*	84.69	85.35	85.02

[a] CCKS: China Conference on Knowledge Graph and Semantic Computing.

[b] CNER: clinical named entity recognition.

[c] char: character.

[d] Data not available.

[e] AT: adversarial training.

[f] LSTM: long short-term memory.

[g] CRF: conditional random field.

[h] FS: fully shared.

[i] TL: transfer learning.

[j] The best experimental results are italicized.

Discussion

Overview

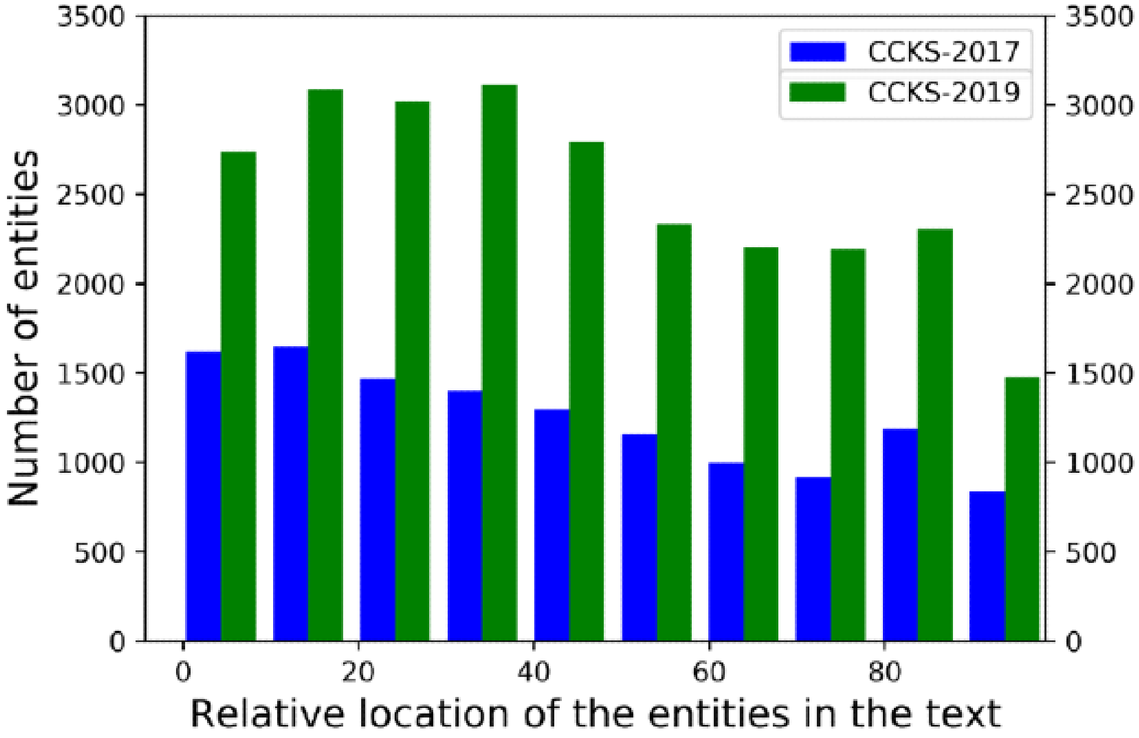
By comparing the experimental results, we notice that our method has excellent performance on the Chinese CNER task, which surpassed the character-based and word-based methods. In the future, we will conduct ablation experiments to further explore the influence of the lattice LSTM structure and ELMo embedding on the model performance.

Dataset Analysis

First, we analyzed the two Chinese CNER datasets. Figure 3 shows the distribution of the relative locations of clinical entities in the training set of the two datasets.

From the figure, we can intuitively observe that the distribution of entity locations in the two datasets is similar and relatively uniform; however, the distribution of entities from the CCKS-2019 CNER dataset is obviously more sparse than that of the CCKS-2017 CNER dataset. This indicates that the CCKS-2019 dataset labels were relatively unbalanced and there were more *outside* labels, which explains the reason why the results from the same models using CCKS-2017 CNER dataset were superior to those using the CCKS-2019 CNER dataset. Meanwhile, Tables 2 and 3 showed that there were very few image entities and laboratory entities in the test set—34 and 37, respectively—compared with the training set from the CCKS-2019 CNER dataset. This means that the distribution of labels in the test set and training set from the CCKS-2019 CNER dataset was quite different, which is another reason for the weaker performance by the model when using the CCKS-2019 CNER dataset.

Figure 3. Distribution of relative locations of entities in two Chinese clinical named entity recognition (CNER) datasets. CCKS: China Conference on Knowledge Graph and Semantic Computing.



Effectiveness of the Lattice LSTM Model

The comparison of the results of the standard lattice LSTM model and the character-based and word-based methods from using the two datasets is shown in Table 6. From the table, we observe that the performance of the standard lattice LSTM model surpassed that of the char-BiLSTM-CRF and word-BiLSTM-CRF (char LSTM) models. Compared with the better-performing word-BiLSTM-CRF (char LSTM) model, the performance of the model using the lattice LSTM on CCKS-2017 CNER dataset improved by 0.84%; the performance on the CCKS-2019 CNER dataset significantly improved by 1.29%. Although the word-BiLSTM-CRF (char LSTM) and lattice LSTM models used the same word embedding and lexicon, the word-BiLSTM-CRF (char LSTM) model first uses the lexicon for word segmentation, which imposes a hard restriction on the use of its subsequences, while the lattice LSTM model is free to consider lexicon words. This provides evidence that the lattice LSTM model can dynamically integrate potential word information, is superior to the character-based and word-based methods, and can achieve excellent performance in solving the Chinese CNER problem.

Table 6. Comparison of results between character-based or word-based methods and the lattice long short-term memory (LSTM) model on two datasets.

Model	CCKS[a]-2017 CNER[b] dataset			CCKS-2019 CNER dataset		
	Precision, %	Recall, %	F1 score, %	Precision, %	Recall, %	F1 score, %
Char[c]-BiLSTM[d]-CRF[e] (baseline)	88.86	86.78	87.81	81.67	80.01	80.83
Word-BiLSTM-CRF (char LSTM[f])	*89.86* [g]	87.34	88.58	83.58	82.21	82.89
Lattice-LSTM-CRF	89.66	*89.18*	*89.42*	*85.11*	*83.27*	*84.18*

[a] CCKS: China Conference on Knowledge Graph and Semantic Computing.

[b] CNER: clinical named entity recognition.

[c] char: character.

[d] BiLSTM: bidirectional long short-term memory.

[e] CRF: conditional random field.

[f] LSTM: long short-term memory.

[g] The best experimental results are italicized.

Effectiveness of ELMo Embedding

Table 7 shows the comparative results of different types of character embedding that were added to the lattice LSTM model using the two CNER datasets. The first line is the standard lattice LSTM model, and the second line is an embedding with equal dimensions and random initialization. It can be seen that there were slight improvements on both datasets, which may be due to the increase in parameters. In the third line, the character embedding trained by the GloVe tool [13] was added, and the F1 scores on the two datasets reached 89.70% and 84.62%, respectively, which shows that the addition of domain-specific character embedding is effective. The performance of the ELMo-lattice-LSTM-CRF (ML [many languages]) model, with pretrained ELMo representation for multiple languages [37,46], was slightly reduced compared to the standard lattice-LSTM-CRF model. This is likely because the pretrained ML model was trained on the general field corpus, so there was the problem of semantic inaccuracy.

Table 7. Comparison of different types of character embedding added to the lattice long short-term memory (LSTM) model using two clinical named entity recognition (CNER) datasets.

Model	CCKS[a]-2017 CNER[b] dataset			CCKS-2019 CNER dataset		
	Precision, %	Recall, %	F1 score, %	Precision, %	Recall, %	F1 score, %
Lattice-LSTM[c]-CRF[d]	89.66	89.18	89.42	85.11	83.27	84.18
Random-lattice-LSTM-CRF	88.79	*90.32* [e]	89.55	85.10	83.65	84.37
GloVe[f]-lattice-LSTM-CRF	89.63	89.77	89.70	*85.32*	83.90	84.62
ELMo[g]-lattice-LSTM-CRF (ML[h])	89.90	88.69	89.29	82.23	84.09	83.15
ELMo-lattice-LSTM-CRF	*90.20*	90.06	*90.13*	84.69	*85.35*	*85.02*

[a] CCKS: China Conference on Knowledge Graph and Semantic Computing.

[b] CNER: clinical named entity recognition.

[c] LSTM: long short-term memory.

[d] CRF: conditional random field.

[e] The best experimental results are italicized.

[f] GloVe: Global Vectors for Word Representation.

[g] ELMo: Embeddings from Language Models.

[h] ML: many languages.

The experimental results show that our proposed method was the best among all the methods, and it exceeded the standard lattice LSTM model by 0.71% and 0.84% on two datasets, respectively. These results demonstrate that the pretrained ELMo embedding trained on the medical corpus can further improve the performance of the model. After adding the pretrained ELMo embedding, the model used character information and weighted potential word information in sentences through the lattice LSTM structure; the model also obtained the domain-specific contextualized character representations, so as to obtain the rich semantic information of the EMRs, which is conducive to improving the performance of the model in the Chinese CNER task.

Error Analysis

We carried out error analysis on each entity category and on the reasons for misclassification. As shown in Table 8, we compared the results of our method with those of the char-BiLSTM-CRF model and the word-BiLSTM-CRF (char LSTM) model with respect to various entity categories: disease, image, laboratory, operation, drug, and anatomy. Since the distribution of results was similar, only the results of the CCKS-2019 CNER dataset are used for illustration.

Table 8. Comparison of the results regarding each entity category when using the China Conference on Knowledge Graph and Semantic Computing (CCKS)-2019 clinical named entity recognition (CNER) dataset.

Model	F1 scores for each entity category, %						
	Disease	Image	Laboratory	Operation	Drug	Anatomy	All
Char[a]-BiLSTM[b]-CRF[c]	80.23	77.75	74.41	*83.61* [d]	88.74	80.25	80.83
Word-BiLSTM-CRF (char LSTM[e])	81.45	80.56	77.41	81.54	91.86	*84.56*	82.89
ELMo[f]-lattice-LSTM-CRF	*83.66*	*85.23*	*78.28*	82.12	*97.05*	83.79	*85.02*

[a] char: character.

[b] BiLSTM: bidirectional long short-term memory.

[c] CRF: conditional random field.

[d] The best experimental results are italicized.

[e] LSTM: long short-term memory.

[f] ELMo: Embeddings from Language Models.

From the table, our method showed a significant improvement regarding image and drug entities, with F1 scores 4.67% and 5.19% higher than the previous best results; in particular, the F1 score for the drug entity reached 97.05%. Through analysis, we determined that the improvement of image entities was mainly due to the fact that image entities are mostly compound words in Chinese CNER, such as “心脏彩超” (color Doppler ultrasound of the heart), “腹部彩超” (color Doppler ultrasound of the abdomen), and “肝脏彩超” (color Doppler ultrasound of the liver). For instance, “心脏彩超” is often divided into two parts: the anatomy entity “心脏” (heart) and the image entity “彩超” (color Doppler ultrasound). In the drug entity, single characters in terms such as “奥沙利铂” (oxaliplatin) and “希罗达” (Xeloda) are almost meaningless or even interfere with semantic understanding. Lattice LSTM improves the accuracy by constructing a medical domain lexicon and dynamically integrating word information. However, we noticed that all the methods did not perform well regarding the laboratory entity. This may be because laboratory entities are more complex than other entity types, in which mixed representations occur more often, such as “ca74-2,” “间接coombs试验” (indirect Coombs test), and “g6pd活性试验” (glucose-6-phosphate dehydrogenase [G6PD] activity test); in addition, entities can be too short, such as “氯” (chlorine), “hb,” and “ph.” This is still a great challenge for the research of Chinese CNER; it is also the direction in which future research is heading.

Conclusions

By introducing the lattice LSTM model and a variant ELMo language model, this paper proposes a new Chinese CNER deep learning method. Our approach allows the model to coordinate the use of the character information and potential word information and takes advantage of contextualized character presentations, so as to make full use of EMR information. Finally, we used the CRF layer to capture the dependency between adjacent labels. We constructed a series of experiments on two Chinese CNER datasets to evaluate the performance of the model. The results showed that the ELMo-lattice-LSTM-CRF model that we proposed achieved excellent results, with F1 scores of 90.13% and 85.02% on the two datasets, respectively, which exceeded the performance of the standard lattice-LSTM-CRF model and achieved a competitive system. Overall, the results show that our approach for Chinese CNER is effective and can be used in future research. In future work, we will further generalize our model to improve its applicability and apply it to other small datasets through transfer learning methods.

Acknowledgments

This work was supported by the Zunyi Science and Technology Program (Nos. 遵市科合HZ字[2019]1号, 遵市科合HZ字[2019]22号, and 遵市科合HZ字[2020]43号), the Huichuan District Science and Technology Program (遵汇科合201711), and the Master's Fund of Zunyi Medical University (F-974).

Conflicts of Interest

None declared.

References

1. Wasserman RC. Electronic medical records (EMRs), epidemiology, and epistemology: Reflections on EMRs and future pediatric clinical research. Acad Pediatr 2011;11(4):280-287 [FREE Full text] [doi: 10.1016/j.acap.2011.02.007] [Medline: 21622040]
2. Zhang Y, Wang X, Hou Z, Li J. Clinical named entity recognition from Chinese electronic health records via machine learning methods. JMIR Med Inform 2018 Dec 17;6(4):e50 [FREE Full text] [doi: 10.2196/medinform.9965] [Medline: 30559093]

3. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009 Oct;42(5):760-772 [FREE Full text] [doi: 10.1016/j.jbi.2009.08.007] [Medline: 19683066]
4. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 2017 Jul 15;33(14):i37-i48 [FREE Full text] [doi: 10.1093/bioinformatics/btx228] [Medline: 28881963]
5. Jauregi Unanue I, Zare Borzeshi E, 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]
6. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016 Presented at: 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; San Diego, CA; June 12-17, 2016 p. 260-270 URL: https://www.aclweb.org/anthology/N16-1030.pdf [doi: 10.18653/v1/n16-1030]
7. Zeng D, Sun C, Lin L, Liu B. LSTM-CRF for drug-named entity recognition. Entropy (Basel) 2017 Jun 17;19(6):283 [FREE Full text] [doi: 10.3390/e19060283]
8. Ma X, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016 Presented at: 54th Annual Meeting of the Association for Computational Linguistics; August 7-12, 2016; Berlin, Germany p. 1064-1074 URL: https://www.aclweb.org/anthology/P16-1101.pdf [doi: 10.18653/v1/p16-1101]
9. Sahu SK, Anand A. Unified neural architecture for drug, disease, and clinical entity recognition. In: Agarwal B, Balas VE, Jain L, Poonia RC, Sharma M, editors. Deep Learning Techniques for Biomedical and Health Informatics. London, UK: Academic Press; 2020:1-19.
10. Wang Q, Zhou Y, Ruan T, Gao D, Xia Y, He P. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. J Biomed Inform 2019 Apr;92:103133 [FREE Full text] [doi: 10.1016/j.jbi.2019.103133] [Medline: 30818005]
11. Zhang Y, Yang J. Chinese NER using lattice LSTM. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers). 2018 Presented at: 56th Annual Meeting of the Association for Computational Linguistics; July 15-20, 2018; Melbourne, Australia p. 1554-1564 URL: https://www.aclweb.org/anthology/P18-1144.pdf [doi: 10.18653/v1/p18-1144]
12. Mikolov T, Sutskever I, Chen K. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 27th Conference on Neural Information Processing Systems 2013 (NIPS 2013). 2013 Presented at: 27th Conference on Neural Information Processing Systems 2013 (NIPS 2013); December 5-10, 2013; Lake Tahoe, NV p. 3111-3119 URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
13. Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: Association for Computational Linguistics (ACL); 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25-29, 2014; Doha, Qatar p. 1532-1543 URL: https://www.aclweb.org/anthology/D14-1162.pdf [doi: 10.3115/v1/d14-1162]
14. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K. Deep contextualized word representations. In: Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018 Presented at: 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 1-6, 2018; New Orleans, LA p. 2227-2237 URL: https://www.aclweb.org/anthology/N18-1202.pdf [doi: 10.18653/v1/n18-1202]
15. Fukuda K, Tamura A, Tsunoda T, Takagi T. Toward information extraction: Identifying protein names from biological papers. Pac Symp Biocomput 1998:707-718 [FREE Full text] [Medline: 9697224]
16. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: Extraction of Drugs, Genes and Relations from the biomedical literature. Pac Symp Biocomput 2000:517-528 [FREE Full text] [doi: 10.1142/9789814447331_0049] [Medline: 10902199]
17. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
18. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006 Jul 26;6:30 [FREE Full text] [doi: 10.1186/1472-6947-6-30] [Medline: 16872495]
19. Luo G, Huang X, Lin CY. Joint entity recognition and disambiguation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 Presented at: 2015 Conference on Empirical Methods in Natural Language Processing; September 17-21, 2015; Lisbon, Portugal p. 879-888 URL: https://www.aclweb.org/anthology/D15-1104.pdf [doi: 10.18653/v1/d15-1104]
20. Passos A, Kumar V, McCallum A. Lexicon infused phrase embeddings for named entity resolution. In: Proceedings of the 18th Conference on Computational Natural Language Learning. 2014 Presented at: 18th Conference on Computational

Natural Language Learning; June 26-27, 2014; Baltimore, MA p. 78-86 URL: https://www.aclweb.org/anthology/W14-1609.pdf [doi: 10.3115/v1/w14-1609]

21. Asahara M, Matsumoto Y. Japanese named entity extraction with redundant morphological analysis. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. 2003 Presented at: 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology; May 27-June 1, 2003; Edmonton, Canada p. 8-15 URL: https://dl.acm.org/doi/pdf/10.3115/1073445.1073447 [doi: 10.3115/1073445.1073447]
22. McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4. 2003 May 4 Presented at: 7th Conference on Natural Language Learning at HLT-NAACL 2003; May 31-June 1, 2003; Edmonton, Canada p. 188-191 URL: https://dl.acm.org/doi/pdf/10.3115/1119176.1119206 [doi: 10.3115/1119176.1119206]
23. Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. 2005 Presented at: 43rd Annual Meeting of the Association for Computational Linguistics; June 25-30, 2005; Ann Arbor, MI p. 363-370 URL: https://dl.acm.org/doi/pdf/10.3115/1219840.1219885 [doi: 10.3115/1219840.1219885]
24. Collobert R, Weston J, Bottou L. Natural language processing (almost) from scratch. J Mach Learn Res 2011;12:2493-2537 [FREE Full text]
25. Cocos A, Fiks AG, Masino AJ. Deep learning for pharmacovigilance: Recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. J Am Med Inform Assoc 2017 Jul 01;24(4):813-821. [doi: 10.1093/jamia/ocw180] [Medline: 28339747]
26. Announcement of data release and call for participation. Fourth i2b2/VA shared-task and workshop: Challenges in natural language processing for clinical data. i2b2. 2010. URL: https://www.i2b2.org/NLP/Relations/ [accessed 2020-08-25]
27. Uzuner Ö, 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]
28. Feng Y, Ying-Ying C, Gen-Gui Z. Intelligent recognition of named entities in electronic medical records [article in Chinese]. Chin J Biomed Eng 2011;30(2):256-262 [FREE Full text]
29. Xia Y, Wang Q. Clinical named entity recognition: ECUST in the CCKS-2017 shared task 2. In: Proceedings of the Evaluation Task at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2017). 2017 Presented at: Evaluation Task at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2017); August 26-29, 2017; Chengdu, China p. 43-48 URL: http://ceur-ws.org/Vol-1976/paper08.pdf
30. Turian J, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010 Presented at: 48th Annual Meeting of the Association for Computational Linguistics; July 11-16, 2010; Uppsala, Sweden p. 384-394 URL: https://www.aclweb.org/anthology/P10-1040.pdf
31. Jiang Z, Li L, Huang D. Training word embeddings for deep learning in biomedical text mining tasks. In: Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2015 Presented at: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); November 9-12, 2015; Washington, DC p. 625-628 URL: https://ieeexplore.ieee.org/document/7359756 [doi: 10.1109/bibm.2015.7359756]
32. Melamud O, Goldberger J, Dagan I. context2vec: Learning generic context embedding with bidirectional LSTM. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. 2016 Presented at: 20th SIGNLL Conference on Computational Natural Language Learning; August 7-12, 2016; Berlin, Germany p. 51-61 URL: https://www.aclweb.org/anthology/K16-1006.pdf [doi: 10.18653/v1/k16-1006]
33. Peters ME, Ammar W, Bhagavatula C, Power R. Semi-supervised sequence tagging with bidirectional language models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017 Presented at: 55th Annual Meeting of the Association for Computational Linguistics; July 30-August 4, 2017; Vancouver, Canada p. 1756-1765 URL: https://www.aclweb.org/anthology/P17-1161.pdf [doi: 10.18653/v1/p17-1161]
34. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2-7, 2019; Minneapolis, MN p. 4171-4186 URL: https://www.aclweb.org/anthology/N19-1423.pdf [doi: 10.18653/v1/N19-1423]
35. club.xywy.com. URL: http://club.xywy.com/ [accessed 2020-08-25]
36. China National Knowledge Infrastructure (CNKI). URL: https://www.cnki.net/ [accessed 2020-08-25]
37. Che W, Liu Y, Wang Y, Zheng B, Liu T. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In: Proceedings of the Computational Natural Language Learning (CoNLL) 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. 2018 Presented at: Computational Natural Language Learning (CoNLL) 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies; October 31-November 1, 2018; Brussels, Belgium p. 55-64 URL: https://www.aclweb.org/anthology/K18-2005.pdf [doi: 10.18653/v1/K18-2005]

38. roberta_zh. GitHub. URL: https://github.com/brightmart/roberta_zh [accessed 2020-08-25]
39. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly optimized BERT pretraining approach. arXiv. 2019 Jul 26. URL: https://arxiv.org/pdf/1907.11692.pdf [accessed 2020-08-30]
40. Wu J, Hu X, Zhao R, Ren F, Hu M. Clinical named entity recognition via bi-directional LSTM-CRF model. In: Proceedings of the Evaluation Task at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2017). 2017 Presented at: Evaluation Task at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2017); August 26-29, 2017; Chengdu, China p. 31-36 URL: http://ceur-ws.org/Vol-1976/paper06.pdf
41. jieba. GitHub. URL: https://github.com/fxsjy/jieba [accessed 2020-08-25]
42. Li Z, Zhang Q, Liu Y, Feng D, Huang Z. Recurrent neural networks with specialized word embedding for Chinese clinical named entity recognition. In: Proceedings of the Evaluation Task at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2017). 2017 Presented at: Evaluation Task at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2017); August 26-29, 2017; Chengdu, China p. 55-60 URL: http://ceur-ws.org/Vol-1976/paper10.pdf
43. pytorch. GitHub. URL: https://github.com/pytorch [accessed 2020-08-25]
44. Zhao S, Cai Z, Chen H, Wang Y, Liu F, Liu A. Adversarial training based lattice LSTM for Chinese clinical named entity recognition. J Biomed Inform 2019 Nov;99:103290. [doi: 10.1016/j.jbi.2019.103290] [Medline: 31557528]
45. Li N, Luo L, Ding Z, Song Y, Yang Z, Lin H. DUTIR at the CCKS-2019 Task 1: Improving Chinese clinical named entity recognition using stroke ELMo and transfer learning. In: Proceedings of the 4th China Conference on Knowledge Graph and Semantic Computing (CCKS 2019). 2019 Presented at: 4th China Conference on Knowledge Graph and Semantic Computing (CCKS 2019); August 24-27, 2019; Hangzhou, China URL: https://conference.bj.bcebos.com/ccks2019/eval/webpage/pdfs/eval_paper_1_1_3.pdf
46. ELMoForManyLangs. GitHub. URL: https://github.com/HIT-SCIR/ELMoForManyLangs [accessed 2020-08-25]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BiLM: bidirectional language model
BiLSTM: bidirectional long short-term memory
BIO: beginning, inside, outside
CCKS: China Conference on Knowledge Graph and Semantic Computing
char: character
CNER: clinical named entity recognition
CNN: convolutional neural network
CRF: conditional random field
ELMo: Embeddings from Language Models
EMR: electronic medical record
G6PD: glucose-6-phosphate dehydrogenase
GloVe: Global Vectors for Word Representation
i2b2: Informatics for Integrating Biology and the Bedside
LSTM: long short-term memory
ML: many languages
NER: named entity recognition
NLP: natural language processing
SVM: support vector machine

Edited by C Lovis; submitted 04.05.20; peer-reviewed by P Ding, M Wang; comments to author 28.05.20; revised version received 22.07.20; accepted 03.08.20; published 04.09.20.

Please cite as:
Li Y, Wang X, Hui L, Zou L, Li H, Xu L, Liu W
Chinese Clinical Named Entity Recognition in Electronic Medical Records: Development of a Lattice Long Short-Term Memory Model With Contextualized Character Representations
JMIR Med Inform 2020;8(9):e19848
URL: http://medinform.jmir.org/2020/9/e19848/
doi:10.2196/19848
PMID:32885786

Original Paper

An Intelligent Mobile-Enabled System for Diagnosing Parkinson Disease: Development and Validation of a Speech Impairment Detection System

Liang Zhang[1*], PhD; Yue Qu[2*], MS; Bo Jin[2], PhD; Lu Jing[3], MS; Zhan Gao[4], PhD; Zhanhua Liang[3], PhD

[1]International Business College, Dongbei University of Finance and Economics, Dalian, China

[2]School of Innovation and Entrepreneurship, Dalian University of Technology, Dalian, China

[3]Department of Neurology, The First Affiliated Hospital of Dalian Medical University, Dalian, China

[4]Beijing Haoyisheng Cloud Hospital Management Technology Ltd, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Lu Jing, MS
Department of Neurology
The First Affiliated Hospital of Dalian Medical University
No.222 Zhongshan Road
Dalian, 116011
China
Phone: 86 18098876262
Email: jinglu131129@126.com

Abstract

Background: Parkinson disease (PD) is one of the most common neurological diseases. At present, because the exact cause is still unclear, accurate diagnosis and progression monitoring remain challenging. In recent years, exploring the relationship between PD and speech impairment has attracted widespread attention in the academic world. Most of the studies successfully validated the effectiveness of some vocal features. Moreover, the noninvasive nature of speech signal–based testing has pioneered a new way for telediagnosis and telemonitoring. In particular, there is an increasing demand for artificial intelligence–powered tools in the digital health era.

Objective: This study aimed to build a real-time speech signal analysis tool for PD diagnosis and severity assessment. Further, the underlying system should be flexible enough to integrate any machine learning or deep learning algorithm.

Methods: At its core, the system we built consists of two parts: (1) speech signal processing: both traditional and novel speech signal processing technologies have been employed for feature engineering, which can automatically extract a few linear and nonlinear dysphonia features, and (2) application of machine learning algorithms: some classical regression and classification algorithms from the machine learning field have been tested; we then chose the most efficient algorithms and relevant features.

Results: Experimental results showed that our system had an outstanding ability to both diagnose and assess severity of PD. By using both linear and nonlinear dysphonia features, the accuracy reached 88.74% and recall reached 97.03% in the diagnosis task. Meanwhile, mean absolute error was 3.7699 in the assessment task. The system has already been deployed within a mobile app called No Pa.

Conclusions: This study performed diagnosis and severity assessment of PD from the perspective of speech order detection. The efficiency and effectiveness of the algorithms indirectly validated the practicality of the system. In particular, the system reflects the necessity of a publicly accessible PD diagnosis and assessment system that can perform telediagnosis and telemonitoring of PD. This system can also optimize doctors' decision-making processes regarding treatments.

(JMIR Med Inform 2020;8(9):e18689) doi:10.2196/18689

KEYWORDS

Parkinson disease; speech disorder; remote diagnosis; artificial intelligence; mobile phone app; mobile health

Introduction

Parkinson disease (PD) is a long-term degenerative disorder of the central nervous system that mainly affects the motor system. In the early stages, the symptoms include tremor; rigidity; slowness of movement; and difficulty with walking, talking, thinking, or completing other simple tasks. Dementia becomes common in the later stages of the disease. More than a third of patients have experienced depression and anxiety [1]. Other symptoms include sensory and sleep problems. In 2017, PD affected more than 10 million people worldwide, making it the second-most common neurological condition after Alzheimer disease. Currently, there is no cure for PD [2]. Accurate diagnosis, prognosis, and progression monitoring remain nontrivial.

As reported in previous work [3,4], approximately 90% of patients with PD develop voice and speech disorders during the course of the disease, which can have a negative impact on functional communication, thus leading to a decline in the quality of life [5]. Reduced volume (ie, hypophonia), reduced pitch range (ie, monotone), and difficulty with the articulation of sounds or syllables (ie, dysarthria) are the most common speech problems [6]. At the same time, many patients gradually dislike communication because of their own language barriers, which will cause more serious speech disorders and then form a vicious circle. Note that the speech signal–based test is noninvasive and can be self-administered. Hence, it has been regarded as a promising approach in PD diagnosis, evaluation, and progression monitoring, especially in the telediagnosis and telemonitoring medical fields.

In this work, we built a publicly accessible real-time system to efficiently diagnose and assess the severity of PD via speech signal analysis. The most relevant works can be found in Lahmiri et al [7] and Wroge et al [8]. They utilize similar machine learning algorithms as those based on previously proposed audio features [9-13]; however, their work neither considered severity assessment of PD nor made a publicly accessible app that allows for real-time mobile-aided PD diagnosis or evaluation, which is actually a trend and even a necessity in the current 4G and future 5G era for telediagnosis and telemonitoring. For instance, the outbreak of coronavirus disease 2019 (COVID-19) highlights the importance of intelligent and accurate telehealth during disease epidemics.

More specifically, our system first collects the speech signals of the subjects and then utilizes speech signal processing techniques to extract a variety of speech impairment features; it further utilizes advanced machine learning algorithms to diagnose PD and analyze the disease severity. In our work, in the speech signal feature-extraction stage, we utilized many traditional and novel methods to obtain clinically meaningful voice signal features, such as jitter, fine-tuning, recurrence period density entropy, pitch period entropy, signal-to-noise ratio, harmonics-to-noise ratio (HNR), and the mel frequency cepstral coefficient [9-11]. We regarded the PD diagnosis task as a classification problem and then utilized classical algorithms (eg, support vector machine [SVM] and artificial neural network [ANN]) to perform diagnosis. We formed the PD severity assessment task into a regression problem, with the Unified Parkinson Disease Rating Scale (UPDRS) score as the dependent variable; the UPDRS is the most widely employed scale for tracking PD symptom progression. Various regression algorithms (eg, support vector regression [SVR] and least absolute shrinkage and selection operator [LASSO] regression) were tested. We then obtained the most suitable model by comparing and blending different algorithms. In the end, we developed a mobile phone app for our system to realize remote diagnosis, severity evaluation, and progression monitoring of PD, which will significantly reduce detection and prevention costs.

The main structure of this paper is divided into four parts: (1) description of the methods used in our system: data collection, data preprocessing, feature extraction of speech signals, classification, and regression problem formulation, (2) analysis of our experimental results, (3) system description of our mobile app, and (4) final discussion.

Methods

Data Collection

The speech signal data used in the experiment came from two sources:

1. One part of the dataset came from the open data platform from the University of California Irvine (UCI) Machine Learning Repository, where three sets of parkinsonian speech data with different characteristics were obtained.
2. The other part of the dataset was collected in collaboration with the Department of Neurology, the First Affiliated Hospital of Dalian Medical University, China. The data recorded the voice signals of patients with PD.

In practice, the collected pronunciation content needs to be short and reflect the patient's speech disorder to a certain extent. On one hand, considering the need for different languages, dialects, and accents as well as unclear pronunciations, we adopted the continuous pronunciation method. Meanwhile, the control of the vocal cords and airflow is also weakened due to the weakening control of the pronunciation system of the nervous system. On the other hand, since the relationship between the vibration of the vocal cords and the speech disorder is relatively strong, the vowels can better reflect the degree of speech impairment [6,11,14]. Another fact is that the basic vowels in different regions of the world are very similar, so it is more reasonable to use vowels. The vowels used here are the five long vowels with the following English phonetic symbols: [ɑ:], [:], [i:], [:], and [u:]; the subjects are required to pronounce them repeatedly. The collected syllables are shown in Table 1.

Table 1. Collected syllables.

International phonetic symbol	Duration (seconds)
[ɑː]	3
[ɜ :]	3
[iː]	3
☒	3
[uː]	3

The UPDRS [15] is the most commonly used severity indicator in clinical studies of PD. It is evaluated via filling out a form, which requires considerable medical expertise, so it is difficult for patients to perform self-testing using this scale. That explains why we need automatic and artificial intelligence–powered prediction tools. We collected the UPDRS score as the dependent variable in our regression task. At present, UPDRS version 3.0 is the most widely used version, and it can be divided into four parts:

1. Mentation, behavior, and mood, including a total of four questions (16 points).
2. Activities of daily living, including a total of 13 questions (52 points).
3. Motor examination, including a total of 14 questions (108 points).
4. Treatment complications, including a total of 11 questions (23 points).

In summary, UPDRS version 3.0 has a total of 42 questions and the highest score is 199 points. The higher the UPDRS score, the more serious the PD is. The third item, *motor examination*, can reflect the severity of speech disorder. In practice, when collecting the data, the doctor is required to evaluate the total UPDRS score as well as the value of the motor examination score.

Note that the first part of the data is open source, and we can easily download this data from UCI's official website. Therefore, the datasets were mainly used to train the machine learning models and verify the validity of the dysphonia features, all of which have been integrated in our app system. This part of the data will be introduced in detail in the Results section. The second part of the data requires us to work closely with local hospitals—we collected PD patients' vocal data in a local hospital; the data collection table is shown in Multimedia Appendix 1. The Data Preprocessing section that follows describes how we processed the second part of the data, which has been implemented as a function in our app. We then extracted the dysphonia features, which have also been integrated as a function in our app system. Moreover, we tested them with machine learning models, which have been trained based on the first part of data, and achieved good results in the PD diagnosis task. Until now, the number of collected Chinese speech signals is still not big enough to train an effective model. Therefore, our model within our app system was trained by the first part of the data: the first and third datasets were used in the diagnosis task and severity assessment task, respectively. However, this app is continuously collecting new data, including positive and negative samples. As the amount of data increases in the future, we will utilize advanced technology, such as transfer learning, to realize PD diagnosis and severity evaluation for people in various regions.

Data Preprocessing

The initially collected voice signals cannot be directly used; some preprocessing was required. This operation removed some of the interference factors and paved the way for subsequent feature extraction. The formats of different audio files were unified into the WAV file format, with 44,100 Hz sampling frequency and two channels. These audio files were then uploaded into the back-end server for storage.

The first step of data preprocessing is *sampling frequency conversion*, that is, resampling, which can uniformly record the speech frequency and reduce the amount of calculation by down-clocking. In our work, only one channel of the speech signal (ie, the left channel) is reserved, and then the sampling frequency is converted to 10 kHz.

The second step is *pre-emphasis*. Since the low-frequency part of speech signals tends to contain noise, we performed pre-emphasis to filter out the low frequencies and improve the resolution of the high-frequency part of speech signals. In our work, a first-order, finite impulse response, high-pass digital filter was used to achieve pre-emphasis [16]. The transfer function is defined in equation 1 of Figure 1. In equation 1, a is the pre-emphasis coefficient; generally, $0.9 < a < 1.0$. Let $x(n)$ denote the voice sample value at time n. After the pre-emphasis processing, the result is $y(n) = x(n) – ax(n–1)$, where a=0.9375.

Figure 1. Equations 1-10. FN: false negative; FP: false positive; MAE: mean absolute error; MSE: mean square error; RMSE: root mean square error; TN: true negative; TP: true positive.

$$H(z) = 1 - az^{-1} \quad (1)$$

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \quad (2)$$

$$R_{xx}(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k) \quad (3)$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (4)$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{F1} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (7)$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |T_i - P_i| \quad (8)$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} (T_i - P_i)^2 \quad (9)$$

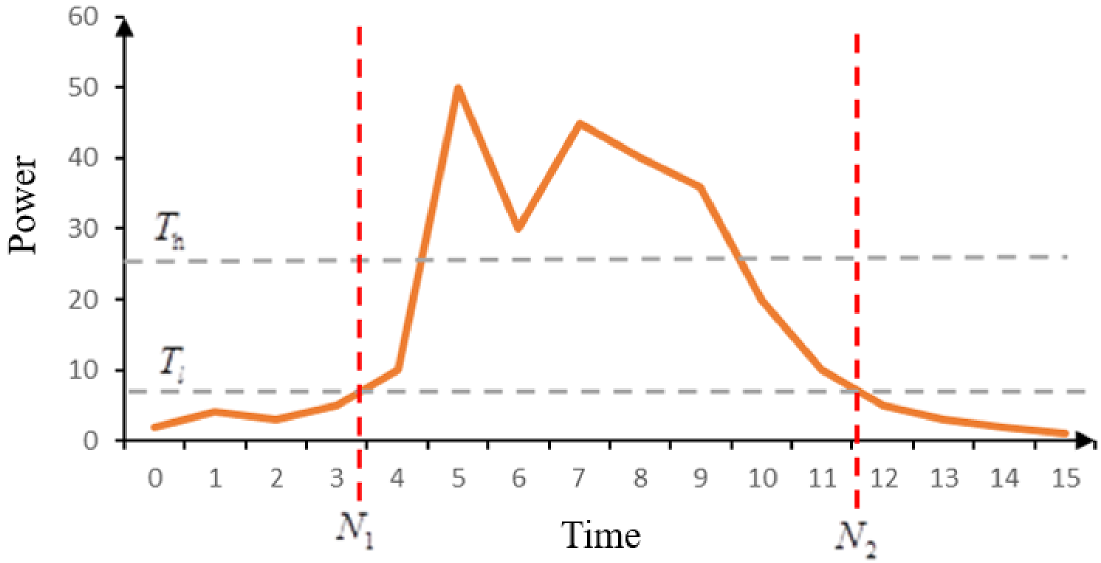
$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (T_i - P_i)^2} \quad (10)$$

The third step is *windowing and framing*. The speech signal was divided into some shorter signal segments (ie, frames) for processing, which is the framing process, such that the signal can be treated as stationary in the short-time window. In practice, to reduce the impact of segmenting on the statistical properties of the signal, we applied windowing to the temporal segments. The frame width in our work was set as 25 milliseconds long, the frame shift was 10 milliseconds long, and the Hamming window was leveraged as the window function.

The fourth step is *silent discrimination*. Because there is no guarantee that the collected audio files will always have sound, it is necessary to filter out the blank periods of those sounds. Therefore, silent discrimination, also known as voice endpoint detection, was required. A common solution is to use double-threshold methods [17], which are based on the principles of short-time energy, short-term average amplitude, and short-time zero-crossing rate. In our work, for the sake of simplicity and algorithm efficiency, we utilized only short-term energy as the principle for the double-threshold method. The definition of short-term average energy is shown Figure 1, equation 2.

For illustration, as is shown in Figure 2, we let T_h and T_l denote the upper and lower thresholds, respectively. The voiced part must have a section above T_h. The endpoint energy of the voiced part is equal to T_l. N_1 is the starting point, N_2 is the ending point, and *w* is the Hamming window. The fifth step is *fundamental frequency extraction*. The fundamental frequency refers to the lowest and theoretically strongest frequency in the sound, which reflects the vibration frequency of the sound source. In our work, we adopted the most widely used autocorrelation method to extract the fundamental frequency.

Figure 2. Principle of the double-threshold method. N_1: starting point; N_2: ending point; T_h: upper threshold; T_l: lower threshold.



The short-term autocorrelation function is defined in Figure 1, equation 3. We need to obtain the first positive peak point, $R_{xx}(k_f)$, after crossing the zero point in sequence $R_{xx}(k)$, and $1/k_f$ is the extracted fundamental frequency.

Note that the audio files may be mixed with unknown noise, which can cause a sudden jump at some points. These points are called wild points or outliers. Therefore, it is necessary to initially remove the wild points. We first calculated the average value of the fundamental frequency of the audio and then deleted the point that was too far from the average value.

Dysphonia Features

In 2012, Tsanas et al summarized 132 features of speech impairments [11]. Considering the speed requirement of the real-time system, the selected model cannot use all of the features. The final selected features [18-23] are illustrated in Table 2.

Table 2. Dysphonia features.

Classification and dysphonia features	Description
Pitch [18] (fundamental frequency)	
F_0_mean	Mean of pitch
F_0_max	Max of pitch
F_0_min	Min of pitch
F_0_median	Median of pitch
F_0_std	SD of pitch
Jitter [18] (pitch period perturbation)	
Jitter	Jitter
Jitter_abs	Absolute jitter
Jitter_PPQ5	5 adjacent points' jitter
Jitter_rap	3 adjacent points' jitter
Jitter_ddp	Difference of 3 adjacent points' jitter
Shimmer [18] (amplitude perturbation)	
Shimmer	Shimmer: percentage
Shimmer_dB	Shimmer: decibels (dB)
Shimmer_APQ5	5 adjacent points' shimmer
Shimmer_APQ3	3 adjacent points' shimmer
Shimmer_dda	Difference of 3 adjacent points' shimmer
Shimmer_APQ11	11 adjacent points' shimmer
Harmonics-to-noise ratio (HNR) and noise-to-harmonics ratio (NHR) [19]	
HNR_mean	Mean of HNR
HNR_std	SD of HNR
NHR_mean	Mean of NHR
NHR_std	SD of NHR
Nonlinear feature	
DFA	Detrended fluctuation analysis [20]
RPDE	Recurrence period density entropy [21]
D2	Correlation dimension [22]
PPE	Pitch period entropy [23]

Problem Formulation

Diagnosis

Because the predicted value in PD diagnosis is discrete and binary, it can be regarded as a two-category classification problem. This paper chose the following classical classification algorithms: (1) SVM, (2) ANN, (3) Naive Bayes, and (4) logistic regression.

