JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 10 (2022), Issue 7 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Review

Viewpoint

Original Papers

Review

Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review

Carina Nina Vorisek[1], MD, MSc; Moritz Lehne[1], PhD; Sophie Anne Ines Klopfenstein[1,2], MD; Paula Josephine Mayer[1]; Alexander Bartschke[1], MA; Thomas Haese[1], BSc; Sylvia Thun[1], MD, PhD

[1]Core Facility Digital Medicine and Interoperability, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany

[2]Institute for Medical Informatics, Charité – Universitätsmedizin Berlin, Berlin, Germany

Corresponding Author:
Carina Nina Vorisek, MD, MSc
Core Facility Digital Medicine and Interoperability
Berlin Institute of Health at Charité – Universitätsmedizin Berlin
Anna-Louisa-Karsch-Str. 2
Berlin, 10178
Germany
Phone: 49 30 450 543049
Email: carina-nina.vorisek@bih-charite.de

Abstract

Background: The standard Fast Healthcare Interoperability Resources (FHIR) is widely used in health information technology. However, its use as a standard for health research is still less prevalent. To use existing data sources more efficiently for health research, data interoperability becomes increasingly important. FHIR provides solutions by offering resource domains such as "Public Health & Research" and "Evidence-Based Medicine" while using already established web technologies. Therefore, FHIR could help standardize data across different data sources and improve interoperability in health research.

Objective: The aim of our study was to provide a systematic review of existing literature and determine the current state of FHIR implementations in health research and possible future directions.

Methods: We searched the PubMed/MEDLINE, Embase, Web of Science, IEEE Xplore, and Cochrane Library databases for studies published from 2011 to 2022. Studies investigating the use of FHIR in health research were included. Articles published before 2011, abstracts, reviews, editorials, and expert opinions were excluded. We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and registered this study with PROSPERO (CRD42021235393). Data synthesis was done in tables and figures.

Results: We identified a total of 998 studies, of which 49 studies were eligible for inclusion. Of the 49 studies, most (73%, n=36) covered the domain of clinical research, whereas the remaining studies focused on public health or epidemiology (6%, n=3) or did not specify their research domain (20%, n=10). Studies used FHIR for data capture (29%, n=14), standardization of data (41%, n=20), analysis (12%, n=6), recruitment (14%, n=7), and consent management (4%, n=2). Most (55%, 27/49) of the studies had a generic approach, and 55% (12/22) of the studies focusing on specific medical specialties (infectious disease, genomics, oncology, environmental health, imaging, and pulmonary hypertension) reported their solutions to be conferrable to other use cases. Most (63%, 31/49) of the studies reported using additional data models or terminologies: Systematized Nomenclature of Medicine Clinical Terms (29%, n=14), Logical Observation Identifiers Names and Codes (37%, n=18), International Classification of Diseases 10th Revision (18%, n=9), Observational Medical Outcomes Partnership common data model (12%, n=6), and others (43%, n=21). Only 4 (8%) studies used a FHIR resource from the domain "Public Health & Research." Limitations using FHIR included the possible change in the content of FHIR resources, safety, legal matters, and the need for a FHIR server.

Conclusions: Our review found that FHIR can be implemented in health research, and the areas of application are broad and generalizable in most use cases. The implementation of international terminologies was common, and other standards such as the Observational Medical Outcomes Partnership common data model could be used as a complement to FHIR. Limitations such as the change of FHIR content, lack of FHIR implementation, safety, and legal matters need to be addressed in future releases to expand the use of FHIR and, therefore, interoperability in health research.

(JMIR Med Inform 2022;10(7):e35724) doi:10.2196/35724

KEYWORDS

Fast Healthcare Interoperability Resources; FHIR; interoperability; health research; health care; health information technology; research; clinical research; public health; epidemiology

Introduction

Within the current COVID-19 pandemic, there was a broad realization of the currently limited data collection processes and how powerful the exchange of scientific data could be if interoperability between health care and research was provided [1]. Although there was a large amount of data in the health care ecosystem, there was lack of data that adheres to Findable, Accessible, Interoperable, and Reusable [2] principles for users to find, use, analyze, and share data on COVID-19. This applies specifically to academic health research where the lack of interoperability between health care and research often inhibits the use of existing data sources for research. Commonly, the data collections of health research are stored in decentralized, autonomous data infrastructures which requires integration into common frameworks to enable centralized search and access.

However, processing national and cross-national scientific data across different institutions and software systems requires international standards and terminologies: the Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP) common data model (CDM) is used in observational research, whereas the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Standard (ODM) is used specifically for the exchange of data within clinical trials [3]. CDISC is providing standards such as standardized raw data sets (Study Data Tabulation Model; SDTM), also considered a CDM, as well as standardized analysis data sets models. Further established standards are the terminologies Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) and Logical Observation Identifiers, Names, and Codes (LOINC). SNOMED CT is the most comprehensive clinical health care terminology worldwide providing more than 350,000 concepts, whereas LOINC is a standard for laboratory tests and clinical observations. One of the latest emerging standards for the exchange of health data is the standard Fast Healthcare Interoperability Resources (FHIR).

FHIR is a standard used in health information technology introduced in 2011 by the Standard Developing Organization Health Level Seven International (HL7). FHIR is based on previous HL7 standards (HL7 versions 2 and 3 and Clinical Document Architecture) and combines their advantages with established modern web technologies such as a Representational State Transfer (REST) architecture; application programming interface (API), XML, and JSON formats; and authorization tools (Open Authorization). In FHIR, all exchangeable content is defined by distinct basic building blocks—referred to as resources—which define the content and structure of information and can refer to each other using reference mechanisms [4].

The base FHIR specification serves as a foundation providing basic resources, frameworks, APIs, and a platform in which different solutions can be implemented [5]. To cover information not included in the basic resources, FHIR provides a built-in extension mechanism and can be adapted for specific use cases while ensuring interoperability. Additional rules and constraints within resources can be defined in profiles. Therefore, FHIR covers various domains of health care with its resources and can be used for different purposes and in various contexts and workflows.

With regard to health research, there is still a lack of use of international standards when exchanging data between health care and research institutions. However, there have been recent regulative and legislative changes promoting standards and interoperability in health care [6-8]. In addition, there are initiatives of HL7 promoting FHIR's use in health research, such as the Vulcan HL7 FHIR Accelerator aiming to connect clinical research and health care, the MedMorph project aiming to advance public health by using standards such as FHIR, and the collaboration of HL7 and OHDSI on a single common data model [9-11]. As many research platforms and modern data management systems, such as the Extensible Neuroimaging Archive Toolkit open-source imaging informatics platform, use extensible REST APIs [12,13], FHIR may be the new standard to fill the interoperability gap in health research with its REST architecture. Existing reviews on FHIR investigate the general use of FHIR in digital health [14] or its use in electronic health records [15]. However, to the best of our knowledge, the use of FHIR in health research has not been systematically investigated. Therefore, the aim of our study was to provide a systematic review of existing literature to determine the current state of use cases, implementation, goals, and limitations of FHIR in health research.

Methods

Protocol, Registration, and Ethical Considerations

This systematic review was conducted in accordance with the (PRISMA) Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines [16]. The review was registered with the International Prospective Register of Systematic Reviews (PROSPERO; CRD42021235393) [17]. As data originated from published studies, ethical approval for this study was not requested.

Inclusion and Exclusion Criteria

We included studies investigating the use of FHIR in health research. We did not focus on particular patient populations, interventions, control groups, or outcomes, except the use of FHIR in health research. Details on inclusion and exclusion criteria are presented in Textbox 1.

Textbox 1. Inclusion and exclusion criteria for paper review.

Inclusion criteria

- Studies focusing on the use of FHIR in health care research
- Original papers published in peer-reviewed journals in English
- Studies with publication dates no earlier than 2011

Exclusion criteria

- Studies focusing on the use of FHIR in electronic health records, mobile and web apps, decision support, and data protection or security
- General overviews on FHIR
- Comments, books, editorials, or reviews
- Language other than English
- Studies conducted before 2011

Information Sources and Search Strategy

A comprehensive literature search was performed through the PubMed/MEDLINE, Embase, Web of Science, IEEE Xplore, and Cochrane Library databases. In addition, citation tracking and reference list checking were performed. The goal of the search strategy was to retrieve all relevant studies related to our research question published between 2011 and 2022. Search terms were therefore relatively broad to make sure that all potentially relevant studies were identified. Search terms used for the database searches were "FHIR" and "Fast Healthcare Interoperability Resources." Information on the detailed search strategy for each database is provided as an appendix to this review (Multimedia Appendix 1). The search was conducted on February 26, 2022.

Study Selection and Data Collection Process

Study selection included 2 screening levels: (1) screening of titles and abstracts of all studies identified in the literature search and (2) full-text review of studies that had not been excluded in the first step. Review at the first stage of screening was performed independently by 2 authors (ML and SAIK) using the Rayyan web app [18]. Remaining disagreements were resolved by a third author (CNV). Further full-text screening at the second stage and data extraction were performed by 6 authors (CNV, ML, SAIK, PJM, AB, and TH), and disagreements of at least two authors at this stage were resolved by the last author (ST).