Severity Assessment

Because the predicted value (ie, the UPDRS score) is continuous in the assessment of the severity of speech impairment in PD, it can be seen as a regression problem. This paper chose the following classical regression algorithms: (1) SVR, (2) linear regression, and (3) LASSO regression.

Results

Overview

We should initially introduce some indicators to evaluate the quality of the algorithms. First, for a two-category classification problem, there are usually the following classification results, as seen in Table 3.

Table 3. Classification confusion matrix.

Class	Predictive class	Predictive negative class
Actual positive class	True positive (TP)	False negative (FN)
Actual negative class	False positive (FP)	True negative (TN)

Then, the indicators are generally employed to evaluate the classification effect (see Figure 1, equations 4-7). The accuracy represents the proportion of subjects who are classified correctly out of the total number of subjects; precision indicates the proportion of real patients who are predicted to be sick; recall indicates the proportion of patients who are predicted to be sick; and the F1 value is the harmonic mean of the accuracy rate and the recall rate. In our PD diagnosis task, if a normal user is detected to be sick, the impact is usually not large, since we can continue to check the result using various clinical methods. However, if a model fails to detect PD, the impact is relatively large. Hence, the most important indicator is the recall rate.

Second, for a regression problem, if the total number of samples is *N*, the true value of the *i*-th sample is T_i, and the predicted value is P_i, then the indicators in equations 8-10 (see Figure 1) are available. Among the indicators, mean absolute error (MAE) measures the average magnitude of the errors in a set of predictions, without considering their direction; mean square error (MSE) and root mean square error (RMSE) are quadratic scoring rules that also measure the average magnitude of the error. However, both MSE and RMSE give a relatively higher weight to large errors. As a result, they are more useful when large errors are particularly undesirable.

According to the characteristics of the dataset, different experiments were performed on the three kinds of datasets downloaded from UCI. The characteristics of these datasets are shown in Table 4.

Table 4. Characteristics of three datasets from the University of California Irvine.

Data characteristics	Dataset 1	Dataset 2	Dataset 3
Creation date (year/month/day)	2008/06/26	2014/06/12	2009/10/29
Number of subjects			
Parkinson disease	23	48	42
Non-Parkinson disease	8	20	0
Number of records (ie, samples)	195	1208	5875
Number of features	22	26	18
Task	Classification	Classification and regression	Regression

All results are based on experiments with 5-fold cross validation. To evaluate our models' efficiency and effectiveness, for the PD diagnosis (ie, classification task), the ratio of the training set to the validation set was 4:1 in the first two datasets. We then used a dataset collected from a local hospital as the test dataset. The data collection table is shown in Multimedia Appendix 1. We collected a dataset that included 14 PD patients and 30 non-PD patients in total. For the PD severity evaluation (ie, regression task), the ratio of training set to the validation set to the testing set was 4:1:1 in the third dataset. The testing results are shown in the following paragraphs.

For the first set of data [9], we conducted classification experiments according to a combination of linear and nonlinear features; the final result is shown in Table 5.

Table 5. Classification results for the first set of data.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Support vector machine	*88.74* [a]	88.89	*97.03*	92.55
Logistic regression	85.71	89.97	91.32	90.18
Neural network (single layer)	88.68	91.16	94.26	92.45
Neural network (double layer)	88.63	92.55	93.38	*92.71*
Naive Bayes	69.24	*96.02*	62.37	75.21

[a] Italics represent the highest values.

We can see that the combination of linear and nonlinear features for the diagnosis of PD patients is feasible and effective. The SVM algorithm achieved higher accuracy and recall rate, and the Naive Bayes algorithm had the worst effect. According to the previous discussion, the recall rate is the most important indicator. At the same time, considering the speed requirement of the mobile app, our system finally leveraged the SVM algorithm to perform the PD patient diagnosis. From Multimedia Appendix 2, we can see that these features have small *P* values, especially for the nonlinear features, which statistically show the effectiveness of these features.

For the second set of data [24], we conducted classification experiments using only linear features, and the final result is demonstrated in Table 6.

Table 6. Classification results for the second set of data.

Algorithm	Accuracy, %	Precision, %	Recall, %	F1 score, %
Support vector machine	66.71	66.37	*83.71* [a]	73.98
Logistic regression	66.56	67.68	79.08	72.84
Neural network (single layer)	*70.78*	71.13	81.54	*75.89*
Neural network (double layer)	70.29	*71.45*	80.81	75.40
Naive Bayes	59.36	61.80	73.78	67.19

[a] Italics represent the highest values.

It can be clearly seen that using only linear features for PD diagnosis brings about a poor model performance, which is consistent with the conclusion from Tsanas et al [11] that feeding linear features into speech models is not very satisfactory. Meanwhile, some researchers claimed that nonlinear features are more effective [23], and another PD speech dataset analysis study [24] also obtained similar results. In particular, our experimental results showed that the SVM algorithm achieved a relatively high recall rate.

For the third set of data (ie, regression) [25], we tested multiple regression algorithms on the third dataset. The final result is illustrated in Table 7.

Table 7. Regression results on the third dataset.

Algorithm	Mean absolute error	Mean square error	Root mean square error
Linear regression	8.0786	95.1344	9.7494
Support vector machine	*3.7699* [a]	*34.1202*	*5.8357*
Least absolute shrinkage and selection operator	8.0687	91.1600	9.7452

[a] Italics represent the best values.

Experimental results showed that both linear and nonlinear features contribute to the severity assessment of PD patients. Among regression algorithms, the SVR algorithm achieved the best performance on each indicator, and the prediction results of LASSO and linear regressions were not much different; the reason for this is that LASSO regression is actually a variant of linear regression. Hence, the system finally adopted SVR as the severity evaluation method.

In particular, we selected the best results from each algorithm and observed the degree of fit. Figures 3-5 show the fitting results of the aforementioned three methods. In each figure, the upper graph is the degree of fitting of the training set and the lower graph is the degree of fitting of the test set; the red line is the predicted value and the blue line is the true value. It can be seen from these three figures that SVR fits the best.

As we know, LASSO can perform feature selection [26] by setting the feature weights to zero. The five characteristics most relevant to the value are shown in Table 8.

Figure 3. Linear regression fitting. The red line is the predicted value and the blue line is the true value. UPDRS: Unified Parkinson's Disease Rating Scale.

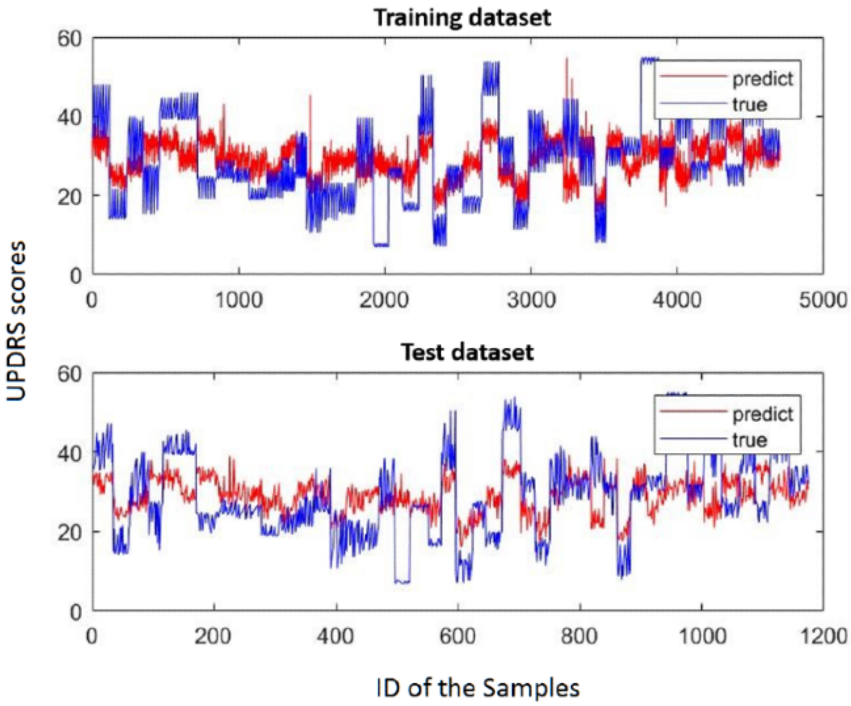


Figure 4. Support vector regression (SVR) fitting. The red line is the predicted value and the blue line is the true value. UPDRS: Unified Parkinson's Disease Rating Scale.

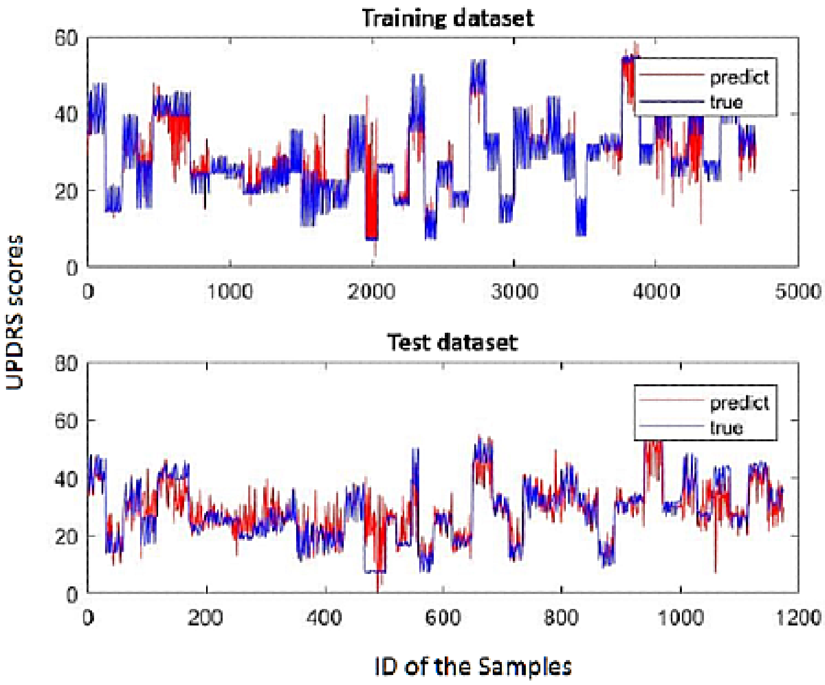


Figure 5. Least absolute shrinkage and selection operator (LASSO) fitting. The red line is the predicted value and the blue line is the true value. UPDRS: Unified Parkinson's Disease Rating Scale.

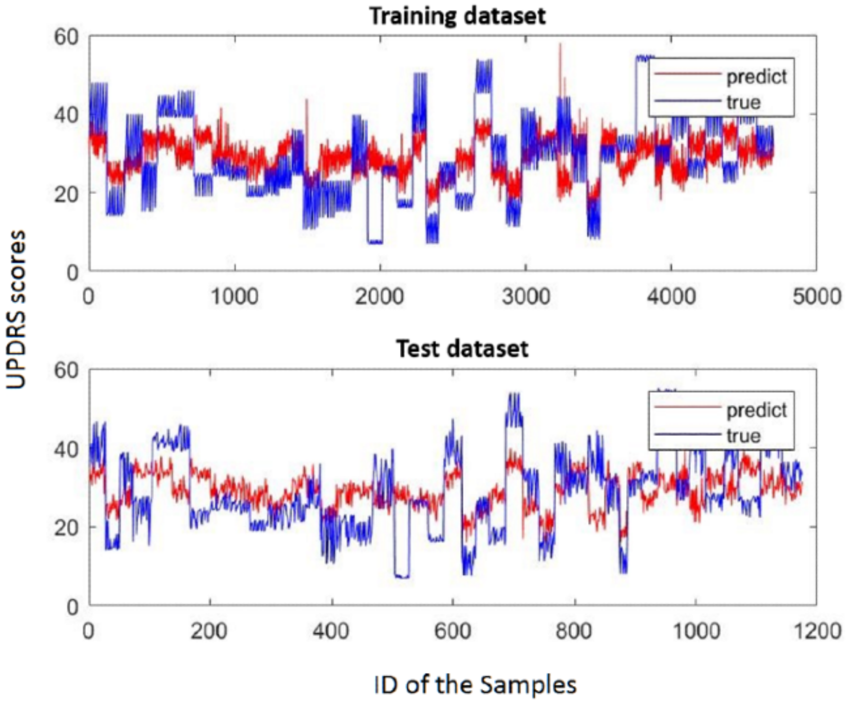


Table 8. Top five principal characteristics.

Feature	Corresponding weighted value
Age	2.84
Harmonics-to-noise ratio mean	–2.66
Absolute jitter	–2.18
Detrended fluctuation analysis	2.14
Pitch period entropy	1.51

It can be considered that these five characteristics are highly correlated with the UPDRS score. Age itself is highly related to PD, and the rest of the characteristics have three nonlinear features—HNR mean is also a nonlinear feature—indicating the importance of nonlinear features. Gender also explains why the regression result of the second set of data was relatively poor.

From Multimedia Appendix 2, we see that these features all have small *P* values—the features of the Jitter series may be a bit higher than others—which proves that we need these features for our system.

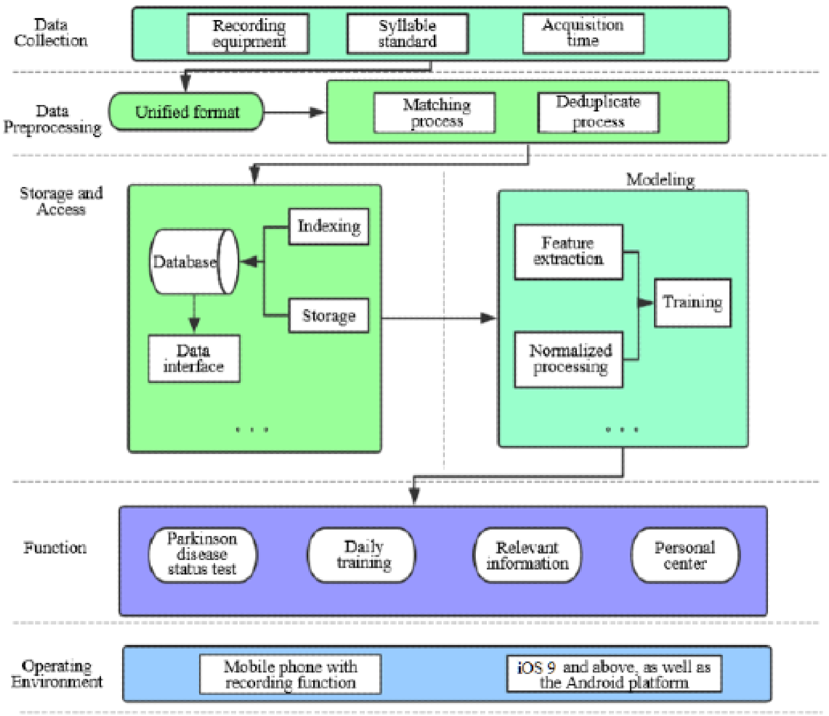
In summary, the PD speech detection system uses SVM and SVR for PD speech diagnosis and severity assessment, respectively.

System

System Overview

Figure 6 shows the architecture of our app system—called the No Pa app—including voice signal collection, data preprocessing, data storage and access, and signal modeling. At its core, the PD diagnosis model is SVM trained by the first set of data and the PD severity assessment model is SVR trained by the third set of data. Meanwhile, Figure 6 displays the four key functions and the operating environment in the application layer.

Figure 6. Architecture overview of the No Pa app system.

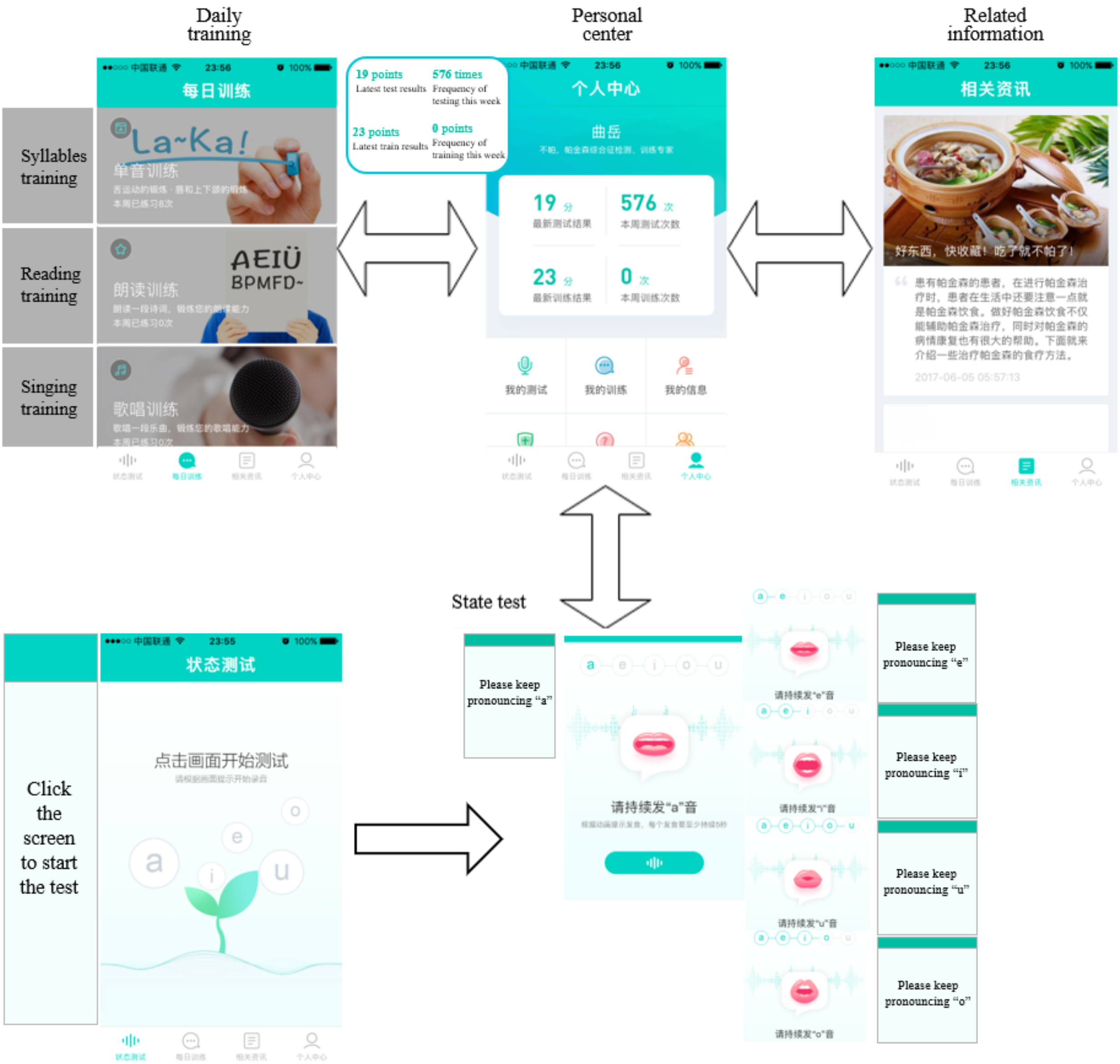


The Main Function

Android and iOS versions of the No Pa mobile app are currently available online. The app includes four functions—state test, daily training, related information, and personal center—which are shown as follows (see Figure 7 for a few screen captures):

1. State test: the subject pronounces five long vowels according to the voice guidance, and each long vowel sound lasts for 5 seconds. Then, our system will calculate the current speech impairment severity status.
2. Daily training: the daily training function aims to improve subjects' speech impairment status by encouraging them to speak. It includes monophonic training, reading training, and singing training. Monophonic training includes the user's pronunciation training according to some specific single syllables; during reading training, the user reads ancient poetry; and singing training improves the user's daily training interest via singing songs. Note that each training function will give a corresponding feedback score according to our speech signals model. However, since the calculation is not based on the five long vowels, the scores may not be accurate, but it is acceptable since our aim is to attract subjects' attention to daily training in speaking.
3. Related information: this function provides users with some advice about PD and physical health.
4. Personal center: this function helps the user view their testing history and some personal information.

Figure 7. Screen captures from the No Pa app showing four functional modules.


Daily training
Syllables training
Reading training
Singing training
每日训练
单音训练
朗读训练
歌唱训练
Personal center
个人中心
19 points Latest test results
576 times Frequency of testing this week
23 points Latest train results
0 points Frequency of training this week
我的测试
我的训练
我的信息
Related information
相关资讯
好东西，快收藏！吃了就不怕了！
State test
状态测试
点击画面开始测试
Click the screen to start the test
请持续发"a"音
Please keep pronouncing "a"
Please keep pronouncing "e"
Please keep pronouncing "i"
Please keep pronouncing "u"
Please keep pronouncing "o"
状态测试
每日训练
相关资讯
个人中心

Back-End Configuration

The back-end server of the No Pa app is the Alibaba Cloud Server. Its configuration is as follows: 4-core central processing unit (CPU), 8 GB RAM, 64-bit Ubuntu system, and 200 GB disk space.

Algorithm Acceleration

The original system's computational cost can range from 20 to 30 seconds without any acceleration techniques. Experimental results showed that autocorrelation calculation is the most time-consuming unit, so the C++ programming language was used to accelerate the autocorrelation calculation. To speed up the system, we adopted MEX (MATLAB executable) technology [27] as the acceleration scheme. In the end, the computational cost for predicting UPDRS scores was compressed from 20 seconds to only about 1 second. This response time is acceptable for an app.

Guide and Interaction

For better a user experience, we provided voice-guided navigation that can offer step-by-step instructions. Meanwhile, considering that PD patients may suffer from hand tremors, we designed big buttons in this app. Moreover, if they do not click the recording function button or the system fails to record an effective sound, the system will give them a reminder.

Discussion

Principal Findings

Traditionally, PD patients need to be diagnosed by physical examination. We can now use a mobile app to help conduct straightforward and rapid detection. For PD patients or healthy people, instant detection and consistent monitoring of disease conditions are extremely important. For doctors, the app can be used as a decision-support tool to provide assistance in treatment and diagnosis.

We have built this mobile app by embedding a voice-oriented system. At the core of the system are machine learning algorithms. Experimental results showed that SVM and SVR achieved the best performance for the diagnosis (ie, classification task) and severity evaluation (ie, regression task) of PD, respectively. The recall rate of the classification task can reach 97.03% (ie, the patient's recognition ability), and the absolute average error of the regression task can reach 3.7699, which is acceptable since the value of UPDRS scores range from 0 to 199.

Finally, we will summarize the contributions of our work. We have built a voice-oriented system that can remotely and conveniently diagnose PD. The system first collects a user’s five long vowels and then efficiently extracts dysphonia features, such that machine learning algorithms can be applied to the classification or regression of PD-related tasks. First, the system has been integrated into an app for public use. Second, our experiments have validated the effectiveness of voice signal–related features proposed by mainstream studies. Third, our system incorporates voice signal collection, feature extraction, and an algorithm interface, which can be regarded as a standard open-source platform for new algorithm development in voice signal–oriented disease identification tasks.

Comparison With Prior Work

There have been various studies utilizing vocal features for PD diagnosis or severity evaluation. More specifically, Lahmiri et al [7] proposed a study about diagnosing PD based on dysphonia measures. They chose the same dataset as our first dataset and their results are similar to ours. However, our method achieved a higher recall value on this dataset. Wroge et al [8] also focused on PD diagnosis by speech signal analysis. After some speech signal processing, they extracted two groups of features—Audio-Visual Emotion recognition Challenge (AVEC) [12] and Geneva Minimalistic Acoustic Parameter Set (GeMAPS) features [13]—which were then fed into some machine learning models. However, their feature extraction process relied on some existing tools, which are not easily integrated into an app. In particular, their work needs to extract 1262 features while our work only extracts 24 features. Moreover, the accuracy of their results based on SVM and ANN were both lower than ours. Similar work that is based on the above features can be found in Tracy et al [28]. Deep learning methods have also been leveraged to learn patterns from vocal feature sets [29]. However, their model lacks explanations due to the inherent nature of deep learning models and achieves an inferior performance compared with our model. Moreover, besides PD diagnosis, our system realizes PD severity evaluation, which may be more helpful for patients and doctors.

Limitations

Our data were collected from healthy people and patients with PD from Dalian, China; the quantity of data is still not big enough. In the future, we plan to collect more disease-related data from different regions worldwide to improve the generalization of the model. At the same time, we will use deep learning methods to study the speech signals of patients with PD to avoid cumbersome manual extraction of speech signals.

Acknowledgments

We are deeply indebted to people from the Department of Neurology, the First Affiliated Hospital of Dalian Medical University, China, who supported us with data collection and offered us a lot of professional advice. In addition, we owe our thanks to all volunteers, who allowed us to collect their voice data with patience, and all users of the No Pa app. This study was supported by the National Key R&D Program of China (2018YFC0116800), the National Natural Science Foundation of China (No. 61772110), the China Education and Research Network (CERNET) Innovation Project (NGII20170711), the Program of Introducing Talents of Discipline to Universities (Plan 111) (No. B20070), and the Liaoning Provincial Hospital Reform Focus on Diagnosis and Treatment in Clinical Departments: Parkinson's Disease and Related Movement Disorders in Individualized Diagnosis and Neural Control Platform Construction Project (LNCCC-C06-2015).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data collection table.
[XLSX File (Microsoft Excel File), 10 KB - medinform_v8i9e18689_app1.xlsx]

Multimedia Appendix 2
P values for linear and nonlinear features.
[XLSX File (Microsoft Excel File), 13 KB - medinform_v8i9e18689_app2.xlsx]

References

1. Kano O, Ikeda K, Cridebring D, Takazawa T, Yoshii Y, Iwasaki Y. Neurobiology of depression and anxiety in Parkinson's disease. Parkinsons Dis 2011;2011:143547 [FREE Full text] [doi: 10.4061/2011/143547] [Medline: 21687804]

2. Davie CA. A review of Parkinson's disease. Br Med Bull 2008;86:109-127 [FREE Full text] [doi: 10.1093/bmb/ldn013] [Medline: 18398010]
3. Ho AK, Iansek R, Marigliani C, Bradshaw JL, Gates S. Speech impairment in a large sample of patients with Parkinson's disease. Behav Neurol 1999;11(3):131-137 [FREE Full text] [doi: 10.1155/1999/327643]
4. Mahler LA, Ramig LO, Fox C. Evidence-based treatment of voice and speech disorders in Parkinson disease. Curr Opin Otolaryngol Head Neck Surg 2015;23(3):209-215 [FREE Full text] [doi: 10.1097/moo.0000000000000151]
5. Miller N, Noble E, Jones D, Burn D. Life with communication changes in Parkinson's disease. Age Ageing 2006 May;35(3):235-239. [doi: 10.1093/ageing/afj053] [Medline: 16540492]
6. Fox CM, Morrison CE, Ramig LO, Sapir S. Current perspectives on the Lee Silverman Voice Treatment (LSVT) for individuals with idiopathic Parkinson disease. Am J Speech Lang Pathol 2002 May;11(2):111-123 [FREE Full text] [doi: 10.1044/1058-0360(2002/012)]
7. Lahmiri S, Dawson DA, Shmuel A. Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures. Biomed Eng Lett 2018 Feb;8(1):29-39 [FREE Full text] [doi: 10.1007/s13534-017-0051-2] [Medline: 30603188]
8. Wroge TJ, Özkanca Y, Demiroglu C, Si D, Atkins DC, Ghomi RH. Parkinson's disease diagnosis using machine learning and voice. In: Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB). 2018 Presented at: IEEE Signal Processing in Medicine and Biology Symposium (SPMB); December 1, 2018; Philadelphia, PA p. 1-7. [doi: 10.1109/spmb.2018.8615607]
9. Little M, McSharry P, Hunter E, Spielman J, Ramig L. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. Nat Precedings 2008 Sep 12:1-27 [FREE Full text] [doi: 10.1038/npre.2008.2298.1]
10. Guo PF, Bhattacharya P, Kharma N. Advances in detecting Parkinson's disease. In: Proceedings of the 2nd International Conference on Medical Biometrics. 2010 Jun Presented at: 2nd International Conference on Medical Biometrics; June 28-30, 2010; Hong Kong, China p. 306-314 URL: https://doi.org/10.1007/978-3-642-13923-9_33 [doi: 10.1007/978-3-642-13923-9_33]
11. Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng 2012 May;59(5):1264-1271 [FREE Full text] [doi: 10.1109/tbme.2012.2183367]
12. Valstar M, Schuller B, Smith K, Eyben F, Jiang B, Bilakhia S, et al. AVEC 2013: The continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '13). 2013 Presented at: 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '13); October 21-25, 2013; Barcelona, Spain p. 3-10 URL: https://dl.acm.org/doi/10.1145/2512530.2512533 [doi: 10.1145/2512530.2512533]
13. Eyben F, Scherer KR, Schuller BW, Sundberg J, Andre E, Busso C, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. IEEE Trans Affect Comput 2016 Apr 1;7(2):190-202. [doi: 10.1109/taffc.2015.2457417]
14. Tsanas A. Accurate Telemonitoring of Parkinson's Disease Symptom Severity Using Nonlinear Speech Signal Processing and Statistical Machine Learning [doctoral thesis]. Oxford, UK: University of Oxford; 2012 Jun. URL: https://ora.ox.ac.uk/objects/uuid:2a43b92a-9cd5-4646-8f0f-81dbe2ba9d74/download_file?file_format=pdf&safe_filename=DPhil%2Bthesis_post_viva_v8.pdf&type_of_work=Thesis [accessed 2012-12-30] [WebCite Cache ID ora.ox.ac.uk/objects/uuid:2a43b92a-9cd5-4646-8f0f-81dbe2ba9d74]
15. Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The Unified Parkinson's Disease Rating Scale (UPDRS): Status and recommendations. Mov Disord 2003 Jul;18(7):738-750. [doi: 10.1002/mds.10473] [Medline: 12815652]
16. Cohen I, Benesty J, Gannot S, editors. Speech Processing in Modern Communication: Challenges and Perspectives. Berlin, Germany: Springer-Verlag; 2010.
17. Sakhnov K, Verteletskaya E, Simak B. Dynamical energy-based speech/silence detector for speech enhancement applications. In: Proceedings of the World Congress on Engineering (WCE 2009). Volume I. 2009 Jul Presented at: World Congress on Engineering (WCE 2009); July 1-3, 2009; London, UK p. 801 URL: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=6AACCA0D6A768D38D72BF9E501C027C8?doi=10.1.1.149.3572&rep=rep1&type=pdf
18. Boersma P, Weenink D. Praat: Doing phonetics by computer. Phonetic Sciences, University of Amsterdam. 2002. URL: http://webcache.googleusercontent.com/search?q=cache:EwTtYfw9KBAJ:fon.hum.uva.nl/praat/+&cd=1&hl=en&ct=clnk&gl=ca [accessed 2020-09-02]
19. Ferrer CA, González E, Hernández-Díaz ME. Evaluation of time and frequency domain-based methods for the estimation of harmonics-to-noise-ratios in voice signals. In: Proceedings of the Iberoamerican Congress on Pattern Recognition (CIARP 2006). 2006 Nov Presented at: Iberoamerican Congress on Pattern Recognition (CIARP 2006); November 14-17, 2006; Cancun, Mexico p. 406-415 URL: https://doi.org/10.1007/11892755_42 [doi: 10.1007/11892755_42]
20. Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL. Mosaic organization of DNA nucleotides. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 1994 Feb;49(2):1685-1689. [doi: 10.1103/physreve.49.1685] [Medline: 9961383]

21. Kim H, Eykholt R, Salas JD. Nonlinear dynamics, delay times, and embedding windows. Physica D 1999 Mar;127(1-2):48-60 [FREE Full text] [doi: 10.1016/s0167-2789(98)00240-1]
22. Janjarasjitt S, Scher MS, Loparo KA. Nonlinear dynamical analysis of the neonatal EEG time series: The relationship between sleep state and complexity. Clin Neurophysiol 2008 Aug;119(8):1812-1823. [doi: 10.1016/j.clinph.2008.03.024] [Medline: 18486543]
23. Little MA, McSharry PE, Roberts SJ, Costello DA, Moroz IM. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Biomed Eng Online 2007;6(1):23 [FREE Full text] [doi: 10.1186/1475-925x-6-23]
24. Sakar BE, Isenkul ME, Sakar CO, Sertbas A, Gurgen F, Delil S, et al. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. IEEE J Biomed Health Inform 2013 Jul;17(4):828-834 [FREE Full text] [doi: 10.1109/jbhi.2013.2245674]
25. Tsanas A, Little MA, McSharry PE, Ramig LO. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE Trans Biomed Eng 2010 Apr;57(4):884-893 [FREE Full text] [doi: 10.1109/tbme.2009.2036000]
26. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 2018 Dec 05;58(1):267-288 [FREE Full text] [doi: 10.1111/j.2517-6161.1996.tb02080.x]
27. Eshkabilov S. MEX files, C/C++, and standalone applications. In: Beginning MATLAB and Simulink: From Novice to Professional. New York, NY: Apress; 2019:259-273.
28. Tracy JM, Özkanca Y, Atkins DC, Hosseini Ghomi R. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. J Biomed Inform 2020 Apr;104:103362. [doi: 10.1016/j.jbi.2019.103362] [Medline: 31866434]
29. Gunduz H. Deep learning-based Parkinson’s disease classification using vocal feature sets. IEEE Access 2019;7:115540-115551 [FREE Full text] [doi: 10.1109/ACCESS.2019.2936564]

Abbreviations

ANN: artificial neural network
AVEC: Audio-Visual Emotion recognition Challenge
CERNET: China Education and Research Network
COVID-19: coronavirus disease 2019
CPU: central processing unit
GeMAPS: Geneva Minimalistic Acoustic Parameter Set
HNR: harmonics-to-noise ratio
LASSO: least absolute shrinkage and selection operator
MAE: mean absolute error
MEX: MATLAB executable
MSE: mean square error
PD: Parkinson disease
RMSE: root mean square error
SVM: support vector machine
SVR: support vector regression
UCI: University of California Irvine
UPDRS: Unified Parkinson Disease Rating Scale

Edited by G Eysenbach; submitted 12.03.20; peer-reviewed by J Lei, C Yu; comments to author 20.04.20; revised version received 06.07.20; accepted 26.07.20; published 16.09.20.

Please cite as:
Zhang L, Qu Y, Jin B, Jing L, Gao Z, Liang Z
An Intelligent Mobile-Enabled System for Diagnosing Parkinson Disease: Development and Validation of a Speech Impairment Detection System
JMIR Med Inform 2020;8(9):e18689
URL: http://medinform.jmir.org/2020/9/e18689/
doi:10.2196/18689
PMID:32936086

Original Paper

Chronological Age Assessment in Young Individuals Using Bone Age Assessment Staging and Nonradiological Aspects: Machine Learning Multifactorial Approach

Ana Luiza Dallora[1], PhD; Ola Kvist[2], MD; Johan Sanmartin Berglund[1], PhD, MD; Sandra Diaz Ruiz[2], PhD, MD; Martin Boldt[3], PhD; Carl-Erik Flodmark[4], PhD, MD; Peter Anderberg[1], PhD

[1]Department of Health, Blekinge Institute of Technology, Karlskrona, Sweden
[2]Department of Pediatric Radiology, Karolinska University Hospital, Stockholm, Sweden
[3]Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden
[4]Department of Clinical Sciences, Lund University, Lund, Sweden

Corresponding Author:
Peter Anderberg, PhD
Department of Health
Blekinge Institute of Technology
Valhallavägen 1
Karlskrona, 371 41
Sweden
Phone: 46 0734223736
Email: peter.anderberg@bth.se

Abstract

Background: Bone age assessment (BAA) is used in numerous pediatric clinical settings as well as in legal settings when entities need an estimate of chronological age (CA) when valid documents are lacking. The latter case presents itself as critical as the law is harsher for adults and granted rights along with imputability changes drastically if the individual is a minor. Traditional BAA methods have drawbacks such as exposure of minors to radiation, they do not consider factors that might affect the bone age, and they mostly focus on a single region. Given the critical scenarios in which BAA can affect the lives of young individuals, it is important to focus on the drawbacks of the traditional methods and investigate the potential of estimating CA through BAA.

Objective: This study aims to investigate CA estimation through BAA in young individuals aged 14-21 years with machine learning methods, addressing the drawbacks of research using magnetic resonance imaging (MRI), assessment of multiple regions of interest, and other factors that may affect the bone age.

Methods: MRI examinations of the radius, distal tibia, proximal tibia, distal femur, and calcaneus were performed on 465 men and 473 women (aged 14-21 years). Measures of weight and height were taken from the subjects, and a questionnaire was given for additional information (self-assessed Tanner Scale, physical activity level, parents' origin, and type of residence during upbringing). Two pediatric radiologists independently assessed the MRI images to evaluate their stage of bone development (blinded to age, gender, and each other). All the gathered information was used in training machine learning models for CA estimation and minor versus adult classification (threshold of 18 years). Different machine learning methods were investigated.

Results: The minor versus adult classification produced accuracies of 0.90 and 0.84 for male and female subjects, respectively, with high recalls for the classification of minors. The CA estimation for the 8 age groups (aged 14-21 years) achieved mean absolute errors of 0.95 years and 1.24 years for male and female subjects, respectively. However, for the latter, a lower error occurred only for the ages of 14 and 15 years.

Conclusions: This study investigates CA estimation through BAA using machine learning methods in 2 ways: minor versus adult classification and CA estimation in 8 age groups (aged 14-21 years), while addressing the drawbacks in the research on BAA. The first achieved good results; however, for the second case, the BAA was not precise enough for the classification.

(JMIR Med Inform 2020;8(9):e18846) doi:10.2196/18846

KEYWORDS

chronological age assessment; bone age; skeletal maturity; machine learning; magnetic resonance imaging; radius; distal tibia; proximal tibia; distal femur; calcaneus

Introduction

Background

Skeletal maturity is a radiological concept that refers to the stage of bone development in an individual [1]. This maturation process occurs gradually in the growth plates and is measured by the degree of mineralization of the bone along with its size and shape [1]. Bone age (BA) is a closely related concept in which age is estimated based on the degree of skeletal maturity of an individual [2].

The estimation of the BA of an individual, or bone age assessment (BAA), is performed in numerous clinical settings involving diagnosis and time of treatment of orthopedics, orthodontics, endocrinology, growth disorders, and estimations of final height [3]. In these cases, the BA of an individual is assessed by medical professionals and compared with their chronological age (CA). If they are found to be relatively advanced or retarded, appropriate actions are taken by the medical professionals.

BAA is also performed outside the clinical setting when legal entities need an estimation of the CA of an individual for judicial decisions when valid documents are lacking. This refers to cases regarding adoption, criminal proceedings, and pedopornography judicial issues as well as in determining age fraud in youth sports competitions [4-7]. Furthermore, with the upsurge of immigration due to the rise of worldwide conflicts, another critical scenario in which BAA is applied concerns the determination of an individual being minor in the absence of valid or trustworthy documents. This is the case of numerous young asylum seekers who are given special rights granted by the United Nations Convention on the Rights of the Child, regarding reception, health care, and education [8,9].

From these examples, it is possible to assume that, especially regarding legal standpoints, BAA is a crucial tool for making high stake decisions that have the potential to greatly affect individuals' lives.

Traditional BAA

The traditional methods for BAA are based on the appearance of growth plates through the analysis of diaphysis (primary ossification centers) and epiphysis (secondary ossification centers), where cartilage tissue gradually turns into bone tissue during the process of bone development. A process that ceases when the diaphysis and epiphysis are fused, indicating that the growth plate is ossified [1].

The most common procedures for BAA are the Greulich-Pyle (GP) and Tanner-Whitehouse (TW) methods. Both of these methods assess radiographic images of the hand and wrist areas as these are regions of interest (ROIs) with a large number of ossification centers aggregated in a small area that can easily have images taken from.

The GP method [10] attributes BA by comparing the radiograph image of the individual being assessed to the nearest reference image in a hand and wrist atlas in terms of bone development. The TW method [11] is a scoring system that evaluates the ulna, radius, carpals, and 13 short bones of the hand. Scores are attributed to these regions based on the stage of bone development, which ranges from A to I. The scores are then aggregated in a total score that is converted into the BA.

Having been developed in the 30s and 50s, the GP and TW methods, respectively, conveyed groundbreaking developments in numerous clinical settings and are still heavily employed for BAA purposes to this day.

Other Proposed BAA Methods

The field of BAA evolved as the GP and TW methods were proposed, exploring new ROIs with different ossification timings. This section summarizes the proposed studies regarding BAA in various ROIs.