Data Extraction and Analysis

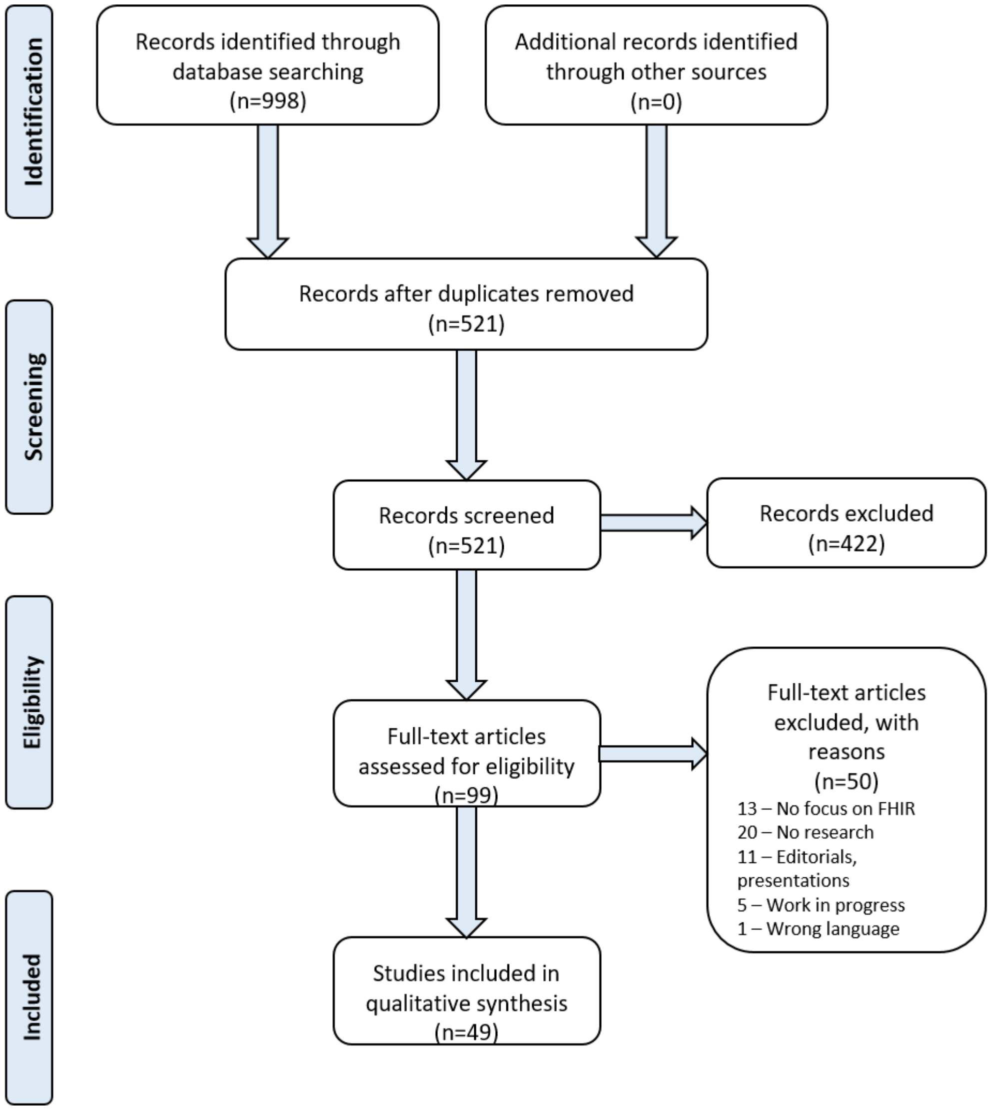
Data synthesis was conducted in tables and figures. For categorical variables, simple and relative frequencies and proportions were used. To identify the networks of coauthors, we also performed a network analysis that investigated, for all authors of the included studies, whether they were coauthors in a study. Results were visualized in a network graph. We did not assess bias in studies due to the lack of quantitative tools applicable to technical papers on standards. All analyses were done with R statistical software (version 4.0.5; R Foundation for Statistical Computing) [19] and the *tidyverse* packages [20]. All data and analyses scripts are provided in a GitHub repository [21].

Results

Study Selection and Extraction

A total of 998 articles were identified through the database searches (344 from MEDLINE, 359 from Embase, 201 from Web of Science, 84 from IEEE Xplore, and 10 from Cochrane Library). No additional records were identified through citation tracking and reference list checking. We excluded 477 duplicates and 422 articles that did not meet the inclusion criteria or met the exclusion criteria. Among the 99 full-text articles assessed for eligibility, an additional 50 studies were excluded. Finally, 49 [1,12,22-68] articles met the inclusion criteria and were included in the systematic review (Figure 1). Details on the exclusion reasons for the full-text evaluation can be found in Multimedia Appendix 2, and the exclusion reasons for the abstract evaluation can be found in the GitHub repository [21].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for identifying articles eligible for inclusion. FHIR: Fast Healthcare Interoperability Resources.



Characteristics of Included Studies

Publication dates ranged from 2016 to 2022 with the median in 2020. Of the 49 included studies, 73% (n=36) were published between 2020 and 2022. The increase of publications from 2020 onward is visualized in Figure 2, showing the temporal trend of all FHIR publications identified in the databases with the search terms "FHIR" OR "Fast Healthcare Interoperability Resources" and the number of publications included into the analysis per year.

The results of the network analysis of coauthorships are shown in Figure 3. Of a total of 256 authors, most (85%, n=217) appeared only once in the included studies, and no author occurred more than 6 times within the included studies. Most coauthorship networks were restricted to individual studies, with occasional connections between networks (ie, authors having published studies with different groups of coauthors).

Of the 49 studies, the majority were conducted in Germany (47%, n=23) [12,26,28-31,34,35,40-42,45-47,52,53,56-58,60,62, 63,69], the United States (27%, n=13) [22,25,36,44,48-50,61, 64-66,68,70], and Australia (6%, n=3) [1,43,67]. The remaining studies were performed in Austria (2%, n=1) [32], Canada (2%, n=1) [24], France (2%, n=1) [51], Greece (2%, n=1) [59], Japan (2%, n=1) [27], Pakistan (2%, n=1) [38], Spain (2%, n=1) [55], Switzerland (2%, n=1) [39], Taiwan (2%, n=1) [23], and the United Kingdom (2%, n=1) [37].

Figure 2. Number of publications per year (all: all FHIR publications identified in the databases with the search terms “FHIR” OR “Fast Healthcare Interoperability Resources”; included: studies included in this review). FHIR: Fast Healthcare Interoperability Resources.

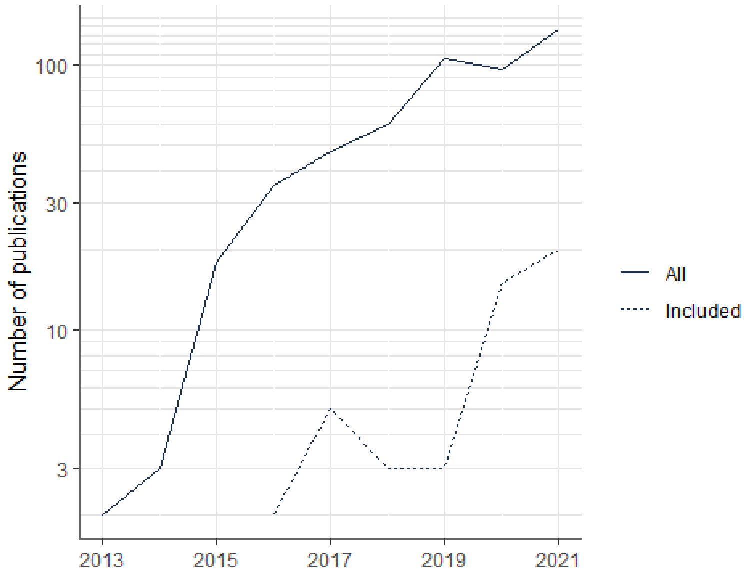


Figure 3. Network of coauthorships. Each point represents an author. Point size and color indicate the number of publications of this author (between 1 and 6). Lines indicate that authors have coauthored at least one paper together. Line thickness represents the number of coauthorships.

Research Domain and Area of FHIR Application

Of the 49 studies, most (73%, n=36) studies covered the research domain of clinical research, of which 10 (20%) studies were clinical trials [22,29-31,36,39,43,56,65,66]; 3 (6%) studies focused on solutions in public health and epidemiology [38,40,64], and the remaining studies did not specify their research domain (20%, n=10; Figure 4) [24,32,41,42,45-47,50,63,69]. The included studies used FHIR for the standardization of data (41%, n=20) [23,26,30,34,41,45-48,51-53,57-60,63,66,67,70], data capture (29%, n=14) [1,12,22,24,27,35-37,43,44,55,61,64,65], recruitment (14%, n=7) [28,29,31,32,49,56,62], analysis (12%, n=6) [25,38,42,50,68,69], and consent management (4%, n=2; Table 1) [39,40]. Details on the included studies are presented in Table 2.

Figure 4. Number of studies according to research domain.

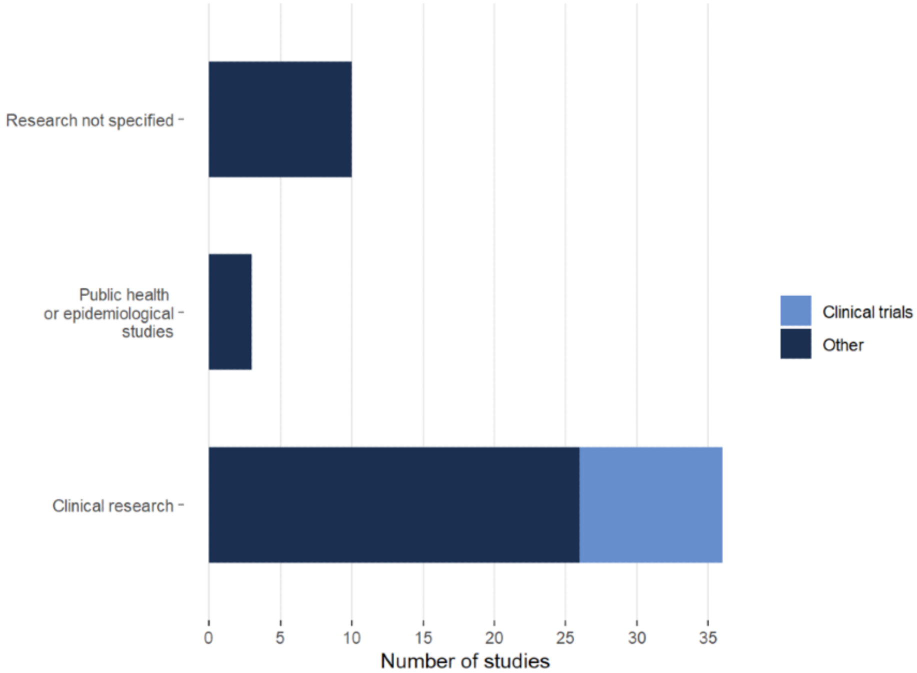


Table 1. Numbers of studies according to area of FHIR application, medical specialty, and international standard.