Newer hand and wrist studies on BAA include the Gilsanz and Ratib [1] digital hand atlas and the Fels method [12]. The first is composed of artificially created reference images that represent the average development of 29 classes of subjects aged from 0 to 18 years. The Fels method [12] is a statistical method that provides a relative measure of the BA and standard error that takes into consideration the distribution of chronological ages in the study's sample with BA similar to the individual being assessed. It is based on 98 indicators of bone maturity (ossification, radiopaque densities, bony projection, shape changes, and ossification of epiphysis).

Clavicle staging systems observe one or both sides of the medial clavicular epiphysis. The method proposed by Kreitner et al [13] presents 4 stages of ossification of the medial clavicular epiphysis, in which the last stage may have an epiphyseal scar visible. Schmeling et al [9] proposed 5 stages of ossification, but the last stage was only achieved when the epiphyseal scar was not apparent. Kellinghaus et al [14] built on the Schmeling et al [9] staging by applying subclassifications for the second and third stages. These studies report complete ossification of this growth plate around the ages of 26 to 27 years.

Knee studies proposed staging systems that also vary on subscales on specific stages and the appearance of the epiphyseal scar in the last stage. O'Connor et al [15] proposed 5 stages of ossification of the distal femur, proximal tibia, and proximal fibula epiphysis (the epiphyseal scar may be visible in the last stage). Dedouit et al [16] proposed 5 stages of ossification of the distal femur and proximal tibia epiphysis, assessing the appearance of cartilage signal intensity with magnetic resonance imaging (MRI). Krammer et al [17] proposed 5 stages of ossification of the distal femur epiphysis, with subclassifications on the second and third stages, with the last stage achieved only when the epiphyseal scar is no longer visible. This method also makes use of MRI images. Knee studies usually argue that a subject is younger or older than the age of 18 years.

Studies on foot ROIs are usually concerned with younger ages. Ekizoglu et al [18] proposed a staging system for the foot ROI that shows complete ossification in the ages between 12 and 16 years.

Not very much is explored in the literature, the arm ROI was studied in the proximal humerus epiphysis by Ekizoglu et al [19] employing a scoring system based on Schmeling et al [9] and Kellinghaus et al [14] on MRI images. This study points out the earliest ages for the last stage of ossification at 17 and 18 years.

Drawbacks in Assessing Chronological Age Using BAA Methods

In the lack of valid or trustworthy documents, BAA is currently employed as a valuable tool for legal entities to evaluate CA with regard to important legal ages. Nevertheless, it is possible to identify several drawbacks of the largely employed GP and TW methods as well as recently proposed methods, regarding the use of BAA for CA determination:

- They almost exclusively employ medical imaging techniques that expose the individual to ionizing radiation, such as radiographs, which raises grave ethical issues especially with regard to exposing minors to radiation for nontherapeutic purposes.
- They only focus on the physical appearance of the growth plates, not including other information that might possibly affect bone development [20].
- They mostly focus on a single ROI, which in the vast majority of cases is the hand area [20].

The first drawback can be addressed by the employment of MRI technology, which is already present in some of the mentioned knee and arm studies. Besides being a radiation-free modality of medical imaging, it also allows the manipulation of contrast to highlight different tissue types [21]. The epiphyseal plate consists of cartilage tissue, which is mainly composed of collagen fiber protein. Collagen has a 3D structure of fibers that, in MRI images, is shown as zones of different intensities, giving it a multilaminar appearance. It is known that the structure of cartilage changes in terms of the number of laminae and thickness in the course of bone development [22]. Hence, contrary to radiographs that highlight the bone, the MRI technology might have the potential to offer better visualization of growth plates, thus being an interesting radiation-free modality of medical imaging for BAA.

To address the second drawback, the methods for assessing BA should investigate factors that may play a role in the process of bone development and ossification of growth plates, that is, BMI [20,23], pubertal growth [24], physical activity [25], ethnicity [8,20,26], and socioeconomic factors [8], which are often overlooked [20].

Addressing the third drawback could be done by employing multiple ROIs. When it comes to estimations of CA, most of the BAA studies, especially methods that propose stages of maturity for set ROIs, follow an approach of identifying the minimum age in which the ossification of the growth plate is completed for a particular ROI. These studies usually focus on a single age of legal importance, which varies significantly between countries, with ages ranging from 14 to 21 years [13]. Using multiple ROIs may provide more information about more ages.

An additional drawback that is specific to the GP and TW methods is that they are based on data collected from subjects of average and upper socioeconomic classes in the 30s and 50s, respectively. Hence, these methods may not reflect secular trends that nowadays point to higher height and earlier puberty [27], which could affect the accuracy of the methods. For the TW method, an update released in 2001 (TW3) revised the calculation of the BA from the attributed scores to address this problem [28].

Machine Learning for BAA

From the presented drawbacks, it is noticeable that the BAA research could benefit from methods that are able to aggregate multiple pieces of information (ie, multiple ROIs and factors) in a systematic way. A technology that is able to work in this setting is machine learning (ML), which is already widely employed in diverse medical fields, such as diabetes, cancer, cardiology, mental health, and the analysis of clinical text data [29,30]. ML consists of various types of algorithms that are able to learn how to perform a task from a set of examples while improving its performance based on its experience in carrying out a particular task. It builds a model that encapsulates the knowledge to perform the task; then, in light of new data, the model is able to correctly perform the learned task within an acceptable measure of performance [31].

ML algorithms have already been employed in various models for assessing the BA of an individual. A recent systematic literature review on BAA with ML methods [20] showed that the research is heavily focused on models that make use of a single ROI, the hand in most cases, having radiographs as the choice of imaging technology and do not usually consider other factors that could play a role in bone development [20]. The most notable, commercially available ML BAA system is the BoneXpert [32], which performs an automatic radiograph analysis based on the GP and TW methods. However, it covers the age range of 2 to 17 years and leaves out important legal ages.

Objectives of the Study

Given the importance of the assessment of CA through BAA in numerous scenarios and its potential ways of affecting the lives of young individuals, it is important to focus on the drawbacks of the methods currently in use and investigate the potential of BAA in estimating CA. Thus, the objectives of this study are as follows:

- To investigate the extent to which ML models can aid in CA estimation through BAA in young individuals aged 14 to 21 years.
- To investigate whether ML models can aid in the determination of minors through BAA, considering the threshold of 18 years, in young individuals aged 14 to 21 years.
- To address the drawbacks in the research on CA estimation from BAA, with regard to using radiation-free medical

imaging technology, the assessment of multiple ROIs and other factors that may play a role in bone development.

Methods

Overview

To train the CA estimation ML models proposed in this paper, MRI images of the wrist, knee, and foot were taken from volunteer subjects and assessed by radiologists to evaluate their stage of bone development. The 5 growth zones considered in this study were calcaneus, distal tibia, proximal tibia, distal femur, and radius. Each growth zone was assessed separately and blinded to gender and age.

Before the examination, the subjects had their height and weight measured for the BMI calculation and were asked to answer a questionnaire to gather information on their physical activity level, parents' origin, type of residence during upbringing, and a self-assessed Tanner Scale of pubertal growth [33,34].

All radiological and nonradiological data gathered were used to train binary and multiclass classifiers. For the binary classifier, the individuals in the sample were divided into minors or adults, with a threshold of 18 years, and the classification followed into discriminating individuals into 1 of the 2 classes. The multiclass classifier aims to classify an individual into 1 of the 8 classes defined by age groups ranging from 14 to 21 years.

The remainder of this section details the population, data used in the experiments, statistical analysis, and procedures for model building in the experiments.

Recruitment

This study prospectively conducted MRI examinations of 938 healthy subjects (465 males and 473 females) aged between 14 and 21 years (inclusive), during 2017 and 2018. The participants of the study had images taken from the knee, foot, and wrist in the same examination session. Additionally, the weight and height of each participant were also collected to calculate the BMI.

The following criteria were used to determine participation in the study:

- Inclusion criteria: the participants should have been born in Sweden, where the study was conducted, and have a birth certificate verified by the Swedish national authorities.
- Exclusion criteria: a history of bilateral fractures or trauma near the regions of assessment, a history of chronic disease or the use of long-term medications, noncompliance during the examination, having resided outside Sweden for more than six consecutive months, or past or current pregnancy (all female subjects were tested).

Data Privacy and Study Ethics

The study was conducted in accordance with the Declaration of Helsinki and was approved by the Central Ethical Review Board in Stockholm (diary numbers: 2017/4-31/4, 2017/1184-32, 2017/1773-32). Written informed consent was obtained from all subjects and legal guardians (in the case of subjects aged younger than 18 years). All data were anonymized and stratified by age and gender.

Population

A total of 455 male and 467 female subjects constituted the final sample (Table 1). After the MRI examinations and assessment of images by radiologists, 10 male and 6 female subjects were removed from the study's sample because they had the assessment of one or more ROI missing. The missing values for the assessment by the radiologists could be due to one of the following reasons: movement artifact, error in the sequence that made the image nongradable, likely trauma in the region of assessment, and missing MRI examination in one or more ROIs.

Table 1. Demographics of the final sample.

Demographics	Age group								Total
	14	15	16	17	18	19	20	21	
Number of female subjects	59	58	57	60	59	57	57	60	467
Number of male subjects	58	56	60	58	53	58	53	59	455

Data and Data Collection Procedures

The data used to train the classifiers were the radiologists' assessment of the calcaneus, distal tibia, proximal tibia, distal femur, and radius growth zones; the additional information gathered before the examination was physical activity level, parents' origin, type of residence during upbringing, and a self-assessed Tanner Scale of pubertal growth and BMI. The following section details the data and procedures for collection.

MRI Examinations

MRI examinations were performed to capture images of the calcaneus, distal tibia, proximal tibia, distal femur, and radius growth plates of the subjects participating in the study. All MRI examinations were conducted within 6 months of the subjects' birthday date on 1.5-T whole-body MRI scanners with dedicated hand, knee, and ankle coils. The examinations were performed on the nondominant side of the knee, hand, and foot, save when past fracture or trauma had taken place near the region. In these cases, the dominant side was imaged. The images of all ROIs were taken in the same examination session.

The examinations were carried out at 2 sites. Site 1 used Magnetom Avanto Fit (Siemens Healthcare GmbH) and Achieva (Philips Healthcare) whole-body scanners, and Site 2 used a Signa (GE Healthcare) whole-body scanner. All examinations followed the same protocol, which included a T2 sequence with cartilage dedicated exposure. The settings were 256×256 pixel resolution and 160×160 mm field of view.

Assessment of Magnetic Resonance Images

The assessment of the MRI images was performed independently by 2 radiologists with 3 and 30 years of experience in pediatric radiology, who were blinded to the age and gender of the participants. A third radiologist with 13 years of experience in pediatric radiology assessed the images when the first 2 radiologists could not reach a final agreement about the stage.

The staging system used to assess MRI images is a version of the staging methods proposed by Dedouit et al [16] and Kellinghaus et al [14] with minor modifications. This staging is defined as follows:

- Stage 1: Continuous, stripe-like, cartilage signal intensity is present between the metaphysis and epiphysis with a thickness greater than 1.5 mm with a multilaminar appearance.
- Stage 2: Continuous cartilage signal intensity is present between the metaphysis and epiphysis with a thickness greater than 1.5 mm with increased signal intensity but without a multilaminar appearance.
- Stage 3: Continuous cartilage signal intensity is present between the metaphysis and epiphysis with a thickness of less than 1.5 mm with increased signal intensity.
- Stage 4a: Noncontinuous cartilage signal intensity. A hazy area involving one-third or less of the growth plate is present between the metaphysis and epiphysis, representing the epiphyseal-metaphyseal fusion.
- Stage 4b: Noncontinuous cartilage signal intensity. A hazy area involving between one-third and two-third of the growth plate is present between the metaphysis and epiphysis, representing epiphyseal-metaphyseal fusion.
- Stage 4c: Noncontinuous cartilage signal intensity A hazy area involving more than two-thirds of the growth plate is present between the metaphysis and epiphysis, representing epiphyseal-metaphyseal fusion.
- Stage 5: The epiphyseal cartilage fused completely with or without an epiphyseal scar in all MRI slices.

Body Mass Index

The BMI was calculated using the measures of the participants' weight *w* and height *h*, as in the following equation 1 [35]:

Data characteristics regarding the calculated BMI for the subjects are shown in the Multimedia Appendix 1.

Questionnaire Information

Additional information from the participants was gathered by a questionnaire given to them at the examination session. The information gathered by the questionnaire refers to the variables "Residence," "Physical Activity," "Parent Origin," and "Tanner Scale," shown in Table 2, which summarizes all input and output variables considered for building the models. Data characteristics regarding the data collected by the questionnaire are shown in the Multimedia Appendix 1.

Table 2. Summary of the input and output variables considered in the model building.

Variable	Description	Values
Input variables		
Radius	Radiologists' assessments of the Radius growth zone	Stage 1; Stage 2; Stage 3; Stage 4a; Stage 4b; Stage 4c; Stage 5
Distal femur	Radiologists' assessments of the distal femur growth zone	Stage 1; Stage 2; Stage 3; Stage 4a; Stage 4b; Stage 4c; Stage 5
Proximal tibia	Radiologists' assessments of the proximal tibia growth zone	Stage 1; Stage 2; Stage 3; Stage 4a; Stage 4b; Stage 4c; Stage 5
Distal tibia	Radiologists' assessments of the distal tibia growth zone	Stage 1; Stage 2; Stage 3; Stage 4a; Stage 4b; Stage 4c; Stage 5
Calcaneus	Radiologists' assessments of the calcaneus growth zone	Stage 1; Stage 2; Stage 3; Stage 4a; Stage 4b; Stage 4c; Stage 5
BMI	Body mass index of the participant, calculated as in the equation (1)	Numeric
Residence	Type of residence the participant lives in (or lived during upbringing)	Rented; owned
Physical activity	The participants' daily level of activity	Highly inactive; inactive; little active; active; highly active
Parent origin	Origin of the participants' parents, regarding if they were born outside Sweden or not	No foreign-born parents; one foreign-born parent; both foreign-born parents
Tanner scale	Self-assessed Tanner Scale for pubertal growth [33,34]	Stage 1; Stage 2; Stage 3; Stage 4; Stage 5
Output variables		
Minor	Characterizes the participant as being a minor or not, regarding the threshold of 18 years. This is the output variable for the binary classification models	Yes; no
Age	Regards the age group which the participant belongs to. This is the output variable for the multi-class classification models	14; 15; 16; 17; 18; 19; 20; 21

Statistical Analyses

The Cohen kappa coefficient [36] and percent of agreement [37] were calculated to measure the interobserver agreement between the pediatric radiologists in all investigated ROIs. Statistical analyses were performed using SPSS Statistics (version 24; IBM Corp).

Model Building

CA Estimation Models

In this study, various ML algorithms were investigated to build classifiers to discriminate subjects into minor (positive class) or adults (negative class) and classifiers to classify subjects into 1 of 8 age groups (14 to 21 years). Models for male and female subjects were built separately.

Data Preprocessing

The data used to build the models consisted of the radiologists' assessment of the 5 growth zones, following the aforementioned stages, the questionnaire information, and the calculated BMI. These data presented missing values that were handled by the K-nearest neighbor (KNN) multiple imputations. This technique finds K complete entries that are the closest to an incomplete entry (ie, contains missing data) and fills its missing values with the mean (in the case of numeric variables) or the most frequent one (in the case of categorical variables) [38]. In this study, the number of nearest neighbors K for the KNN imputation was set to 1. The motivation for this choice is based on literature findings that advise limiting K as a way to preserve the original variability of the data, reducing the risk of entries having few neighbors that are too distant from each other [39]. There is also a risk of increasing the influence of noise in the data with a small K, but as in the data set of this study, the highest rate of imputed instances was 1.9%; this influence was considered to be not very relevant. The distance used by the KNN multiple imputation technique was the Gower distance [40]. The number of imputed instances for each variable in both male and female subsets is shown in Table 3.

Table 3. Number of imputed instances and percentage over the male and female data sets.

Variable	Male data set, n (%)	Female data set, n (%)
Radiologists' assessments of the radius, distal femur, proximal tibia, distal tibia, calcaneus	0 (0)	0 (0)
Residence	1 (0.2)	3 (0.6)
Physical activity	9 (1.9)	6 (1.2)
Tanner Scale	3 (0.6)	1 (0.2)
BMI	0 (0)	0 (0)
Parents origin	0 (0)	3 (0.6)

ML Algorithms

The choice of the ML algorithms explored in this study was based on the summary of the evidence of a recently published systematic literature review (SLR) on the application of ML for BAA [20]. This SLR points out that the studies proposing BAA classifiers employ algorithms of the following categories: artificial neural networks, support vector machines, Bayesian networks, decision trees, and K-nearest neighbors. An additional search was conducted in the literature (Scopus, PubMed, and Web of Science), after the search date of the mentioned SLR [20] (February 2019) to look for additional algorithms, but no new categories were found to be added to the list.

Another motivation for this choice of ML algorithms is that it also guarantees a diversified list of classifiers that make use of different types of learning techniques, such as rule-based, instance-based, Bayesian inference, kernel, and perceptron learners. We referred to the following book by Kuhn and Johnson [41] for the specific algorithms and implementations used in this study.

Therefore, the choice of ML algorithms for the experiments of this study includes decision tree, random forest, multilayer perceptron, support vector machines, naïve Bayes, and K-nearest neighbors.

Experimental Setup

All experiments were performed using a stratified, nested cross-validation [42]. In this approach, in each iteration, one fold of the outer cross-validation is used for testing and the remaining 4 are used in an inner cross-validation to tune the algorithm's hyperparameters. This was done to obtain a more reliable estimate of the error as the test fold in each outer iteration is not used to execute performance optimization [43]. It is also worth noting that the data splits were performed in a stratified manner, which means that the classes' proportions in each split are kept the same as in the total sample. In the experiments of this study, a five-fold outer, three-fold inner stratified nested cross-validation was performed. The reduced number of folds in the inner cross-validation was employed to avoid having a low number of subjects to represent each class in the folders, due to the high number of classes in the multiclass classification problem. Additionally, before each inner cross-validation, a grid search was performed to find suitable hyperparameters for each of the selected ML algorithms. The hyperparameters for each selected algorithm are listed in Table 4. The ML experiments were conducted in the R framework with the *caret* package. The default versions of the algorithms were used.

Table 4. Configuration of the R algorithms included in the experiment.

ML[a] algorithm	R implementation	Tuning parameters
Decision tree	*rpart*	*cp*
Random forest	*rf*	*mtry*
Multi-layer perceptron	*mlp*	*size*
Support vector machines	*svmRadial*	*Sigma, C*
Naïve Bayes	*nb*	*fL, usekernel, adjust*
K-nearest neighbors	*knn*	*k*

[a] ML: machine learning.

Model Evaluation Metrics

The performance metrics used to evaluate the models were as follows: mean absolute error (MAE), root mean squared error (RMSE), accuracy, precision, recall, and area under the curve (AUC), as in Gaudette and Japkowicz [44] and Sokolova and Lapalme [45] guidelines for ordinal multiclass classification. For the binary classification models, all but MAE and RMSE are used. The SDs for each metric are also reported.

The MAE represents the mean of the absolute difference between the estimated age output of the classifier and the correct CA of the subject, over all examples. The RMSE gives more weight to larger errors compared with MAE, which tends to prefer fewer errors overall. The MAE and RMSE are calculated as follows in equations 2 and 3, respectively:

Where n is the number of samples, is the estimated age, and y is the CA of the subject.

For the remaining evaluation metrics, considering l the number of classes, we define the following:

- True-positives (TP): Entries predicted to be in class C_l actually in class C_l.
- False-positives (FP): Entries predicted to be in C_l but are not actually in class C_l.
- True-negatives (TN): Entries not predicted to be in C_l and are not actually in class C_l.
- False-negatives (FN): Entries not predicted to be in C_l, but are actually in class C_l.

The accuracy, precision, recall, and AUC for binary classification are calculated as follows:

In the case of the multiclass classification, these are calculated as the average of the metrics calculated for each class C_l (macro averaging) [45]. The AUC metric is calculated by averaging pairwise comparisons, as proposed by Hand and Till [46].

General results are given for the ML algorithms in terms of the mean and SDs of each of the performance metrics for the outer cross-validation test sets. In-depth results are given to the best performing models.

Results

Interobserver Agreement

The kappa Cohen coefficient was calculated to evaluate the agreement between the 2 observers' assessments of the MRI images. The results indicated substantial agreement according to the general guidelines [47] for all of the assessed ROIs: 0.77 for the calcaneus, 0.65 for the distal femur, 0.72 for the distal tibia, 0.73 for the proximal tibia, and 0.67 for the radius.

The percent agreement for the assessed ROI was as follows: 94.2% for the calcaneus, 80.8% for the distal femur, 90.6% for the distal tibia, 86.8% for the proximal tibia, and 79.4% for the radius. These results show that the radiologists agreed on a stage in the vast majority of cases.

Results of the Growth Plate Assessments

The results of the assessments of the calcaneus, distal tibia, proximal tibia, distal femur, and radius for male and female subjects are shown in detail in Multimedia Appendices 2 and 3, respectively.

In all of the assessed growth plates, for both sexes, stages 1 and 2 were not evidenced. Few instances of stage 3 were observed on male subjects on the calcaneus and radius growth plates, accounting for 2 and 15 cases, respectively. In female subjects, stage 3 was evidenced in only 2 cases for the radius growth plate.

The female subjects' results show that for all assessed growth plates, nearly all or most of the sample was already in the last stage of ossification (stage 5): 94.6% of the calcaneus, 90.8% of the distal tibia, 81.6% of the proximal tibia, 74.5% of the distal femur, and 65.5% of the radius cases. These numbers moderately change for male subjects, accounting for 80.4% of the calcaneus, 70.1% of the distal tibia, 57.6% of the proximal tibia, 54.9% of the distal femur, and 47.4% of the radius cases.

Table 5 shows the proportion within each age group of subjects who had all of the growth plates considered in this study already in stage 5. This table shows that female subjects had all growth plates fused 2 years before the male subjects. For female subjects, from the age of 19 years, all subjects of the sample already have all of the growth plates fused, although for male subjects, the same happens from the age of 21 years.

Table 5. Numbers and percentages (over each age group) of subjects with all of the growth plates in stage 5, for male and female subjects.

Characteristic	Female subjects, n (%)	Male subjects, n (%)
Age group (years)		
14	2 (3.3)	0 (0)
15	8 (13.7)	0 (0)
16	23 (40.3)	3 (5)
17	44 (73.3)	13 (22.4)
18	53 (89.8)	31 (58.4)
19	57 (100)	50 (86.2)
20	57 (100)	50 (94.3)
21	60 (100)	59 (100)
Total	304 (65.1)	206 (45.2)

Results for the Classification of Minors Versus Adults

A threshold of 18 years was used to determine adulthood in the classification of minors versus adults, which is the case in many European countries. MAE and RMSE were not used as performance metrics in this case because for classifications they only make sense in the context of an ordinal classification. The results for the male subjects' binary classifiers in terms of the mean and SD of the performance metrics on the outer cross-validation test sets are shown in Table 6.

Table 6. Mean performance metrics and respective SDs (in years) for the classification of minor versus adults for the male subjects.

Types	Accuracy, mean (SD)	AUC[a], mean (SD)	Precision, mean (SD)	Recall, mean (SD)
Decision tree	0.90 (0.02)	0.90 (0.02)	0.86 (0.04)	0.96 (0.03)
Random forest	0.90 (0.01)	0.90 (0.01)	0.87 (0.03)	0.94 (0.04)
Support vector machines	0.90 (0.02)	0.90 (0.02)	0.87 (0.04)	0.93 (0.07)
Multi-layer perceptron	0.82 (0.17)	0.82 (0.16)	0.79 (0.16)	0.95 (0.04)
K-nearest neighbors	0.87 (0.02)	0.87 (0.02)	0.84 (0.03)	0.92 (0.03)
Naïve bayes	0.73 (0.04)	0.74 (0.04)	0.65 (0.03)	1.00 (0.00)

[a] AUC: area under the curve.

The decision tree, random forest, and support vector machine algorithms had very similar results in general, presenting no significant difference between them. The random forest algorithm was chosen in terms of the best combination of precision and recall, but in practical settings, there are no differences between these algorithms. Table 7 shows the random forest results for each of the outer cross-validation test sets. The average model was chosen in terms of median accuracy, which was 0.90. Between Models 1 and 4, Model 1 was chosen for better recall in classifying minors. The optimized hyperparameter given by the grid search for Model 1 was *mtry=2* (number of candidate variables at each tree split).

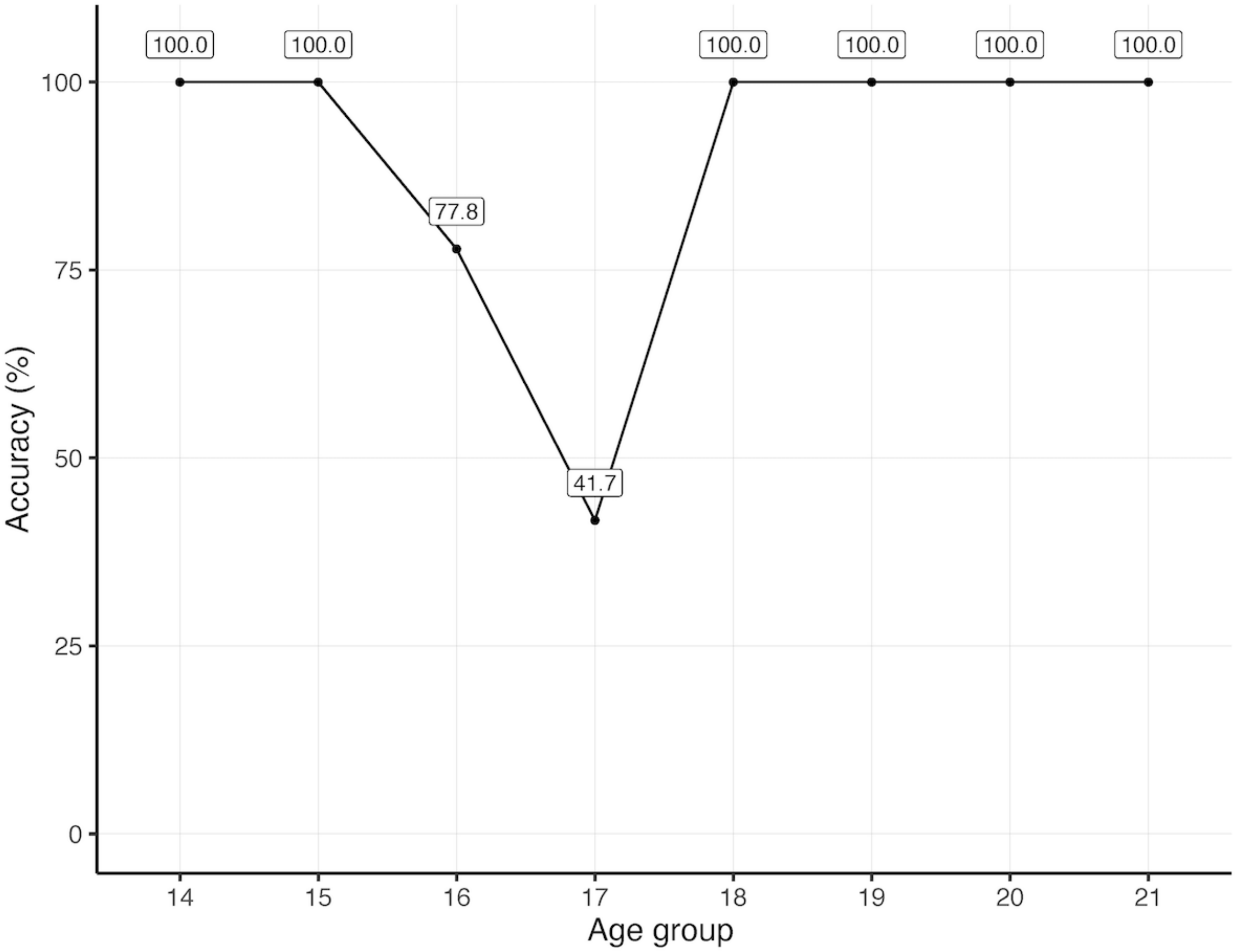
Table 7. Performance results for the Random Forest algorithm on each of the outer cross-validation test sets, for the male sample.

Model	Accuracy	AUC[a]	Precision	Recall—minors	Recall—adults
1 (median)	0.90	0.90	0.83	1.00	0.80
2	0.89	0.89	0.87	0.91	0.87
3	0.89	0.89	0.87	0.96	0.83
4	0.90	0.90	0.90	0.89	0.91
5	0.92	0.92	0.89	0.96	0.89

[a] AUC: area under the curve.

Figure 1 presents the results achieved by Model 1 per age group. It is important to note that even with low accuracy results for the age of 17 years (41.7%), the model still minimizes the error of classifying minors as adults, achieving a recall of 100% for this classification.

Figure 1. Accuracy per age group for the minor versus adults classification model, for male subjects.



In the case of the female subjects, the decision tree, random forest, and multi-layer perceptron algorithms presented very similar results (Table 8), which do not present a relevant significant difference between them. The chosen algorithm for the female subject case was the random forest algorithm for the best combination of performance measures. Table 9 shows the results for the random forest algorithm in each of the outer cross-validation test sets. Except for Model 1, there was essentially no relevant variation between models, and in practical settings, they can be considered equal. Thus, Model 2 was chosen as the average model. The optimized hyperparameter given by the grid search for Model 1 was mtry=6.

Table 8. Mean performance metrics and respective SDs (in years) for the classification of minor versus adults for the female subjects.

Types	Accuracy, mean (SD)	AUC[a], mean (SD)	Precision, mean (SD)	Recall, mean (SD)
Decision tree	0.82 (0.02)	0.82 (0.02)	0.74 (0.02)	0.97 (0.01)
Random forest	0.83 (0.02)	0.83 (0.01)	0.76 (0.02)	0.97 (0.01)
Support vector machines	0.81 (0.04)	0.81 (0.04)	0.75 (0.04)	0.92 (0.05)
Multi-layer perceptron	0.82 (0.02)	0.82 (0.02)	0.75 (0.02)	0.95 (0.04)
K-nearest neighbors	0.78 (0.06)	0.78 (0.06)	0.73 (0.06)	0.87 (0.08)
Naïve bayes	0.67 (0.03)	0.67 (0.02)	0.60 (0.02)	1.00 (0.00)

[a] AUC: area under the curve.

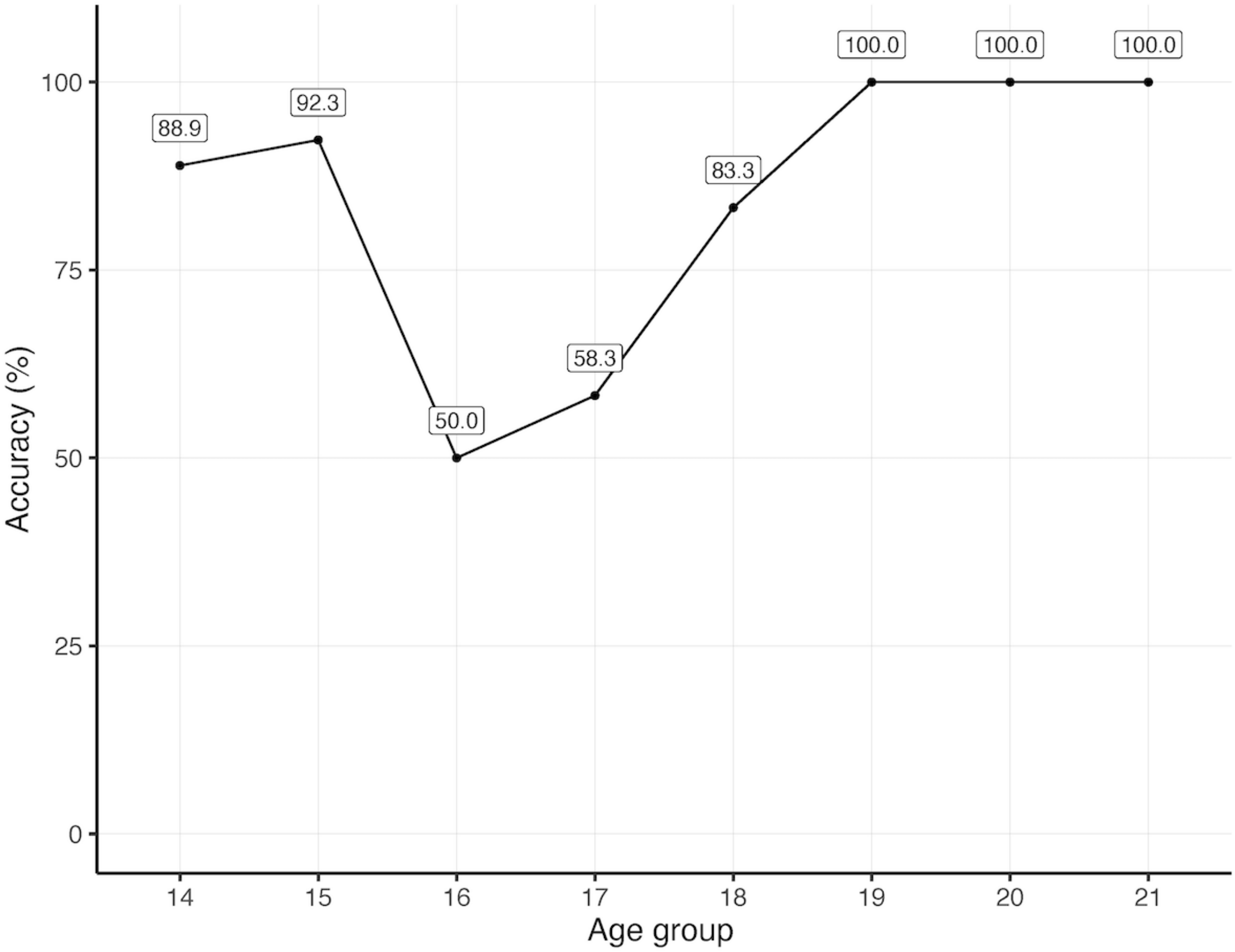
Table 9. Performance results for the random forest algorithm on each of the outer cross-validation test sets, for the female sample.

Model	Accuracy	AUC[a]	Precision	Recall—minors	Recall—adults
1	0.81	0.81	0.73	0.96	0.66
2 (median)	0.84	0.84	0.77	0.96	0.72
3	0.84	0.84	0.77	0.98	0.70
4	0.84	0.84	0.78	0.96	0.72
5	0.84	0.84	0.77	0.98	0.70

[a] AUC: area under the curve.

The accuracies per age group are shown in the graph in Figure 2. The model achieved lower accuracies for the ages of 16 and 17 years (50.0% and 58.3%, respectively), but as in the case of the male subjects, the model minimizes the worst type of error, which is the misclassification of minors, achieving a high recall of 96%.

Figure 2. Accuracy per age group for the minor versus adults classification model, for female subjects.



Results for the CA Estimation Models

The CA estimation models are multiclass classifiers that aim to classify subjects in 1 of the 8 age groups (from 14 to 21 years). Table 10 shows the results for the male subjects' models in terms of the mean and SDs of the performances on the outer cross-validation test sets. The best performing algorithm in the male case was the multilayer perceptron (MLP), which achieved the best MAE (0.98 years), mean RMSE (1.32 years), and mean precision (0.65 years) values, in addition to having the second-best values of mean accuracy and mean AUC.

Table 10. Mean (SD) of the performance metrics for the male subjects' classification models.

Algorithms	MAE[a] (years), mean (SD)	Accuracy, mean (SD)	RMSE[b] (years), mean (SD)	AUC[c], mean (SD)	Precision, mean (SD)	Recall, mean (SD)
Decision tree	1.28 (0.13)	0.32 (0.03)	1.78 (0.17)	0.81 (0.02)	0.49 (0.06)	0.81 (0.11)
Random forest	1.04 (0.07)	0.34 (0.03)	1.44 (0.13)	0.85 (0.01)	0.57 (0.09)	0.73 (0.14)
Support vector machine	1.03 (0.09)	0.34 (0.03)	1.43 (0.08)	0.85 (0.01)	0.52 (0.09)	0.67 (0.12)
Multi-layer perceptron	0.98 (0.08)	0.33 (0.02)	1.32 (0.13)	0.84 (0.01)	0.65 (0.27)	0.61 (0.31)
K-nearest neighbor	1.16 (0.11)	0.30 (0.04)	1.57 (0.15)	0.82 (0.03)	0.59 (0.10)	0.59 (0.10)
Naïve bayes	1.07 (0.10)	0.29 (0.02)	1.39 (0.19)	0.81 (0.01)	0.57 (0.06)	0.58 (0.21)

[a] MAE: mean absolute error.

[b] RMSE: root mean squared error.

[c] AUC: area under the curve.

The performances of the MLP algorithm on each of the outer cross-validation test sets are shown in Table 11. The average model was chosen in terms of the median MAE, which corresponds to Model 1, with a value of 0.95 years. The optimized hyperparameter given by the grid search for the average MLP model was *size=27* (number of units in the hidden layer). The average model was chosen to select an algorithm that would not be overly optimistic in its estimation.

Table 11. Performance results for the multi-layer perceptron algorithm on each of the outer cross-validation test sets, for the male sample.

Model	MAE[a] (years)	Accuracy	AUC[b]	RMSE[c] (years)	Recall	Precision
1 (median)	0.95	0.33	0.83	1.29	0.91	0.48
2	1.08	0.30	0.85	1.40	073	0.35
3	0.89	0.32	0.84	1.17	0.17	1.00
4	0.91	0.33	0.83	1.23	0.83	0.59
5	1.05	0.35	0.84	1.49	0.42	0.83

[a] MAE: mean absolute error.

[b] AUC: area under the curve.

[c] RMSE: root mean squared error.

The results for the chosen model, discriminated by age groups, are shown in Table 12. The model shows lower errors for the younger and older ages of the age spam considered in the study. In addition, the model has a clear trend of overestimating the ages of the male subjects in general. Thus, even with an MAE of 0.95 years, the model is limited to its capacity to classify individuals from the age of 16 years. From the age of 19 years, the model tends to classify all subjects as 20 years old as nearly all subjects of these ages have all growth plates on stage 5.

Table 12. Mean absolute error and SD for the average male model.

Measure	Age group							
	14	15	16	17	18	19	20	21
MAE[a] (years)	0.18 (SD 0.60)	0.82 (SD 0.90)	1.25 (SD 1.44)	1.91 (SD 1.56)	1.50 (SD 1.45)	1.00 (SD 0.60)	0.00 (SD 0.00)	0.92 (SD 0.29)

[a] MAE: mean absolute error.

Table 13 shows the results for the CA estimation models for female subjects in terms of the mean and standard deviations of the performances on the outer cross-validation test sets. In the case of the female subjects, the best performing algorithm was the support vector machine (SVM), which achieved the best MAE (1.21 years), mean accuracy (0.32), mean RMSE (1.68 years), and mean AUC (0.80).

Table 13. Mean (SD) of the performance metrics, for the female subjects' classification models.

Algorithms	MAE[a] (years), mean (SD)	Accuracy, mean (SD)	RMSE[b] (years), mean (SD)	AUC[c], mean (SD)	Precision, mean (SD)	Recall, mean (SD)
Decision tree	1.31 (0.09)	0.28 (0.02)	1.78 (0.13)	0.80 (0.02)	0.56 (0.05)	0.82 (0.09)
Random forest	1.29 (0.10)	0.30 (0.03)	1.77 (0.10)	0.79 (0.02)	0.59 (0.13)	0.74 (0.17)
Support vector machine	1.21 (0.06)	0.32 (0.04)	1.68 (0.06)	0.80 (0.01)	0.55 (0.07)	0.71 (0.11)
Multi-layer perceptron	1.36 (0.24)	0.30 (0.02)	1.85 (0.37)	0.77 (0.02)	0.60 (0.11)	0.63 (0.22)
K-nearest neighbors	1.41 (0.12)	0.30 (0.02)	1.96 (0.12)	0.76 (0.03)	0.55 (0.07)	0.61 (0.18)
Naïve bayes	1.74 (0.23)	0.22 (0.02)	2.23 (0.27)	0.65 (0.03)	0.58 (0.06)	0.82 (0.09)

[a] MAE: mean absolute error.

[b] RMSE: root mean squared error.

[c] AUC: area under the curve.

Table 14 shows the performance results for each of the outer cross-validation test sets for the SVM algorithm. For the case of the female subjects, the median resulted in an MAE of 1.24 years, which pertained to Models 1 and 2. Model 1 was chosen as the average model for presenting the best accuracy between the two. The optimized parameter given by the grid search for the average SVM model was *sigma=0.0421* (kernel parameter) *and C=4* (penalty parameter).

The MAE results per age group are shown in Table 15. As in the male subjects' case, the female model also overestimates the ages of female subjects in general, but with higher MAE and standard deviations.

Table 14. Performance results for the support vector machine algorithm on each of the outer cross-validation test sets, for the female sample.

Model	MAE[a] (years)	Accuracy	AUC[b]	RMSE[c] (years)	Recall	Precision
1 (median)	1.24	0.37	0.79	1.75	0.75	0.56
2	1.24	0.27	0.80	1.67	0.55	0.67
3	1.25	0.33	0.78	1.70	0.75	0.50
4	1.11	0.32	0.81	1.58	0.67	0.53
5	1.20	0.32	0.81	1.72	0.83	0.83

[a] MAE: mean absolute error.

[b] AUC: area under the curve.

[c] RMSE: root mean squared error.

Table 15. Mean absolute error and standard deviation for the male median model

Measure	Age group							
	14	15	16	17	18	19	20	21
MAE[a] (years), mean (SD)	0.42 (0.79)	1.42 (1.93)	1.17 (1.79)	2.17 (2.49)	1.75 (1.88)	1.09 (1.43)	0.90 (1.27)	1.00 (1.34)

[a] MAE: mean absolute error.

Discussion

Principal Findings

This paper presents experiments with the estimation of CA and classification of minors versus adults (on the threshold of 18 years) of male and female subjects using ML algorithms. To build the models, 2 radiologists assessed the stage of bone development of the calcaneus, distal tibia, proximal tibia, distal femur, and radius growth plates of 455 male and 467 female volunteer subjects (922 subjects in total) from MRI images. Additional variables were also used to build the models: BMI, physical activity level, parents' origin, type of residence during upbringing, and self-assessed Tanner Scale of pubertal growth. The methodology adopted in the study aimed at addressing the drawbacks of the BAA methods that are employed in CA estimation for legal scenarios.

From the stage assessments of the MRI images, we could infer that female subjects mature earlier than male subjects regarding the bone development of the knee, wrist, and foot, which is in line with prior studies [1,17,18,48,49]. In this study, the first age in which the whole sample had all fused growth plates (stage

5) was 19 years for female (467/467, 100%) and 21 years for male (455/455, 100%) subjects.

Another important point to be discussed with regard to the stage assessments is that the female sample had cases that had all of the considered growth plates already fused since the age of 14 years, spamming throughout all ages considered in the study (14 to 21 years). Since the assessment of stage 5, unlike the other stages, requires that all of the slices from the MRI examination to present a fused growth plate, even if there is a degree of misassessment, it would still mean that these cases would display a well-advanced level of maturation in all of these ages, implying a high degree of biological variation in the female sample with regard to BA. Additionally, in total, 65.5% (304/467) of the female sample consisted of cases in which the subjects presented all growth plates already in stage 5, which means that for classification purposes, for more than half of the sample, the estimation of CA would depend only on the additional factors (self-assessed Tanner Scale, BMI, residence type, physical activity, and parents origin), which were not enough to discriminate between age groups. This hindered the performance of classifiers, especially the CA estimation models. The same phenomenon occurred for the male sample, which also negatively affected the performance of the classifiers, but to a lesser degree, as 45.2% (206/455) of the sample had all growth plates of stage 5, from the age of 16 to 21 years.

The minors versus adults classification achieved good accuracy results for both male (90%) and female subjects (84%). These models portrayed a drop in the performance for the ages of 16 and 17; however, the recalls regarding the correct classification of minors were very high in both male and female cases (100% and 96%, respectively). This is important because the problem of minors versus adults classification is asymmetric as the misclassification of minors for adults in a judicial scenario is much more problematic than the inverse. In most cases, the application of the law is harsher for adults, and imputability, along with granted rights, can drastically change between these groups.

The CA estimation models achieved MAEs of 0.95 years and 1.24 years for male and female subjects, respectively. However, a look at a depth of the models showed that for both male and female models, only the ages of 14 and 15 years achieved acceptable MAE values. It could be argued that for the ages of 16 to 21 years, the estimation of a precise CA based on stages of bone development of the calcaneus, distal tibia, proximal tibia, distal femur, and radius growth plates would be somewhat unfit for male individuals and very unfit for female individuals. Furthermore, we could argue that staging may not offer a precise enough measure for the estimation of the CA of individuals of the ages considered in this study.

Compared with dental age, height, and age at menarche, BA is still the most reliable biological indicator for assessing maturation in young individuals [50], but it may not be a strong predictor of CA. BAA was conceived to be used in conjunction with CA to evaluate the maturation of an individual that can be delayed or advanced due to various factors that may include hormonal disorders, and chronic illnesses [8].

Regarding the agreement of the radiologists on the assessment of the growth plates' stage of development, substantial agreement was achieved, which is a satisfactory result as there is a lack of guidelines for BAA using MRI in the research. In addition, the individuals employed in the assessment of the MRI images were specialized pediatric radiologists with experience in BAA.

From a methodological point of view, this study employed a nested cross-validation approach that aims to avoid reporting overly optimistic results that could be derived from a *lucky* test set.

Comparison With Prior Work

Most of the studies in the area of BAA that employ ML algorithms aim to build automatic approaches for estimating BA and evaluating BA given by radiologists [20]. The biggest initiative for proposing automated approaches in this direction was the Radiological Society of North America (RSNA) 2018 Bone Age Challenge [51]. This challenge provided a database of circa 12,000 radiographs of subjects from 0 to 19 years, labeled with the BA given by radiologists, following the GP method. Although the first places achieved MAEs of 4.26, 4.35, and 4.38 months, these results are not comparable with our results because the aim of our study was to estimate CA, and the RSNA challenge goal was to propose models for predicting the BA given by radiologists [51]. In addition, it is worth mentioning that however large the sample provided for the challenge, only 0.74% (94/12,612) of the sample consisted of 18- and 19-year-old subjects, which are important legal ages.

For studies that employ BA concepts to predict the CA of subjects, there are studies by Dallora et al [52] and Stern et al [53]. Both employ MRI as the medical imaging of choice, and most importantly, they are not based on traditional BAA to make their predictions of CA. They employ deep learning technology, which is able to learn the important features in the images and then perform regression or classification [54]. The reasoning behind using deep learning to interpret images and learn features is that it is difficult for humans to translate image features into descriptive means, and it is easy to lose information on the process. On the other hand, this problem has a reduced risk of occurring with algorithms able to analyze images pixel by pixel [55]. Dallora et al [52] used knee MRI images and achieved an MAE of 0.793 years for male subjects in the range of 14 to 20 years, and 0.988 years for female subjects in the range of 14 to 19 years. Stern et al [53] used MRI images of the hand and achieved an MAE of 0.82 years for male subjects in the range of 13 to 19 years. A previous study by Stern et al [56] proposed a deep learning multifactorial approach that used MRI volumes of the hand, clavicle, and teeth to estimate the CA of male subjects aged 13 to 25 years, achieving an MAE of 1.01 years. The study by Tang et al [57] used MRI for CA estimation in adolescents from 12 to 17 years, which leaves out the legal age of 18, using artificial neural networks. This was also a multifactorial approach that considered the subjects' height, weight, and bone marrow composition intensity quantified by MRI and TW3 assessment, achieving a mean disparity (comparison between the mean CA for all subjects and the mean estimated age for all subjects) of 0.1 years. This study also demonstrated that the BA given by the TW3 method was consistently lower than that of the subjects' CA. The study by Hillewig et al [58] investigated a multiple ROI approach that considered primarily the radiologists' assessment of MRI images of the clavicle, but also the assessment of x-rays of the hand and wrist area, with the aim of determining whether an individual is younger or older than 18 years, considering a sample of subjects from 16 to 26 years. It was evidenced that the clavicle assessment in stage IV (according to the Schmeling et al [9] and Kreitner et al [13] staging systems) was particularly important for age determination; however, in cases where staging is challenging for radiologists, the assessment of the hand and wrist area is essential.

Limitations

Regarding limitations of the study, it could be argued that due to the high number of classes in the multiclass classification, the sample size in each class would not be large enough to build a generalizable model. However, to address this issue, we employed methods to ensure that the model would not overfit and for not choosing the most overly optimistic choice given by the nested cross-validation. In addition, during data collection, we ensured a uniform number of subjects in each class to guarantee a balanced data set.

The selected ROI for this work took into consideration the stress levels for the minors and young adult subjects with regard to the MRI examination. Hence, the clavicle and arm were not considered because it would require the subjects to go head in the MRI machine, which could cause discomfort and stress to the young subjects due to loud noises and small enclosed spaces. In addition, the clavicle has a high risk of producing moving artifacts due to breathing movements. On the practical side, the examination time was on average 15 min, and the inclusion of these 2 regions would take approximately double the time.

Conclusions

This paper presented models for CA estimation and minors versus adults classification (on a threshold of 18 years) using ML algorithms. The models were trained with radiologists assessment of the calcaneus, distal tibia, proximal tibia, distal femur, and radius; and additional information regarding physical activity level, parents' origin, type of residence during upbringing, and a self-assessed Tanner Scale of pubertal growth. The models proposed for the classification of minor versus adults produced accuracies of 90% and 84% for male and female subjects, respectively, with very high recalls for the classification of minors. However, for the chronological age estimation for the 8 age groups, ranging from to 14 to 21, the variables in the model did not turn out to be precise enough for estimating the exact CA, only showing acceptable values of MAE for the ages of 14 and 15 years.

Future research should focus on applying deep learning technology for the estimation of CA using multiple ROIs.

Acknowledgments

The authors would like to express their greatest appreciation to the participants and staff who participated in our study. This work was supported by the National Board of Health and Welfare of Sweden (Socialstyrelsen). The funding source had no involvement in the study design, data collection, analysis, interpretation, or reporting of this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Data characteristics of the variables collected through the questionnaire and BMI.

[PDF File (Adobe PDF File), 96 KB - medinform_v8i9e18846_app1.pdf]

Multimedia Appendix 2

Results from the assessment of magnetic resonance images of male subjects.

[PDF File (Adobe PDF File), 25 KB - medinform_v8i9e18846_app2.pdf]

Multimedia Appendix 3

Results from the assessment of magnetic resonance images of female subjects.

[PDF File (Adobe PDF File), 24 KB - medinform_v8i9e18846_app3.pdf]

References

1. Gilsanz V, Ratib O. -. Hand Bone Age: A Digital Atlas Of Skeletal Maturity. Springer Science & Business Media; 2005:A.
2. Mansourvar M, Ismail MA, Herawan T, Raj RG, Kareem SA, Nasaruddin FH. Automated bone age assessment: motivation, taxonomies, and challenges. Comput Math Methods Med 2013;2013:391626 [FREE Full text] [doi: 10.1155/2013/391626] [Medline: 24454534]
3. Satoh M. Bone age: assessment methods and clinical applications. Clin Pediatr Endocrinol 2015 Oct;24(4):143-152 [FREE Full text] [doi: 10.1297/cpe.24.143] [Medline: 26568655]

4. Dvorak J, George J, Junge A, Hodler J. Application of MRI of the wrist for age determination in international U-17 soccer competitions. Br J Sports Med 2007 Aug;41(8):497-500 [FREE Full text] [doi: 10.1136/bjsm.2006.033431] [Medline: 17347314]
5. Dvorak J, George J, Junge A, Hodler J. Age determination by magnetic resonance imaging of the wrist in adolescent male football players. Br J Sports Med 2007 Jan;41(1):45-52 [FREE Full text] [doi: 10.1136/bjsm.2006.031021] [Medline: 17021001]
6. Schmidt S, Vieth V, Timme M, Dvorak J, Schmeling A. Examination of ossification of the distal radial epiphysis using magnetic resonance imaging. New insights for age estimation in young footballers in FIFA tournaments. Sci Justice 2015 Mar;55(2):139-144. [doi: 10.1016/j.scijus.2014.12.003] [Medline: 25754000]
7. Cunha E, Baccino E, Martrille L, Ramsthaler F, Prieto J, Schuliar Y, et al. The problem of aging human remains and living individuals: a review. Forensic Sci Int 2009 Dec 15;193(1-3):1-13. [doi: 10.1016/j.forsciint.2009.09.008] [Medline: 19879075]
8. Hjern A, Brendler-Lindqvist M, Norredam M. Age assessment of young asylum seekers. Acta Paediatr 2012 Jan;101(1):4-7. [doi: 10.1111/j.1651-2227.2011.02476.x] [Medline: 21950617]
9. Schmeling A, Schulz R, Reisinger W, Mühler M, Wernecke K, Geserick G. Studies on the time frame for ossification of the medial clavicular epiphyseal cartilage in conventional radiography. Int J Legal Med 2004 Feb;118(1):5-8. [doi: 10.1007/s00414-003-0404-5] [Medline: 14534796]
10. Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist. Am J Med Sci 1959;238(3):393. [doi: 10.1097/00000441-195909000-00030]
11. Ehrenberg AS, Tanner JM, Whitehouse RH, Marshall WA, Healy MJ, Goldstein H. Assessment of skeletal maturity and prediction of adult height (TW2 method). Appl Stat 1977;26(1):80. [doi: 10.2307/2346874]
12. Nahhas RW, Sherwood RJ, Chumlea WC, Duren DL. An update of the statistical methods underlying the FELS method of skeletal maturity assessment. Ann Hum Biol 2013;40(6):505-514 [FREE Full text] [doi: 10.3109/03014460.2013.806591] [Medline: 23992229]
13. Kreitner K, Schweden FJ, Riepert T, Nafe B, Thelen M. Bone age determination based on the study of the medial extremity of the clavicle. Eur Radiol 1998;8(7):1116-1122. [doi: 10.1007/s003300050518] [Medline: 9724422]
14. Kellinghaus M, Schulz R, Vieth V, Schmidt S, Pfeiffer H, Schmeling A. Enhanced possibilities to make statements on the ossification status of the medial clavicular epiphysis using an amplified staging scheme in evaluating thin-slice CT scans. Int J Legal Med 2010 Jul;124(4):321-325. [doi: 10.1007/s00414-010-0448-2] [Medline: 20354711]
15. O'Connor JE, Bogue C, Spence LD, Last J. A method to establish the relationship between chronological age and stage of union from radiographic assessment of epiphyseal fusion at the knee: an Irish population study. J Anat 2008 Feb;212(2):198-209 [FREE Full text] [doi: 10.1111/j.1469-7580.2007.00847.x] [Medline: 18179475]
16. Dedouit F, Auriol J, Rousseau H, Rougé D, Crubézy E, Telmon N. Age assessment by magnetic resonance imaging of the knee: a preliminary study. Forensic Sci Int 2012 Apr 10;217(1-3):232.e1-232.e7. [doi: 10.1016/j.forsciint.2011.11.013] [Medline: 22153621]
17. Krämer JA, Schmidt S, Jürgens KU, Lentschig M, Schmeling A, Vieth V. Forensic age estimation in living individuals using 3.0 T MRI of the distal femur. Int J Legal Med 2014 May;128(3):509-514. [doi: 10.1007/s00414-014-0967-3] [Medline: 24504560]
18. Ekizoglu O, Hocaoglu E, Can IO, Inci E, Aksoy S, Bilgili MG. Magnetic resonance imaging of distal tibia and calcaneus for forensic age estimation in living individuals. Int J Legal Med 2015 Jul;129(4):825-831. [doi: 10.1007/s00414-015-1187-1] [Medline: 25904076]
19. Ekizoglu O, Inci E, Ors S, Kacmaz IE, Basa CD, Can IO, et al. Applicability of T1-weighted MRI in the assessment of forensic age based on the epiphyseal closure of the humeral head. Int J Legal Med 2019 Jan;133(1):241-248. [doi: 10.1007/s00414-018-1868-7] [Medline: 29804276]
20. Dallora AL, Anderberg P, Kvist O, Mendes E, Diaz Ruiz S, Sanmartin Berglund J. Bone age assessment with various machine learning techniques: a systematic literature review and meta-analysis. PLoS One 2019;14(7):e0220242 [FREE Full text] [doi: 10.1371/journal.pone.0220242] [Medline: 31344143]
21. Crema MD, Roemer FW, Marra MD, Burstein D, Gold GE, Eckstein F, et al. Articular cartilage in the knee: current MR imaging techniques and applications in clinical practice and research. Radiographics 2011;31(1):37-61 [FREE Full text] [doi: 10.1148/rg.311105084] [Medline: 21257932]
22. Gründer W. MRI assessment of cartilage ultrastructure. NMR Biomed 2006 Nov;19(7):855-876. [doi: 10.1002/nbm.1092] [Medline: 17075962]
23. de Simone M, Farello G, Palumbo M, Gentile T, Ciuffreda M, Olioso P, et al. Growth charts, growth velocity and bone development in childhood obesity. Int J Obes Relat Metab Disord 1995 Dec;19(12):851-857. [Medline: 8963351]
24. Cutler GB. The role of estrogen in bone growth and maturation during childhood and adolescence. J Steroid Biochem Mol Biol 1997 Apr;61(3-6):141-144. [Medline: 9365183]
25. Mirtz T, Chandler JP, Eyers CM. The effects of physical activity on the epiphyseal growth plates: a review of the literature on normal physiology and clinical implications. J Clin Med Res 2011 Feb 12;3(1):1-7 [FREE Full text] [doi: 10.4021/jocmr477w] [Medline: 22043265]

26. Ontell FK, Ivanovic M, Ablin DS, Barlow TW. Bone age in children of diverse ethnicity. AJR Am J Roentgenol 1996 Dec;167(6):1395-1398. [doi: 10.2214/ajr.167.6.8956565] [Medline: 8956565]
27. Karlberg J. Secular trends in pubertal development. Horm Res 2002;57(Suppl 2):19-30. [doi: 10.1159/000058096] [Medline: 12065922]
28. Tanner J, Whitehouse R, Cameron N. Assessment of skeletal maturity and prediction of adult height (TW3 Method). 3rd edn. WB Saunders: London; 2001.
29. Shen J, Zhang CJ, Jiang B, Chen J, Song J, Liu Z, et al. Artificial intelligence versus clinicians in disease diagnosis: systematic review. JMIR Med Inform 2019 Aug 16;7(3):e10010 [FREE Full text] [doi: 10.2196/10010] [Medline: 31420959]
30. Triantafyllidis AK, Tsanas A. Applications of machine learning in real-life digital health interventions: review of the literature. J Med Internet Res 2019 Apr 5;21(4):e12286 [FREE Full text] [doi: 10.2196/12286] [Medline: 30950797]
31. Mitchell TM. Machine learning and data mining. Commun ACM 1999 Nov;42(11):30-36 [FREE Full text] [doi: 10.1145/319382.319388]
32. Thodberg H, Kreiborg S, Juul A, Pedersen K. The BoneXpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging 2009 Jan;28(1):52-66. [doi: 10.1109/TMI.2008.926067] [Medline: 19116188]
33. Marshall WA, Tanner JM. Variations in pattern of pubertal changes in girls. Arch Dis Child 1969 Jun;44(235):291-303 [FREE Full text] [doi: 10.1136/adc.44.235.291] [Medline: 5785179]
34. Marshall WA, Tanner JM. Variations in the pattern of pubertal changes in boys. Arch Dis Child 1970 Feb;45(239):13-23 [FREE Full text] [doi: 10.1136/adc.45.239.13] [Medline: 5440182]
35. Keys A, Fidanza F, Karvonen MJ, Kimura N, Taylor HL. Indices of relative weight and obesity. Int J Epidemiol 2014 Jun;43(3):655-665. [doi: 10.1093/ije/dyu058] [Medline: 24691951]
36. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 2016 Jul 2;20(1):37-46. [doi: 10.1177/001316446002000104]
37. Bobicev V, Sokolova M. Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective. RANLP 2017;97:- [FREE Full text] [doi: 10.26615/978-954-452-049-6_015]
38. Zhang S. Nearest neighbor selection for iteratively kNN imputation. J Syst Softw 2012 Nov;85(11):2541-2552 [FREE Full text] [doi: 10.1016/j.jss.2012.05.073]
39. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak 2016 Jul 25;16(Suppl 3):74 [FREE Full text] [doi: 10.1186/s12911-016-0318-z] [Medline: 27454392]
40. Gower JC. A general coefficient of similarity and some of its properties. Biometrics 1971 Dec;27(4):857. [doi: 10.2307/2528823]
41. Kuhn M, Johnson K. Applied Predictive Modeling. Springer, New York, NY 2013:-. [doi: 10.1007/978-1-4614-6849-3]
42. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform 2014 Mar 29;6(1):10 [FREE Full text] [doi: 10.1186/1758-2946-6-10] [Medline: 24678909]
43. Wainer J, Cawley G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. arXiv cs. LG 2018;Available:- [FREE Full text]
44. Gaudette L, Japkowicz N. Evaluation Methods for Ordinal Classification. Advances in Artificial Intelligence. Springer Berlin Heidelberg. pp. 207? 2009:210. [doi: 10.1007/978-3-642-01818-3_25]
45. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag 2009 Jul;45(4):427-437. [doi: 10.1016/j.ipm.2009.03.002]
46. Hand D, Till R. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn 2001;45:186. [doi: 10.1023/A:1010920819831]
47. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
48. Margalit A, Cottrill E, Nhan D, Yu L, Tang X, Fritz J, et al. The spatial order of physeal maturation in the normal human knee using magnetic resonance imaging. J Pediatr Orthop 2019 Apr;39(4):e318-e322. [doi: 10.1097/BPO.0000000000001298] [Medline: 30451813]
49. O'Connor JE, Coyle J, Bogue C, Spence LD, Last J. Age prediction formulae from radiographic assessment of skeletal maturation at the knee in an Irish population. Forensic Sci Int 2014 Jan;234:188.e1-188.e8. [doi: 10.1016/j.forsciint.2013.10.032] [Medline: 24262807]
50. Cox L. The biology of bone maturation and ageing. Acta Paediatr Suppl 1997 Nov;423:107-108. [doi: 10.1111/j.1651-2227.1997.tb18386.x] [Medline: 9401555]
51. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, et al. The RSNA pediatric bone age machine learning challenge. Radiology 2019 Feb;290(2):498-503 [FREE Full text] [doi: 10.1148/radiol.2018180736] [Medline: 30480490]
52. Dallora AL, Berglund JS, Brogren M, Kvist O, Diaz Ruiz S, Dübbel A, et al. Age assessment of youth and young adults using magnetic resonance imaging of the knee: a deep learning approach. JMIR Med Inform 2019 Dec 5;7(4):e16291 [FREE Full text] [doi: 10.2196/16291] [Medline: 31804183]

53. Štern D, Payer C, Urschler M. Automated age estimation from MRI volumes of the hand. Med Image Anal 2019 Dec;58:101538 [FREE Full text] [doi: 10.1016/j.media.2019.101538] [Medline: 31400620]
54. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. IEEE Access 2018;6:9375-9389 [FREE Full text] [doi: 10.1109/ACCESS.2017.2788044]
55. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. IEEE J Biomed Health Inform 2017 Jan;21(1):4-21. [doi: 10.1109/JBHI.2016.2636665] [Medline: 28055930]
56. Stern D, Payer C, Giuliani N, Urschler M. Automatic age estimation and majority age classification from multi-factorial MRI data. IEEE J Biomed Health Inform 2019 Jul;23(4):1392-1403. [doi: 10.1109/JBHI.2018.2869606] [Medline: 31059459]
57. Tang FH, Chan JL, Chan BK. Accurate age determination for adolescents using magnetic resonance imaging of the hand and wrist with an artificial neural network-based approach. J Digit Imaging 2019 Apr;32(2):283-289 [FREE Full text] [doi: 10.1007/s10278-018-0135-2] [Medline: 30324428]
58. Hillewig E, Degroote J, van der Paelt T, Visscher A, Vandemaele P, Lutin B, et al. Magnetic resonance imaging of the sternal extremity of the clavicle in forensic age estimation: towards more sound age estimates. Int J Legal Med 2013 May;127(3):677-689. [doi: 10.1007/s00414-012-0798-z] [Medline: 23224029]

Abbreviations

AUC: area under the curve
BA: bone age
BAA: bone age assessment
CA: chronological age
GP: Greulich-Pyle
KNN: K-nearest neighbors
MAE: mean absolute error
MLP: multilayer perceptron
MRI: magnetic resonance imaging
RMSE: root mean squared error
ROI: region of interest
RSNA: Radiological Society of North America
SLR: systematic literature review
SVM: support vector machines
TW: Tanner-Whitehouse

Edited by G Eysenbach; submitted 23.03.20; peer-reviewed by M Sokolova, D Carvalho, P Kocbek, B Brumen; comments to author 12.06.20; revised version received 06.08.20; accepted 13.08.20; published 21.09.20.

Please cite as:
Dallora AL, Kvist O, Berglund JS, Ruiz SD, Boldt M, Flodmark CE, Anderberg P
Chronological Age Assessment in Young Individuals Using Bone Age Assessment Staging and Nonradiological Aspects: Machine Learning Multifactorial Approach
JMIR Med Inform 2020;8(9):e18846
URL: http://medinform.jmir.org/2020/9/e18846/
doi:10.2196/18846
PMID:32955457

Corrigenda and Addenda

Correction: Good News and Bad News About Incentives to Violate the Health Insurance Portability and Accountability Act (HIPAA): Scenario-Based Questionnaire Study

Joana Gaia[1], PhD; Xunyi Wang[2], PhD; Chul Woo Yoo[3], PhD; G Lawrence Sanders[1], PhD

[1]State University of New York at Buffalo, Buffalo, NY, United States

[2]Hankamer School of Business, Baylor University, Waco, TX, United States

[3]Florida Atlantic University, Boca Raton, FL, United States

Corresponding Author:
G Lawrence Sanders, PhD
State University of New York at Buffalo
325G Jacobs
Buffalo, NY,
United States
Phone: 1 7166452373
Email: mgtsand@buffalo.edu

Related Article:

Correction of: https://medinform.jmir.org/2020/7/e15880/

(JMIR Med Inform 2020;8(9):e24243) doi:10.2196/24243

In “Good News and Bad News About Incentives to Violate the Health Insurance Portability and Accountability Act (HIPAA): Scenario-Based Questionnaire Study” (JMIR Med Inform 2020;8(7):e15880) the authors noted two errors.

In the original article, the incorrect state was listed for the affiliation of author Chul Woo Yoo. The original affiliation was:

> *Florida Atlantic University, Boca Raton, NY, United States*

The corrected affiliation is:

> *Florida Atlantic University, Boca Raton, FL, United States*

As well, the scenario descriptions in Textbox 1 were not complete. The original descriptions for the scenarios were:

> *Scenario 1: Nurse’s aide, no personal context*
>
> *Suppose you are a nurse’s aide at a hospital, and you earn US $30,000 per year. A friend asks you to get them some information on a patient you have been caring for. What amount of money would you receive to make this acceptable?*
>
> *Scenario 2: Doctor, no personal context*
>
> *Suppose you are a doctor at a hospital, and you earn US $200,000 per year. A very close friend asks you to access patient information to help them in an upcoming legal battle. What amount of money would you receive to make this acceptable?*

The complete descriptions for the scenarios are:

> ***Scenario 1: Nurse’s aide, no personal context***
>
> *Suppose you are a nurse’s aide at a hospital and you earn US $30,000 per year. A friend asks you to get them some information on a patient you have been caring for. What amount of money would you receive to make this acceptable?*
>
> ***Scenario 2: Doctor, no personal context***
>
> *Suppose you are a doctor at a hospital and you earn US $200,000 per year. A very close friend asks you to access patient information to help them in an upcoming legal battle. What amount of money would you receive to make this acceptable?*
>
> ***Scenario 3: Insurance local celebrity, no personal context***
>
> *Suppose you work for an insurance company and make US $60,000 per year. A relative asks you to get insurance data on a famous local celebrity from the organization you work for. What amount of money would you receive to make this acceptable?*
>
> ***Scenario 4: Your mother needs an experimental treatment, personal context***
>
> *Your mother has just been diagnosed with a rare condition that causes kidney failure and is fatal if untreated. This condition can be treated, but the treatment is still considered experimental and is therefore not covered by health insurance, nor is it eligible for any type of financial assistance. The*

treatment is available both nationally and internationally and costs US $100,000. A media outlet approaches you to get information about a famous politician and offers to pay you US $100,000 for that information. This money can save your mother's life. Would you accept the payment from the media outlet and give the money to your mother?

Scenario 5: Best friend needs air medical transportation, personal context

Your best friend has been in an all-terrain vehicle accident in a rural area of Kansas. He or she has life-threatening injuries and needs air medical transportation to receive lifesaving medical care. The medical air evacuation is not covered by insurance and costs US $50,000. Your best friend will not survive ground transportation or local medical care. A media outlet offers you US $50,000 to obtain the health care records of a famous reality television star. This money can save your best friend's life. Would you accept the payment from the news outlet to give the money to your best friend?

Each scenario also included the following question:

What do you think is the likelihood of getting caught if you accept the money?

Extremely unlikely (0%)

Moderately unlikely (7%)

Slightly unlikely (25%)

Neither likely nor unlikely (50%)

Slightly likely (75%)

Moderately likely (93%)

Extremely likely (100%)

The correction will appear in the online version of the paper on the JMIR Publications website on September 15, 2020, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 10.09.20; this is a non–peer-reviewed article;accepted 10.09.20; published 15.09.20.

Please cite as:
Gaia J, Wang X, Yoo CW, Sanders GL
Correction: Good News and Bad News About Incentives to Violate the Health Insurance Portability and Accountability Act (HIPAA): Scenario-Based Questionnaire Study
JMIR Med Inform 2020;8(9):e24243
URL: http://medinform.jmir.org/2020/9/e24243/
doi:10.2196/24243
PMID:32931445

Original Paper

A Smartphone App to Manage Cirrhotic Ascites Among Outpatients: Feasibility Study

Patricia Bloom[1], MD; Thomas Wang[2], MD; Madeline Marx[1], BA; Michelle Tagerman[2], MPH, MSc; Bradley Green[1], BA; Ashwini Arvind[1], MBBS; Jasmine Ha[1], BA; Judith Bloom[1], NP; James M Richter[1], MA, MD

[1]Department of Gastroenterology, Massachusetts General Hospital, Boston, MA, United States

[2]Department of Medicine, Massachusetts General Hospital, Boston, MA, United States

Corresponding Author:
Patricia Bloom, MD
Department of Gastroenterology
Massachusetts General Hospital
55 Fruit St
Boston, MA, 02114
United States
Phone: 1 844 233 0433
Email: ppbloom@med.umich.edu

Abstract

Background: Ascites is a common, painful, and serious complication of cirrhosis. Body weight is a reliable proxy for ascites volume; therefore, daily weight monitoring is recommended to optimize ascites management.

Objective: This study aims to evaluate the feasibility of a smartphone app in facilitating outpatient ascites management.

Methods: In this feasibility study, patients with cirrhotic ascites requiring active management were identified in both inpatient and outpatient settings. Patients were provided with a Bluetooth-connected scale, which transmitted weight data to a smartphone app and then via the internet to an electronic medical record (EMR). Weights were monitored every weekday. In the event of a weight change of ≥5 lbs in 1 week, patients were called and administered a short symptom questionnaire, and providers received an email alert. The primary outcomes of this study were the percentage of enrolled days during which weight data were successfully transmitted to an EMR and the percentage of weight alerts that prompted responses by the provider.

Results: In this study, 25 patients were enrolled: 12 (48%) were male, and the mean age was 58 (SD 13; range 35-81) years. A total of 18 (72%) inpatients were enrolled. Weight data were successfully transmitted to an EMR during 71.2% (697/979) of the study enrollment days, with technology issues reported on 16.5% (162/979) of the days. Of a total of 79 weight change alerts fired, 41 (52%) were triggered by weight loss and 38 (48%) were by weight gain. Providers responded in some fashion to 66 (84%) of the weight alerts and intervened in response to 45 (57%) of the alerts, for example, by contacting the patient, scheduling clinic or paracentesis appointments, modifying the diuretic dose, or requesting a laboratory workup. Providers responded equally to weight increase and decrease alerts (*P*=.87). The staff called patients a mean of 3.7 (SD 3.5) times per patient, and the number of phone calls correlated with technology issues (*r*=0.60; *P*=.002). A total of 60% (15/25) of the patients chose to extend their participation beyond 30 days. A total of 17 patient readmissions occurred during the study period, with only 4 (24%) related to ascites.

Conclusions: We demonstrated the feasibility of a smartphone app to facilitate the management of ascites and reported excellent rates of patient and provider engagement. This innovation could enable early therapeutic intervention, thereby decreasing the burden of morbidity and mortality among patients with cirrhosis.

(JMIR Med Inform 2020;8(9):e17770) doi:10.2196/17770

KEYWORDS

health care innovation; ascites; telemedicine; health care delivery; technology; mobile phone

Introduction

Ascites represents a major burden for patients with cirrhosis. Cirrhotic ascites is associated with poor health-related quality of life [1], hospital admissions [2-4], high cost of care [4-6], and increased mortality [3,7]. For decades, body weight has been identified as a useful proxy for ascites volume, but accurate weight monitoring at home has been difficult. In fact, monitoring weight is central to expert guidelines for ascites management [8,9] and treatment trials [10-12]. Weight changes signal a change in ascites volume and may provoke laboratory testing for renal injury, modification of diuretic dose, and large-volume paracentesis (LVP). Failure to recognize early signs of increasing ascites or overdiuresis has long been recognized as a preventable cause of ascites-related readmissions [6]. Timely transmission of accurate weight data from patients to their hepatology providers may allow for early intervention and prevent readmissions.

Technology represents a promising tool to facilitate the management of ascites by increasing the quality and quantity of patient-provider communication about weight data. In a recent interview study of patients with an early readmission for decompensated cirrhosis, the majority stated that they would use a smartphone to manage their condition, particularly if it was able to transmit weight data to their provider [13]. Retrospective and survey studies suggest that programs with enhanced outpatient care can improve outcomes for patients with ascites [14,15].

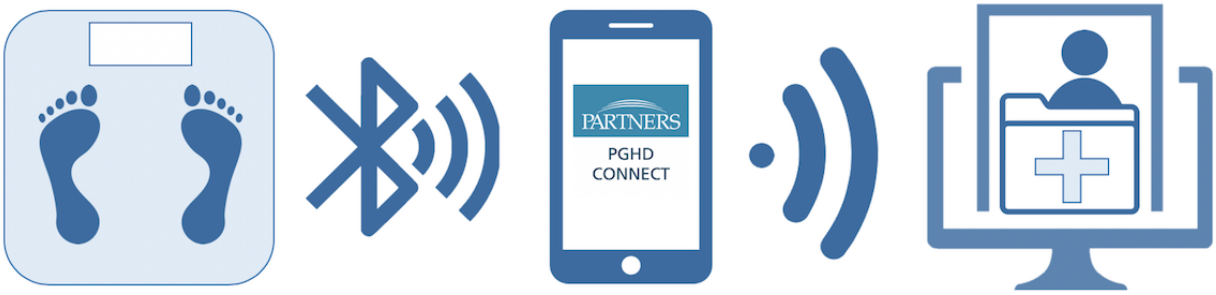
We created a simple telemonitoring program, in which patient weight data are communicated daily to an electronic medical record (EMR) via a Bluetooth-connected scale and a smartphone app, and alerts for significant weight changes are emailed to the hepatology provider. In this study, we assessed the feasibility of the telemonitoring program for ascites management. Specifically, we evaluated whether patients would regularly weigh themselves and whether providers would respond to weight alerts.

Methods

Study Design

We conducted a feasibility study of an outpatient weight monitoring program for patients requiring active management of their cirrhotic ascites. Patients were consented, instructed in the use of the app, and provided with a Bluetooth-connected scale to use at home (Figure 1), which transmitted weight data to a smartphone app and then to the EMR. Weights were monitored every weekday, and significant weight changes prompted an email alert to hepatology providers. At the end of enrollment, patients, caregivers, and hepatology providers were interviewed for feedback. Written informed consent was obtained from patients, and a study fact sheet was given to caregivers and hepatology providers, who provided verbal consent to participate. This study was approved by the Partners HealthCare Institutional Review Board.

Figure 1. Flow of weight information. Weight data are collected from the Bluetooth-connected scale, transmitted via a Bluetooth connection to the PGHDConnect app, and then via the internet to the electronic medical record.



The Technology

Eligible patients were assisted in downloading and registering for a no-cost smartphone app and were given a no-cost Bluetooth-connected scale (A&D UC-352BLE digital scale). Digital scale weights were transmitted via a Bluetooth connection to the Partners Patient-Generated Health Data Connect (PGHD*Connect*) app and then transmitted securely via the internet to the Partners *e*Care EMR (Epic). The PGHD*Connect* app is currently used for clinical care at Partners HealthCare and has been deemed Health Insurance Portability and Accountability Act (HIPAA)–compliant and secure by the Partners HealthCare information services team.

Study Population

We enrolled patients receiving active management of their cirrhotic ascites, as this population may benefit most from an outpatient weight monitoring system. First, we performed daily screening of the inpatient hepatology consult census to identify patients with a clinical diagnosis of cirrhosis and requiring active management of ascites during their admission, including therapeutic paracentesis, diuretic hold or titration, or treatment of renal or electrolyte dysfunction resulting from ascites management. After several months, when hepatology providers were increasingly aware of the study, we began enrolling both inpatient and outpatient referrals from hepatology providers and stopped active screening of the inpatient census.

Patients were approached for consent and enrollment if deemed appropriate by their primary outpatient hepatology provider. Of importance, patients were required to own a smartphone and be able to stand for daily weighing to enroll, as these are essential requirements for the program to function. The diagnosis of cirrhosis and ascites was confirmed by the hepatology provider. Patients with poorly controlled hepatic encephalopathy and severe ongoing cognitive dysfunction were excluded. Other inclusion criteria included aged 18 years or older, English speaking, and capacity to provide informed consent.

Upon enrollment, patients were asked if they had a caregiver who would likely assist them in using the app and scale. A study fact sheet was provided to the patient, caregiver, and hepatology providers.

Study Procedures

Once qualified patients provided informed consent, the study staff assisted them in downloading the PGHD*Connect* app and pairing the app with their scale. Nonphysician study staff (MM, MT, BD, and AA) monitored the patients' weights daily in the EMR. In the event of a weight change of >5 lbs in 1 week, these study staff called and administered a short symptom questionnaire. The study staff then emailed providers a summary of the weight change event and answers to the short symptom questionnaire. Study staff also called patients if no weight data appeared in the EMR and troubleshot technology issues. Finally, the study staff tracked providers' responses to weight change alerts. Weights were monitored every weekday for 4 weeks, though patients were allowed to enroll for shorter or longer durations. The program was paused during hospitalizations and resumed upon discharge.

Patients were enrolled between January and October 2019. At the end of the enrollment, a semistructured interview with patients, caregivers, and hepatology providers was conducted to obtain qualitative data through open-ended questions. Study staff (MM, MT, BG, and AA) audio-recorded an exit interview with the patients and, if available, their caregivers. A physician investigator (PB) performed all exit interviews with hepatology providers. Chart review was performed on all enrolled patients to obtain the following data: demographic factors, etiology and severity of liver disease, and diagnoses at index admission and readmission.

Statistical Analysis

The primary outcomes of this study were the percentage of enrolled days during which weight data were successfully transmitted to the EMR and the percentage of weight alerts that prompted a response by the provider. Our team predetermined that receiving weight data in the EMR on 50% of the days enrolled would signal feasibility, as weight data are not required every single day for optimal ascites management. In addition, we determined that providers responding to at least 50% of weight alerts would constitute an adequate threshold for feasibility.

Descriptive data were summarized as mean (SD; continuous) or presented as proportions (categorical). Missing data were accounted for by adjusting the denominator. Significance testing was performed using the Student *t* test for continuous variables and Fisher exact test for categorical variables.

Exit Interview

The purpose of the exit interviews was to evaluate facilitators and barriers to the intervention, to explore the root causes of outpatient ascites management failures, and to explore the desired features of a digital ascites management tool. We created a semistructured interview to target these themes for patients, caregivers, and hepatology providers (Multimedia Appendix 1). The interviewees were asked open-ended questions and were asked for further clarification when needed. The themes were generated based on the principle of qualitative research reflexivity: by reflecting on clinical experience, literature review, and preconceptions to reduce bias in interviewing and analysis [16]. The exit interviews were piloted on the authors and edited in an iterative fashion.

Qualitative Data Analysis

Interview transcripts were imported into NVivo 11.0 (QSR International). Two investigators (PB and TW) iteratively read and coded interview transcripts for themes [17]. The principle of grounded theory was applied: as themes emerged from the data, specific lines of text were coded into themes [18]. The analysts then jointly compared codes, resolved discrepancies, and developed a taxonomy of themes. Themes were refined until saturation was reached, with a final taxonomy of 13 themes. This final taxonomy was applied to all transcripts by the 2 analysts, with a kappa agreement of 84%.

Institutional Structure

This study was conducted at a single urban academic liver transplant center. Patients with cirrhosis were admitted to a hospital medicine service with hepatology or gastroenterology consultation. The majority of the patients were locals and lived within an hour's drive from the hospital, although some patients were transferred from other parts of New England.

Results

Patient Characteristics

After screening 151 consecutive adult patients admitted with cirrhotic ascites, 18 patients from the inpatient setting and 7 patients from the outpatient setting were enrolled (Figure 2). The 25 enrolled patients had a mean age of 58 years (SD 13; range 35-81 years), mean model for end-stage liver disease (MELD) score of 15.8 (SD 5.9), and 12 (48%) were men. The etiology of cirrhosis was alcohol-related in 11 (44%), nonalcoholic steatohepatitis in 9 (36%), and viral infection in 3 (12%). Of the 25 patients, 5 (20%) patients were on 1 diuretic at enrollment, 17 (68%) were on 2 diuretics, and 3 (12%) were on no diuretics (Table 1).