Area	Studies (N=49), n (%)
FHIR[a] application	
Standardization of data	20 (41)
Data capture	14 (29)
Recruitment	7 (14)
Analysis	6 (12)
Consent management	2 (4)
Medical specialty	
Generic approach	27 (55)
Infectious disease	8 (16)
Oncology	6 (12)
Genomics	4 (8)
Pulmonary hypertension	1 (2)
Neuroimaging research	1 (2)
Genomic cancer medicine	1 (2)
Environmental health	1 (2)
International standard	
Other	21 (43)
None	18 (37)
LOINC[b]	14 (29)
SNOMED CT[c]	18 (37)
ICD-10[d]	9 (18)
OMOP[e]	6 (12)

[a]FHIR: Fast Healthcare Interoperability Resources.
[b]LOINC: Logical Observation Identifiers Names and Codes.
[c]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.
[d]ICD-10: International Classification of Diseases 10th Revision.
[e]OMOP: Observational Medical Outcomes Partnership.

Table 2. Characteristics of studies.

Source, year	Country	Item mapped to FHIR[a]	Objective for FHIR use	FHIR resources
Banach et al [56], 2021	Germany	Medical and demographic data from free-text eligibility criteria	Estimation of the number of potentially eligible patients for planning multicenter trials based on free-text criteria and using a consented data set based on FHIR	—[b]
Bauer et al [1], 2020	Australia	Questionnaire	Ontology-based standard questionnaire for linking genomic data with clinical outcomes	Questionnaire
Bialke et al [40], 2018	Germany	Modular consent templates	Support improvement for consent definition and consent documentation	Consent
Bild et al [28], 2020	Germany	Informed consent template	Cross-site interoperability layer for representing the validity of data use policies derived from signed informed consent templates and regulatory framework	Consent and Patient
Brandt et al [71], 2021	United States	Phenotype definitions from the Phenotype Knowledgebase repository	Repository of structured phenotype definitions for automation of cohort identification.	Patients, Encounter, Procedure, Medication orders, Condition, and Observation
Cheng et al [44], 2021	United States	EHR[c] Data	Seamless data exchange between the REDCap[d] research electronic data capture and any EHR system with a FHIR API[e]	Patient, Observation, AllergyIntolerance,MedicationOrder, and Condition
Deppenwieset al [57], 2021	Germany	Oncology data	Provide a transformation tool from oncology data XML files to FHIR for oncological data to enable clinical research	Medication, MedicationStatement, and Procedure
Eapen et al [24], 2019	Canada	Electronic form components	Management, editing, and rendering of electronic forms in the form of an open-source framework	Questionnaire and QuestionnaireResponse
Fischer et al [35], 2020	Germany	Common data set from a German pulmonary hypertension registry	Feasibility of HL7[f] FHIR Bundle and XSLT[g] as a generic ETL[h] process to populate an OMOP[i] CDM[j]	Patient, Encounter, and Observation
Garza et al [61], 2020	United States	Concomitant medications, demographics, eligibility, labs, medical history, therapeutic area–specific, procedure, encounters, vital signs, other, administrative, questionnaires, and study drug administration	Developing and implementing a systematic mapping approach for evaluating HL7 FHIR standard coverage in multicenter clinical trials.	Observation, Patient, Specimen, Encounter, Diagnostic Report, and Condition
González- Castro et al [55], 2021	Spain	Clinical patient data (from EHR) and patient-generated data	Collection and aggregation of survivorship data (use cases colon cancer and breast cancer)	Patient, Condition, Observation, MedicationStatement, Encounter, and Procedure
Gruendner et al [69], 2020	Germany	Clinical patient data	Analysis within and across institutions	—
Gruendner et al [42], 2021	Germany	Metadata	Developing a Metadata Schema based on FHIR to gather metadata on clinical, epidemiological, and public health studies; elevate data FAIRness[k]; and widen analysis possibilities across health research domains	ResearchStudy, Questionnaire, and DocumentReference
Guérin et al [51], 2021	France	Clinical and omics data in oncology	Improve and accelerate retrospective and prospective clinical and genomic data sharing in oncology	MolecularSequence and Observation
Gulden et al [31], 2018	Germany	Eligibility criteria of clinical trials	Recruitment of patients for clinical trials using eligibility criteria	Condition and Patient
Gulden et al [30], 2021	Germany	Clinical trial data	Multisite clinical trial registry	ResearchStudy
Hong et al [25], 2017	United States	Ovarian cancer data	Support of clinical statistics and analysis leveraging standardized data exchange and access services based on FHIR	Patient, Observation, Condition, and Procedure

Source, year	Country	Item mapped to FHIR[a]	Objective for FHIR use	FHIR resources
Hund et al [53], 2021	Germany	Process data	Developing a framework to enable standardized, shared processes using Business Process Model and Notation and FHIR for arbitrary biomedical research	ActivityDefinition, Binary, Bundle, CodeSystem, Endpoint,Group, NamingSystem, Organization, Practitioner, PractitionerRole, ResearchStudy, StructureDefinition, Subscription, and Task
Jiang et al [70], 2017	United States	Clinical research data	Development and assessment of a consensus-based approach for harmonizing the OHDSI[l] CDM with HL7 FHIR	Observation
Kilintzis et al [59], 2022	Greece	Clinical information from in-ICU[m] COVID-19 patients	Fusion of clinical information with chest sounds and imaging of COVID-19 ICU patients	Media
Klopfenstein et al [41], 2021	Germany	Metadata of clinical, epidemiological and public health studies	Developing a Metadata Schema based on FHIR to gather metadata on clinical, epidemiological, and public health studies; elevate data FAIRness; and widen analysis possibilities across health research domains	ResearchStudy, Questionnaire, and DocumentReference
Khalique and Khan [38], 2017	Pakistan	EHR	Analysis or mining of EHR data and contextual information to assess the population's health	—
Khvastova et al [12], 2020	Germany	Open-source research platform (XNAT[n])	Feasibility study for the full integration of FHIR into XNAT	Patient
Lackerbauer et al [32], 2019	Austria	Informed consent or questionnaires	Automated verification of answers	Questionnaire and QuestionnaireResponse
Lambarki et al [58], 2021	Germany	Oncology data	Use and apply a harmonized FHIR-based modular data set in a federated data platform for translational cancer research to store data in a structured manner and enable data transfer	Condition, Observation, Procedure, MedicationStatement, Patient, Organization, Specimen, ClinicalImpression, Encounter, and ServiceRequest
Lee et al [23], 2020	Taiwan	IPS[o]	FHIR-based global infectious disease surveillance and case-tracking model	MedicationStatement, Medication, AllergyIntolerance, Condition, Immunization, Procedure, Organization, Observation, CarePlan, and Location
Lenert et al [50], 2021	United States	Clinical data	Availability of data for research	Patient, Encounter, Condition, Procedure, Observation, MedicationRequest, and MedicationAdministration
Leroux et al [67], 2017	Australia	Data model	Mapping CDISC[p] ODM[q] to FHIR	Patient, Observation, EpisodeOfCare, Encounter, QuestionnaireResponse, Questionnaire, and CarePlan
Majeed et al [60], 2021	Germany	General patient information, encounter, or visit related information; individual data points; observations; measurements; and surveys	Developing a generic ETL framework to process patient data into FHIR and enable data integration in a single central data warehouse as a prerequisite for translational research	Patient, Observation, and Encounter
Metke-Jimenez et al [43], 2019	Australia	REDCap forms	Data export from REDCap into FHIR resources	Encounter, Observation, Condition, and Patient
Peng et al [52], 2021	Germany	Genomic Variant Cell Format data	Coverage of Variant Cell Format data in OMOP CDM with and without using FHIR as intermediate layer	MolecularSeqeunce, Patient, and Condition

Source, year	Country	Item mapped to FHIR[a]	Objective for FHIR use	FHIR resources
Pfiffner et al [22], 2016	United States	ResearchKit data	Patient-reported outcomes	Contract, Questionnaire, QuestionnaireResponse, Patient, and Observation
Reinecke et al [29], 2020	Germany	Patient ID lists	Data-driven recruitment of patients for clinical trials, storage of patient lists, and generation of notifications	List
Rinaldi et al [45], 2021	Germany	Microbiology data	Standardization of clinical data from patient care and medical research in the field of infection control	DiagnosticReport, Observation, Specimen, and ServiceRequest
Rinaldi et al [47], 2021	Germany	OpenEHR Template	Mapping infection control related data across 3 different standards—OpenEHR, FHIR, and OMOP CDM—to maximize analysis capabilities	DiagnosticReport, Observation, Specimen, ServiceRequest, and Encounter
Sass et al [26], 2020	Germany	COVID-19 data	Standardized data model	Patient, Consent, Observation, Condition, Procedure, Encounter, Medication, and MedicationStatement
Sass et al [46], 2021	Germany	Medication chapter of the German Procedure Classification and Identification of Medicinal Products–compliant medication terminology	Representation of structured medication data	Patient, Procedure, MedicationStatement, and Medication
Tanaka et al [27], 2020	Japan	SS-MIX2[r]	Mapping electronic medical record items between SS-MIX2 and HL7 FHIR	Patient, Encounter, Condition, AllergyIntolerance, Observation, Specimen, ServiceRequest, MedicationRequest, and MedicationDispense
Ulrich et al [34], 2016	Germany	Metadata or CRF[s]	Metadata repository	Questionnaire
Wagholikar et al [36], 2017	United States	Common data model demographics, laboratory results, and diagnoses	Clinical apps sharing via a platform	—
Wang et al [48], 2021	United States	FDA[t]'s Adverse Event Reporting System data	Potential use of FHIR for postmarket safety surveillance for drug products	AdverseEvent
Weber et al [39], 2020	Switzerland	Electronic consent form	Designing of a FHIR-based eConsent app for Android and evaluation of acceptance	Contract
Wettstein et al [62], 2021	Germany	Clinical data	Using FHIR for automated and distributed feasibility queries to find available cohort sizes across institutions	Group, ResearchStudy, and Task
Wettstein et al [63], 2021	Germany	Medical routine data	HL7 FHIR version R4 is used to define the necessary communication messages as well as process input and output variables.	Group, ResearchStudy, and Task
Wu et al [37], 2018	United Kingdom	EHR data and unstructured documents	Semantic search system for obtaining clinical insights from unstructured clinical notes	Patient and DocumentReference
Xu et al [64], 2020	United States	Data set of patients with "asthma-like" conditions	Impact of airborne pollutant exposures on asthma (research question)	—
Zong et al [65], 2020	United States	Colorectal cancer report	Automatic population of eCRFs in colorectal clinical cancer trials	Questionnaire and QuestionnaireResponse
Zong et al [66], 2021	United States	Colorectal cancer data model	Framework for capturing common data elements from CRFs and FHIR resources to identify clinical information needs	DiagnosticReport and Observation

Source, year	Country	Item mapped to FHIR[a]	Objective for FHIR use	FHIR resources
Zong et al [68], 2020	United States	EHR	Discovery of genotype-phenotype associations	Condition, and Observation

[a]FHIR: Fast Healthcare Interoperability Resources.
[b]Not available.
[c]EHR: electronic health record.
[d]REDCap: Research Electronic Data Capture.
[e]API: application programming interface.
[f]HL7: Health Level Seven International.
[g]XSLT: Extensible Stylesheet Language Transformations.
[h]ETL: Extract-Transform-Load.
[i]OMOP: Observational Medical Outcomes Partnership.
[j]CDM: common data model.
[k]FAIR: Findable, Accessible, Interoperable, and Reusable.
[l]OHDSI: Observational Health Data Sciences and Informatics
[m]ICU: intensive care unit.
[n]XNAT: Extensible Neuroimaging Archive Toolkit.
[o]IPS: International Patient Summary.
[p]CDISC: Clinical Data Interchange Standards Consortium.
[q]ODM: Operational Data Model.
[r]SS-MIX2: Standardized Structured Medical Information Exchange2.
[s]CRF: Case Report Form.
[t]FDA: U.S. Food and Drug Administration.

Study Objectives

In terms of medical specialty, most (55%, 27/49) of the studies [24,27-32,34,36-42,44,46,48,49,53,56,60-63,67,70] were using a generic approach—implementable in any kind of specialty (Table 2). Of the remaining studies, 16% (8/49) use cases focused on infectious disease [1,22,23,26,45,47,50,59], whereas 12% (6/49) focused on oncology [25,55,57,58,65,66] and 8% (4/49) on genomics [43,52,68,69]. Further medical specialties were environmental health (2%, 1/49) [64], genomic cancer medicine (2%, 1/49) [51], neuroimaging research (2%, 1/49) [12], and pulmonary hypertension (2%, 1/49) [35]. Despite studies implementing FHIR in specific use cases, 55% (12/22) of the studies [1,12,22,23,25,35,43,50,52,58,64,69] reported generic solutions conferrable to other use cases. Details on study objectives with regards to FHIR use can be found in Table 2 and Multimedia Appendix 3.

International Standards

Among the 49 studies, 37% (n=18) did not report on or use additional standards or terminologies [12,22-24,27,28,30-32,38,39,48,50,55,57,64,66,69]. SNOMED CT [1,25,26,35,37,43,45-47,51,55,56,65,70] and LOINC [25,26,35,37,42-45,47,49,51,55,56,58,61,65,68,70] were reported to be used by 29% (n=14) and 37% (n=18) of the studies, respectively; 18% (n=9) of the studies used International Classification of Diseases 10th Revision [25,26,35,37,49,51,58,65,68] and 12% (n=6) used OMOP CDM [26,29,35,47,52,60]; and 43% (n=21) of the studies used additional standards which were categorized under "Other" (Table 1) [26,34-37,40,42,43,45-47,49,51,56,58-60,62,63,67,70]. The implemented FHIR resources by each study are listed in Table 2; 5 (10%) studies did not precisely list their FHIR resources used [36,38,56,64,69]. Information on the FHIR version used was provided by 45% (n=22) of the studies [22,23,25,26,28,30,32,35,40,42,48,49,57,59,60,66,68,70], which can be found in Multimedia Appendix 4.

Limitations of FHIR Use

With regard to the limitations of FHIR use, Bild et al [28], Lackerbauer et al [32], and Metke-Jimenez et al [43] reported the possible content changes of new versions of FHIR resources. Generalizability was a concern in the studies of Khalique et al [38] and Zong et al [65]. The need for a FHIR server [69] and the requirement for a protocol for deidentification [1] were additional limitations. Reinecke et al [29] had not tested the exchange of data between locations and therefore could not provide information in terms of use and results of their prototype. Wagholikar et al [36] implemented a limited subset of FHIR resources in their platform and therefore the filtering of FHIR resources using complex query formats was not supported. In terms of electronic consents, safety and legal matters were major concerns [39]. Zong et al [68], investigating the discovery of genotype-phenotype associations, reported the lack of information on differences in genetic data as well as extra mapping efforts since the data were from multiple sources. In addition, there was a lack of resources preventing the demonstration of use in the study. Generalizability was also a concern in this study in terms of exploring the FHIR framework within other variants and noncancer phenotypes in future work.

Discussion

Principal Findings

This systematic review summarizes the current state of use cases implementing FHIR in health research. As FHIR was developed in 2011, we included studies from 2011 to 2022 and found that half of studies were published between 2020 and 2022, displaying an increased use of FHIR in the past years. Interestingly, the first publication of our included studies emerged in 2016, indicating a 5-year latency between the publication of the FHIR standard and the publication of studies addressing its use in health research. Germany and the United States were the countries with the highest number of publications, which might be due to recent regulatory measurements and initiatives: in the United States, the 21st Century Cures Act requires the use of FHIR for health data; and in Germany, the medical informatics initiative aiming to close the gap between research and health care used FHIR in their already established use cases. Our network analysis showed that authorships were dispersed relatively equally across studies, not dominated by individual research groups or authors.

Most studies aimed to primarily standardize their data for health research and reported using additional international standards and terminologies. Within studies using FHIR for data capture, the FHIR resource "Questionnaire" was often used. Further areas of FHIR use were analysis, recruitment, and consent management. The literature shows that fast and efficient patient screening for clinical trial recruitment support systems is important, and there is a current lack of standards and interoperability of in these systems, as well as with regard to eligibility criteria [72].

The majority of studies followed a more generic approach rather than implementing FHIR for a specific use case. The studies establishing use cases focused on infectious diseases, specifically COVID-19, as well as genomics, oncology, and imaging—which are all specialties more advanced in terms of digitalization. Among these use cases, only a small number of studies reported limited generalizability of their results.

Though provided by FHIR specifically for research, resources out of the domains "Public Health & Research" and "Evidence-Based Medicine" were used in only 4 studies. A recently published study investigated the feasibility of the FHIR resource "ResearchStudy" in a metadata registry for COVID-19 research and found that there was a need for the use of extensions on more than 20% of the data items [41]. However, the resources "ResearchStudy" and "ResearchSubject" are currently under revision and will likely be tailored more to researchers' needs when released with FHIR version R5 in 2022 [73].

Our analysis found that FHIR was used as a complement to other standards. Studies reporting on terminologies mostly used SNOMED CT and LOINC, both terminologies supported by FHIR within its value sets. There were 6 studies that used FHIR in addition to OMOP CDM, a standard widely used in observational research. Using OMOP CDM, a recommended way of transforming and transferring data from existing databases—Extract-Transform-Load tools are used for each source separately. To connect multiple heterogeneous databases, FHIR can be used as an intermediate format for local data extraction [35]. Reinecke et al [29] also extended the OMOP CDM with FHIR to exchange electronic health record data to connect the CDM to several health care systems. However, there were also limitations as Leroux et al [67] mapped CDISC SDTM and FHIR and found that CDISC SDTM's use of controlled terminology is inhibiting semantic interoperability solutions; FHIR uses semantic standards accepted in health care that are usually precoordinated (eg, SNOMED CT and LOINC), whereas CDISC SDTM uses only controlled terminology in postcoordination. Therefore, there would be the need for sponsors to translate terminologies used within systems. Leroux et al [67] proposed the new FHIR resources "ClinicalStudyPlan" and "ClinicalStudyData"—equivalent to ODM "Study" and "ClinicalData" elements—which could overcome the semantic incompatibility. However, mappings with data transformation may lead to information loss and errors; therefore, developing ODM toward FHIR would be preferable, and the draft of ODM version 2.0 already includes better support for FHIR [74,75]. In addition, HL7 and CDISC have jointly released a mapping implementation guide to help transform FHIR content into CDISC Clinical Data Acquisition Standards Harmonization Implementation Guide or SDTM Implementation Guide data sets. [76,77].