Figure 2. Patient screening flowchart.

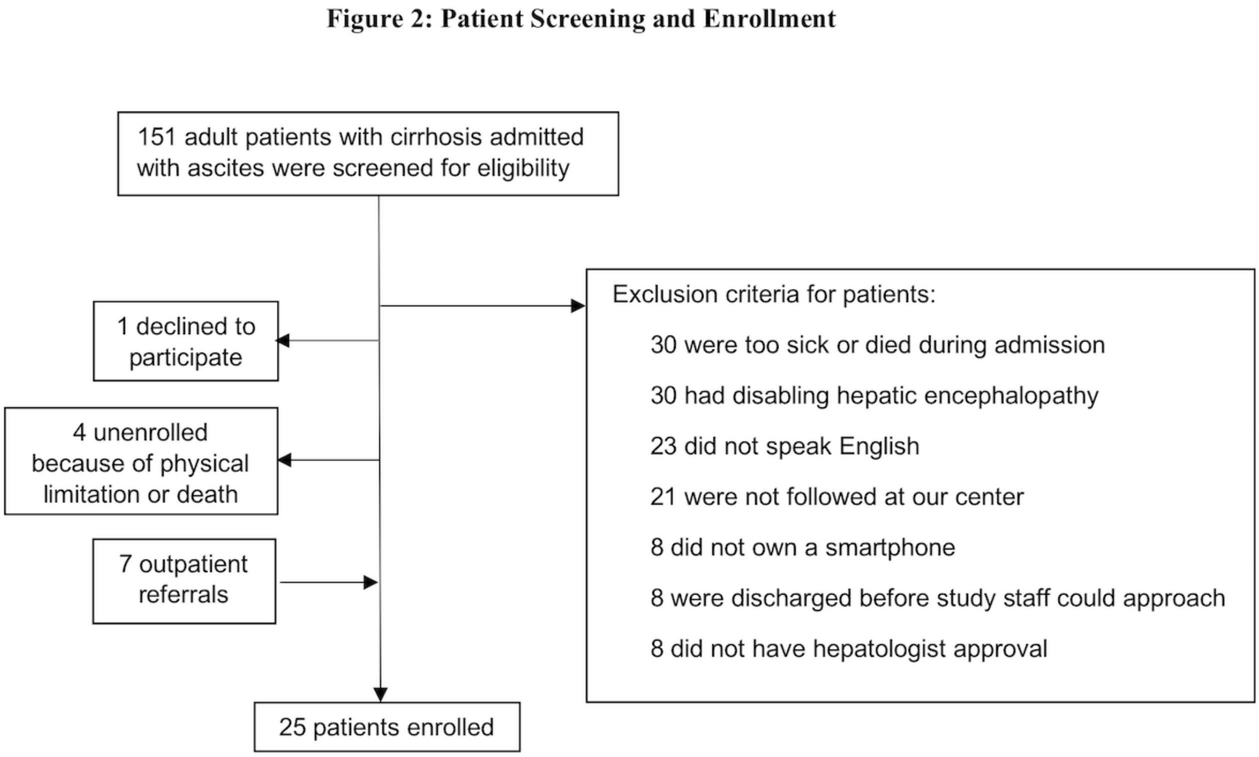


Figure 2: Patient Screening and Enrollment
151 adult patients with cirrhosis admitted with ascites were screened for eligibility
Exclusion criteria for patients:
30 were too sick or died during admission
30 had disabling hepatic encephalopathy
23 did not speak English
21 were not followed at our center
8 did not own a smartphone
8 were discharged before study staff could approach
8 did not have hepatologist approval
1 declined to participate
4 unenrolled because of physical limitation or death
7 outpatient referrals
25 patients enrolled

Table 1. Patient characteristics and enrollment.

Patient characteristics	Total cohort (N=25)	Inpatient enrollees (n=18)	Outpatient enrollees (n=7)	*P* value[a]
Age (years), mean (SD)	57.6 (12.8)	60.1 (9.9)	51.3 (17.6)	.12
Male, n (%)	48	39	71	.20
Etiology of cirrhosis, n (%)				
Alcohol	44	44	43	.91
Nonalcoholic steatohepatitis	36	39	29	.91
Viral	12	11	14	.91
Other	8	6	14	.91
Model for end-stage liver disease, mean (SD)	15.8 (5.9)	16.7 (6.5)	13.3 (2.9)	.20
Diuretics, n (%)				
Furosemide alone	12	11	14	.80
Spironolactone alone	8	11	0	.80
Furosemide and spironolactone	68	61	86	.80
None	12	17	0	.80
Diagnostic paracentesis during admission, n (%)	N/A[b]	61	N/A	N/A
Therapeutic paracentesis during admission, n (%)	N/A	50	N/A	N/A

[a] Comparing inpatient and outpatient subgroups.

[b] N/A: not applicable.

Weight Data Transmission

Patients were enrolled in the ascites monitoring program for 979 total patient-days (Table 2). Weight data were successfully transmitted into the EMR on 697 (71.1%) days. On 162 days (16.5% of the enrollment days), weight data did not appear in the EMR, and patients reported a likely issue with technology. The remote nature of this intervention limited the ability of study staff to precisely ascertain the cause of each technology issue, but the most common culprits included the scale not pairing to the phone, weight data not transmitting from the phone into the EMR, and spontaneous disconnections of

Bluetooth or the internet. Updates were made to the app to improve data transmission from older smartphone operating systems, which reduced technology issues over the course of the study.

Patients were more likely to weigh themselves in the morning, with 593 weights transmitted before noon, as opposed to 104 transmitted after noon. The percentage of days with weight data in the morning was not associated with phone calls (*P*=.33), technology issues (*P*=.46), alerts fired (*P*=.58), or admissions (*P*=.96).

The location of enrollment, inpatient or outpatient, did not lead to differences in the number of days with weight data transmitted (*P*=.29) or the number of calls required (*P*=.51).

Table 2. Smartphone app outcomes by the patient.

Patient	Days enrolled	Days with weight in EMR[a]	Calls for no weight in EMR	Days of technology issues	Weights before/after noon	Alerts fired	Weight increase alerts	Weight decrease alerts	Provider responded to alert	Provider intervened after alert	Admissions while enrolled
1	30	10	13	21	6/4	1	0	1	1	0	1
2	25	25	0	0	13/12	4	0	4	3	2	1
3	28	27	1	2	25/2	2	2	0	2	1	0
4	34	25	7	0	24/1	0	0	0	0	0	1
5	32	28	1	1	27/1	1	0	1	0	0	1
6	45	27	4	0	26/1	2	2	0	2	2	6
7	16	9	7	7	3/6	1	1	0	1	1	0
8	12	7	0	0	3/4	2	1	1	1	1	1
9	28	28	0	0	28/0	1	0	1	1	1	0
10	43	29	4	0	28/1	3	2	1	3	2	0
11	44	33	3	0	17/16	0	0	0	0	0	0
12	35	33	1	0	32/1	1	1	0	0	0	2
13	62	25	5	32	25/0	1	0	1	1	1	0
14	51	3	9	45	3/0	0	0	0	0	0	0
15	61	19	7	37	17/2	3	1	2	2	1	2
16	42	42	0	0	42/0	8	5	3	8	3	1
17	28	10	5	3	8/2	2	1	1	2	2	0
18	72	67	1	0	63/4	16	9	7	12	7	0
19	87	80	7	0	53/27	13	8	5	11	11	0
20	27	27	0	0	27/0	6	5	1	5	3	0
21	43	28	7	9	22/6	2	1	1	2	0	0
22	13	3	5	0	2/1	0	0	0	0	0	1
23	58	54	3	2	46/8	3	2	1	3	1	0
24	28	26	0	0	23/3	4	0	4	3	3	0
25	35	32	2	3	30/2	3	0	3	3	3	0
Total, n (%)[b]	979 (100)	697 (71.1)	92 (9.3)	162 (16.5)	593/104 (60.5/10.6)	79 (8.0)	41 (4.1)	38 (3.8)	66 (6.7)	45 (4.5)	17[c]

[a] EMR: electronic medical record.

[b] Percentage of event occurrence per the total number of days enrolled.

[c] Not applicable.

Weight Change Alerts

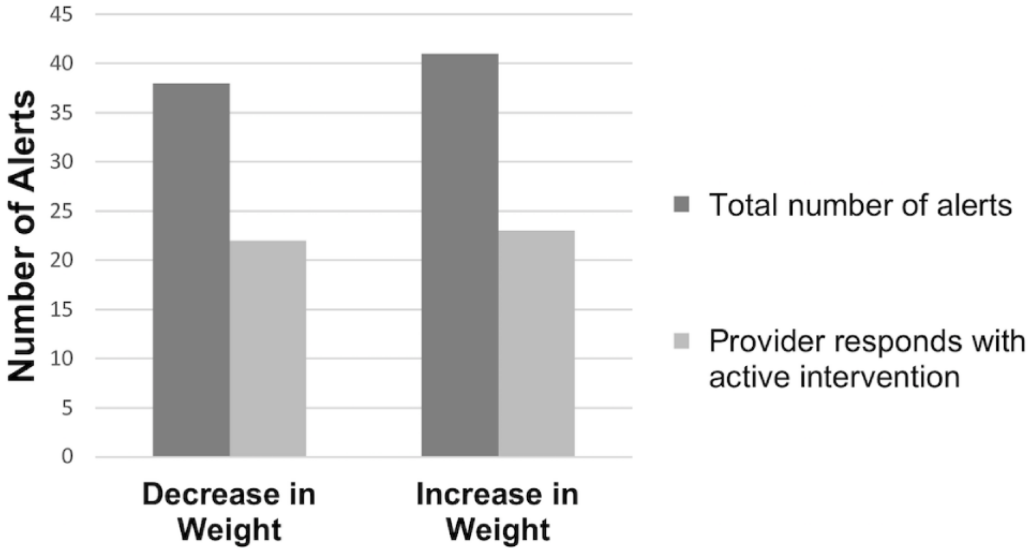
There were 79 weight alerts emailed to hepatology providers during the study period, meaning that a notable weight change occurred on 8.0% (79/979) of the days that patients were enrolled in the program. The weight alerts were evenly divided between alerts for weight increase (41/79 alerts, 52%) and weight decrease (38/79 alerts, 48%). Of the 25 patients, 10 (40%) patients prompted both weight increase and weight decrease alerts during the study period, 7 (28%) patients prompted only weight decrease alerts, 4 (16%) patients prompted only weight increase alerts, and 4 (16%) patients prompted no alerts.

Providers responded in some fashion to 66 of the 79 (84%) alerts and responded with an active intervention to 45 (57%) alerts. Active interventions included communicating with the patient, ordering testing, adjusting diuretics, ordering a paracentesis, or scheduling a follow-up appointment. Provider responses not characterized as *active intervention* included an acknowledgment of the alert and, in many cases, an explanation as to why the weight change was expected. Providers were equally likely to intervene on a weight increase alert as they were to a weight decrease alert (Figure 3; *P*=.87).

For every weight increase alert, patients were asked by the study staff about shortness of breath, early satiety, and a tense abdomen. For every weight decrease alert, patients were asked about dizziness and nausea, vomiting, or diarrhea. For all alerts, patients were queried about diuretic compliance. For 24 of the 79 weight alerts, the patient could not be reached on that day to report symptoms. With weight increase alerts, 14 of the 28 patients (50%) reported shortness of breath, 12 reported (43%) early satiety, and 14 (50%) reported a tense abdomen. With 27 weight decrease alerts, 8 (30%) reported dizziness and 10 (37%) reported nausea, vomiting, or diarrhea. Patients reported a diuretic compliance 84% (46/55) of the time. In the Fisher exact test, reporting shortness of breath (*P*=.16), early satiety (*P*=.09), tense abdomen (*P*=.06), dizziness (*P*=.83), and nausea, vomiting, or diarrhea (*P*=.75) were not associated with the nature of provider response. Compliance with diuretics was also not associated with the nature of the provider response (*P*=.07).

Patients underwent a total of 13 paracenteses during the study period. Of the 38 weight decrease alerts, 10 (26%) occurred within 2 days of a paracentesis. Of the 41 weight increase alerts, 15 (37%) were followed by a referral for paracentesis within 7 days.

Figure 3. Provider response by weight alert type (*P*=.87).



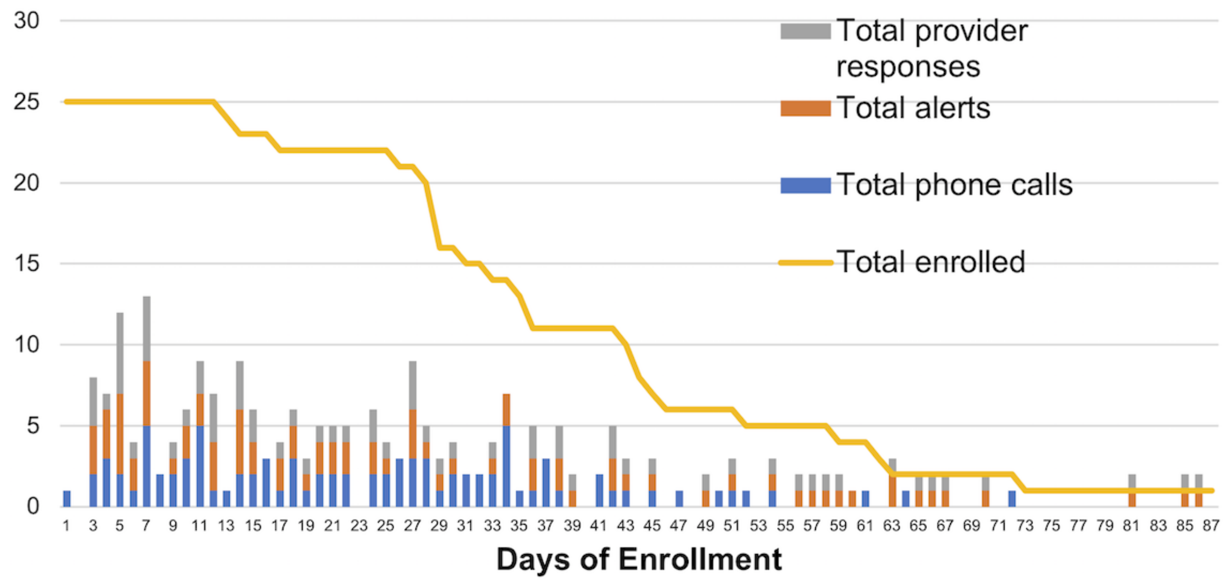
Engagement

A total of 92 phone calls were made by the study staff when weight data did not appear in the EMR, with a mean of 3.7 calls (SD 3.5) per patient and a range of 0 to 13 calls per patient. The number of phone calls correlated with days of technology issues (*r*=0.60; *P*=.002), perhaps because calls were prompted by a lack of weight data in the EMR. The number of phone calls was not correlated with the number of hospital admissions during the study period (*P*=.88), MELD score (*P*=.59), or the number of weight alerts (*P*=.27).

Providers responded to 84% (66/79) of the alerts, with a response rate range of 0% to 100%. There was no trend of providers responding to alerts more frequently at the beginning of enrollment as compared with at the end of enrollment (Figure 4). There were 12 hepatology providers with patients enrolled in this program, including 7 attendings, 3 fellows, and 2 nurse practitioners. Nurse practitioners and fellows work with attendings but were the primary responders to alerts in this program. Using the Fisher exact test, we found a significant difference in responsiveness to alerts by provider type, with fellows performing an active intervention to 75% (24/32) of the alerts, nurse practitioners to 59% (10/17) of the alerts, and attendings to 37% (11/30) of the alerts (*P*=.03).

By default, patients were enrolled for 28 days but could remain in the program longer if both the patient and the hepatology provider desired longer enrollment. Out of the 25 patients, 15 (60%) patients extended their participation for over 30 days, and 6 of those 15 (24%) were enrolled for over 50 days. On the other hand, 3 of the 25 (12%) patients chose to withdraw from the program within 25 days. The reasons for early withdrawal included technology issues, being too ill to stand on the scale, and being nonresponsive to phone calls.

Figure 4. Calls and alert responsiveness during enrollment.



Patient, Caregiver, and Hepatology Provider Feedback

Eighteen patients, 10 hepatology providers, and 4 caregivers provided a detailed exit interview. Patients felt that the program was easy (17/18, 94%), looked at their weights in the smartphone app (15/18, 88%), and preferred the smartphone app to another digital tool such as a computer (94%). Patients and caregivers found benefits of the program to increase connectedness to providers, a better sense of ascites status, ease of the program, and increased adherence to weight tracking at home (Table 3 and Textbox 1). Patients and caregivers who experienced technology issues were frustrated by those problems, yet some still found the overall program beneficial.

Table 3. Patient and caregiver exit interview responses: quantitative results from patient exit interviews (n=18).

Questions[a]	Response, n (%)	Total, N
Facilitators and barriers		
Was the program easy?	Yes, 17 (94)	18
Was it difficult to remember daily weights?	Yes, 2 (11)	18
Did you look at the app during the program?	Yes, 15 (88)	17
Did you have technical difficulties?	Yes, 8 (44)	18
Root causes		
Can you name your diuretics?	Yes, 11 (65)	17
In the last week, have you missed your diuretics?	Yes, 2 (12)	16
Have you had difficulty paying for medications?	Yes, 2 (12)	16
Is weight monitoring important for ascites management?	Yes, 18 (100)	16
Have you consumed alcohol recently?	Yes, 1 (6)	17
Will you stay somewhere else in the next month?	Yes, 2 (12)	16
Desires features of the ultimate digital tool		
Do you prefer another device type (eg, computer) over a smartphone?	Yes, 1 (6)	18
Was the timing of phone calls a problem?	Yes, 0 (0)	18

[a] Abbreviated for ease of interpretation. For the full interview script, see Multimedia Appendix 1.

Textbox 1. Themes and representative quotes from patient and caregiver exit interviews.

Benefit of program: connectedness to providers

- “I just feel better knowing that my doctor is aware of my weights on a daily basis.”
- “It was awesome in that I was in constant communication with my doctor about what was going on in terms of my weight and how to proceed. Do we need more diuretics … do we need a paracentesis?”

Benefit of program: better sense of ascites status

- “If I leave it to memory, I only remember yesterday’s [weight]. The program gives me a whole history, so I can look back 5 days, 7 days, to see if there were any real fluctuations.”
- “It made me be more aware of my sodium intake.”

Ease of program

- “I really liked that it was something I could do every morning, and I could see the ascites was going away.”

Other benefits

- “Where before I might have had 90% compliance, now I have 100% compliance.”
- “I don’t think that she would’ve had all the problems that she’s had, if she would’ve had this scale a long time ago. I mean, it seems like a simple thing, but for someone with this problem, it’s a huge deciding factor. It really is.”

Challenges

- “It was kind of difficult for me because I’m not very savvy on cellphones.”
- “Was a great idea and all that, but it’s very frustrating when you try to set it up and it doesn’t work.”

Root causes of ascites mismanagement

- “They told me that if I have an uncomfortable feeling, a hard stomach or difficulty breathing, [I should] call them to make an appointment. The only problem I had with that is I don’t know which doctor to get in touch with.”
- “He’s the problem… trying to keep him away from salt and especially processed meats.”
- “I’ve been out of work… Sometimes I go without my medications a lot.”

Desired features of future tool

- “Phone app is good for me.”
- “I prefer something of this nature on my laptop.”

Hepatology providers generally found the program easy and helpful (Textbox 2). They enjoyed regular access to accurate weight data, the content of weight alerts, and the dialogue the program created between the provider and the patient. Many providers noted a small increase in work required to respond to weight alerts, but most did not find this burdensome. A few providers described features of the ideal patient to enroll in this program: ascites is symptomatic, the patient is motivated to engage with the medical system and improve health, and the patient is relatively compensated medically outside of ascites (see Multimedia Appendix 2 for expanded quotes).

Textbox 2. Themes and representative quotes from hepatology provider exit interviews.

Positives of the program

- “I like having access to the weight measurements on a regular basis. It’s a lot easier than asking him to weigh himself and transmit it back to me. It definitely changed the clinical management.”
- “It allows you to keep people out of the hospital.”

Challenges of the program

- “It seemed like a number of the folks that I had, there were these weird technical issues with the scale, the Bluetooth, whatever it was.”
- “If we’re doing people who are in the hospital [and] going home, their diet changes dramatically. I don’t know that capturing their weight necessarily accurately reflects their fluid status alone. I think it’s their nutritional status also.”

How the smartphone app helps patients

- “XX is a patient who is completely new to ascites… and was just starting on diuretics. The weight tracking program actually helped her see the progress the diuretics were making… The program actually allowed us to have a dialogue about how she was supposed to lose weight.”

Features of the ideal patient enrollee

- “I think it may work better for just outpatient. And maybe starting off with a cohort that’s less sick… I think that it may be more beneficial in patients who you’re just starting on diuretics and patients who have kind of stable nutritional status, stable—not the truly decompensated cirrhotics.”
- “XX had a particularly unstable weight. His ascites was highly symptomatic. He lived a relatively long distance from the hospital. And not only did it provide us useful information, he personally liked the idea of being engaged and it gave him some sense of control over his own care and body.”

Desired features of future program

- “I think because I had multiple patients involved, the emails, occasionally, it seemed like they were coming frequently but I think that’s just because I would get a separate email per patient. So, I think if this was to be potentially rolled out and people had multiple patients involved, if they could maybe be grouped but I don’t know whether that’s possible.”

Hospital Admissions During the Study Period

The 25 patients enrolled in this program were admitted on 17 occasions during the study period. Of the 17 admissions, 4 (24%) were related to ascites or ascites treatment, and 12% (3/25) of patients had an ascites-related admission while enrolled. A total of 4 (24%) admissions were for gastrointestinal bleeding or anemia, 3 (18%) for infection, 2 (12%) for hepatic encephalopathy, and 4 (24%) for reasons unrelated to liver disease. Of the 17 admissions, 5 (29%) occurred within 7 days after a weight alert, 15 (88%) were among inpatient enrollees, and 2 (12%) were among outpatient enrollees.

Discussion

Principal Findings

A simple telemonitoring system for patients with cirrhotic ascites was able to transmit weight data into the EMR on >50% of the days, and hepatology providers responded to >50% of the weight alerts, thus meeting our predetermined threshold for feasibility. Our system of a Bluetooth-connected scale and a smartphone app transmitted weight data into the EMR on 71.2% (697/979) of the days that patients were enrolled. Providers responded in some fashion to 84% (66/79) of weight change alerts and responded actively with an intervention to 57% (45/79) of alerts.

Weight change alerts appeared to correlate with ascites and influenced ascites management. Approximately one-third of the weight increase alerts were followed by an LVP. Providers responded equally to weight increase and weight decrease alerts and did not appear to respond less frequently over the course of patient enrollment. It is notable that symptoms reported with the alert did not significantly influence the nature of the provider’s response. A larger prospective study will be needed to further evaluate their utility. Only 12% (3/25) of patients were readmitted for ascites while enrolled in this program. More experience with the system or refinement of the alert criteria may improve effectiveness. Although there was no comparison group in this study, other cohorts have found 13.8% of such patients readmitted within 30 days and 25% within 90 days [2,19].

Patients and providers remained engaged throughout the program. Patients continued to transmit weight data, even at the end of their enrollment. In fact, 60% (15/25) of the patients extended their participation beyond 30 days and 24% (6/25) beyond 50 days. A few patients terminated the program early, but mainly because they became too ill to participate. Similarly, providers continued to respond to alerts throughout the enrollment period. Weight change alerts were fired on only 8% of the days that patients were enrolled; this low alert frequency likely contributed to the high rate of provider responsiveness. Finally, we were able to stop proactively screening the inpatient census, because of the increasing provider referrals during the study.

The main perceived benefits by patients and caregivers were increased connectedness to providers and a better sense of ascites status. Hepatology providers likewise enjoyed the easy dialogue with patients facilitated by the program.

Limitations

Technology issues impaired weight data transmission, provoked phone calls, and impacted patient experience. Patients reported some form of technology issue on 16.5% (162/979) of the days enrolled. At times, this involved an unpaired scale and a smartphone app, connectivity problems, or the need for a software update. Some of these issues were resolved with software updates, and technology issues decreased over the course of the study. Patients and caregivers who experienced technology issues expressed frustrations in their exit interviews, yet most still found the overall program beneficial.

Hepatology provider interviews revealed that certain patients may benefit from this program more than others. They described the ideal enrollee as having symptomatic ascites, motivated to engage with the medical system and improve health, and be relatively compensated outside of their ascites. Although patients with other active medical issues, such as gastrointestinal bleeding or malnutrition, may be at higher risk for poor outcomes, their other medical issues may influence their weight and therefore the efficacy of this program.

We suspect that there are several reasons why this telemonitoring program was feasible at our center. First, we had access to a smartphone app that was able to securely transmit weight data from the patient's home into our EMR. Second, the app allowed both patients and providers to be immediately aware of accurate weight trends. Third, the program was at no cost to the patients. Fourth, hepatology providers needed to exert little effort to enroll their patients, and the program generated high-yield information (weight alerts) for their attention, on only 8.1% (79/979) of the days that patients were enrolled. Finally, the telemonitoring program directly addressed a core challenge of ascites management: lack of accurate, timely weight data reaching hepatology providers.

Telemonitoring programs are not a stand-alone solution for ascites management. The program required study staff to monitor weights, make phone calls, formulate alerts, facilitate easy enrollment, and ask field questions. We suspect that the program would not have been feasible without this larger infrastructure surrounding the app.

Rigorous evidence supporting the efficacy of mobile health interventions in chronic disease is lacking [20]. Future studies on telemonitoring interventions should be based on lessons learned in feasibility studies such as this one, assessing efficacy using validated methods, and assessing the cost-effectiveness of performing such interventions on a larger scale [21].

Conclusions

A smartphone-based telemonitoring system was feasible for the management of cirrhotic ascites. Future studies are required to assess the efficacy of such a program in reducing hospital admissions and improving patient and provider experience.

Acknowledgments

The authors wish to acknowledge the training and technological support provided by the PGHD*Connect* team, Karen Chase and Stephen Michael Chaisson. This project was supported by a Partners Health Care Center of Expertise in Quality and Safety grant and by a Massachusetts General Hospital Department of Medicine Innovation Program Spark grant, awarded to PB.

Authors' Contributions

PB was responsible for planning, conducting, analyzing, and writing the manuscript. PB is also the corresponding author of the study. MM, TW, MT, AA, and BG were responsible for conducting and preparing the study report. TW is also responsible for performing the study analysis. JH, JB, and RC were responsible for preparing the study report. JR was responsible for planning, conducting, and preparing the study report.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Exit interviews.
[DOCX File , 16 KB - medinform_v8i9e17770_app1.docx]

Multimedia Appendix 2
Expanded quotes.
[DOCX File , 16 KB - medinform_v8i9e17770_app2.docx]

References

1. Macdonald S, Jepsen P, Alrubaiy L, Watson H, Vilstrup H, Jalan R. Quality of life measures predict mortality in patients with cirrhosis and severe ascites. Aliment Pharmacol Ther 2019 Feb;49(3):321-330 [FREE Full text] [doi: 10.1111/apt.15084] [Medline: 30585338]

2. Tapper EB, Halbert B, Mellinger J. Rates of and reasons for hospital readmissions in patients with cirrhosis: a multistate population-based cohort study. Clin Gastroenterol Hepatol 2016 Aug;14(8):1181-8.e2. [doi: 10.1016/j.cgh.2016.04.009] [Medline: 27085758]
3. Scaglione SJ, Metcalfe L, Kliethermes S, Vasilyev I, Tsang R, Caines A, et al. Early hospital readmissions and mortality in patients with decompensated cirrhosis enrolled in a large national health insurance administrative database. J Clin Gastroenterol 2017 Oct;51(9):839-844. [doi: 10.1097/MCG.0000000000000826] [Medline: 28383303]
4. Shaheen A, Nguyen HH, Congly SE, Kaplan GG, Swain MG. Nationwide estimates and risk factors of hospital readmission in patients with cirrhosis in the United States. Liver Int 2019 May;39(5):878-884. [doi: 10.1111/liv.14054] [Medline: 30688401]
5. di Pascoli M, Ceranto E, de Nardi P, Donato D, Gatta A, Angeli P, et al. Hospitalizations due to cirrhosis: clinical aspects in a large cohort of Italian patients and cost analysis report. Dig Dis 2017;35(5):433-438. [doi: 10.1159/000458722] [Medline: 28245467]
6. Volk ML, Tocco RS, Bazick J, Rakoski MO, Lok AS. Hospital readmissions among patients with decompensated cirrhosis. Am J Gastroenterol 2012 Feb;107(2):247-252 [FREE Full text] [doi: 10.1038/ajg.2011.314] [Medline: 21931378]
7. Planas R, Montoliu S, Ballesté B, Rivera M, Miquel M, Masnou H, et al. Natural history of patients hospitalized for management of cirrhotic ascites. Clin Gastroenterol Hepatol 2006 Nov;4(11):1385-1394. [doi: 10.1016/j.cgh.2006.08.007] [Medline: 17081806]
8. Runyon BA. Introduction to the revised American association for the study of liver diseases practice guideline management of adult patients with ascites due to cirrhosis 2012. Hepatology 2013 Apr;57(4):1651-1653. [doi: 10.1002/hep.26359] [Medline: 23463403]
9. European Association for the Study of the Liver. EASL clinical practice guidelines on the management of ascites, spontaneous bacterial peritonitis, and hepatorenal syndrome in cirrhosis. J Hepatol 2010 Sep;53(3):397-417. [doi: 10.1016/j.jhep.2010.05.004] [Medline: 20633946]
10. Tapper E, Baki J, Hummel S, Lok A. Design and rationale of a randomized-controlled trial of home-delivered meals for the management of symptomatic ascites: the SALTYFOOD trial. Gastroenterol Rep (Oxf) 2019 Apr;7(2):146-149 [FREE Full text] [doi: 10.1093/gastro/goz005] [Medline: 30976428]
11. Bellos I, Kontzoglou K, Psyrri A, Pergialiotis V. Tolvaptan response improves overall survival in patients with refractory ascites: a meta-analysis. Dig Dis 2020;38(4):320-328 [FREE Full text] [doi: 10.1159/000503559] [Medline: 31578028]
12. Ginès P, Wong F, Watson H, Milutinovic S, del Arbol LR, Olteanu D, HypoCAT Study Investigators. Effects of satavaptan, a selective vasopressin V(2) receptor antagonist, on ascites and serum sodium in cirrhosis with hyponatremia: a randomized trial. Hepatology 2008 Jul;48(1):204-213. [doi: 10.1002/hep.22293] [Medline: 18508290]
13. Bloom PP, Marx M, Wang TJ, Green B, Ha J, Bay C, et al. Attitudes towards digital health tools for outpatient cirrhosis management in patients with decompensated cirrhosis. BMJ Innov 2020 Jan 14;6(1):18-25. [doi: 10.1136/bmjinnov-2019-000369]
14. Siddique SM, Lane-Fall M, McConnell MJ, Jakhete N, Crismale J, Porges S, et al. Exploring opportunities to prevent cirrhosis admissions in the emergency department: a multicenter multidisciplinary survey. Hepatol Commun 2018 Mar;2(3):237-244 [FREE Full text] [doi: 10.1002/hep4.1141] [Medline: 29507899]
15. Hudson B, Round J, Georgeson B, Pring A, Forbes K, McCune CA, et al. Cirrhosis with ascites in the last year of life: a nationwide analysis of factors shaping costs, health-care use, and place of death in England. Lancet Gastroenterol Hepatol 2018 Feb;3(2):95-103. [doi: 10.1016/S2468-1253(17)30362-X] [Medline: 29150405]
16. Malterud K. Qualitative research: standards, challenges, and guidelines. Lancet 2001 Aug 11;358(9280):483-488. [doi: 10.1016/S0140-6736(01)05627-6] [Medline: 11513933]
17. Giacomini MK, Cook DJ. Users' guides to the medical literature: XXIII. Qualitative research in health care B. What are the results and how do they help me care for my patients? Evidence-based medicine working group. J Am Med Assoc 2000 Jul 26;284(4):478-482. [doi: 10.1001/jama.284.4.478] [Medline: 10904512]
18. Bradley EH, Curry LA, Devers KJ. Qualitative data analysis for health services research: developing taxonomy, themes, and theory. Health Serv Res 2007 Aug;42(4):1758-1772 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00684.x] [Medline: 17286625]
19. Seraj SM, Campbell EJ, Argyropoulos SK, Wegermann K, Chung RT, Richter JM. Hospital readmissions in decompensated cirrhotics: factors pointing toward a prevention strategy. World J Gastroenterol 2017 Oct 7;23(37):6868-6876 [FREE Full text] [doi: 10.3748/wjg.v23.i37.6868] [Medline: 29085229]
20. Bashi N, Fatehi F, Fallah M, Walters D, Karunanithi M. Self-management education through mhealth: review of strategies and structures. JMIR Mhealth Uhealth 2018 Oct 19;6(10):e10771 [FREE Full text] [doi: 10.2196/10771] [Medline: 30341042]
21. Marcolino MS, Oliveira JA, D'Agostino M, Ribeiro AL, Alkmim MB, Novillo-Ortiz D. The impact of mhealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018 Jan 17;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]

Abbreviations

EMR: electronic medical record
LVP: large-volume paracentesis
MELD: model for end-stage liver disease
PGHDConnect: patient-generated health data connect

Edited by C Lovis, G Eysenbach; submitted 12.01.20; peer-reviewed by J Louissaint, RA Karin, H Mehdizadeh; comments to author 15.03.20; revised version received 29.03.20; accepted 30.03.20; published 02.09.20.

Please cite as:
Bloom P, Wang T, Marx M, Tagerman M, Green B, Arvind A, Ha J, Bloom J, Richter JM
A Smartphone App to Manage Cirrhotic Ascites Among Outpatients: Feasibility Study
JMIR Med Inform 2020;8(9):e17770
URL: https://medinform.jmir.org/2020/9/e17770
doi:10.2196/17770
PMID:32876581

Original Paper

Treatment of Left Ventricular Circulation Disorder: Application of Echocardiography Information Data Monitoring

Yulong Chen[1], MD; Jianxia Du[1], BD; Xiao Sun[1], BD; Qiancheng Li[1], BD; Ming Qin[1], MD; Qian Xiao[2], BD; Mark Bryan[3], PhD

[1]Department of Ultrasound, Xuzhou Children’s Hospital, Xuzhou Medical University, Xuzhou, China
[2]Department of Emergency, Xuzhou Central Hospital, Xuzhou, China
[3]Dipartimento di Biomedicina, Universita di Torino, Torino, Italy

Corresponding Author:
Qian Xiao, BD
Department of Emergency
Xuzhou Central Hospital
199 Jiefang South Road, Quanshan District
Xuzhou, 221000
China
Phone: 86 0516 83956400
Email: xiaoqianleo968@163.com

Abstract

Background: Cardiac hypertrophy induced by pressure overload is one of the important causes of heart failure and sudden cardiac death. At present, there are few studies on the outcome of left ventricular hypertrophy and left ventricular function after complete pressure load removal.

Objective: This study aims to better simulate the changes of left ventricular structure and function during the process of left ventricular pressure overload and deloading, and to explore the application of echocardiography in it.

Methods: In this study, healthy male (BALB/C) mice were used as research objects to establish an ascending aorta constriction model, to carry out echocardiographic and hemodynamic examinations, to establish an ascending aorta deconstriction model in mice, and to carry out echocardiographic and hemodynamic examinations.

Results: Compared with the sham operation group, the left ventricular end-systolic diameter (LVESD), left ventricular end-diastolic diameter (LVEDD), interventricular septal (IVS), and left ventricular posterior wall (LVPW) in the constriction operation group were significantly increased (*P*=.02, *P*=.02, *P*=.02, and *P*=.02, respectively). LVESD, LVEDD, IVS, and LVPW in the early and late constriction groups were significantly decreased, and the degree of decrease in the early group was greater than that in the late group; compared with the sham operation group, left ventricular diastolic pressure in the constriction operation group increased significantly at 9 and 15 weeks after operation (*P*=.03). Left ventricular systolic pressure at 15 weeks after operation decreased to a certain extent but was higher than that of the sham operation group (*P*=.02). The maximal rate of the increase of left ventricular pressure at 3 weeks, 9 weeks, and 15 weeks after operation decreased significantly (*P*=.03, *P*=.02, and *P*=.02, respectively).

Conclusions: In this study, the ascending aorta coarctation model and descending aorta coarctation model were successfully established, which verifies the value of echocardiography information data monitoring in the treatment of left ventricular circulation disorders and the evaluation of surgical treatment.

(JMIR Med Inform 2020;8(9):e19110) doi:10.2196/19110

KEYWORDS

echocardiography information; data monitoring; hemodynamics; left ventricular circulation disorder

Introduction

The diastolic process of the heart is an important part of the cardiac cycle. If abnormal physiological function occurs in this part, it will lead to left ventricular diastolic dysfunction, which will lead to the decline of left ventricular compliance and relaxation, leading to heart disease. Hypertension and aortic stenosis account for a large proportion of adult heart disease in China. Their common characteristics are that they lead to left ventricular pressure overload and then left ventricular hypertrophy [1,2]. Usually, surgical replacement of the stenosed aortic valve and control of blood pressure can reverse left ventricular hypertrophy and improve the survival rate of patients. However, the recovery of cardiac function in some patients with aortic stenosis after operation is not ideal [3]. In addition, some surveys show that about 15%-20% of patients with hypertension without left ventricular hypertrophy will develop left ventricular hypertrophy during treatment [4]. Although patients with left ventricular hypertrophy receive adequate antihypertensive treatment, the incidence of cardiovascular adverse events is still higher than those without left ventricular hypertrophy [5]. Therefore, it is inappropriate to treat left ventricular hypertrophy inhibitors only. At present, most of the literature about left ventricular hypertrophy focuses on its pathogenesis, molecular mechanism, and the changes of left ventricular function after intervention with pressure overload [6,7]. However, there are few studies on the outcome of left ventricular hypertrophy and left ventricular function after complete removal of pressure load. The evaluation of left ventricular function is the most important index in the diagnosis of various cardiovascular diseases and the evaluation of drug treatment effect. It can predict whether patients have heart failure, and it plays an important role in the clinic [8]. Therefore, it is necessary to establish an ideal left ventricular hypertrophy model and its reversal model, and to explore the characteristics of left ventricular structure and function during the process of left ventricular pressure overload-unloading through effective diagnostic means.

Studies have shown that echocardiography can evaluate the characteristics of left ventricular thickness, left ventricular volume, and wall motion, and it is a widely used auxiliary examination method for judging left ventricular diastolic function in the clinic [9]. The heart rate of mice is generally between 300-500 BPM, and the ventricular cavity is small, so ordinary echocardiography cannot obtain a clear image [10]. In recent years, as the ultrasound technology develops, especially the appearance of high-frequency probes, the reliability of mouse heart function ultrasound evaluation has been improved. Wang et al [11] explored the evaluation value of echocardiography in the structure and function of the mouse myocardial infarction model. In addition, they believed that echocardiography can accurately and sensitively detect the location and severity of myocardial infarction, providing an important basis for clinical diagnosis and treatment of myocardial infarction [11]. It can be seen that echocardiography has a broad application prospect in the evaluation of mouse heart function. In this study, the whole heart function of the animal model is evaluated by echocardiography, and the left ventricular blood circulation is evaluated.

In summary, to establish ascending aorta constriction animal models and ascending aorta deconstriction animal models, and to verify the application value of echocardiography in evaluating the whole heart function of animal models, mice were taken as the research objects to explore the application of echocardiography in the diagnosis and treatment of left ventricular blood circulation disorders in mice so as to provide a basis for left ventricular hypertrophy. Reversal research provides an excellent animal model and provides a reference for the monitoring of coronary heart disease and heart failure.

Methods

Experimental Animal and Groups

A total of 60 healthy male BALB/c mice (XXX Animal Center) aged 6-8 weeks and weighing about 18-25 g were selected as experimental subjects. All animals were fed in cages with national standard rodent feed, with 4 in each cage. Mice could eat and drink freely. There was no significant difference in body weight between groups. The feeding environment had natural light. The room temperature was controlled at 20-26 °C, and the humidity was controlled at 40%-50%. The mice were fed adaptively for 2 weeks. Animal handling and experimental procedures conformed to the National Laboratory Animal standards and were approved by the ethical committee.

All mice were randomly divided into 6 groups, 10 in each group, namely, sham operation group, 3-week constriction operation group, 9-week constriction operation group, 15-week constriction operation group, early deconstriction group, and late deconstriction group.