With regard to limitations using FHIR, there were certain drawbacks reported by the included studies such as the possible change in the content of different versions of FHIR resources, safety, legal matters, and the need for a FHIR server. Not all studies tested the use of FHIR in practice and, therefore, could not provide results on the actual FHIR implementation.

Limitations

One limitation of our study is the lack of quality evaluation due to missing established tools for evaluating technical papers on standardization in health care. For technical evaluations, structured information on additional standards, software, and FHIR version was missing in several studies. Therefore, our analysis on additional used standards might be biased as half of the studies did not report on using other international standards or terminologies. In addition, there were studies that did not list their FHIR resources clearly or at all. We aimed to guarantee an optimal systematic review process targeting academic peer-reviewed literature that is available in English; however, limitations remained as we may have missed relevant studies that were not published in the target language. Furthermore, we assumed that the published literature provides a surplus on successful FHIR initiatives because, in general, unsuccessful initiatives tend to stay unpublished [78]. Thus, our review may suffer from publication bias. In addition, this study investigated studies with a clear focus on FHIR in health research. However, there might be research projects using FHIR without FHIR being the central message or included in title and abstract.

Conclusions

To the best of our knowledge, this is the first systematic review investigating the use of FHIR in health research. It was shown that FHIR has been successfully implemented in clinical, public health, and epidemiological research at the stages of recruitment and consent management, data capture, and standardization as well as analysis of patient data. The implementation of international terminologies such as SNOMED CT and LOINC is common and, together with the REST API, stands out in comparison with other health research standards. Other standards such as OMOP CDM were used as a complement to FHIR in some studies, and a future aim could be the development of an infrastructure for the seamless integration and communication of health information across different standards. This approach is reinforced by the current development of collaborations of different Standards Developing Organizations such as OHDSI and FHIR and the improved support of FHIR in combination with CDISC. Resources of the domain "Public Health & Research" and "Evidence-Based Medicine" were rarely used and could further elevate interoperability in health research, specifically after their modifications in FHIR version R5. However, this approach will need to address current limitations but could, if successfully implemented, elevate digitalized health research.

Acknowledgments

This work was done as part of the National Research Data Infrastructure for Personal Health Data Consortium [79]. We gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (German Research Foundation; NFDI 13/1).

Conflicts of Interest

ST is the vice chair of Health Level Seven Deutschland. The remaining authors declare no other conflicts of interests.

Multimedia Appendix 1
Search strategy for each database.
[DOCX File , 32 KB - medinform_v10i7e35724_app1.docx]

Multimedia Appendix 2
Reasons for the exclusion of full-text evaluation.
[DOCX File , 61 KB - medinform_v10i7e35724_app2.docx]

Multimedia Appendix 3
Word cloud showing the keywords of the main FHIR use objectives of the studies. FHIR: Fast Healthcare Interoperability Resources.
[PNG File , 31 KB - medinform_v10i7e35724_app3.png]

Multimedia Appendix 4
Additional information on the included studies.
[DOCX File , 43 KB - medinform_v10i7e35724_app4.docx]

References

1. Bauer DC, Metke-Jimenez A, Maurer-Stroh S, Tiruvayipati S, Wilson LOW, Jain Y, et al. Interoperable medical data: The missing link for understanding COVID-19. Transbound Emerg Dis 2021 Jul;68(4):1753-1760 [FREE Full text] [doi: 10.1111/tbed.13892] [Medline: 33095970]
2. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
3. ODM-XML. Clinical Data Interchange Standards Consortium. URL: https://www.cdisc.org/standards/data-exchange/odm [accessed 2021-06-17]
4. Resourcelist - FHIR v4.3.0. Health Level Seven International. URL: https://www.hl7.org/fhir/resourcelist.html [accessed 2021-12-20]
5. Index - FHIR v4.3.0. Health Level Seven International. URL: https://hl7.org/FHIR/index.html [accessed 2021-09-03]
6. Neues Bündnis für mehr Interoperabilität im Gesundheitswesen. Ärzte Zeitung. 2020 Aug 26. URL: https://www.aerztezeitung.de/Wirtschaft/Neues-Buendnis-fuer-mehr-Interoperabilitaet-im-Gesundheitswesen-412313.html [accessed 2021-08-30]
7. Policies and technology for interoperability and burden reduction. Centers for Medicare & Medicaid Services. URL: https://www.cms.gov/Regulations-and-Guidance/Guidance/Interoperability/index [accessed 2021-08-30]
8. Craven J. FDA drafts data standards guidance for RWD. Regulatory Focus. 2021 Oct 22. URL: https://www.raps.org/news-and-articles/news-articles/2021/10/fda-drafts-data-standards-guidance-for-rwd [accessed 2022-04-22]
9. Vulcan. Health Level Seven International. URL: https://www.hl7.org/vulcan/ [accessed 2022-04-22]
10. MedMorph Research Data Exchange Content IG. Health Level Seven International. URL: http://build.fhir.org/ig/HL7/fhir-medmorph-research-dex-ig/index.html [accessed 2022-04-22]

11. HL7 International and OHDSI announce collaboration to provide single common data model for sharing information in clinical care and observational research. Observational Health Data Science and Informatics. URL: https://www.ohdsi.org/ohdsi-hl7-collaboration/ [accessed 2022-04-22]
12. Khvastova M, Witt M, Essenwanger A, Sass J, Thun S, Krefting D. Towards interoperability in clinical research - enabling FHIR on the open-source research platform XNAT. J Med Syst 2020 Jul 09;44(8):137 [FREE Full text] [doi: 10.1007/s10916-020-01600-y] [Medline: 32642856]
13. Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. Neuroinformatics 2007;5(1):11-34. [doi: 10.1385/ni:5:1:11] [Medline: 17426351]
14. Lehne M, Luijten S, Vom Felde Genannt Imbusch P, Thun S. The use of FHIR in digital health - a review of the scientific literature. Stud Health Technol Inform 2019 Sep 03;267:52-58. [doi: 10.3233/SHTI190805] [Medline: 31483254]
15. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. Correction: The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021 Aug 17;9(8):e32869 [FREE Full text] [doi: 10.2196/32869] [Medline: 34403353]
16. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
17. Vorisek C, Klopfenstein S, Lehne M, Haese T, Bartschke A, Thun S. Use of Fast Healthcare Interoperability Resources (FHIR) in health care research data – a systematic review. National Institute for Health Research. 2021 Mar 13. URL: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=235393 [accessed 2021-03-15]
18. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
19. The R Project for Statistical Computing. R Foundation for Statistical Computing. URL: https://www.r-project.org./ [accessed 2021-06-17]
20. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. JOSS 2019 Nov;4(43):1686. [doi: 10.21105/joss.01686]
21. FHIR-Research-Review. GitHub. URL: https://github.com/BIH-CEI/FHIR-Research-Review [accessed 2021-12-20]
22. Pfiffner PB, Pinyol I, Natter MD, Mandl KD. C3-PRO: Connecting ResearchKit to the Health System Using i2b2 and FHIR. PLoS One 2016;11(3):e0152722 [FREE Full text] [doi: 10.1371/journal.pone.0152722] [Medline: 27031856]
23. Lee H, Kung H, Lee Y, Chao JC, Udayasankaran JG, Fan H, et al. Global infectious disease surveillance and case tracking system for COVID-19: development study. JMIR Med Inform 2020 Dec 22;8(12):e20567 [FREE Full text] [doi: 10.2196/20567] [Medline: 33320826]
24. Eapen BR, Costa A, Archer N, Sartipi K. FHIRForm: an open-source framework for the management of electronic forms in healthcare. Stud Health Technol Inform 2019;257:80-85. [doi: 10.3233/978-1-61499-951-5-80] [Medline: 30741177]
25. Hong N, Prodduturi N, Wang C, Jiang G. Shiny FHIR: an integrated framework leveraging Shiny R and HL7 FHIR to empower standards-based clinical data applications. Stud Health Technol Inform 2017;245:868-872 [FREE Full text] [doi: 10.3233/978-1-61499-830-3-868] [Medline: 29295223]
26. Sass J, Bartschke A, Lehne M, Essenwanger A, Rinaldi E, Rudolph S, et al. The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. BMC Med Inform Decis Mak 2020 Dec 21;20(1):341 [FREE Full text] [doi: 10.1186/s12911-020-01374-w] [Medline: 33349259]
27. Tanaka K, Yamamoto R. Implementation of a secured cross-institutional data collection infrastructure by applying HL7 FHIR on an existing distributed EMR storages. Stud Health Technol Inform 2020 Jun 26;272:155-158. [doi: 10.3233/SHTI200517] [Medline: 32604624]
28. Bild R, Bialke M, Buckow K, Ganslandt T, Ihrig K, Jahns R, et al. Towards a comprehensive and interoperable representation of consent-based data usage permissions in the German medical informatics initiative. BMC Med Inform Decis Mak 2020 Jun 05;20(1):103 [FREE Full text] [doi: 10.1186/s12911-020-01138-6] [Medline: 32503529]
29. Reinecke I, Gulden C, Kümmel M, Nassirian A, Blasini R, Sedlmayr M. Design for a modular clinical trial recruitment support system based on FHIR and OMOP. Stud Health Technol Inform 2020 Jun 16;270:158-162. [doi: 10.3233/SHTI200142] [Medline: 32570366]
30. Gulden C, Blasini R, Nassirian A, Stein A, Altun FB, Kirchner M, et al. Prototypical clinical trial registry based on Fast Healthcare Interoperability Resources (FHIR): design and implementation study. JMIR Med Inform 2021 Jan 12;9(1):e20470 [FREE Full text] [doi: 10.2196/20470] [Medline: 33433393]
31. Gulden C, Mate S, Prokosch HU, Kraus S. Investigating the capabilities of FHIR search for clinical trial phenotyping. Stud Health Technol Inform 2018;253:3-7. [doi: 10.3233/978-1-61499-896-9-3] [Medline: 30147028]
32. Lackerbauer AM, Krauss O, Helm E. Automated verification of structured questionnaires using HL7. Stud Health Technol Inform 2019;258:11-15. [doi: 10.3233/978-1-61499-959-1-11] [Medline: 30942704]
33. Jenkins KJ, Gauvreau K, Newburger JW, Spray TL, Moller JH, Iezzoni LI. Consensus-based method for risk adjustment for surgery for congenital heart disease. J Thorac Cardiovasc Surg 2002 Jan;123(1):110-118. [doi: 10.1067/mtc.2002.119064] [Medline: 11782764]