Establishment of Ascending Aorta Coarctation Model in Mice

In addition to sham-operated mice, other mice were intraperitoneally injected with 2% tribromethanol at a dose of 25 mg/kg (Shanghai Jingke Chemical Technology Co, Ltd, China). After successful anesthesia, the mice were fixed in supine position on the operating table (Beijing Semiconductor Equipment Factory, China). During the operation, if the depth of anesthesia was not enough, 2% tribromethanol could be added, but the anesthesia could not be too deep. Hair was cut off from the neck and upper chest of mice, and iodophor (Beijing Baioubowei Biotechnology Co, Ltd., China) was used to disinfect them. Plastic tubes suitable for size were inserted into the trachea through the mouth and connected to a small animal ventilator (Shenzhen Reward Life Technology Co, Ltd, China) to assist mice in breathing. The skin was cut from the upper sternum to the medial side of the external jugular vein, and the subcutaneous fascia was separated to expose the superficial muscles of the neck. The muscles were cut and divided into two sides along the median line to expose the trachea. In mice, the thymus was pulled apart, and the surrounding tissues of ascending aorta were separated so that the aortic arch was fully exposed. The aortic arch was separated and the 4-0 Prolene line (Shanghai Yuyan Scientific Instruments Co, Ltd, China) was used to circumvent and tie the aortic arch. We were careful not to ligate and leave space for gaskets and needles. The 27G needle was placed in the para-aortic junction, and then the gasket was inserted to tighten the junction rapidly. After ligation, needles needed to be removed quickly. Ligation should be done accurately and quickly, keeping in situ as far as possible to prevent damage to the ascending aortic adventitia. After ligation, penicillin (100,000 units, Shijiazhuang Best Pharmaceutical and Chemical Co, Ltd, China) should be sprayed into the chest cavity. The thymus glands were seamless, the thoracic glands were closed layer by layer, and the skin was sutured. In the sham operation group, the ascending aorta was separated without ligation, and the other procedures were the same.

Echocardiographic Examination of Ascending Aorta Coarctation in Mice

Echocardiography was performed in mice before and 3, 9, and 15 weeks after ascending aortic coarctation. Fasting was required 4-6 hours before the examination. The mice were given intraperitoneal injection of 2% tribromethanol at a dose of 15 mg/kg and fixed on the examination table in supine position to remove the hair on the chest, upper abdomen, and the roots of both upper limbs. First, a sodium sulfide solution (8%, Shanxi Xinchengshun Chemical Co, Ltd, China) was used to rinse and depilate. Clean water was then used for flushing. Attention was paid to check whether the area's hair was completely removed. The VIVID7 Echocardiograph Diagnostic Instrument (GE Company, United States) was used for echocardiographic examination in mice. The probe frequency was 12 MHz, and the observation sites were bilateral thoracic cavity and subxiphoid process of mice. The indicators included left ventricular end-systolic diameter (LVESD), left ventricular end-diastolic diameter (LVEDD), interventricular septal (IVS), and left ventricular posterior wall (LVPW).

Hemodynamic Study of Ascending Aorta Coarctation in Mice

The mice were intraperitoneally injected with 2% tribromethanol at a dose of 25 mg/kg and fixed on the operating table in supine position. Hair was cut off from the neck and upper chest. Sodium sulfide ethanol solution was used to rinse and depilate hair and then disinfected with iodophor. The aseptic cave towel was laid, and a longitudinal incision about 2 cm long was made along the middle of the neck. Skin and subcutaneous fascia were cut to expose neck muscles. The sternohyoid muscles were bluntly separated from each other to expose the trachea. The right muscular layer was sutured and fixed on the operating table by a needle with thread No 0. In the groove of the trachea and right sternohyoid muscle, the internal carotid artery was found and separated, and two silk threads were pierced. The distal end of the right internal carotid artery was ligated and slightly pulled outward by the filament near the central end. The Pre-Chong Heparin (Shanghai Kanglang Biotechnology Co, Ltd, China) No 24 intravenous indwelling needle (Anhui Hongzhong Medical Devices Co, Ltd, China) was used to puncture arteries. After the puncture, the needle core was extracted and connected with the preflushed heparin pressure transducer (Beijing Zhishu Duobao Biotechnology Co, Ltd, China). Internal carotid artery pressure was observed on a multifunctional physiological recorder (ML870, Powerlab, Australia). The silk thread was released, and we inserted a venous indwelling needle into the heart. During the operation, attention was paid to the change of pressure on the display. If the pressure suddenly increased, it indicated that it had entered the left ventricle. Silk thread was used to tie a knot outside the blood vessel and fix the intravenous indwelling needle to prevent blood leakage. However, it should be noted that knotting should not be too tight. Once the connection was successful and the waveform was stable, it was necessary to start recording data, including left ventricular diastolic pressure (LVDP), left ventricular systolic pressure (LVSP), maximal rate of the increase of left ventricular pressure (+dp/dt Max), and maximal rate of the decrease of left ventricular pressure (-dp/dt Max).

Establishment of Ascending Aorta Deconstriction Mouse Model

At 3 and 9 weeks after ascending aorta coarctation, ascending aorta coarctation was performed on mice in the early and late decoarctation groups. The mice were intraperitoneally injected with 2% tribromethanol at a dose of 25 mg/kg and fixed on the operating table in supine position. The hair was cut off from the neck and upper chest. It was necessary to rinse and depilate with sodium sulfide ethanol solution and then disinfect with iodophor. Small animal ventilators were connected to assist mice in breathing, exposing the trachea. The thoracotomy was performed between the two or three ribs, and the thymus was opened to expose the aortic arch. The original ligation line was found. Down the ligation line, the gasket was separated. Penicillin (100,000 units) was sprayed in the chest cavity. The thymus was not seamed. It was then necessary to close the chest layer by layer and suture the skin.

Echocardiographic Examination of Ascending Aorta Decoarctation Model in Mice

Echocardiography was performed in mice at 3, 9, and 15 weeks after deconstriction. The mice were intraperitoneally injected with 2% tribromethanol at a dose of 15 mg/kg and fixed on the examination table in supine position. Hair was removed from the chest, upper abdomen, and the roots of both upper limbs. It was necessary to rinse with sodium sulfide solution and then rinse with water. The VIVID7 echocardiographic diagnostic instrument was used to perform echocardiographic examination in mice. Detection indicators included LVESD, LVEDD, IVS, and LVPW.

Hemodynamic Study of Ascending Aorta Decoarctation in Mice

The mice were intraperitoneally injected with 2% tribromethanol at a dose of 25 mg/kg and fixed on the operating table in supine position. Hair was cut off from the neck and upper chest. It was first necessary to rinse and depilate with sodium sulfide ethanol solution and then disinfect with iodophor. The aseptic cave towel was laid. Along the middle of the neck, an incision about 2 cm in length was made to expose the trachea. A No 0 thread was used to suture and fix the right muscular layer on the operating table. In the groove of the trachea and right sternohyoid muscle, the internal carotid artery was found and separated, and two silk threads were pierced. The distal end of the right internal carotid artery was ligated and slightly pulled outward by the filament near the central end. A preflushed heparin 24 venous indwelling needle was used to puncture the artery. After the puncture was completed, the needle core needed to be pulled out and connected with the pressure transducer for preflushing heparin. The internal carotid artery pressure was observed on a multifunctional physiological recorder. To loosen the silk thread, a venous indwelling needle was inserted into the heart. During the operation, it was necessary to pay attention to the pressure changes on the display. If the pressure suddenly increased, it indicated that it had entered the left ventricle. Silk thread was used to tie a knot outside the blood vessel to fix the intravenous indwelling needle to prevent blood leakage, but it should be noted that the knot should not be too tight. After the connection was successful and the waveform was stabilized, data were recorded, including LVDP, LVSP, +dp/dt max, -dp/dt max.

Statistical Methods

SPSS 26.0 (IBM Corp) statistical software was used to analyze the data. The measurement data was expressed as mean and standard deviation. The mean of the two samples was compared by the *t* test. The counting data was expressed as incidence n (%). The comparison used the chi-square test, and the difference was statistically significant when *P*<.05.

Results

Echocardiographic Results of Ascending Aorta Coarctation Mouse Model and Ascending Aorta Decoarctation Mouse Model

The results of echocardiography in ascending aorta coarctation mice and ascending aorta decoarctation mice are shown in Figures 1 and 2. From the ascending aorta coarctation model in mice, it can be seen that, compared with the sham operation group, the LVESD, LVEDD, IVS, and LVPW of the coarctation operation group increased significantly with statistical significance (*P*=.02, *P*=.02, *P*=.02, and *P*=.02, respectively). In the constriction group, compared with the preoperative group, LVESD, LVEDD, IVS, and LVPW in mice increased significantly 3 weeks after the operation with statistical significance (*P*=.01, *P*=.01, *P*=.02, and *P*=.01, respectively), suggesting that the mice had centripetal hypertrophy. Compared with 3 weeks after the operation, the LVESD, LVEDD, and IVS of mice increased significantly at 9 weeks after the operation with statistical significance (*P*=.02, *P*=.01, and *P*=.02, respectively), while the increase of LVPW was not significant, suggesting that the mice have eccentric hypertrophy (Figure 2A). Compared with 9 weeks after operation, LVESD and LVEDD in mice increased significantly at 15 weeks after the operation, and the difference was statistically significant (*P*=.03 and *P*=.03. respectively), suggesting that heart failure occurs in mice.

Figure 1. Echocardiographic results of ascending aorta coarctation mouse model and ascending aorta decoarctation mouse model (A: LVESD; B: LVEDD; C: IVS; D: LVPW). IVS: interventricular septal; LVEDD: left ventricular end-diastolic diameter; LVESD: left ventricular end-systolic diameter; LVPW: left ventricular posterior wall; W: weeks.

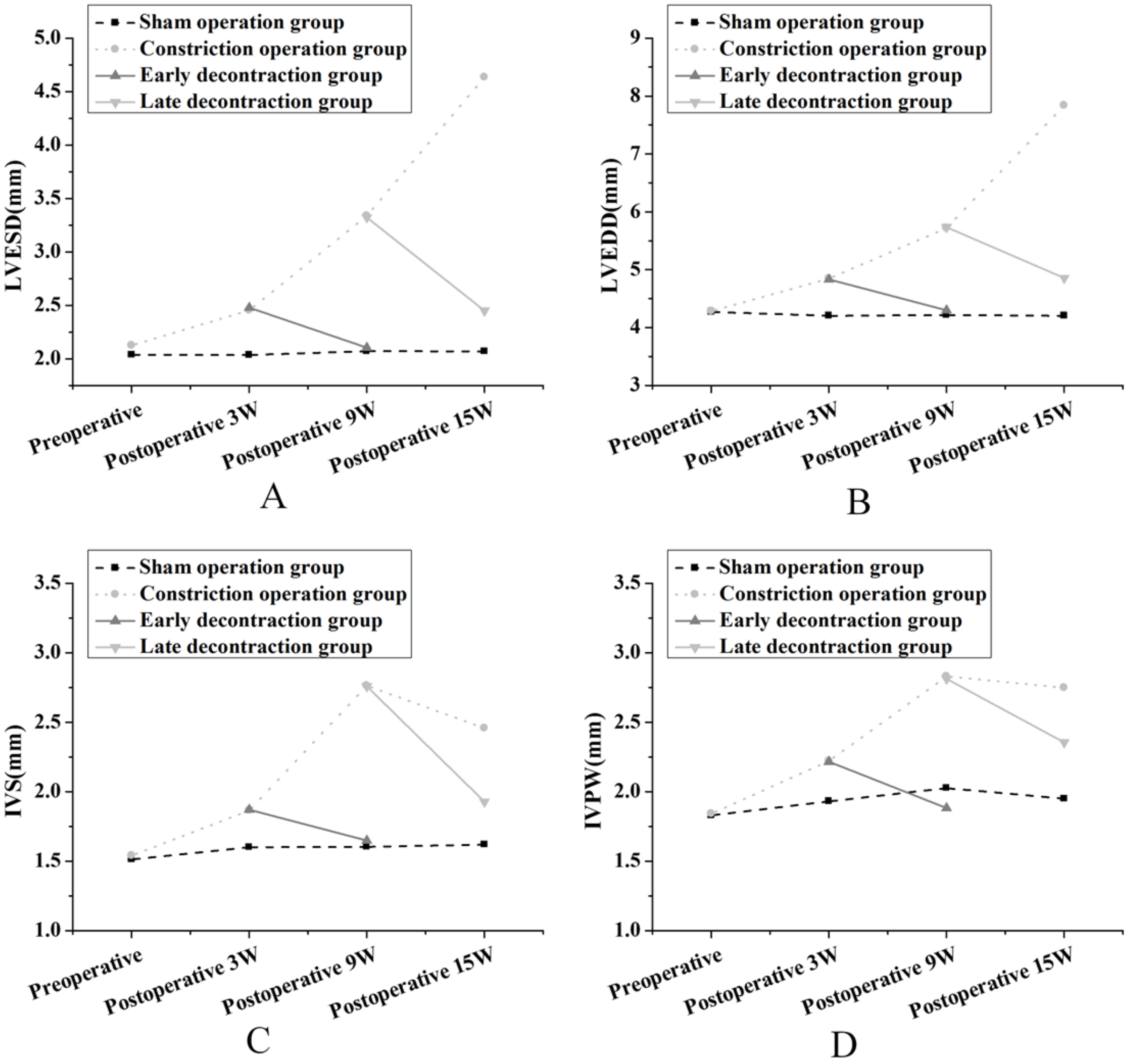
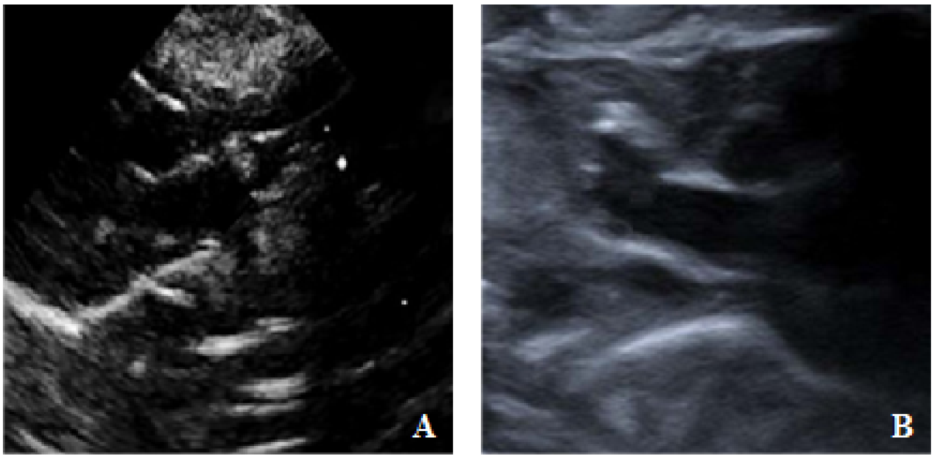


Figure 2. Echocardiography of mice (A: ascending aorta constriction mouse model; B: ascending aorta deconstriction mouse model).



From the model of ascending aorta deconstriction mice, it can be seen that, compared with the operation group, the left ventricular function of the early deconstriction group and the late deconstriction group were reversed, the blood circulation disorder was improved, and the effect of the early deconstriction group was more obvious than that of the late deconstriction group. In the early deconstriction group, the LVESD, LVEDD, IVS, and LVPW of mice decreased significantly at the 9th week and tended toward the initial level (Figure 2B). In the delayed decoarctation group, the levels of LVESD, LVEDD, IVS, and LVPW decreased significantly at the 15th week but could not return to the initial level, suggesting the persistence of ventricular hypertrophy and the persistence of left ventricular circulation disorders.

Hemodynamic Results of Ascending Aorta Coarctation Mice Model and Ascending Aorta Decoarctation Mice Model

The hemodynamic results of the ascending aorta coarctation mice and ascending aorta decoarctation mice are shown in Table 1. From the ascending aorta coarctation model in mice, it can be seen that, compared with the sham operation group, there was no difference in LVDP in the coarctation group 3 weeks after operation. At 9 weeks and 15 weeks after the operation, the LVDP of mice in the constriction group increased significantly (*P*=.03 and *P*=.03, respectively). Compared with the sham-operated group, the LVSP in the constriction group increased significantly at 3 weeks and 9 weeks after the operation (*P*=.03 and *P*=.02, respectively). At 15 weeks after the operation, the LVSP of the constriction group decreased to a certain extent but was higher than that of the sham operation group (*P*=.02). Compared with the sham-operated group, the -dp/dt max in the constriction group decreased significantly at 3 weeks, 9 weeks, and 15 weeks (*P*=.03, *P*=.02, and *P*=.02, respectively). The results of hemodynamics were consistent with those of echocardiography.

Table 1. Hemodynamic results of ascending aorta constriction model and ascending aorta deconstriction model in mice.

Period	LVDP[a] (mmHg)	LVSP[b] (mmHg)	+dp/dt max[c] (mmHg/sec)	-dp/dt max[d] (mmHg/sec)
Sham operation group	4.8	125	6400	4700
3-week constriction surgery group	4.9	168	5000	3950
9-week constriction surgery group	8.1	175	5200	3500
15-week constriction surgery group	14.8	172	4200	3400
Early deconstriction group	4.9	120	6000	4100
Late deconstriction group	6.3	141	5600	4080

[a] LVDP: left ventricular diastolic pressure.

[b] LVSP: left ventricular systolic pressure.

[c] +dp/dt max: maximal rate of the increase of left ventricular pressure.

[d] -dp/dt max: maximal rate of the decrease of left ventricular pressure.

In the ascending aorta deconstriction mice model, compared with the sham-operated mice, the differences of LVDP, LVSP, and +dp/dt max in the early deconstriction mice were not statistically significant, while -dp/dt max was significantly lower (*P*=.03). Compared with the constriction operation group, the LVDP and LVSP in the early deconstriction group were significantly lower, and the +dp/dt max and -dp/dt max were significantly higher. It was observed that the left ventricular function of ascending aorta deconstriction mice was restored, and the blood circulation disorder had been improved. Compared with the constriction operation group, the LVDP and LVSP of mice in the delayed deconstriction group decreased significantly (*P*=.03 and *P*=.02), and the +dp/dt max and -dp/dt max increased significantly, with statistical significance (*P*=.03 and *P*=.04, respectively). The results of hemodynamics were consistent with those of echocardiography.

Discussion

Result Analysis

In the past, most of the animal models of left ventricular hypertrophy under pressure overload were established by ligation of the abdominal aorta. This model is easy to operate and has a high survival rate, but the modeling time is too long, sometimes even more than 1 year [12]. In recent years, some scholars have proposed models of aortic arch and ascending aorta coarctation, but the models need the support of endotracheal intubation and artificial ventilation conditions [13]. However, ligation of the aortic arch can increase the pressure of the right brain and its limbs, increase the collateral circulation through the head and arm, and then reduce the afterload, making the modeling a failure [14].

To establish an ideal model of left ventricular hypertrophy and its reversion, and simulate the changes of left ventricular structure and function during the process of left ventricular pressure overload and deload, the healthy male BALB/c mice were taken as the research objects, and the models of ascending aorta constriction mice and ascending aorta deconstriction mice were established. To evaluate the success of the model, echocardiography and hemodynamics were performed. Echocardiography has the advantages of a noninvasive and simple operation, which can reflect the left ventricular function and wall thickness. The heart rate of mice was fast and the ventricular cavity was small, and M-mode ultrasound can make the imaging clearer and provide accurate indexes such as the diameter of the ventricular cavity and the thickness of the ventricular wall.

The results of this study showed that the LVESD, LVEDD, IVS, and LVPW of the mice in the operation group increased significantly compared with those before operation (*P*=.01, *P*=.01, *P*=.02, and *P*=.01, respectively), suggesting that there was centripetal hypertrophy in the mice. At this time, the increased myocardial contractility can fully compensate for the increase of pressure load. The previously mentioned indexes showed a further upward trend in the 6th and 9th weeks after operation. At the 9th week after operation, the LVESD, LVEDD, and IVS of the mice in the constriction group increased significantly, while the LVPW increased slightly, suggesting that the mice appeared to have centrifugal hypertrophy. At 15 weeks after the operation, the LVESD and LVEDD of the mice in the constriction group increased significantly, indicating that the mice had heart failure. The detection of left ventricular pressure can directly prove the success of the aortic coarctation model and the degree of coarctation. It was found that the left ventricular pressure increased significantly after ascending aortic coarctation, which indicated that the model was constructed successfully. In the early group, IVS and LVPW decreased to nearly the initial level 3 weeks after the removal of systole, and LVESD and LVEDD recovered completely 6 weeks after the removal of systole. There were also differences in left ventricular function between the late group and the early group, suggesting that the reversal of left ventricular hypertrophy was related to the duration.

Research Reliability

The results of hemodynamics were consistent with those of echocardiography, which further proved the reliability of echocardiography. Hendrikx et al [15] showed that in the mouse model of myocardial infarction, echocardiography could provide reliable measurement of left ventricular function with high accuracy. It showed that echocardiography had important application value in the evaluation of mouse heart function, which was consistent with the results of this study.

Based on the model of ascending aorta coarctation, the model of ascending aorta decoarctation in mice was established, and the accuracy and reliability of the modeling method were verified. Echocardiography can accurately measure the left ventricular function of mice, which has important application value in the diagnosis of left ventricular blood circulation disorder and the evaluation of surgical treatment effect.

Conclusion

The application of echocardiography in the diagnosis and treatment of left ventricular blood circulation disorders in mice was explored. Through research, it was found that echocardiography can accurately measure the left ventricular function of mice and has important application value in the diagnosis of left ventricular blood circulation disorders and the evaluation of surgical treatment effect. It provides an excellent animal model for the study of left ventricular hypertrophy and its reversion, and provides a reference for the monitoring of coronary heart disease and heart failure. However, there are some deficiencies in the research process, such as the small number of experimental animals, which leads to a certain degree of deviation in the results. Therefore, in the later research process, the number of experimental animals will be further increased to make the results more valuable for reference.

Conflicts of Interest

None declared.

References

1. Wulin G, Guohua D, Tong Z, Dongxue B, Chunhua L, Xiaojing S, et al. Tonifying Qi and activating blood circulation in terms of Traditional Chinese Medicine: their effects in patients with myocardial infarction. J Tradit Chin Med 2018 Oct;38(5):726-732. [doi: 10.1016/s0254-6272(18)30911-7]
2. Cho I, Chang H, Heo R, Kim I, Sung JM, Chang B, et al. Association of thoracic aorta calcium score with left ventricular hypertrophy and clinical outcomes in patients with severe aortic stenosis after aortic valve replacement. Ann Thorac Surg 2017 Jan;103(1):74-81. [doi: 10.1016/j.athoracsur.2016.05.039] [Medline: 27440307]
3. Burgos PFM, Luna Filho B, de Assis Costa F, Bombig MTN, de Souza D, Bianco HT, et al. Electrocardiogram performance in the diagnosis of left ventricular hypertrophy in hypertensive patients with left bundle branch block. Arq Bras Cardiol 2017 Jan;108(1):47-52 [FREE Full text] [doi: 10.5935/abc.20160187] [Medline: 27992034]
4. El Saiedi SA, Mira MF, Sharaf SA, Al Musaddar MM, El Kaffas RMH, AbdelMassih AF, et al. Left ventricular diastolic dysfunction without left ventricular hypertrophy in obese children and adolescents: a Tissue Doppler Imaging and Cardiac Troponin I Study. Cardiol Young 2017 Aug 07;28(1):76-84. [doi: 10.1017/s1047951117001627]
5. Cariou E, Bennani Smires Y, Victor G, Robin G, Ribes D, Pascal P, Toulouse Amyloidosis Research Network. Diagnostic score for the detection of cardiac amyloidosis in patients with left ventricular hypertrophy and impact on prognosis. Amyloid 2017 Jun;24(2):101-109. [doi: 10.1080/13506129.2017.1333956] [Medline: 28553897]
6. Cho H, Choi HJ, Kang HG, Ha IS, Cheong HI, Han KH, et al. Influence of the method of definition on the prevalence of left-ventricular hypertrophy in children with chronic kidney disease: data from the Know-Ped CKD study. Kidney Blood Press Res 2017;42(3):406-415 [FREE Full text] [doi: 10.1159/000478867] [Medline: 28689198]
7. Viana Gonçalves IC, Cerdeira CD, Poletti Camara E, Dias Garcia JA, Ribeiro Pereira Lima Brigagão M, Bessa Veloso Silva R, et al. Tempol improves lipid profile and prevents left ventricular hypertrophy in LDL receptor gene knockout (LDLr-/-) mice on a high-fat diet. Rev Port Cardiol 2017 Sep;36(9):629-638 [FREE Full text] [doi: 10.1016/j.repc.2017.02.014] [Medline: 28826937]
8. Kitamura M, Amano Y, Takayama M, Shibuya J, Matsuda J, Sangen H, et al. Usefulness of non-anteroseptal region left ventricular hypertrophy using cardiac magnetic resonance to predict repeat alcohol septal ablation for refractory obstructive

hypertrophic cardiomyopathy. Am J Cardiol 2017 Jul 01;120(1):124-130. [doi: 10.1016/j.amjcard.2017.03.248] [Medline: 28483204]

9. Tian JP, Wang J, Tian XK, Du FH, Wang T. The impact of visit-to-visit systolic blood pressure variability on residual renal function and left ventricular hypertrophy in peritoneal dialysis patients. Turk J Med Sci 2018 Apr 30;48(2):279-285. [doi: 10.3906/sag-1704-92] [Medline: 29714440]
10. Conn NK, Schwarz KQ, Borkholder DA. In-Home cardiovascular monitoring system for heart failure: comparative study. JMIR mHealth uHealth 2019 Jan 18;7(1):e12419 [FREE Full text] [doi: 10.2196/12419] [Medline: 30664492]
11. Wang J, Tan WJ, Li X, Zhang GP, Huang JY, Chen XH, et al. [High-frequency echocardiography for assessment of regional wall motion abnormality and cardiac function in mice with myocardial infarction]. Nan Fang Yi Ke Da Xue Xue Bao 2017 Aug 20;37(8):1014-1021 [FREE Full text] [Medline: 28801279]
12. Takahashi M, Kinugawa S, Takada S, Kakutani N, Furihata T, Sobirin MA, et al. The disruption of invariant natural killer T cells exacerbates cardiac hypertrophy and failure caused by pressure overload in mice. Exp Physiol 2020 Mar 18;105(3):489-501. [doi: 10.1113/EP087652] [Medline: 31957919]
13. Eckert M, Volmerg JS, Friedrich CM. Augmented reality in medicine: systematic and bibliographic review. JMIR mHealth uHealth 2019 Apr 26;7(4):e10967 [FREE Full text] [doi: 10.2196/10967] [Medline: 31025950]
14. Liu L, Duan S, Zhang Y, Wu Y, Zhang L. Initial experience of the synchronized, real-time, interactive, remote transthoracic echocardiogram consultation system in rural China: longitudinal observational study. JMIR Med Inform 2019 Jul 08;7(3):e14248 [FREE Full text] [doi: 10.2196/14248] [Medline: 31287062]
15. Hendrikx G, Bauwens M, Wierts R, Mottaghy F, Post M. Left ventricular function measurements in a mouse myocardial infarction model. Nuklearmedizin 2018 Mar 06;55(03):115-122. [doi: 10.3413/nukmed-0776-15-11]

Abbreviations

IVS: interventricular septal
LVDP: left ventricular diastolic pressure
LVEDD: left ventricular end-diastolic diameter
LVESD: left ventricular end-systolic diameter
LVPW: left ventricular posterior wall
LVSP: left ventricular systolic pressure
+dp/dt max: maximal rate of the increase of left ventricular pressure
-dp/dt max: maximal rate of the decrease of left ventricular pressure

Edited by K Kalemaki, Z Du, H Li; submitted 03.04.20; peer-reviewed by Z Liu, Y Zeng, X Yang; comments to author 17.04.20; revised version received 30.04.20; accepted 01.05.20; published 16.09.20.

Please cite as:
Chen Y, Du J, Sun X, Li Q, Qin M, Xiao Q, Bryan M
Treatment of Left Ventricular Circulation Disorder: Application of Echocardiography Information Data Monitoring
JMIR Med Inform 2020;8(9):e19110
URL: http://medinform.jmir.org/2020/9/e19110/
doi:10.2196/19110
PMID:32936076

Original Paper

Nomogram for Predicting COVID-19 Disease Progression Based on Single-Center Data: Observational Study and Model Development

Tao Fan[1*], PhD, MD; Bo Hao[1*], PhD, MD; Shuo Yang[1*], MD; Bo Shen[1*], PhD, MD; Zhixin Huang[1*], PhD, MD; Zilong Lu[1], MD; Rui Xiong[1], MD; Xiaokang Shen[1], MD; Wenyang Jiang[1], PhD, MD; Lin Zhang[1], MD; Donghang Li[1], MD; Ruyuan He[1], MD, PhD; Heng Meng[1], MD, PhD; Weichen Lin[1], MD, PhD; Haojie Feng[1], MD; Qing Geng[1], PhD, MD

Renmin Hospital, Wuhan University, Wuhan, China

[*]these authors contributed equally

Corresponding Author:
Qing Geng, PhD, MD
Renmin Hospital
Wuhan University
238 Jiefang Road
Wuhan, 430060
China
Phone: 27 88041911 880419
Email: gengqingwhu@whu.edu.cn

Abstract

Background: In late December 2019, a pneumonia caused by SARS-CoV-2 was first reported in Wuhan and spread worldwide rapidly. Currently, no specific medicine is available to treat infection with COVID-19.

Objective: The aims of this study were to summarize the epidemiological and clinical characteristics of 175 patients with SARS-CoV-2 infection who were hospitalized in Renmin Hospital of Wuhan University from January 1 to January 31, 2020, and to establish a tool to identify potential critical patients with COVID-19 and help clinical physicians prevent progression of this disease.

Methods: In this retrospective study, clinical characteristics of 175 confirmed COVID-19 cases were collected and analyzed. Univariate analysis and least absolute shrinkage and selection operator (LASSO) regression were used to select variables. Multivariate analysis was applied to identify independent risk factors in COVID-19 progression. We established a nomogram to evaluate the probability of progression of the condition of a patient with COVID-19 to severe within three weeks of disease onset. The nomogram was verified using calibration curves and receiver operating characteristic curves.

Results: A total of 18 variables were considered to be risk factors after the univariate regression analysis of the laboratory parameters (P<.05), and LASSO regression analysis screened out 10 risk factors for further study. The six independent risk factors revealed by multivariate Cox regression were age (OR 1.035, 95% CI 1.017-1.054; P<.001), CK level (OR 1.002, 95% CI 1.0003-1.0039; P=.02), CD4 count (OR 0.995, 95% CI 0.992-0.998; P=.002), CD8 % (OR 1.007, 95% CI 1.004-1.012, P<.001), CD8 count (OR 0.881, 95% CI 0.835-0.931; P<.001), and C3 count (OR 6.93, 95% CI 1.945-24.691; P=.003). The areas under the curve of the prediction model for 0.5-week, 1-week, 2-week and 3-week nonsevere probability were 0.721, 0.742, 0.87, and 0.832, respectively. The calibration curves showed that the model had good prediction ability within three weeks of disease onset.

Conclusions: This study presents a predictive nomogram of critical patients with COVID-19 based on LASSO and Cox regression analysis. Clinical use of the nomogram may enable timely detection of potential critical patients with COVID-19 and instruct clinicians to administer early intervention to these patients to prevent the disease from worsening.

(JMIR Med Inform 2020;8(9):e19588) doi:10.2196/19588

KEYWORDS

coronavirus disease 2019; COVID-19; risk factors; nomogram

Introduction

Background

COVID-19 is a respiratory illness that is caused by the novel virus SARS-CoV-2; it was first reported in December 2019 in Wuhan, Hubei Province, China [1-5]. Although governments of countries worldwide have called for and taken relevant measures to stop the spread of the disease, the epidemic has not been effectively controlled [6-8]. Symptoms of COVID-19 range from mild cough to pneumonia; patients may even be asymptomatic [3,9]. There is evidence that this disease can be spread from person to person [10]. Whole genome sequencing showed that SARS-CoV-2 is a beta coronavirus that is similar to human severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV). This new coronavirus evolved from SARS-CoV and MERS-CoV and requires enhanced surveillance and further investigation [11]. Like several other coronaviruses, SARS-CoV-2 initially causes mild or moderate symptoms in most patients [12,13]. To date, only a small percentage of patients with SARS-CoV-2 infection have developed severe pneumonia. Although the average incubation period of SARS-CoV-2 in the human body is 14 days, some patients progress rapidly to respiratory failure once they become infected. Recognition of the risk factors that promote COVID-19 progression and early intervention of this disease may prevent its exacerbation. Understanding these factors is also very important to reduce the proportion of critically ill patients and to improve the cure rate.

Study Goals

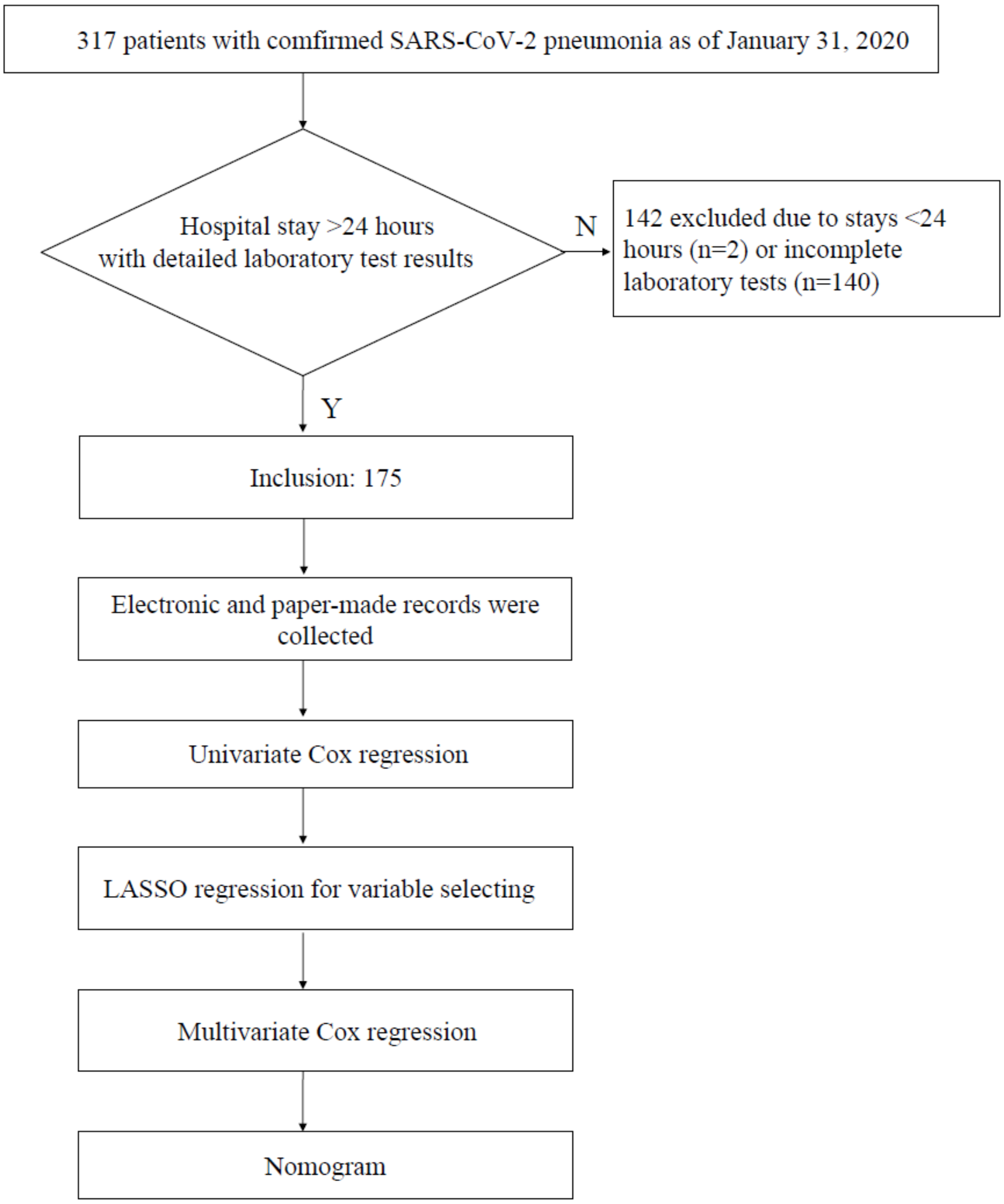
The aim of this study was to summarize the epidemiological and clinical characteristics of 175 patients with SARS-CoV-2 infection who were hospitalized in Renmin Hospital of Wuhan University from January 1 to January 31, 2020. The study also aimed to explore the independent risk factors in COVID-19 progression and accurately assess the incidence of severe SARS-CoV-2 infection. In addition, a new risk-predictive model was established to screen out potential critical patients for early intervention.

Methods

Patient Recruitment

For this retrospective single-center study, 175 patients who were hospitalized from January 1 to January 31, 2020, in Renmin Hospital of Wuhan University were enrolled. Patients who were hospitalized for less than 24 hours or who lacked detailed laboratory test results were excluded (Figure 1). All the patients in this study were diagnosed according to the World Health Organization (WHO) interim guidance [14]. All the patients with COVID-19 were diagnosed using a reverse transcriptase–polymerase chain reaction (RT-PCR) assay for SARS-CoV-2 according to the Pneumonitis Diagnosis and Treatment Plan for SARS-CoV-2 Infection (Trial Version 5) issued by the National Health Commission of the People's Republic of China (NHCPRC). The test was performed using specific steps described previously [9]. All the clinical features, radiological characteristics, clinical laboratory results, and outcome data of the 175 patients were obtained from electronic medical records. Detailed patient information was collected, including past medical history, current medical history, laboratory findings, imaging data, treatment measures, symptoms, signs, and immune function test results. Follow-up was initiated from suspicion of infection or confirmed diagnosis to the time when the patient's condition became severe to the time of discharge or to January 31, 2020. We recorded the baselines of these tests, with the first value being within three days of onset admission.

Figure 1. Study flowchart. LASSO: least absolute shrinkage and selection operator.


317 patients with comfirmed SARS-CoV-2 pneumonia as of January 31, 2020
Hospital stay >24 hours with detailed laboratory test results
N
142 excluded due to stays <24 hours (n=2) or incomplete laboratory tests (n=140)
Y
Inclusion: 175
Electronic and paper-made records were collected
Univariate Cox regression
LASSO regression for variable selecting
Multivariate Cox regression
Nomogram

Statistical Analysis

Frequency rates were used to describe categorical variables, and means were used to describe continuous variables. Differences between groups were tested using the chi-square test, *t* test, or Mann-Whitney U test. Univariate analysis, least absolute shrinkage and selection operator (LASSO) regression, and multivariate Cox regression were used to screen for independent risk factors. The statistical analyses were performed using GraphPad Prism 8 (GraphPad Software). Nomogram, calibration, and receiver operating characteristic (ROC) curves were established using R version 3.6.1 (R Project). The risk score was calculated using multivariate Cox regression. The cutoff values of independent risk factors were based on the maximum Youden index. A 2-sided α <.05 was considered to indicate statistical significance.

Results

Patient Characteristics

All 175 patients in this study were confirmed to be infected with SARS-CoV-2. The average age of the 175 patients was 46.9 years (SD 17.33). Severe or critical patients were significantly older than mild or moderate patients (*P*<.001). Critical patients had a wider range of lung lesions than milder patients (*P*<.001). Chest tightness was more common in severe or critical patients than in mild or moderate patients (*P*=.001). Table 1 shows comparisons of the significant characteristics between the two groups of patients enrolled in the study (*P*<.05).

Table 1. Baseline characteristics of the patients infected with SARS-CoV-2 (N=175).

Participant characteristics	Values by disease severity				*P* value
	Mild (n=27)	Moderate (n=72)	Severe (n=18)	Critical (n=58)	
Sex, n (%)					.64
Female	13 (48.1)	41 (56.9)	11 (61.1)	28 (48.3)	
Male	14 (51.9)	31 (43.1)	7 (38.9)	30 (51.7)	
Age (years), median (IQR)	31 (29-42)	35 (30-50.75)	51 (37.75-61)	62 (47-71)	<.001
Age (years), n (%)					<.001
≤30	11 (40.7)	21 (29.2)	1 (5.6)	3 (5.2)	
31-40	8 (29.6)	25 (34.7)	4 (22.2)	7 (12.1)	
41-50	4 (14.8)	8 (11.1)	4 (22.2)	6 (10.3)	
51-60	1 (3.7)	13 (18.1)	5 (27.8)	10 (17.2)	
>60	3 (11.1)	5 (6.9)	4 (22.2)	32 (55.2)	
Computed tomography scan, n (%)					<.001
Normal	27 (100)	0 (0)	0 (0)	0 (0)	
One lobe infected	0 (0)	25 (34.7)	6 (33.3)	5 (8.6)	
More than one lobe infected	0 (0)	47 (65.3)	12 (66.7)	53 (91.4)	
Fever, n (%)					.01
Yes	17 (63)	43 (59.7)	14 (77.8)	49 (84.5)	
No	10 (37)	29 (40.3)	4 (22.2)	9 (15.5)	
Dry cough, n (%)					.07
Yes	2 (7.4)	11 (15.3)	3 (16.7)	17 (29.3)	
No	25 (92.6)	61 (84.7)	15 (83.3)	41 (70.7)	
Expectoration, n (%)					.007
Yes	5 (18.5)	28 (38.9)	10 (55.6)	33 (56.9)	
No	22 (81.5)	44 (61.1)	8 (44.4)	25 (43.1)	
Pharyngalgia, n (%)					.25
Yes	4 (14.8)	11 (15.3)	0 (0)	5 (8.6)	
No	23 (85.2)	61 (84.7)	18 (100)	53 (91.4)	
Chest tightness, n (%)					.001
Yes	6 (22.2)	8 (11.1)	7 (38.9)	23 (39.7)	
No	21 (77.8)	64 (88.9)	11 (61.1)	35 (60.4)	
Myalgia, n (%)					.80
Yes	4 (14.8)	6 (8.3)	2 (11.1)	7 (12.1)	
No	23 (85.2)	66 (91.7)	16 (88.9)	51 (87.9)	
Fatigue, n (%)					.62
Yes	8 (29.6)	18 (25)	3 (16.7)	11 (19)	
No	19 (70.4)	54 (75)	15 (83.3)	47 (81)	
Diarrhea, n (%)					.55
Yes	3 (11.1)	5 (6.9)	0 (0)	4 (6.9)	
No	24 (88.9)	67 (93.1)	18 (100)	54 (93.1)	
Headache, n (%)					.08
Yes	6 (22.2)	11 (15.3)	1 (5.6)	3 (5.2)	

Participant characteristics	Values by disease severity				*P* value
	Mild (n=27)	Moderate (n=72)	Severe (n=18)	Critical (n=58)	
No	21 (77.8)	61 (84.7)	17 (94.4)	55 (94.8)	

Laboratory Parameters

The baseline laboratory tests are shown in Table 2. As the patients' conditions worsened, their lymphocyte counts significantly decreased, while their C-reactive protein, lactate dehydrogenase, and creatine kinase (CK) levels increased significantly (Table 2). More importantly, the CD3 (count and ratio), CD4 (count and ratio), CD8 (count and ratio), and CD19 (count and ratio) values of severe or critical patients were lower than those of the mild or moderate patients (Table 2). In addition, severe or critical patients had higher IgG levels than mild or moderate patients (Table 2).