34. Ulrich H, Kock AK, Duhm-Harbeck P, Habermann JK, Ingenerf J. Metadata repository for improved data sharing and reuse based on HL7 FHIR. Stud Health Technol Inform 2016;228:162-166. [doi: 10.3233/978-1-61499-678-1-162] [Medline: 27577363]
35. Fischer P, Stöhr MR, Gall H, Michel-Backofen A, Majeed RW. Data integration into OMOP CDM for heterogeneous clinical data collections via HL7 FHIR bundles and XSLT. Stud Health Technol Inform 2020 Jun 16;270:138-142. [doi: 10.3233/SHTI200138] [Medline: 32570362]
36. Wagholikar KB, Jain R, Oliveira E, Mandel J, Klann J, Colas R, et al. Evolving research data sharing networks to clinical app sharing networks. AMIA Jt Summits Transl Sci Proc 2017;2017:302-307 [FREE Full text] [Medline: 28815145]
37. Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc 2018 May 01;25(5):530-537 [FREE Full text] [doi: 10.1093/jamia/ocx160] [Medline: 29361077]
38. Khalique F, Khan SA. An FHIR-based framework for consolidation of augmented EHR from hospitals for public health analysis. 2017 Presented at: 2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT); September 20-22, 2017; Moscow, Russia p. 1-4. [doi: 10.1109/icaict.2017.8687289]
39. Weber M, Griessbach A, Grossmann R, Blaser J. A FHIr-based eConsent app for the digital hospital. Stud Health Technol Inform 2020 Jun 16;270:3-7. [doi: 10.3233/SHTI200111] [Medline: 32570335]
40. Bialke M, Bahls T, Geidel L, Rau H, Blumentritt A, Pasewald S, et al. MAGIC: once upon a time in consent management-a FHIR tale. J Transl Med 2018 Sep 14;16(1):256 [FREE Full text] [doi: 10.1186/s12967-018-1631-3] [Medline: 30217236]
41. Klopfenstein SAI, Vorisek CN, Shutsko A, Lehne M, Sass J, Löbe M, et al. Fast Healthcare Interoperability Resources (FHIR) in a FAIR metadata registry for COVID-19 research. Stud Health Technol Inform 2021 Nov 18;287:73-77. [doi: 10.3233/SHTI210817] [Medline: 34795084]
42. Gruendner J, Gulden C, Kampf M, Mate S, Prokosch H, Zierk J. A framework for criteria-based selection and processing of Fast Healthcare Interoperability Resources (FHIR) data for statistical analysis: design and implementation study. JMIR Med Inform 2021 Apr 01;9(4):e25645 [FREE Full text] [doi: 10.2196/25645] [Medline: 33792554]
43. Metke-Jimenez A, Hansen D. FHIRCap: Transforming REDCap forms into FHIR resources. AMIA Jt Summits Transl Sci Proc 2019;2019:54-63 [FREE Full text] [Medline: 31258956]
44. Cheng AC, Duda SN, Taylor R, Delacqua F, Lewis AA, Bosler T, et al. REDCap on FHIR: clinical data interoperability services. J Biomed Inform 2021 Sep;121:103871 [FREE Full text] [doi: 10.1016/j.jbi.2021.103871] [Medline: 34298155]
45. Rinaldi E, Saas J, Thun S. Use of LOINC and SNOMED CT with FHIR for microbiology data. Stud Health Technol Inform 2021 May 24;278:156-162. [doi: 10.3233/SHTI210064] [Medline: 34042889]
46. Sass J, Zabka S, Essenwanger A, Schepers J, Boeker M, Thun S. Fast Healthcare Interoperability Resources (FHIR®) representation of medication data derived from German Procedure Classification Codes (OPS) using Identification of Medicinal Products (IDMP) compliant terminology. Stud Health Technol Inform 2021 May 24;278:231-236. [doi: 10.3233/SHTI210074] [Medline: 34042899]
47. Rinaldi E, Thun S. From OpenEHR to FHIR and OMOP data model for microbiology findings. Stud Health Technol Inform 2021 May 27;281:402-406. [doi: 10.3233/SHTI210189] [Medline: 34042774]
48. Wang X, Lehmann H, Botsis T. Can FHIR support standardization in post-market safety surveillance? Stud Health Technol Inform 2021 May 27;281:33-37. [doi: 10.3233/SHTI210115] [Medline: 34042700]
49. Brandt PS, Pacheco JA, Rasmussen LV. Development of a repository of computable phenotype definitions using the clinical quality language. JAMIA Open 2021 Oct;4(4):ooab094 [FREE Full text] [doi: 10.1093/jamiaopen/ooab094] [Medline: 34926996]
50. Lenert LA, Ilatovskiy AV, Agnew J, Rudisill P, Jacobs J, Weatherston D, et al. Automated production of research data marts from a canonical fast healthcare interoperability resource data repository: applications to COVID-19 research. J Am Med Inform Assoc 2021 Jul 30;28(8):1605-1611 [FREE Full text] [doi: 10.1093/jamia/ocab108] [Medline: 33993254]
51. Guérin J, Laizet Y, Le Texier V, Chanas L, Rance B, Koeppel F, et al. OSIRIS: a minimum data set for data sharing and interoperability in oncology. JCO Clin Cancer Inform 2021 Mar;5:256-265 [FREE Full text] [doi: 10.1200/CCI.20.00094] [Medline: 33720747]
52. Peng Y, Nassirian A, Ahmadi N, Sedlmayr M, Bathelt F. Towards the representation of genomic data in HL7 FHIR and OMOP CDM. Stud Health Technol Inform 2021 Sep 21;283:86-94. [doi: 10.3233/SHTI210545] [Medline: 34545823]
53. Hund H, Wettstein R, Heidt CM, Fegeler C. Executing distributed healthcare and research processes - the HiGHmed data sharing framework. Stud Health Technol Inform 2021 May 24;278:126-133. [doi: 10.3233/SHTI210060] [Medline: 34042885]
54. Greiner MV, Beal SJ, Dexheimer J, Krummen K. Evaluating IDENTITY, an automated data sharing platform to improve health outcomes for youth in protective custody. Pediatr 2020 Jul 01;146(1_MeetingAbstract):507-508 [FREE Full text] [doi: 10.1542/peds.146.1MA6.507]
55. González-Castro L, Cal-González VM, Del Fiol G, López-Nores M. CASIDE: A data model for interoperable cancer survivorship information based on FHIR. J Biomed Inform 2021 Dec;124:103953 [FREE Full text] [doi: 10.1016/j.jbi.2021.103953] [Medline: 34781009]