Table 2. Laboratory findings of patients infected with SARS-CoV-2 on admission to hospital (N=175).

Variable	Normal range	Values by disease severity, mean (95% CI)				*P* value
		Mild	Moderate	Severe	Critical	
White blood cell count, $\times 10^9$/L	3.5-9.5	5.16 (4.47 to 5.86)	4.73 (4.39 to 5.07)	5.37 (4.1 to 6.64)	5.22 (4.69 to 5.75)	.35
Neutrophil count, $\times 10^9$/L	1.8-6.3	2.98 (2.35 to 3.61)	2.62 (2.38 to 2.87)	3.95 (2.64 to 5.26)	3.83 (3.31 to 4.35)	<.001
Lymphocyte count, $\times 10^9$/L	1.1-3.2	1.55 (1.33 to 1.77)	1.58 (1.41 to 1.74)	1.02 (0.78 to 1.26)	0.96 (0.82 to 1.09)	<.001
Platelet count, $\times 10^9$/L	125-350	209.2 (182.7 to 235.7)	208.1 (194.4 to 221.7)	211.9 (169.9 to 253.9)	186.9 (171.0 to 202.7)	.19
C-reactive protein, mg/L	0-5	4.99 (1.02 to 8.95)	9.21 (5.87 to 12.55)	31.1 (13.98 to 48.22)	43.24 (32.70 to 53.77)	<.001
Alanine aminotransferase, U/L	9-50	24.93 (15.78 to 34.07)	22.15 (17.26 to 27.05)	29.67 (20.58 to 38.75)	28.64 (23.13 to 34.15)	.28
Aspartate aminotransferase, U/L	15-40	24.85 (20.06 to 29.64)	22.85 (20.40 to 25.30)	29.61 (22.93 to 36.29)	33.24 (28.66 to 37.82)	<.001
Urea, mmol/L	3.1-8.0	3.71 (3.11 to 4.3)	4.05 (3.79 to 4.3)	6.17 (4.6 to 7.74)	5.74 (4.36 to 7.12)	.003
Creatinine, µmol/L	57-97	52.41 (46.13 to 58.68)	54.44 (51.71 to 57.18)	68.06 (54.75 to 81.36)	94.14 (50.26 to 138.0)	.1
Lactate dehydrogenase, U/L	120-250	186.2 (165.1 to 207.2)	193.8 (182.4 to 205.2)	242.1 (201.1 to 283.0)	294.1 (262.7 to 325.5)	<.001
Creatine kinase, U/L	50-310	64.96 (24.26 to 105.7)	71.03 (58.94 to 83.11)	87.67 (45.81 to 129.5)	123.4 (83.10 to 163.7)	.03
CD3 (%)	56-86	72.15 (68.45 to 75.84)	70.8 (68.62 to 72.99)	67.68 (61.38 to 73.99)	58.19 (54.36 to 62.03)	<.001
CD3 count, /µL	723-2737	1124 (923.0 to 1326)	1036 (933.0 to 1140)	658.4 (465.9 to 851.0)	577.1 (460.1 to 694.1)	<.001
CD4 (%)	33-58	41.71 (38.81 to 44.61)	41.19 (39.18 to 43.20)	41.85 (37.44 to 46.26)	31.91 (29.23 to 34.60)	<.001
CD4 count, /µL	404-1612	669.9 (573.6 to 766.3)	610.8 (544.8 to 676.7)	402.7 (288.5 to 517.0)	315 (247.2 to 382.7)	<.001
CD8 (%)	13-39	27.06 (24.62 to 29.51)	25.75 (24.41 to 27.08)	22.64 (18.63 to 26.64)	23.98 (20.97 to 26.98)	.21
CD8 count, /µL	220-1129	444.3 (355.1 to 533.5)	369.8 (328.7 to 410.8)	227.6 (151.7 to 303.4)	237.9 (183.9 to 291.9)	<.001
CD4/CD8 ratio	0.9-2.0	1.64 (1.43 to 1.85)	1.93 (1.46 to 2.4)	2.55 (1.28 to 3.81)	1.63 (1.38 to 1.87)	.17
CD19 (%)	5-22	14.04 (11.53 to 16.55)	13.21 (12.00 to 14.43)	21.82 (15.30 to 28.34)	15.46 (13.39 to 17.54)	<.001
CD19 count, /µL	80-616	210 (169.4 to 250.5)	197.1 (163.0 to 231.2)	178.2 (120.0 to 236.4)	129 (108.4 to 149.6)	.004
CD16+56 (%)	5-26	17.66 (3.458 to 31.86)	13.09 (11.22 to 14.97)	8.92 (6.227 to 11.61)	23.85 (20.21 to 27.49)	<.001
CD16+56 count, /µL	84-724	161.3 (82.20 to 240.3)	183.5 (150.4 to 216.5)	82.78 (55.13 to 110.4)	190 (154.7 to 225.3)	.04
IgG, g/L	8-16	11.83 (10.32 to 13.34)	11.81 (11.01 to 12.62)	18.08 (15.44 to 20.73)	13.79 (12.64 to 14.93)	<.001
IgM, g/L	0.4-3.45	1.19 (1.03 to 1.35)	1.2 (1.07 to 1.33)	1.08 (0.74 to 1.41)	1.13 (1.0 to 1.27)	.79
IgA, g/L	0.76-3.9	1.9 (1.55 to 2.25)	4.49 (–0.42 to 9.39)	1.83 (1.44 to 2.22)	2.23 (1.94 to 2.52)	.71
IgE, IU/mL	<100	71.7 (27.92 to 115.5)	88.89 (41.23 to 136.6)	59.83 (8.444 to 111.2)	84.1 (54.48 to 113.7)	.89

Variable	Normal range	Values by disease severity, mean (95% CI)				*P* value
		Mild	Moderate	Severe	Critical	
Complement C3, g/L	0.81-1.6	0.78 (0.72 to 0.84)	0.83 (0.79 to 0.87)	0.89 (0.81 to 0.98)	0.88 (0.82 to 0.94)	.09
Complement C4, g/L	0.1-0.4	0.2 (0.17 to 0.23)	0.24 (0.22 to 0.27)	0.22 (0.19 to 0.25)	0.28 (0.25 to 0.31)	.003

Screening for Independent Risk Factors and Constructing a Predictive Nomogram

The 175 patients were divided into a Mild group and a Severe group according to disease severity. Patients with mild and moderate illness were included in the Mild Group (nonsevere illness), and patients with severe and critical illness were included in the Severe Group. A total of 18 variables were considered to be risk factors as revealed by the univariate analysis (Multimedia Appendix 1). We performed LASSO Cox regression to further select variables (Figure 2), followed by multivariate Cox regression analysis. Multimedia Appendix 2 shows the results of the multivariate analysis. With older age (odds ratio [OR] 1.035, 95% CI 1.017-1.054); higher levels of blood CK (OR 1.002, 95% CI 1.0003-1.0039), CD8 % (OR 1.007, 95% CI 1.004-1.012), and C3 (OR 6.93, 95% CI 1.945-24.691); and lower levels of CD4 (OR 0.995, 95% CI 0.992-0.998) and CD8 (OR 0.881, 95% CI 0.835-0.931), a patient would be more likely to progress to severe disease within three weeks of disease onset.

Figure 2. (A) The log (λ) values of the 18 parameters as shown using least absolute shrinkage and selection operator (LASSO) coefficient profiles. (B) The most suitable log (λ) values for variable selection based on the LASSO Cox regression.

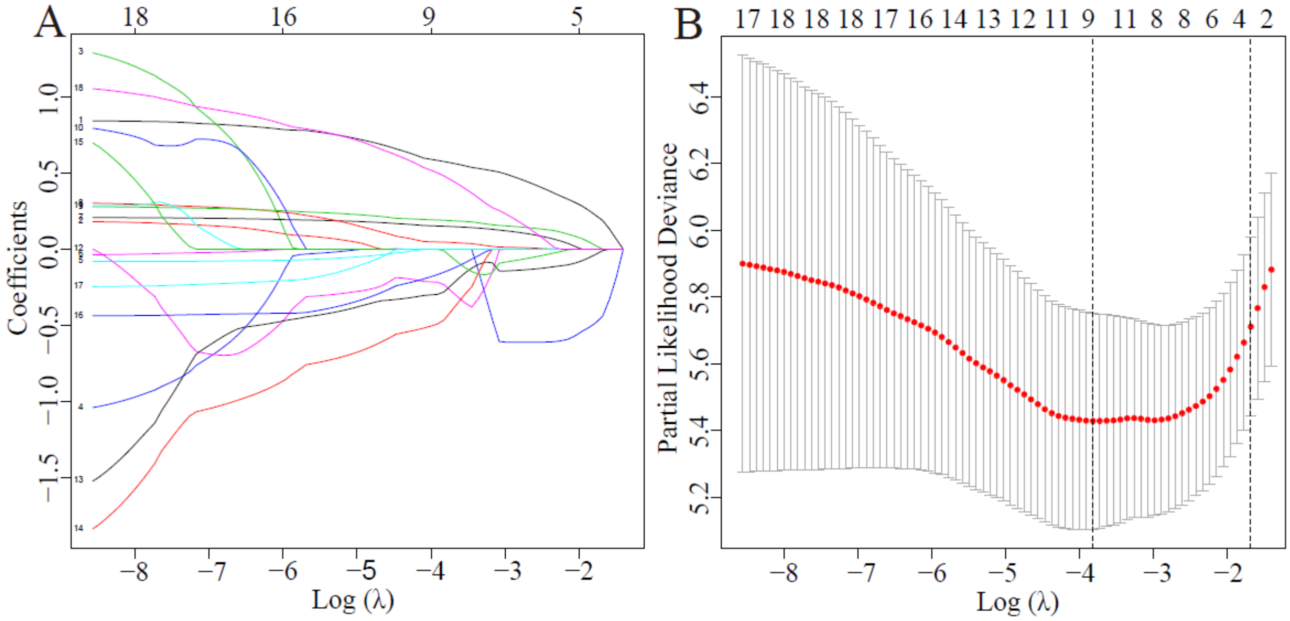
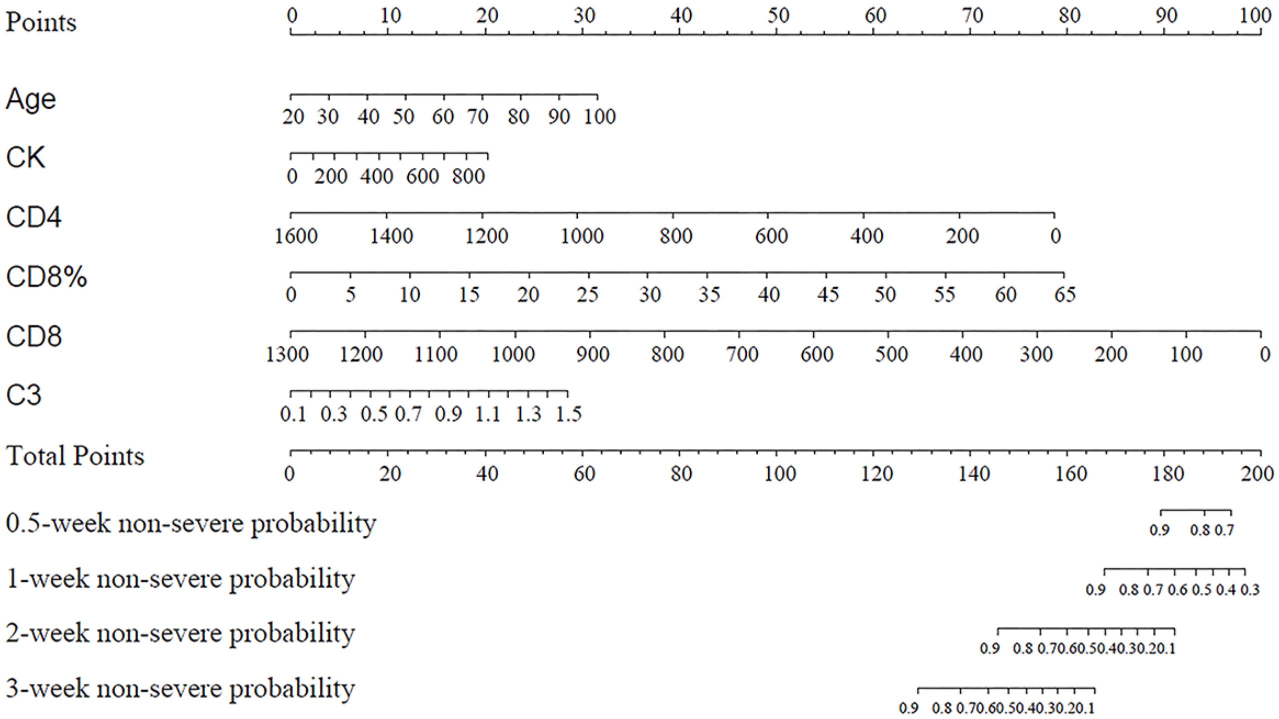


Figure 3 shows the nomogram of the multivariate Cox regression model. All independent risk factors have their own lines in the nomogram, with each receiving a point according to value. The total points are added and match the probability of COVID-19 progression. An example is shown in Multimedia Appendix 3, and the calibration curves are shown in Multimedia Appendix 4. The apparent value is close to the ideal value, which indicates good predictive performance of the model. Multimedia Appendix 5 shows the areas under the ROC curves of the nomogram. The curves proved that this model obtained in the study had good predictive performance for COVID-19 progression within three weeks.

Figure 3. Establishment of the nomogram based on the six independent risk factors resulting from multivariate Cox regression to predict the 0.5-, 1-, 2-, and 3-week nonsevere probabilities for patients with COVID-19 in the developing set. Each selected variable is represented by a line in the figure. According to the value, each variable receives 1 point. The total points are added for each variable and matched with the probability of COVID-19 progression.



Discussion

Principal Results

In this report, we found that age, CK level, CD4 count, CD8 count, CD8 %, and C3 count were the independent risk factors for the progression of COVID-19. In addition, we established a nomogram based on the six independent risk factors to predict the probability of 0.5-, 1-, 2-, and 3-week nonsevere probability.

Since the first case of unexplained pneumonia was reported in the city of Wuhan [15], the disease has spread rapidly worldwide [16,17]. COVID-19 was recognized by the WHO as an international emergency public health event [18]. Current epidemiological studies have shown that the most common symptom of patients with COVID-19 before and after consultation is fever [9,15,19]. In this study, fever was identified in 123/175 patients (70.3%) when they were hospitalized. A total of 76/175 participants in this study (43.4%) were severely or critically ill, which accounts for the much higher rate of severity than that reported by Guan et al [19]. This may be due to the insufficient number of hospital beds and the fact that patients with severe conditions were preferentially admitted. Although the fatality rate of SARS-CoV-2 appears to be lower than that of SARS-CoV or MERS-CoV, the outcomes of SARS-CoV-2 patients are worse once the disease enters the severe stage. A retrospective study showed a 61.5% mortality rate in patients with severe COVID-19 [20]. If patients with high risk factors to progress to severe or critical illness can be screened out in a timely fashion for early intervention, the proportion of severe or critically ill patients and their mortality may be reduced significantly.

In this study, 175 patients were divided into a Mild group (patients with mild and moderate illness) and a Severe group (patients with severe and critical illness). A total of 33 variables were included in this study, including age and clinical laboratory parameters. Indices including age, CD4 count, CD8 count, CD8 %, C3 count, and CK level were filtered out using LASSO and multivariate Cox regression. The indices were considered to be independent risk factors that affect COVID-19 progression. It was reported in a study involving 1099 COVID-19 patients that severe patients were typically seven years older than nonsevere patients (median) and that older patients with COVID-19 were more likely to progress to severe illness [19]. Research has shown that T cells are reduced and eventually fail in COVID-19 patients [21]. The results of this study are similar to the two findings mentioned above.

Because no specific medicine or vaccine has been made available for the treatment of SARS-CoV-2 infection to date [22], it is necessary to predict independent risk factors for the early detection of potential patients with severe COVID-19 and provide early intervention. Based on this research, age, myocardial function, and immune system and complement system function are key factors that impact COVID-19 progression. This study presents a nomogram that may be helpful to clinical physicians. Early intervention and supportive treatment for patients whose age and CK, CD4, CD8 and C3 values are in high-risk ranges may have important significance in reducing the severity and mortality of COVID-19.

Limitations

This study has some limitations. First, only 175 cases were included in the construction of the model that was used to screen for independent risk factors. Second, this is an observational study; therefore, it cannot directly lead to causal conclusions. Third, some patients had severe underlying diseases before becoming infected with the virus; therefore, there may be bias in the calculation of the nonsevere patients' survival times.

Conclusions

The COVID-19 outbreak quickly spread worldwide after it was first discovered in Wuhan. Currently, no specific medicine is available for the treatment of SARS-CoV-2 infection. It is necessary to establish a new predictive model that can be used to screen potential critical patients and provide early intervention. We presented a nomogram, compiled through the use of LASSO regression and multivariate Cox regression, which considers various clinical risk factors in evaluating the probability of COVID-19 progression. This nomogram may help clinical physicians prevent COVID-19 progression.

Acknowledgments

We thank all the patients involved in the study and their families. We thank all the authors for their contributions. This work was supported by the National Natural Science Foundation of China (81770095, 81700093).

Authors' Contributions

TF, ZL, RX, XS, and QG collected the data. TF, BH, BS, SY, ZH, and QG analyzed the data. TF, SY, and QG wrote the manuscript. TF, BH, and QG contributed to the figures. WJ, LZ, DL, RH, HM, WL, HF, and QG contributed to the interpretation of the data.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Univariate Cox regression for independent risk factors in COVID-19 progression.
[DOC File , 47 KB - medinform_v8i9e19588_app1.doc]

Multimedia Appendix 2
Multivariate Cox regression for independent risk factors in COVID-19 progression.
[DOC File , 39 KB - medinform_v8i9e19588_app2.doc]

Multimedia Appendix 3
Instructions for using the nomogram.
[DOC File , 89 KB - medinform_v8i9e19588_app3.doc]

Multimedia Appendix 4
Calibration curves for evaluating the predictive performance of nomogram.
[PNG File , 265 KB - medinform_v8i9e19588_app4.png]

Multimedia Appendix 5
Receiver operating characteristic curves for evaluating the predictive performance of the nomogram.
[PNG File , 138 KB - medinform_v8i9e19588_app5.png]

References

1. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med 2020 Mar 26;382(13):1199-1207 [FREE Full text] [doi: 10.1056/NEJMoa2001316] [Medline: 31995857]
2. Bao Y, Sun Y, Meng S, Shi J, Lu L. 2019-nCoV epidemic: address mental health care to empower society. Lancet 2020 Feb 22;395(10224):e37-e38 [FREE Full text] [doi: 10.1016/S0140-6736(20)30309-3] [Medline: 32043982]
3. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020 Feb 15;395(10223):497-506 [FREE Full text] [doi: 10.1016/S0140-6736(20)30183-5] [Medline: 31986264]
4. Paules CI, Marston HD, Fauci AS. Coronavirus Infections-More Than Just the Common Cold. JAMA 2020 Jan 23;323(8):707-708. [doi: 10.1001/jama.2020.0757] [Medline: 31971553]
5. Parry J. Wuhan: Britons to be evacuated as scientists estimate 44 000 cases of 2019-nCOV in the city. BMJ 2020 Jan 29;368:m351. [doi: 10.1136/bmj.m351] [Medline: 31996342]

6. Ohannessian R, Duong TA, Odone A. Global Telemedicine Implementation and Integration Within Health Systems to Fight the COVID-19 Pandemic: A Call to Action. JMIR Public Health Surveill 2020 Apr 02;6(2):e18810 [FREE Full text] [doi: 10.2196/18810] [Medline: 32238336]
7. Liao Q, Yuan J, Dong M, Yang L, Fielding R, Lam WWT. Public Engagement and Government Responsiveness in the Communications About COVID-19 During the Early Epidemic Stage in China: Infodemiology Study on Social Media Data. J Med Internet Res 2020 May 26;22(5):e18796 [FREE Full text] [doi: 10.2196/18796] [Medline: 32412414]
8. Sesagiri Raamkumar A, Tan SG, Wee HL. Measuring the Outreach Efforts of Public Health Authorities and the Public Response on Facebook During the COVID-19 Pandemic in Early 2020: Cross-Country Comparison. J Med Internet Res 2020 May 19;22(5):e19334 [FREE Full text] [doi: 10.2196/19334] [Medline: 32401219]
9. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA 2020 Feb 07;323(11):1061-1069 [FREE Full text] [doi: 10.1001/jama.2020.1585] [Medline: 32031570]
10. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 2020 Feb 29;395(10225):689-697 [FREE Full text] [doi: 10.1016/S0140-6736(20)30260-9] [Medline: 32014114]
11. Phan LT, Nguyen TV, Luong QC, Nguyen TV, Nguyen HT, Le HQ, et al. Importation and Human-to-Human Transmission of a Novel Coronavirus in Vietnam. N Engl J Med 2020 Feb 27;382(9):872-874 [FREE Full text] [doi: 10.1056/NEJMc2001272] [Medline: 31991079]
12. Mo P, Xing Y, Xiao Y, Deng L, Zhao Q, Wang H, et al. Clinical characteristics of refractory COVID-19 pneumonia in Wuhan, China. Clin Infect Dis 2020 Mar 16:3725 [FREE Full text] [doi: 10.1093/cid/ciaa270] [Medline: 32173725]
13. Han Q, Lin Q, Jin S, You L. Coronavirus 2019-nCoV: A brief perspective from the front line. J Infect 2020 Apr;80(4):373-377 [FREE Full text] [doi: 10.1016/j.jinf.2020.02.010] [Medline: 32109444]
14. Clinical management of severe acute respiratory infection when novel coronavirus (nCoV) infection is suspected: interim guidance. World Health Organization. 2020 Jan 28. URL: https://apps.who.int/iris/handle/10665/330893 [accessed 2020-01-31]
15. Young BE, Ong SWX, Kalimuddin S, Low JG, Tan SY, Loh J, Singapore 2019 Novel Coronavirus Outbreak Research Team. Epidemiologic Features and Clinical Course of Patients Infected With SARS-CoV-2 in Singapore. JAMA 2020 Mar 03;323(15):1488-1494 [FREE Full text] [doi: 10.1001/jama.2020.3204] [Medline: 32125362]
16. Jernigan DB, CDC COVID-19 Response Team. Update: Public Health Response to the Coronavirus Disease 2019 Outbreak - United States, February 24, 2020. MMWR Morb Mortal Wkly Rep 2020 Feb 28;69(8):216-219 [FREE Full text] [doi: 10.15585/mmwr.mm6908e1] [Medline: 32106216]
17. Kraemer MUG, Yang C, Gutierrez B, Wu C, Klein B, Pigott DM, Open COVID-19 Data Working Group, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science 2020 May 01;368(6490):493-497 [FREE Full text] [doi: 10.1126/science.abb4218] [Medline: 32213647]
18. Coronavirus disease (COVID-19) pandemic. World Health Organization. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019 [accessed 2020-03-11]
19. Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, China Medical Treatment Expert Group for Covid-19. Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med 2020 Apr 30;382(18):1708-1720 [FREE Full text] [doi: 10.1056/NEJMoa2002032] [Medline: 32109013]
20. Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med 2020 May;8(5):475-481 [FREE Full text] [doi: 10.1016/S2213-2600(20)30079-5] [Medline: 32105632]
21. Ling Y, Xu S, Lin Y, Tian D, Zhu Z, Dai F, et al. Persistence and clearance of viral RNA in 2019 novel coronavirus disease rehabilitation patients. Chin Med J (Engl) 2020 May 05;133(9):1039-1043 [FREE Full text] [doi: 10.1097/CM9.0000000000000774] [Medline: 32118639]
22. de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol 2016 Aug;14(8):523-534 [FREE Full text] [doi: 10.1038/nrmicro.2016.81] [Medline: 27344959]

Abbreviations

CK: creatine kinase
LASSO: least absolute shrinkage and selection operator
MERS-CoV: Middle East respiratory syndrome coronavirus
NHCPRC: National Health Commission of the People’s Republic of China
OR: odds ratio
ROC: receiver operating characteristic
RT-PCR: reverse transcriptase–polymerase chain reaction
SARS-CoV: severe acute respiratory syndrome coronavirus
WHO: World Health Organization

Edited by G Eysenbach; submitted 24.04.20; peer-reviewed by Y Wu, P Banik; comments to author 12.06.20; revised version received 21.06.20; accepted 10.08.20; published 08.09.20.

Please cite as:
Fan T, Hao B, Yang S, Shen B, Huang Z, Lu Z, Xiong R, Shen X, Jiang W, Zhang L, Li D, He R, Meng H, Lin W, Feng H, Geng Q
Nomogram for Predicting COVID-19 Disease Progression Based on Single-Center Data: Observational Study and Model Development
JMIR Med Inform 2020;8(9):e19588
URL: http://medinform.jmir.org/2020/9/e19588/
doi:10.2196/19588
PMID:32866109

Viewpoint

Applying Blockchain Technology to Address the Crisis of Trust During the COVID-19 Pandemic

Anjum Khurshid[1], MD, PhD

[1] The University of Texas at Austin, Austin, TX, United States

Corresponding Author:
Anjum Khurshid, MD, PhD
The University of Texas at Austin
1701 Trinity Street
Austin, TX, 78712
United States
Phone: 1 5124955225
Email: anjum.khurshid@austin.utexas.edu

Abstract

Background: The widespread death and disruption caused by the COVID-19 pandemic has revealed deficiencies of existing institutions regarding the protection of human health and well-being. Both a lack of accurate and timely data and pervasive misinformation are causing increasing harm and growing tension between data privacy and public health concerns.

Objective: This aim of this paper is to describe how blockchain, with its distributed trust networks and cryptography-based security, can provide solutions to data-related trust problems.

Methods: Blockchain is being applied in innovative ways that are relevant to the current COVID-19 crisis. We describe examples of the challenges faced by existing technologies to track medical supplies and infected patients and how blockchain technology applications may help in these situations.

Results: This exploration of existing and potential applications of blockchain technology for medical care shows how the distributed governance structure and privacy-preserving features of blockchain can be used to create “trustless” systems that can help resolve the tension between maintaining privacy and addressing public health needs in the fight against COVID-19.

Conclusions: Blockchain relies on a distributed, robust, secure, privacy-preserving, and immutable record framework that can positively transform the nature of trust, value sharing, and transactions. A nationally coordinated effort to explore blockchain to address the deficiencies of existing systems and a partnership of academia, researchers, business, and industry are suggested to expedite the adoption of blockchain in health care.

(JMIR Med Inform 2020;8(9):e20477) doi:10.2196/20477

KEYWORDS

blockchain; privacy; trust; contact tracing; COVID-19; coronavirus

Introduction

> *In many ways, it is hard for modern people living in First World countries to conceive of a pandemic sweeping around the world and killing millions of people, and it is even harder to believe that something as common as influenza could cause such widespread illness and death.*
>
> [**Charles River Editors, *The 1918 Spanish Flu Pandemic: The History and Legacy of the World’s Deadliest Influenza Outbreak***]

The COVID-19 Pandemic in the United States and Worldwide

By the end of July 2020, COVID-19 had infected approximately 19 million people worldwide and had caused over 700,000 deaths. The United States is the richest country in the world, with a health expenditure of US $3.5 trillion per year; it has reported the highest number of people infected with COVID-19 (approximately 5 million) as well as the highest number of deaths (>150,000) [1]. Through a lack of reliable data, inability of health care and public health systems to perform active surveillance, inadequate management of needed medical equipment, conflicting information from multiple sources, and limited technology for engagement with patients, the COVID-19 pandemic has clearly demonstrated the failure of existing institutions to protect human health and to avoid widespread suffering.

COVID-19 Is a Crisis of Trust

In the absence of reliable data and accurate information, the suffering due to the COVID-19 crisis has been exacerbated by misinformation, which ranges from warnings of imminent doom to conspiracy theories. The COVID-19 crisis is an information crisis [2]. This crisis has been rightly described as an "infodemic," a term coined in 2002 by Eysenbach [3]. However, the importance of this term has been manifested in this crisis more than ever before by the enormous influence of social media and its role as a source of information for the public about the pandemic [4]. Several months into the greatest public health disaster in a century, with all our technology and information networks, there is still much confusion about the actual prevalence of COVID-19, number of deaths, expected treatments, and best strategies to control the pandemic globally [5,6]. Due to this failure to provide timely, accurate, and reliable information about the infection, the pandemic has worsened the crisis of trust in government institutions and public health agencies [7].

Interestingly, a similar failure of governmental response led to the 2008 financial crisis and precipitated the low trust in government and centralized institutions that has since persisted. Banks, governments, and financial institutions failed the public and left many people exposed and dejected during the financial crisis [8]. This motivated Satoshi Nakamoto, the pseudonym of an unidentified person or group, to write a paper that proposed a new system of establishing trust in financial markets without intermediaries such as governments or banks. This system relied on a distributed ledger technology of peer-to-peer networks and was called blockchain [9]. In 2009, Bitcoin was launched as a decentralized cryptocurrency that bypassed intermediaries such as banks, governments, and large financial institutions and allowed people to transact directly in a secure trust framework [10]. The COVID-19 pandemic promises to inflict even greater hurt and misery to the public than the 2008 financial crisis because it is not only an economic disaster but also a health calamity that has already caused the loss of precious lives worldwide. The role of intermediary organizations has also been unsatisfactory in this situation [11].

Issues of Trust With Existing Institutions

The proponents of the decentralized and distributed system of blockchain technology point to the disadvantages of centralized institutions because these "intermediaries of trust" are slow to respond to changes in the environment, add cost and time to transactions, and adversely affect productivity [12]. Centralized institutions also store data centrally and not only restrict access to these data, thus preventing coordination and efficient sharing of information, but also represent a single point of failure for privacy and security of the information [13]. In 2017, the personal data of more than 143 million customers were exposed through a single breach [14]. According to one source, in the last 10 years, more than 1.4 billion records have been exposed due to government database breaches worldwide [15].

Intermediary institutions are supposed to provide much-needed, trustworthy, and reliable services to society [16]; however, the COVID-19 crisis has exposed the limitations of these institutions with regard to health care [17]. In this time of crisis, both public and private institutions as well as traditional information systems have mostly failed to solve problems related to routine health care delivery [18], including availability of timely data for projections of case numbers [19], identification of high-risk populations [20], tracing contacts of persons with COVID-19 [21], and supply of personal protective equipment (PPE) or inventories of lifesaving drugs [22]. In fact, it has been argued that the number of deaths due to COVID-19 could have been reduced with better access to reliable data [23].

Blockchain and the "Trustless" System

Blockchain has been described as a foundational technology [24] that can dramatically change the paradigm in which social and economic transactions take place. A review of the key characteristics that form the fundamental aspects of blockchain technology may help demonstrate why this technology can be invaluable in addressing some of the issues of mistrust described above. Blockchain technology is based on a "trustless" system, where transactions can be performed among people who do not have any prior relationship yet who can validate the objectivity and principles of the medium in which the transactions occur. This enables transparency of contracts, immutability of data, and accountability of transactions among strangers [25]. Public blockchain networks allow individuals to share their information in complete privacy while maintaining full control of that information. They can also maintain an audit record of each transaction, make it readily available when needed, validate information sources to avoid misinformation, allow tracking of assets as part of the architecture of the network, and provide global connectedness without barriers to flow of information [26,27]. The rules of consensus and validation are transparent, mathematically proven, unbiased, distributed, and objective in nature; these characteristics cannot be ascribed to government or to most key institutions that are handling private information related to the pandemic [28]. The growth of the cryptocurrency market provides proof that the trustless system is a workable solution at a global stage; the COVID-19 pandemic has highlighted the necessity for such a solution [29].

Blockchain in the COVID-19 Pandemic

As a foundational technology that promises to provide new solutions to old problems, blockchain technology is increasingly being applied in innovative ways that are relevant to the challenges created by the COVID-19 pandemic. The failure of existing systems to provide reliable and effective solutions to problems created by this global crisis has highlighted the potential of blockchain applications even more greatly [30]. The crisis has created a unique opportunity to test and develop blockchain-based solutions. It is difficult for health care organizations to implement blockchain technology and adopt its more open, transparent, patient-focused, and robust systems of transaction and information management without more evidence of its effectiveness; however, it is worth testing this technology to develop systems with levels of robustness that current information systems have not been able to achieve. Fortunately, there are already some use cases of blockchain technology that may significantly contribute more effectively to the fight against the COVID-19 pandemic and infodemic in the short run and build capacity to respond to similar health emergencies in the future. We discuss two key examples that relate to medical care directly and where blockchain technology is currently being applied but should be adopted even more widely if proven effective.

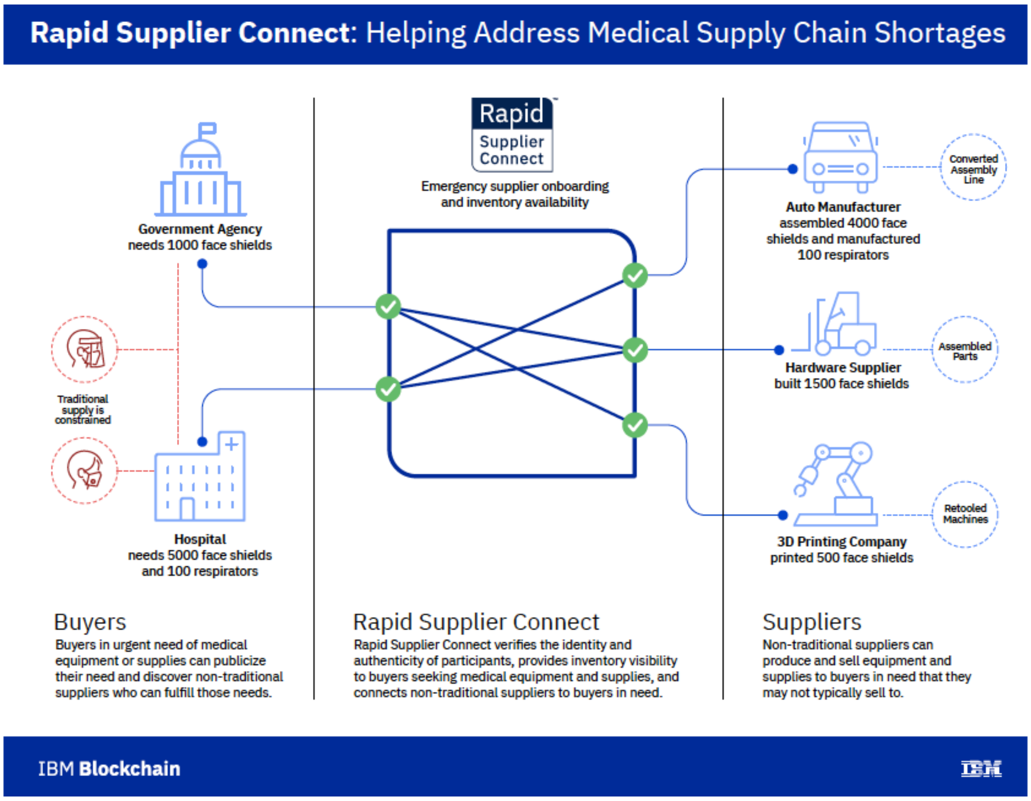
Supply Chains and Blockchain

During the COVID-19 crisis, major supply chain failures have been observed not only for household items such as toilet paper and hand-washing soaps [31] but, more importantly, for PPE and lifesaving ventilators in hospitals and clinics [32]. Blockchain technology provides immutable and distributed ledgers with auditable records, which are ideal for tracking each asset in a supply chain because every actor in the supply chain shares the same information [33]. It is therefore easy to calculate the inventory and the exact stage where assets are in the chain; instant reconciliation can then be achieved without any additional audit or negotiation among the various suppliers and end users. A joint Walmart-IBM project demonstrated how tracking sources of contamination in green vegetables, a task that previously took months, could be achieved within seconds using blockchain [34,35]. Some of the lessons learned from that system are now being applied at the US Food and Drug Agency for counterfeit pharmaceuticals [36]. IBM also designed Rapid Supplier Connect to help with medical supply chains during the COVID-19 pandemic and offered it to health systems and government agencies to help find vendors for medical supplies and PPE [37].

Even during this global crisis, there have been reports of counterfeit medications and poor-quality equipment being sent to organizations and people who are in desperate need of these items [38]. This has led to trust issues in the supply chain [39]. In fact, a report by the Organisation for Economic Co-operation and Development cautioned about increased global trade in fake pharmaceuticals during the COVID-19 pandemic [40]. Blockchain not only provides an efficient way to manage the supply chain but also provides a means to distinguish quality products from counterfeit ones [41]. This is particularly true when items must be moved across international borders, where the levels of information about sources of production and the rules under which the quality checks occur vary greatly. Validation of the quality through peer-to-peer networks such as blockchain can improve trust and decrease unnecessary litigation and disputes [42,43]. An example of such a system is IBM's Trust Your Supplier solution, in which blockchain enables trusted sources of supplier information and digital identity management to reduce the risk of counterfeiting while facilitating onboarding of suppliers and communications between buyers and sellers or suppliers and distributors [44]. As shown in Figure 1, Trust Your Supplier is a permissioned network that limits access to the information on the blockchain and allows for transparency among nodes on the supply chain. Cryptographic security ensures confidentiality of data on the chain, and the immutability of records guarantees that no one party can make changes unilaterally without a consensus.

Another example of the use of blockchain to address the issue of counterfeit drugs is Gcoin (Figure 2) [45].

Figure 1. Schematic of IBM Rapid Supplier Connect [44].



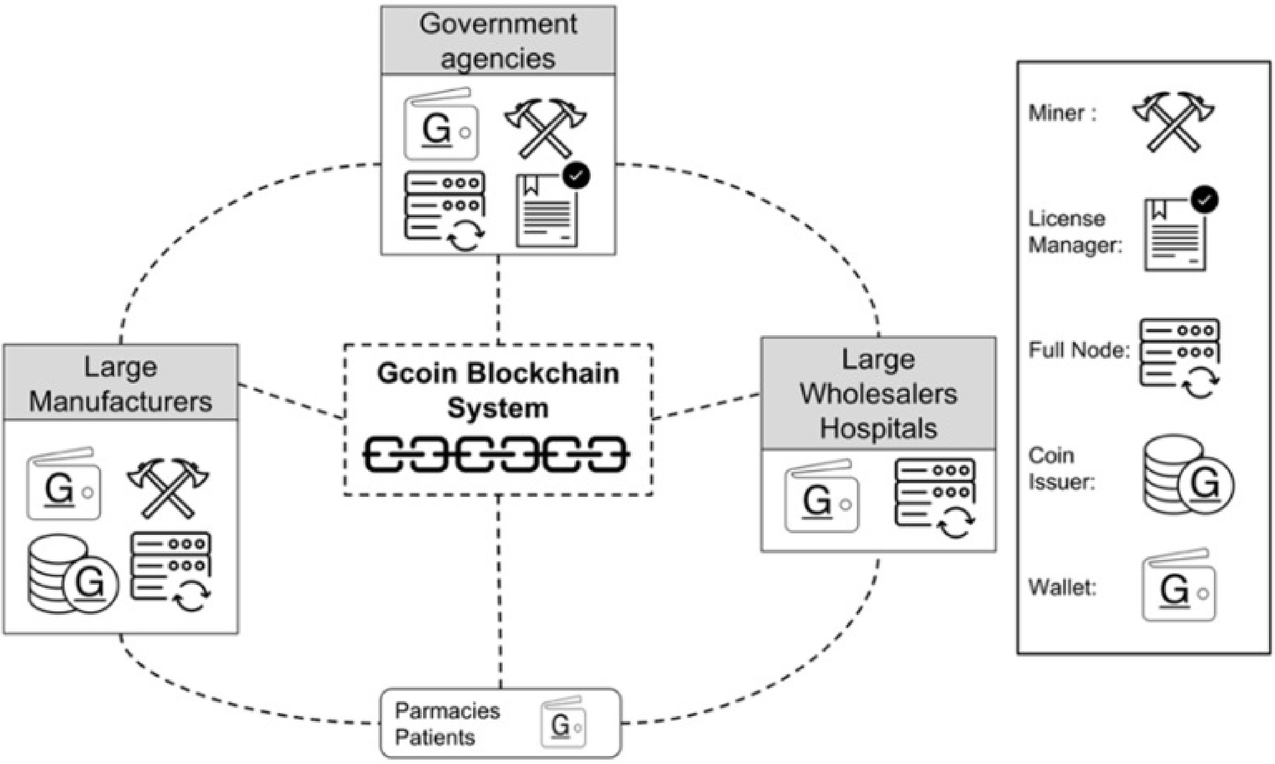
Rapid Supplier Connect: Helping Address Medical Supply Chain Shortages

Buyers
Buyers in urgent need of medical equipment or supplies can publicize their need and discover non-traditional suppliers who can fulfill those needs.