56. Banach A, Ulrich H, Kroll B, Kiel A, Ingenerf J, Kock-Schoppenhauer AK. APERITIF - Automatic Patient Recruiting for Clinical Trials Based on HL7 FHIR. Stud Health Technol Inform 2021 May 27;281:58-62. [doi: 10.3233/SHTI210120] [Medline: 34042705]
57. Deppenwiese N, Delpy P, Lambarki M, Lablans M. ADT2FHIR - a tool for converting ADT/GEKID oncology data to HL7 FHIR resources. Stud Health Technol Inform 2021 Sep 21;283:104-110. [doi: 10.3233/SHTI210547] [Medline: 34545825]
58. Lambarki M, Kern J, Croft D, Engels C, Deppenwiese N, Kerscher A, et al. Oncology on FHIR: A Data Model for Distributed Cancer Research. Stud Health Technol Inform 2021 May 24;278:203-210. [doi: 10.3233/SHTI210070] [Medline: 34042895]
59. Kilintzis V, Beredimas N, Kaimakamis E, Stefanopoulos L, Chatzis E, Jahaj E, et al. CoCross: An ICT Platform Enabling Monitoring Recording and Fusion of Clinical Information Chest Sounds and Imaging of COVID-19 ICU Patients. Healthcare (Basel) 2022 Jan 30;10(2):276 [FREE Full text] [doi: 10.3390/healthcare10020276] [Medline: 35206889]
60. Majeed RW, Stöhr MR, Günther A. HIStream-Import: A Generic ETL Framework for Processing Arbitrary Patient Data Collections or Hospital Information Systems into HL7 FHIR Bundles. Stud Health Technol Inform 2021 May 24;278:75-79. [doi: 10.3233/SHTI210053] [Medline: 34042878]
61. Garza MY, Rutherford M, Myneni S, Fenton S, Walden A, Topaloglu U, et al. Evaluating the coverage of the HL7 FHIR standard to support eSource data exchange implementations for use in multi-site clinical research studies. AMIA Annu Symp Proc 2020;2020:472-481 [FREE Full text] [Medline: 33936420]
62. Wettstein R, Hund H, Kobylinski I, Fegeler C, Heinze O. Feasibility Queries in Distributed Architectures - Concept and Implementation in HiGHmed. Stud Health Technol Inform 2021 May 24;278:134-141. [doi: 10.3233/SHTI210061] [Medline: 34042886]
63. Wettstein R, Hund H, Fegeler C, Heinze O. Data Sharing in Distributed Architectures - Concept and Implementation in HiGHmed. Stud Health Technol Inform 2021 Sep 21;283:111-118. [doi: 10.3233/SHTI210548] [Medline: 34545826]
64. Xu H, Cox S, Stillwell L, Pfaff E, Champion J, Ahalt SC, et al. FHIR PIT: an open software application for spatiotemporal integration of clinical data and environmental exposures data. BMC Med Inform Decis Mak 2020 Mar 11;20(1):53 [FREE Full text] [doi: 10.1186/s12911-020-1056-9] [Medline: 32160884]
65. Zong N, Wen A, Stone DJ, Sharma DK, Wang C, Yu Y, et al. Developing an FHIR-Based Computational Pipeline for Automatic Population of Case Report Forms for Colorectal Cancer Clinical Trials Using Electronic Health Records. JCO Clin Cancer Inform 2020 Mar;4:201-209 [FREE Full text] [doi: 10.1200/CCI.19.00116] [Medline: 32134686]
66. Zong N, Stone DJ, Sharma DK, Wen A, Wang C, Yu Y, et al. Modeling cancer clinical trials using HL7 FHIR to support downstream applications: a case study with colorectal cancer data. Int J Med Inform 2021 Jan;145:104308 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104308] [Medline: 33160272]
67. Leroux H, Metke-Jimenez A, Lawley MJ. Towards achieving semantic interoperability of clinical study data with FHIR. J Biomed Semantics 2017 Sep 19;8(1):41 [FREE Full text] [doi: 10.1186/s13326-017-0148-7] [Medline: 28927443]
68. Zong N, Sharma DK, Yu Y, Egan JB, Davila JI, Wang C, et al. Developing a FHIR-based framework for phenome wide association studies: a case study with a pan-cancer cohort. AMIA Jt Summits Transl Sci Proc 2020;2020:750-759 [FREE Full text] [Medline: 32477698]
69. Gruendner J, Wolf N, Tögel L, Haller F, Prokosch H, Christoph J. Integrating genomics and clinical data for statistical analysis by using GEnome MINIng (GEMINI) and Fast Healthcare Interoperability Resources (FHIR): system design and implementation. J Med Internet Res 2020 Oct 07;22(10):e19879 [FREE Full text] [doi: 10.2196/19879] [Medline: 33026356]
70. Jiang G, Kiefer RC, Sharma DK, Prud'hommeaux E, Solbrig HR. A consensus-based approach for harmonizing the OHDSI common data model with HL7 FHIR. Stud Health Technol Inform 2017;245:887-891 [FREE Full text] [doi: 10.3233/978-1-61499-830-3-887] [Medline: 29295227]
71. Brandt PS, Pacheco JA, Rasmussen LV. Development of a repository of computable phenotype definitions using the clinical quality language. JAMIA Open 2021 Oct;4(4):ooab094 [FREE Full text] [doi: 10.1093/jamiaopen/ooab094] [Medline: 34926996]
72. Köpcke F, Prokosch H. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res 2014 Jul 01;16(7):e161 [FREE Full text] [doi: 10.2196/jmir.3446] [Medline: 24985568]
73. Resourcelist - FHIR v4.2.0. Health Level Seven International. URL: http://hl7.org/fhir/2020Feb/resourcelist.html [accessed 2021-12-21]
74. Kubrick W. CR 3.0 – a manifesto for the next generation of clinical research data standards. Reimagine Research. 2016 Aug 30. URL: https://waynekubick.com/2016/08/30/cr-3-0-a-manifesto-for-the-next-generation-of-clinical-research-data-standards/ [accessed 2021-06-17]
75. Aerts J. Towards a single data exchange standard for use in healthcare and in clinical research. Stud Health Technol Inform 2018;248:55-63. [doi: 10.3233/978-1-61499-858-7-55] [Medline: 29726419]
76. FHIR to CDISC joint mapping implementation guide v1.0. Clinical Data Interchange Standards Consortium. 2021 Sep 01. URL: https://www.cdisc.org/standards/real-world-data/fhir-cdisc-joint-mapping-implementation-guide-v1-0 [accessed 2022-04-22]
77. FHIR to CDISC joint mapping implementation guide 1.0.0 - STU 1. Health Level Seven International. URL: http://hl7.org/fhir/uv/cdisc-mapping/STU1/ [accessed 2022-04-22]

78. Olson CM, Rennie D, Cook D, Dickersin K, Flanagin A, Hogan JW, et al. Publication bias in editorial decision making. JAMA 2002 Jun 05;287(21):2825-2828. [doi: 10.1001/jama.287.21.2825] [Medline: 12038924]
79. National Research Data Infrastructure for Personal Health Data. NFDI4Health Consortium. URL: https://www.nfdi4health.de/ [accessed 2022-06-21]

Abbreviations

API: application programming interface
CDISC: Clinical Data Interchange Standards Consortium
CDM: common data model
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven International
LOINC: Logical Observation Identifiers Names and Codes
ODM: Operational Data Standard
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Observational Medical Outcomes Partnership
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
REST: Representational State Transfer
SDTM: Study Data Tabulation Model
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms

Edited by C Lovis; submitted 29.12.21; peer-reviewed by S Hume, J Hüsers; comments to author 25.02.22; revised version received 22.04.22; accepted 18.05.22; published 19.07.22.

Please cite as:
Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, Thun S
Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review
JMIR Med Inform 2022;10(7):e35724
URL: https://medinform.jmir.org/2022/7/e35724
doi:10.2196/35724
PMID:35852842

Viewpoint

The Power of Patient Engagement With Electronic Health Records as Research Participants

Jeff Pawelek[1], BSc; Katie Baca-Motes[1], MBA; Jay A Pandit[1], MD; Benjamin B Berk[2], MD; Edward Ramos[1,2], PhD

[1]Digital Trials Center, Scripps Research Translational Institute, La Jolla, CA, United States

[2]CareEvolution, Ann Arbor, MI, United States

Corresponding Author:
Edward Ramos, PhD
Digital Trials Center
Scripps Research Translational Institute
3344 N Torrey Pines Ct
Plaza Level
La Jolla, CA, 92037
United States
Phone: 1 858 784 2028
Email: eramos@scripps.edu

Abstract

Electronic health record (EHR) technology has become a central digital health tool throughout health care. EHR systems are responsible for a growing number of vital functions for hospitals and providers. More recently, patient-facing EHR tools are allowing patients to interact with their EHR and connect external sources of health data, such as wearable fitness trackers, personal genomics, and outside health services, to it. As patients become more engaged with their EHR, the volume and variety of digital health information will serve an increasingly useful role in health care and health research. Particularly due to the COVID-19 pandemic, the ability for the biomedical research community to pivot to fully remote research, driven largely by EHR data capture and other digital health tools, is an exciting development that can significantly reduce burden on study participants, improve diversity in clinical research, and equip researchers with more robust clinical data. In this viewpoint, we describe how patient engagement with EHR technology is poised to advance the digital clinical trial space, an innovative research model that is uniquely accessible and inclusive for study participants.

(JMIR Med Inform 2022;10(7):e39145) doi:10.2196/39145

KEYWORDS

electronic health record; EHR; digital health technology; digital clinical trial; underrepresentation; underrepresented in biomedical research; biomedical research

Introduction

The electronic health record (EHR) represents an evolution from static paper-based records to a more portable, interactive, and dynamic medium shared by tools like web-based patient portals and mobile apps. The wealth of information provided in each EHR—such as medical history, medications, diagnoses, treatments, procedures, allergies, laboratory tests, immunizations, hospital admissions, and clinic visits—creates new opportunities. Advancements in how patients access their EHR have resulted in significant expansion in how EHR data are operationalized by providers to inform and deliver care. Further, the patient's ability to access and share their EHR data directly with researchers has opened the door for the research community to glean clinically important information from study cohorts.

The mechanisms by which patients interact with their EHR are in a fluid state of development, and patients expect improved functionality in how they interact with their EHR [1]. It is no surprise that the expanding reach of the digital health system for ambulatory data capture coincides with the expansion of digital clinical research that seeks to leverage these data to gain insights at both individual and population levels. As digital health technologies become more accessible and integrated into daily living, the increasing ubiquity of real-world and real-time health data stands to transform how researchers address questions about health and disease. In this viewpoint, we describe how recent advancements in EHR technology are advancing the digital clinical trial space, an innovative research model that is uniquely accessible, scalable, inclusive, and impactful for study participants and researchers.