Rapid Supplier Connect
Rapid Supplier Connect verifies the identity and authenticity of participants, provides inventory visibility to buyers seeking medical equipment and supplies, and connects non-traditional suppliers to buyers in need.

Suppliers
Non-traditional suppliers can produce and sell equipment and supplies to buyers in need that they may not typically sell to.

Figure 2. Schematic of the Gcoin blockchain system [45].



By maintaining the suppliers and buyers on a shared ledger (ie, blockchain) and by recording all transactions on the chain as immutable records, it becomes much easier to quickly see the origin of the products and also to search for the total supplies of an item without creating a centralized database. Centralized databases are not only difficult to keep up-to-date during crises but may also present easy targets for hacks and breaches. The trustless system of blockchain can significantly help with reducing supply chain failures, particularly as the pandemic enters the stage where vaccines and lifesaving drugs will need to move across international borders to save lives.

Contact Tracing and Blockchain

For a highly infectious disease such as COVID-19, the ability to promptly trace individuals who have been exposed to an infected person is a beneficial public health strategy that can limit the continuing spread of the infection. As the number of COVID-19 cases continues to rise and surges are occurring worldwide, there is increased realization that social distancing and lockdown measures cannot be indefinitely extended. In many cases, compliance with lockdowns has been difficult to enforce, and coercion and significant resource allocation are needed to enforce it [46]. Scenes of police, army, and other government agencies roaming the streets to enforce closure of businesses and lockdowns are ubiquitous on the internet, with many examples of use of violence and threats against citizens [47]. With no other remedy to control the spread of the infection, widespread social distancing and lockdowns are blunt policy levers that are being used by most governments. However, the devastating consequences of this policy on economic activity, social interactions, mental health, and health-seeking behaviors are already visible, leading to the realization that this is not a sustainable strategy. Managing infections at an individual level and taking precautions in close circles that are at risk of infection will be needed to allow some level of normalcy to return. Contact tracing is an important tool in this regard [48].

Many states in the United States and countries worldwide have quickly developed and adopted contact tracing applications since the beginning of the pandemic, with mixed success [49]. As health care delivery systems were overwhelmed with testing and treatment needs, symptom-tracking apps were used to help individuals assess their risk of COVID-19 and their need for testing. Most of these apps require notifications or data sharing with public health or health care organizations to coordinate testing and treatment. With limited supplies of tests, only people with symptoms are asked to be tested. If an individual tests positive for COVID-19, the next steps are tedious public health investigations to identify who else is at risk of being infected by the individual. Traditionally, public health agencies have used contact tracers, who provide this information to potential contacts by telephone or mail and ask them to be tested. This approach is successful when the numbers are not of the magnitude seen in the COVID-19 pandemic. Also, these strategies were developed in the pre–mobile phone era. Due to the relative ease of mobile app development and ubiquitous access to mobile phones, contact tracing apps for COVID-19 are an obvious public health solution.

Countries such as Norway, South Korea, China, Singapore, Germany, and Qatar have developed, encouraged, or enforced the use of contact tracing apps as a public health strategy. Different technologies have been used to provide contact tracing; these technologies use various features built into mobile phone platforms, such as GPS, Bluetooth, Wi-Fi, and quick response (QR) codes. Very quickly, however, individuals and advocacy groups raised concerns regarding data privacy, confidentiality, and security [50]. Simultaneously, numerous reports of hacks, bugs, and misuse of data started to emerge. Some of these apps have been banned or discontinued. This lack of trust is not only reserved for governments but is also the reason that the Google-Apple contact tracing features have been regarded with concern in England and other countries [51]. However, the need for contact tracing and controlling infection by identifying people at risk has not decreased. This is a situation where the trustless system of blockchain technology can provide potential answers on how to balance public health needs with privacy concerns [52].

Blockchain enables information to be collected from individuals without identifying them by using a system of public and private keys [53]. For example, the BeepTrace system uses blockchain to provide encrypted and anonymized personal identification while allowing regulators and health care providers to contact people at risk of infection due to contact with an infected person. The system uses two chains and a public key generated by the government or a public entity to generate location data but also generates a diagnostician key to verify test results. The infected person gives consent to the diagnosing entity, which participates in the blockchain to verify results; however, the government cannot identify the individual. Notifications can be sent to the individual using a separate chain [54].

Previously, the same privacy and data sharing scheme was also proposed in other blockchain-based applications [25]. The key is that through anonymizing and cryptography, a blockchain-based contact tracing app ensures individual privacy while allowing public health departments to contact people who may have been exposed to SARS-CoV-2, the virus that causes COVID-19, through an infected person. These features of security, privacy, trust, transparency, and efficiency are built into the architecture of blockchain and have been difficult to replicate or develop reliably in other applications.

Countries such as Taiwan and South Korea have shown that a robust system of contact tracing can control the spread of infection while allowing normal life to continue for healthy people who are at low risk for infection [55]. However, concerns about privacy and security may limit the implementation of such strategies in different parts of the world, particularly in the United States, which has the highest numbers of cases and deaths [56]. Blockchain technologies that enable individuals to share their personal information in a secure manner with public health agencies without revealing their identity or contributing that information to a centralized government or corporate database may help identify people who come into contact with a patient who has tested positive for COVID-19. This can be achieved through public health agencies or through peer-to-peer notifications, where only the positive status can be shared without sharing other medical or personal data [57]. The capability to track individuals who are positive for COVID-19 and to check their seropositive status for infection may be used as a key tool to enable more responsible reopening of the economy without causing a surge in cases. As we develop vaccinations or develop herd immunity for the infection, blockchain technology may also be used to issue health certifications that can be verified easily by employers and public health agencies to validate the status of an individual [58].

Blockchain technology may be applied in many other aspects related to the long-term fight against the COVID-19 pandemic, such as approval of insurance status within seconds rather than the multiple contacts needed today to verify insurance [59], patient identification at the point of care that does not require filling out multiple forms and carrying documents to appointments with physicians [60], or conducting research without increasing the risk to privacy of individuals [61]. Artificial intelligence [62], the Internet of Things [63], and 3D printing [64] using immutable and verified instructions through blockchain are other technologies that may be greatly helpful in fighting pandemics such as COVID-19 in the future.

Future Considerations

The devastation and suffering caused by the COVID-19 crisis should trigger a resolve to build better systems of data, trust, and transactions to track, respond, and control such pandemics in the future. While blockchain technology holds great promise, and solutions to systemic failures of our current health care, public health, and policy institutions are already being developed, the widespread adoption of this technology requires planning and execution. A national policy agenda to immediately consider how blockchain applications may help enable safe and effective responses to the COVID-19 pandemic will help expedite the acceptance, adoption, and implementation of this technology to improve our flawed systems of health care data and health-related transactions.

Major blockchain and software companies are already in the process of creating a decentralized governance system to create international standards, such as World Wide Web Consortium (W3C) and internet protocols. Hyperledger, Ethereum, BankChain, and R3 are all examples of such consortium-building efforts [65]. Consensus on protocols and rules of business among competitors and collaborators leads to more effective, egalitarian, and implementable rules, which have helped improve the interoperability and scalability of wireless and internet technologies. If governments and large private corporations actively participate and encourage this collaborative global governance rather than considering it a threat to their own hegemonic authority, health systems worldwide will be much better prepared for future health crises such as the COVID-19 pandemic.

Finally, research and development is needed to build and test robust use cases for blockchain applications. University and research institutions should partner with industry and business. Such collaborations are rare and must be established widely to expedite the adoption of blockchain technologies in health. Development and funding of blockchain implementation laboratories in universities and medical schools will help promote such industry-academia partnerships and provide stronger and more reliable evidence to evaluate the impact of blockchain technology in health care. Both health care and blockchain technology require interdisciplinary teams to work together to solve problems. Blockchain laboratories in academic medical centers and public universities can provide platforms for cooperation and creative problem-solving that will help in the fight against pandemics such as COVID-19.

Conclusion

Blockchain technology relies on a distributed, robust, secure, privacy-preserving, and immutable record-keeping framework that can positively transform the nature of trust, value sharing, and transactions. The COVID-19 crisis has highlighted the failure of current systems of trust and data sharing. While this pandemic presents a clear and serious danger to our way of life, it also provides unique opportunities to apply and test new technologies that may help transform our capabilities to fight this pandemic and, in the process, establish a more efficient, democratic, and secure system to respond to future pandemics.

Acknowledgments

The author wants to thank Dr Robert John Bautista for his help in formatting the paper for submission. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

None declared.

References

1. Worldometer. COVID-19 Coronavirus Pandemic. URL: https://www.worldometers.info/coronavirus/ [accessed 2020-04-27]
2. Xie B, He D, Mercer T, Wang Y, Wu D, Fleischmann KR, et al. Global health crises are also information crises: A call to action. J Assoc Inf Sci Technol 2020 Mar 13:A [FREE Full text] [doi: 10.1002/asi.24357] [Medline: 32427189]
3. Eysenbach G. How to Fight an Infodemic: The Four Pillars of Infodemic Management. J Med Internet Res 2020 Jun 29;22(6):e21820 [FREE Full text] [doi: 10.2196/21820] [Medline: 32589589]
4. Zarocostas J. How to fight an infodemic. Lancet 2020 Feb;395(10225):676. [doi: 10.1016/s0140-6736(20)30461-x]
5. Madrigal AC, Meyer R. ‘How Could the CDC Make That Mistake?’. The Atlantic. 2020 May 21. URL: https://www.theatlantic.com/health/archive/2020/05/cdc-and-states-are-misreporting-covid-19-test-data-pennsylvania-georgia-texas/611935/ [accessed 2020-06-07]
6. Shalby C. Serious breakdown in California systems causes inaccurate coronavirus numbers. The Los Angeles Times. 2020 Aug 05. URL: https://www.latimes.com/california/story/2020-08-05/coronavirus-test-results-collecting-hampering-pandemic-response [accessed 2020-08-14]
7. Mystal E. The lesson of the coronavirus? There's no one left to trust. The Nation. 2020 Apr 04. URL: https://www.thenation.com/article/society/coronavirus-reopen-trust/ [accessed 2020-07-31]
8. Earle TC. Trust, confidence, and the 2008 global financial crisis. Risk Anal 2009 Jun;29(6):785-792. [doi: 10.1111/j.1539-6924.2009.01230.x] [Medline: 19392679]
9. Nakamoto S. Bitcoin: A peer-to-peer electronic cash system. Bitcoin.org. 2008. URL: https://bitcoin.org/bitcoin.pdf [accessed 2020-07-26]
10. Böhme R, Christin N, Edelman B, Moore T. Bitcoin: Economics, Technology, and Governance. J Econ Perspect 2015 May 01;29(2):213-238 [FREE Full text] [doi: 10.1257/jep.29.2.213]
11. Tufecki Z. The WHO Shouldn’t Be a Plaything for Great Powers. The Atlantic. 2020 Apr 16. URL: https://www.theatlantic.com/health/archive/2020/04/why-world-health-organization-failed/610063/ [accessed 2020-05-11]
12. Schenker JL. The distribution of trust. The Innovator. 2017 Nov 12. URL: https://innovator.news/blockchain-the-distribution-of-trust-4d36b7bd2508 [accessed 2020-05-11]
13. Atlam HF, Alenezi A, Alassafi MO, Wills GB. Blockchain with Internet of Things: Benefits, Challenges, and Future Directions. IJISA 2018 Jun 08;10(6):40-48. [doi: 10.5815/ijisa.2018.06.05]
14. Bernard TS, Hsu T, Perlroth N, Lieber R. Equifax Says Cyberattack May Have Affected 143 Million in the U.S. The New York Times. 2017 Sep 07. URL: https://www.nytimes.com/2017/09/07/business/equifax-cyberattack.html [accessed 2020-05-11]
15. Tapscott D, Tapscott A. Blockchain solutions in pandemics: A call for innovation and transformation in public health. Blockchain Research Institute. 2020 Apr 07. URL: https://app.hubspot.com/documents/5052729/view/72133013?accessId=54ef98 [accessed 2020-09-10]
16. Freitag M, Bühlmann M. Crafting Trust. Comp Polit Stud 2009 Mar 05;42(12):1537-1566. [doi: 10.1177/0010414009332151]
17. Blumenthal D, Fowler EJ, Abrams M, Collins SR. Covid-19 - Implications for the Health Care System. N Engl J Med 2020 Jul 22. [doi: 10.1056/NEJMsb2021088] [Medline: 32706956]
18. Colglazier EW. Response to the COVID-19 Pandemic: Catastrophic Failures of the Science-Policy Interface. Science & Diplomacy. 2020 Apr 09. URL: http://www.sciencediplomacy.org/editorial/2020/response-covid-19-pandemic-catastrophic-failures-science-policy-interface [accessed 2020-05-11]
19. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]

20. Holmes L, Enwere M, Williams J, Ogundele B, Chavan P, Piccoli T, et al. Black-White Risk Differentials in COVID-19 (SARS-COV2) Transmission, Mortality and Case Fatality in the United States: Translational Epidemiologic Perspective and Challenges. Int J Environ Res Public Health 2020 Jun 17;17(12):4322 [FREE Full text] [doi: 10.3390/ijerph17124322] [Medline: 32560363]
21. Dar AB, Lone AH, Zahoor S, Khan AA, Naaz R. Applicability of Mobile Contact Tracing in Fighting Pandemic (COVID-19): Issues, Challenges and Solutions. SSRN Journal 2020 Apr 01:preprint. [doi: 10.2139/ssrn.3683404]
22. Ranney ML, Griffeth V, Jha AK. Critical Supply Shortages — The Need for Ventilators and Personal Protective Equipment during the Covid-19 Pandemic. N Engl J Med 2020 Apr 30;382(18):e41. [doi: 10.1056/nejmp2006141]
23. Ioannidis JPA. A fiasco in the making? As the coronavirus pandemic takes hold, we are making decisions without reliable data. Stat News. 2020 Mar 17. URL: https://www.statnews.com/2020/03/17/a-fiasco-in-the-making-as-the-coronavirus-pandemic-takes-hold-we-are-making-decisions-without-reliable-data/ [accessed 2020-05-11]
24. Iansiti M, Lakhani KR. The Truth About Blockchain. Harvard Business Review. 2017. URL: https://hbr.org/2017/01/the-truth-about-blockchain [accessed 2020-05-11]
25. Xia Q, Sifah EB, Asamoah KO, Gao J, Du X, Guizani M. MeDShare: Trust-Less Medical Data Sharing Among Cloud Service Providers via Blockchain. IEEE Access 2017;5:14757-14767. [doi: 10.1109/ACCESS.2017.2730843]
26. Kuo TT, Kim HE, Ohno-Machado L. Blockchain distributed ledger technologies for biomedical and health care applications. J Am Med Inform Assoc 2017 Nov 01;24(6):1211-1220 [FREE Full text] [doi: 10.1093/jamia/ocx068] [Medline: 29016974]
27. Tapscott D, Tapscott A. Blockchain Revolution: How the Technology Behind Bitcoin Is Changing Money, Business, and the World. New York, NY: Portfolio; Jan 01, 2016.
28. Barrett J, Jones C, Reed E. Mayors Move to Address Racial Disparity in Covid-19 Deaths. The Wall Street Journal. 2020 Apr 08. URL: https://www.wsj.com/articles/black-hispanic-new-yorkers-account-for-disproportionate-number-of-coronavirus-deaths-11586359447 [accessed 2020-04-11]
29. Ferreira P, Periera E. Contagion Effect in Cryptocurrency Market. J Risk Financial Manag 2019 Jul 10;12(3):115. [doi: 10.3390/jrfm12030115]
30. Wladawsky-Berger I. Blockchain May Offer Solutions to Fighting Covid-19. The Wall Street Journal. 2020 May 01. URL: https://blogs.wsj.com/cio/2020/05/01/blockchain-may-offer-solutions-to-fighting-covid-19/ [accessed 2020-05-11]
31. Manderson L, Levine S. COVID-19, Risk, Fear, and Fall-out. Med Anthropol 2020 Jul;39(5):367-370. [doi: 10.1080/01459740.2020.1746301] [Medline: 32212930]
32. The Lancet. Editorial: COVID-19: protecting health-care workers. Lancet 2020 Mar;395(10228):922. [doi: 10.1016/s0140-6736(20)30644-9]
33. Min H. Blockchain technology for enhancing supply chain resilience. Business Horizons 2019 Jan;62(1):35-45. [doi: 10.1016/j.bushor.2018.08.012]
34. Yiannas F. A New Era of Food Transparency Powered by Blockchain. Innov Technol Gov Glob 2018 Jul;12(1-2):46-56 [FREE Full text] [doi: 10.1162/inov_a_00266]
35. Case Study: How Walmart brought unprecedented transparency to the food supply chain with Hyperledger Fabric. The LINUX Foundation Projects. URL: https://www.hyperledger.org/resources/publications/walmart-case-study [accessed 2020-05-11]
36. FDA takes new steps to adopt more modern technologies for improving the security of the drug supply chain through innovations that improve tracking and tracing of medicines. US Food and Drug Administration. 2019. URL: https://www.fda.gov/news-events/press-announcements/fda-takes-new-steps-adopt-more-modern-technologies-improving-security-drug-supply-chain-through [accessed 2020-07-06]
37. Miliard M. IBM launches blockchain network to bolster medical supply chain during COVID-19. Healthcare IT News. 2020 May 01. URL: https://www.healthcareitnews.com/news/ibm-launches-blockchain-network-bolster-medical-supply-chain-during-covid-19 [accessed 2020-05-11]
38. Su A. Faulty masks. Flawed tests. China's quality control problem in leading global COVID-19 fight. The Los Angeles Times. 2020 Apr 10. URL: https://www.latimes.com/world-nation/story/2020-04-10/china-beijing-supply-world-coronavirus-fight-quality-control [accessed 2020-05-12]
39. Ackland GJ, Chattoe-Brown E, Hamill H, Hampshire KR, Mariwah S, Mshana G. Role of Trust in a Self-Organizing Pharmaceutical Supply Chain Model with Variable Good Quality and Imperfect Information. JASSS 2019;22(2):5. [doi: 10.18564/jasss.3984]
40. OECD/EUIPO. Trade in Counterfeit Pharmaceutical Products. In: Illicit Trade. Paris, France: OECD Publishing; 2020.
41. Sylim P, Liu F, Marcelo A, Fontelo P. Blockchain Technology for Detecting Falsified and Substandard Drugs in Distribution: Pharmaceutical Supply Chain Intervention. JMIR Res Protoc 2018 Sep 13;7(9):e10163 [FREE Full text] [doi: 10.2196/10163] [Medline: 30213780]
42. Haq I, Muselemu O. Blockchain Technology in Pharmaceutical Industry to Prevent Counterfeit Drugs. IJCA 2018 Mar 19;180(25):8-12 [FREE Full text] [doi: 10.5120/ijca2018916579]

43. Kumar R, Tripathi R. Traceability of counterfeit medicine supply chain through Blockchain. 2019 Presented at: 11th International Conference on Communication Systems & Networks (COMSNETS); January 7-11, 2019; Bengaluru, India. [doi: 10.1109/comsnets.2019.8711418]
44. Trust Your Supplier. IBM. URL: https://www.trustyoursupplier.com [accessed 2020-07-06]
45. Tseng JH, Liao YC, Chong B, Liao SW. Governance on the Drug Supply Chain via Gcoin Blockchain. Int J Environ Res Public Health 2018 May 23;15(6) [FREE Full text] [doi: 10.3390/ijerph15061055] [Medline: 29882861]
46. Bodas M, Peleg K. Self-Isolation Compliance In The COVID-19 Era Influenced By Compensation: Findings From A Recent Survey In Israel. Health Aff (Millwood) 2020 Jun;39(6):936-941. [doi: 10.1377/hlthaff.2020.00382] [Medline: 32271627]
47. Mukhopadhyay A. Mukhopadhyay, India: Police under fire for using violence to enforce coronavirus lockdown,. Deutsche Welle. 2020 Mar 28. URL: https://www.dw.com/en/india-police-under-fire-for-using-violence-to-enforce-coronavirus-lockdown/a-52946717 [accessed 2020-05-11]
48. Keeling MJ, Hollingsworth TD, Read JM. Efficacy of contact tracing for the containment of the 2019 novel coronavirus (COVID-19). J Epidemiol Community Health 2020 Oct 23;74(10):861-866 [FREE Full text] [doi: 10.1136/jech-2020-214051] [Medline: 32576605]
49. Singer N. Virus-Tracing Apps Are Rife With Problems. Governments Are Rushing to Fix Them. The New York Times. 2020 Jul 08. URL: https://www.nytimes.com/2020/07/08/technology/virus-tracing-apps-privacy.html [accessed 2020-07-26]
50. Timberg C, Harwell D. Government efforts to track virus through phone location data complicated by privacy concerns. The Washington Post. 2020 Mar 19. URL: https://www.washingtonpost.com/technology/2020/03/19/privacy-coronavirus-phone-data/ [accessed 2020-04-11]
51. Kelion L. NHS rejects Apple-Google coronavirus app plan. BBC News. 2020 Apr 27. URL: https://www.bbc.com/news/technology-52441428 [accessed 2020-05-25]
52. Jones M, Johnson M, Shervey M, Dudley JT, Zimmerman N. Privacy-Preserving Methods for Feature Engineering Using Blockchain: Review, Evaluation, and Proof of Concept. J Med Internet Res 2019 Aug 14;21(8):e13600 [FREE Full text] [doi: 10.2196/13600] [Medline: 31414666]
53. Xu L, Shah N, Chen L, Diallo N, Gao S, Lu Y, et al. Enabling the Sharing Economy: Privacy Respecting Contract based on Public Blockchain. In: BCC '17: Proceedings of the ACM Workshop on Blockchain, Cryptocurrencies and Contracts. 2017 Presented at: BCC 2017: ACM Workshop on Blockchain, Cryptocurrencies and Contracts; April 2, 2017; Abu Dhabi p. 15-21. [doi: 10.1145/3055518.3055527]
54. Xu H, Zhang L, Onireti O, Fang Y, Buchanan WB, Imran MA. BeepTrace: Blockchain-enabled Privacy-preserving Contact Tracing for COVID-19 Pandemic and Beyond. arXiv. Preprint posted online May 21, 2020. [FREE Full text] [doi: 10.13140/RG.2.2.25101.15849/1]
55. Kim MS. Seoul's radical experiment in digital contact tracing. The New Yorker. 2020 Apr 17. URL: https://www.newyorker.com/news/news-desk/seouls-radical-experiment-in-digital-contact-tracing [accessed 2020-05-11]
56. Abeler J, Bäcker M, Buermeyer U, Zillessen H. COVID-19 Contact Tracing and Data Protection Can Go Together. JMIR mHealth uHealth 2020 Apr 20;8(4):e19359 [FREE Full text] [doi: 10.2196/19359] [Medline: 32294052]
57. Hylock RH, Zeng X. A Blockchain Framework for Patient-Centered Health Records and Exchange (HealthChain): Evaluation and Proof-of-Concept Study. J Med Internet Res 2019 Aug 31;21(8):e13592 [FREE Full text] [doi: 10.2196/13592] [Medline: 31471959]
58. Wistrom B. How Blockchain and Immunization Passports Could Help Us Re-Open. Austin Inno. 2020 Apr 21. URL: https://www.americaninno.com/austin/inno-insights/how-blockchain-and-immunization-passports-could-help-us-re-open/ [accessed 2020-05-11]
59. Sinclair S. Chinese insurers tap blockchain to speed coronavirus payouts. Coindesk. 2020 Feb 11. URL: https://www.coindesk.com/chinese-insurers-tap-blockchain-to-speed-up-coronavirus-payouts [accessed 2020-05-11]
60. Morris G, Farnum G, Afzal S, Robinson C, Greene J, Coughlin C. Patient Identification and Matching Final Report. Office of the National Coordinator for Health Information Technology. 2014. URL: https://www.healthit.gov/sites/default/files/patient_identification_matching_final_report.pdf [accessed 2020-05-11]
61. Kuo TT, Ohno-Machado L. ModelChain: Decentralized Privacy-Preserving Healthcare Predictive Modeling Framework on Private Blockchain Networks. arXiv. 2018 Feb 06. URL: https://arxiv.org/abs/1802.01746 [accessed 2020-05-11]
62. Nguyen D, Ding M, Pathirana PN, Seneviratne A. Blockchain and AI-based Solutions to Combat Coronavirus (COVID-19)-like Epidemics: A Survey. TechRxiv. 2020 Apr 14. URL: https://www.techrxiv.org/articles/Blockchain_and_AI-based_Solutions_to_Combat_Coronavirus_COVID-19_-like_Epidemics_A_Survey/12121962/1 [accessed 2020-05-11]
63. Ting DSW, Carin L, Dzau V, Wong TY. Digital technology and COVID-19. Nat Med 2020 Apr;26(4):459-461 [FREE Full text] [doi: 10.1038/s41591-020-0824-5] [Medline: 32284618]
64. Diemer M. Blockchain – Implications and Use Cases for Additive Manufacturing. Frankfurt School Blockchain Center. 2019 Aug. URL: http://explore-ip.com/2019_Blockchain-Additive-Manufacturing.pdf [accessed 2020-05-11]
65. Karkeraa S. Technological Tools. In: Unlocking Blockchain on Azure. Berkeley, CA: Apress; 2020.

Abbreviations

PPE: personal protective equipment
QR: quick response
W3C: World Wide Web Consortium

Edited by G Eysenbach; submitted 19.05.20; peer-reviewed by M Hölbl, C Esposito; comments to author 29.06.20; revised version received 14.08.20; accepted 02.09.20; published 22.09.20.

Please cite as:
Khurshid A
Applying Blockchain Technology to Address the Crisis of Trust During the COVID-19 Pandemic
JMIR Med Inform 2020;8(9):e20477
URL: http://medinform.jmir.org/2020/9/e20477/
doi:10.2196/20477
PMID:32903197

Short Paper

High-Quality Transmission of Cardiotocogram and Fetal Information Using a 5G System: Pilot Experiment

Katsuhiko Naruse[1], MD, PhD; Tomoya Yamashita[2], BA; Yukari Onishi[2], BA; Yuhi Niitaka[2], BA; Fumikage Uchida[3], BEng; Kazuya Kawahata[3], BSc; Mayu Ishihara[3], BSocSci; Hiroshi Kobayashi[1], MD, PhD

[1]Department of Obstetrics and Gynecology, Nara Medical University, Kashihara, Nara, Japan

[2]NTT DoCoMo, Inc, Tokyo, Japan

[3]TOITU Co, Ltd, Tokyo, Japan

Corresponding Author:
Katsuhiko Naruse, MD, PhD
Department of Obstetrics and Gynecology
Nara Medical University
840, Shijo-Cho
Kashihara, Nara, 6348521
Japan
Phone: 81 744223051
Email: naruse@naramed-u.ac.jp

Abstract

Background: A cardiotocogram (CTG) is a device used to perceive the status of a fetus in utero in real time. There are a few reports of its use at home or during emergency transport.

Objective: The aim of this study was to test whether CTG and other perinatal information can be transmitted accurately using an experimental station with a 5G transmission system.

Methods: In the research institute, real-time fetal heart rate waveform data from the CTG device, high-definition video ultrasound images of the fetus, and high-definition video taken with a video camera on a single line were transmitted by 5G radio waves from the transmitting station to the receiving station.

Results: All data were proven to be transmitted with a minimum delay of less than 1 second. The CTG waveform image quality was not inferior, and there was no interruption in transmission. Images of the transmitted ultrasound examination and video movie were fine and smooth.

Conclusions: CTG and other information about the fetuses and pregnant women were successfully transmitted by a 5G system. This finding will lead to prompt and accurate medical treatment and improve the prognosis of newborns.

(JMIR Med Inform 2020;8(9):e19744) doi:10.2196/19744

KEYWORDS

5G communication system; home telecare; online health; telecardiology; telemedicine; teleconsulting

Introduction

A cardiotocogram (CTG) is a device used worldwide in obstetrics to perceive the status of a fetus in the second half of the gestational period in utero in real time. It has been relied upon as a test comparable to fetal ultrasound because a decelerated heart rate detected by this method indicates fatal stress in the fetus, an accelerated heart rate baseline indicates infection in the fetus in utero, and a fair baseline variability indicates the sparing ability of the fetus. However, at present, it is used only in medical care facilities, and there are few reports of its use at home or during emergency transport. In addition, there are many limitations to the prehospital diagnosis that can be made during transport of perinatal diseases [1], and there is an urgent need to develop a device that helps paramedics to recognize obstetric abnormalities.

Recently, wireless/mobile transmission using the next generation of high-speed, high-capacity systems, so-called 5G [2], has been developed in some parts of the world, and it is expected to be applied to the medical field, mainly because of its high speed and good compatibility with cloud computing [3]. It can also be used as a tool for accurately conveying data to medical institutions at home or during patient transfers [3]. However, no research has been conducted yet in the field of obstetrics to verify its utility.

In this pilot study, we tested whether CTG and other perinatal information can be transmitted accurately using an experimental station with a 5G transmission system.

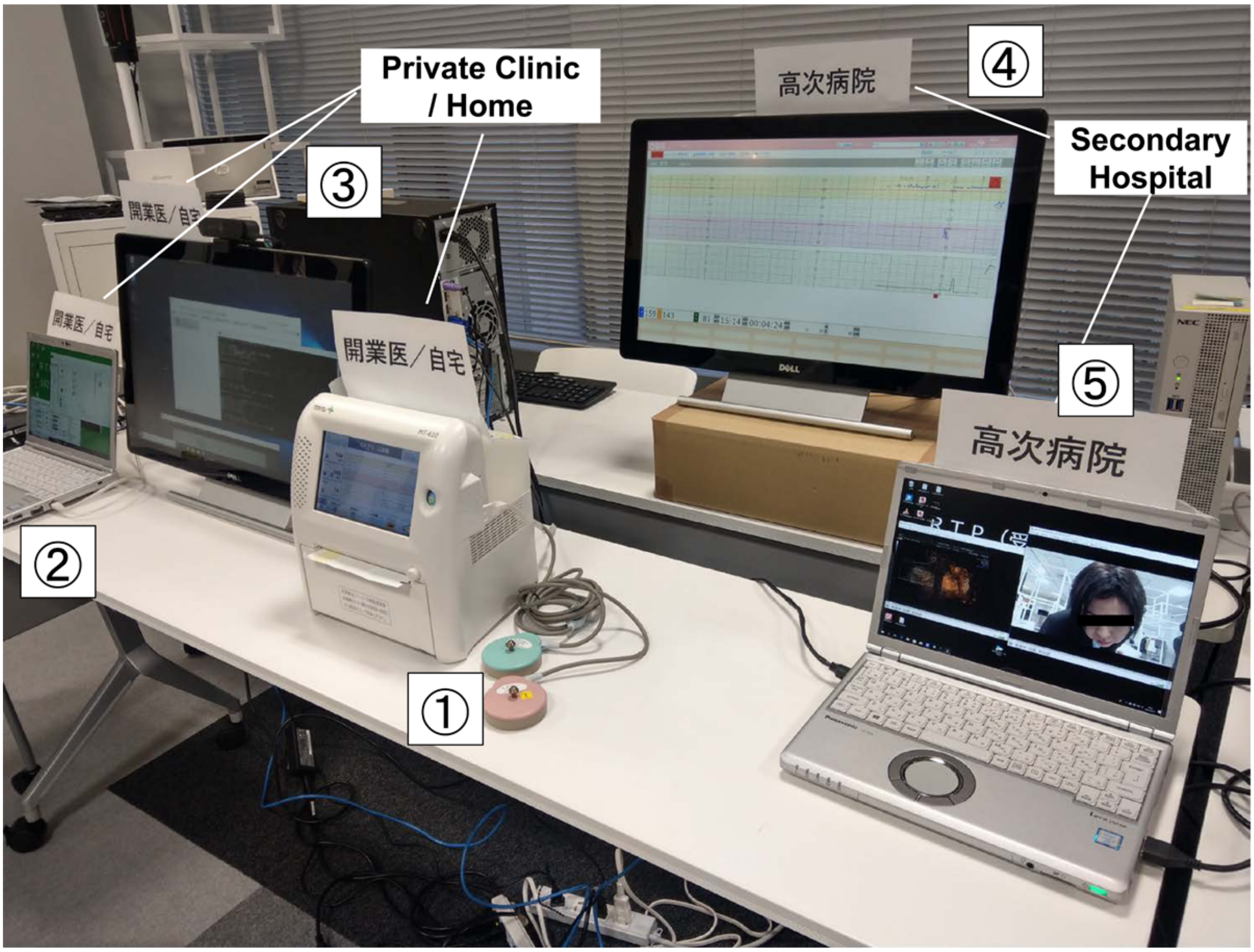
Methods

This study was conducted in December 2019 at the Research Institute of NTT DoCoMo Incorporated (Osaka, Japan). We combined real-time fetal heart rate waveform data of 30 minutes from the MT-610 CTG device (TOITU Co, Ltd), which stores and reproduces anonymized data from actual normal fetuses; high-definition (HD) video ultrasound images of the fetus (Prosound α10, Hitachi Ltd; prerecorded and anonymized, repeat of the 10-minute scan); and HD video taken with a common digital video camera on a single 5G line. All data were transmitted by 4.5 GHz radio waves (selected from the radio wave bands 3.7 GHz, 4.5 GHz, and 28 GHz available for 5G in Japan) from the transmitting station to the receiving station, which is housed at the institute on the same floor of the building. All of this information was reproduced on a computer on the receiving end that simulated a secondary hospital. The CTG waveforms and ultrasound movie images that could be seen on the receiving screen were checked by an obstetrics and gynecology consultant (KN) to confirm if they were acceptable for diagnosis. Another consultant (HK) later checked them again from the stored images. The face of a female model was featured in the HD video, and the complexion and expression of the model could be recognized after the transmission. The model also held a digital clock, so that the delay in data transmission could be defined later.

Results

Connection of the devices using the 5G transmission system from the source side (1, 2, and 3 in Figure 1) and receiver side (4 and 5 in Figure 1) via a radio wave transmitter and receiver on the same floor was successful.

Figure 1. Panoramic view of the experiment. Objects 1-3 are the source side, and objects 4 and 5 are the receiving side of the 5G transmission system. 1: The cardiotocogram (CTG) device (MT-610) transmits the waveform of the fetal heart rate in clinical practice; 2: the computer plays back ultrasound examination videos; 3: the computer plays back movies from a video camera (operation of the real-time self-taken video is checked afterwards); 4: the CTG waveform display is maximized at the receiving end to check visibility; 5: movies of the ultrasound examination (left) and video camera (right) are played simultaneously at the receiving end.



All data were proven to be transmitted with a minimum delay of less than 1 second as per the digital clock in the HD video. The CTG waveform image quality was not inferior, and there was no interruption in transmission. The fetal heart rate was clearly legible, and the variability of the baseline was easy to read. At the same time, the real-time images of the transmitted ultrasound examination were fine and smooth, and no unnatural movements were observed. In addition, the color of the model's face and surrounding movements in the video image were correctly transmitted. There were no unnatural color tones; as a result, we were able to recognize her complexion and expression (Figure 1).

Discussion

To our knowledge, this study is the first to show that CTG and other information about fetuses and pregnant women can be transmitted by a 5G system. A system to monitor CTGs within a medical care facility via an intranet is common, and a system to exchange CTG waveforms between medical institutions via the internet has been put to practical use. However, attempts to send data from a pregnant woman's home to a medical institution via a mobile network or to send data from an ambulance or air ambulance in transit have not been sufficiently carried out, except in one case in 1992 [4]. If fetal information and other data can be transmitted via a mobile network, medical institutions will be able to keep abreast of sudden changes in the condition of the fetus at home or of pregnant women with complications being transported to the hospital, which will lead to prompt and accurate medical treatment and improve the prognosis of the newborn.

5G systems are anticipated to have the potential to significantly improve the speed and stability of transmission in mobile communications. In this study, we found that not only CTG, but also images during ultrasound examinations and high-quality videos of people who mimicked patients could be transmitted without problems. Although our study was performed in a laboratory, and the distance between the transmitter and receiver was not far, recent planning of information technology infrastructure on 5G transmission promises better linkage of devices under both low or high power [5]. In the field of emergency medicine, doctors and paramedics have already started attempting to make accurate diagnoses by using smartphones to take videos, through which consultants at a central hospital can assess the situation in real time [6]. 5G systems are expected to be able to transmit high-quality CTGs as well as more precise ultrasound images of the fetus and HD videos reporting the complexion, expression, behavior, and complaints of pregnant women/patients and the surrounding environment. Developing new solutions for the new era of 5G systems for the home care of pregnant women and emergency transport systems in perinatal care may not only dramatically improve the prognosis for mother and child, but also reduce the burden on health care providers.

Additionally, as predicted by Oleshchuk and Fensli [7], the 5G system will also work with artificial intelligence and big data to enable "self-determined medicine" to make instant diagnoses based on data obtained at home [3]. Issues such as developing new infrastructure for 5G transmission, shorter propagated distance of the radio wave on higher frequency, or data privacy (problems not only for 5G) remain to be solved. However, home monitoring of a fetus with the 5G system is a particularly good application of this new generation of technology, which could create a new future for obstetric care.

Conflicts of Interest

TY, YO, and YN are employed by NTT DOCOMO, Inc, Japan. FU, KK, and MI are employed by TOITU Co, Ltd, Japan. Other authors declare no conflict of interests.

References

1. Jones AE, Summers RL, Deschamp C, Galli RL. A national survey of the air medical transport of high-risk obstetric patients. Air Med J 2001;20(2):17-20. [Medline: 11250614]
2. Chávez-Santiago R, Szydełko M, Kliks A, Foukalas F, Haddad Y, Nolan KE, et al. 5G: The Convergence of Wireless Communications. Wirel Pers Commun 2015;83:1617-1642 [FREE Full text] [doi: 10.1007/s11277-015-2467-2] [Medline: 27076701]
3. Li D. 5G and intelligence medicine-how the next generation of wireless technology will reconstruct healthcare? Precis Clin Med 2019 Dec;2(4):205-208 [FREE Full text] [doi: 10.1093/pcmedi/pbz020] [Medline: 31886033]
4. Poulton TJ, Gutierrez PJ. Fetal monitoring during air medical transport. J Air Med Transp 1992;11(11-12):13, 15-13, 17. [doi: 10.1016/s1046-9095(05)80165-5] [Medline: 10123100]
5. Yin M, Li W, Feng L, Yu P, Qiu X. Emergency Communications Based on Throughput-Aware D2D Multicasting in 5G Public Safety Networks. Sensors (Basel) 2020 Mar 29;20(7) [FREE Full text] [doi: 10.3390/s20071901] [Medline: 32235400]
6. Yanagawa Y, Jitsuiki K, Nagasawa H, Takeuchi I, Madokoro S, Ohsaka H, et al. A Smartphone Video Transmission System for Verification of Transfusion. Air Med J 2019;38(2):125-128. [doi: 10.1016/j.amj.2018.11.012] [Medline: 30898283]
7. Oleshchuk V, Fensli R. Remote Patient Monitoring Within a Future 5G Infrastructure. Wireless Pers Commun 2010 Jul 15;57(3):431-439. [doi: 10.1007/s11277-010-0078-5]

Abbreviations

CTG: cardiotocogram
HD: high definition

Edited by C Lovis; submitted 30.04.20; peer-reviewed by L Sbaffi; comments to author 28.06.20; accepted 28.07.20; published 08.09.20.

Please cite as:
Naruse K, Yamashita T, Onishi Y, Niitaka Y, Uchida F, Kawahata K, Ishihara M, Kobayashi H
High-Quality Transmission of Cardiotocogram and Fetal Information Using a 5G System: Pilot Experiment
JMIR Med Inform 2020;8(9):e19744
URL: http://medinform.jmir.org/2020/9/e19744/
doi:10.2196/19744
PMID:32897237

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org