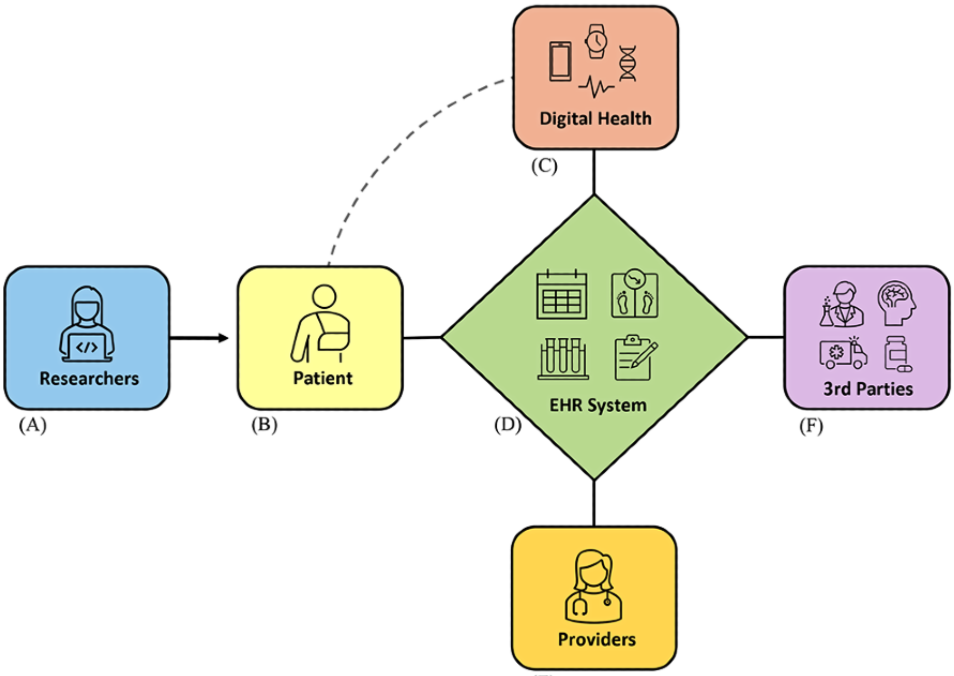
Policy Feeding Progress

Over a decade ago, the Health Information Technology for Economic and Clinical Health (HITECH) Act was enacted to incentivize the "meaningful use" of health information with an emphasis on more widespread adoption of EHRs by hospital systems and medical practices. While there is evidence that the HITECH Act stimulated the uptake of EHR systems across more medical facilities, the development of an interoperable health IT environment using modern internet technology and technical standards was not adequately addressed [2-4]. As a result, subsequent small- and large-scale initiatives have helped unlock some of the more advanced capabilities of EHR connectivity and compatibility.

Harvard Medical School and Boston Children's Hospital created Substitutable Medical Apps and Reusable Technology (SMART), an application programming interface standard that establishes compatibility to allow any EHR-based software application to function with any EHR system, thus equipping hospitals with a broader selection of EHR tools to support ever-changing clinical and business needs. However, the introduction of Fast Healthcare Interoperability Resources (FHIR) proved to be a tipping point and a crucial piece to the elaborate puzzle of health IT infrastructure. The not-for-profit organization Health Level Seven International (HL7) created FHIR, a technical standard that defines how EHR data are accessed and exchanged between different computer systems. Given HL7's robust global community of developers and stakeholders, the FHIR standard gained significant traction within the health IT community [5]. The coupling of the SMART and FHIR standards (known as "SMART on FHIR") is now considered an essential toolkit by hospitals, researchers, and the health IT industry to improve the interoperability of EHR systems [6,7].

More recently, SMART on FHIR resources have been integrated with Epic, Cerner, and other widely used EHR platforms to operationalize clinical decision support tools for providers, such as risk prediction tools for surgical procedures and advanced treatment modalities [8,9]. SMART on FHIR was recently used by Apple for their Health app, which allows users to link their device-generated health data directly to their EHR as a means to consolidate their digital health information [10]. The latest federal mandate, the 21st Century Cures Act, primarily relies on FHIR to expand meaningful *patient use* of EHR systems by incentivizing advancements to patient-facing digital health services and apps, bidirectional sharing of health information between patients and providers, and patient-mediated sharing of EHR with researchers [11,12]. These recent developments have equipped patients with the ability to use their mobile or desktop devices to access physician's notes and laboratory test results, schedule medical appointments, link health and activity monitors, and search for and enroll in clinical trials (Figure 1).

Figure 1. The widespread adoption of SMART on FHIR technical standards has enabled EHR systems to serve as a hub for the secure and efficient exchange of digital health information. (A) Researchers can partner directly with patients to participate in clinical research, and patients can chose to grant permission to researchers to access and use their EHR data; (B) patients can view and manage their EHR through an online patient portal using mobile and desktop devices; (C) patients can link their personal digital health products (eg, fitness trackers, wearable health monitors, at-home genomic tests) to their EHR as a way to centralize various elements of their health information; (D) EHR systems allow patients to schedule appointments with their provider, view provider notes, communicate with their provider, complete routine health surveys, and find opportunities to participate in research; (E) providers enter their clinical notes into their patient's EHR, access external patient-provided digital health information, and work with their patient to ensure critical health information is accurate and current; (F) patients can link their health information from third-party services such as outside providers, imaging centers, laboratories, and pharmacies. EHR: electronic health record; FHIR: Fast Healthcare Interoperability Resources; SMART: Substitutable Medical Apps and Reusable Technology.



The Expanding Role of EHR in Clinical Research

The COVID-19 pandemic put a spotlight on digital health and whether existing technologies were poised to face the unique challenges of a global health crisis forcing remote patient monitoring. Perhaps the most rapid and expansive implementation of digital health during the pandemic was the shift to telehealth, which demonstrated that basic digital technologies were adequate to support the widespread delivery of virtual health care [13]. Other EHR-driven solutions included the ability for patients to create advance care plans, in case of severe illness; new templates to capture COVID-19 test results to inform population-level statistics; and predictive models to stratify risk and inform clinical decision-making for infected patients [14-16]. The integration of these digital health tools—driven in part by large-scale exchange and compilation of patient data—emphasizes the unique role EHR technology plays in addressing complex health problems. For clinical research, increasing EHR adoption provides evermore data to complement health survey and wearable device data, supplement missing data, and reduce participant burden.

The digital nature of EHRs makes them well-suited to be incorporated into decentralized clinical research. In contrast to the traditional, hospital-based paradigm of clinical research, decentralized studies utilize a siteless, patient-centered model that affords study participants the convenience of remote data acquisition (both active and passive) through a combination of mobile apps, wearable devices, electronic surveys, self-collected biosamples, and now—with the advancements described above—patient-mediated EHR connectivity. While the "digital divide" and other barriers still exist, decentralized digital clinical trials can be accessible to a broader spectrum of patient populations—including those who are underrepresented in biomedical research (UBR)—and diminish common barriers and selection biases such as health insurance status, medical provider affiliation, and proximity to an academic medical center [17]. Direct-to-participant recruitment strategies equip researchers with multiple avenues to engage a large pool of potential study participants, thus increasing the sample size and statistical power with which site-based research chronically struggles. Additionally, local institutional review boards offer researchers who are not affiliated with a covered entity (eg, academic medical center) more expedient regulatory pathways to study approval and initiation since HIPAA (Health Insurance Portability and Accountability Act) Privacy Rules are not applicable to noncovered entities [18].

Patient-Mediated EHR Sharing in a Real-World Digital Clinical Trial

In response to the COVID-19 pandemic, the Scripps Digital Trials Center launched the Digital Engagement & Tracking for Early Control & Treatment (DETECT) digital research platform (ClinicalTrial.gov identifier: NCT04336020) [19,20]. DETECT is an observational research effort examining whether individualized changes in heart rate, activity, or sleep—all monitored through the use of a wearable activity tracker—can serve as early indicators of viral infections. DETECT leveraged a remote, siteless research model, and the participant experience allowed for a lightweight entry process. DETECT did not require participants to connect their EHR; however, the study app allowed participants the option to share their EHR data to equip the researchers with additional information for post hoc analyses. Among the entire DETECT cohort (n=40,322), approximately 10% (n=4210) elected to connect their EHR within the study app (enrollment numbers as of March 21, 2022).

DETECT's foundational protocol was designed to support additional substudies aimed at more specific clinical questions. DETECT-At Home Early Alert and Diagnosis (DETECT-AHEAD) is one such substudy that explores the feasibility of an algorithm-driven notification system based on data from wearable sensors, with a specific focus on outcomes from UBR populations. Study participants receive an alert via the study app to perform an at-home COVID-19 test, possibly before they experience symptoms, to rule out infection as soon as possible. For DETECT-AHEAD, sharing of EHR data is a criterion for participation, so all study participants connected their EHR after completing eligibility surveys. The protocol design for DETECT-AHEAD also set goals for enrollment of UBR populations to ensure the substudy cohort was reflective of the US population (Table 1).

In DETECT-AHEAD, the average age of participants was 49.4 years, the male-to-female ratio was 0.68, 15.3% (n=69) of the cohort was under 35 years of age, and racial minorities comprised 30.4% (n=137) of the cohort. DETECT-AHEAD demonstrates that while it is possible to engage a diverse population of participants, more work needs to be done to reach UBR populations that are considered disadvantaged by the digital divide (ie, age ≥65 years, highest education grade <12, annual household income <$10,000) [21]. The methods were performed in accordance with relevant guidelines and regulations and approved by the Scripps Institutional Review Board. All study participants signed an electronic informed consent form.

Table 1. DETECT-AHEAD[a] enrollment numbers in the underrepresented in biomedical research category (self-reported).

Characteristic	Participants (N=450), n (%)
Age (≥65 years)	71 (15.8)
Gender (other)	7 (1.6)
Race	
Hispanic/Latino	45 (10)
Asian	39 (8.7)
Other (non-White)	27 (6)
Black/African American	26 (5.8)
Highest level of education (grades 1-11)	2 (0.4)
Annual household income (<$10,000)	13 (2.9)

[a]DETECT-AHEAD: Digital Engagement & Tracking for Early Control & Treatment – At Home Early Alert and Diagnosis.

Current Challenges and Limitations

While DETECT and DETECT-AHEAD demonstrate how study participants, including those from UBR groups, possess a willingness to share their EHR in a research setting, the research community continues to overcome hurdles to promote the adoption of patient-mediated EHR exchange mechanisms. Perhaps the most notable constraints are the longstanding disparities in universal access to reliable internet service and use of mobile technologies [22,23]. While the digital clinical trial model offers individuals more convenient ways to take part in research and thus fosters inclusivity, its reliance on broadband internet service continues to be a barrier.

Over a quarter (28%) of US adults who live in rural areas do not have broadband internet service, which may partly explain why rural communities interact less with EHRs compared to their urban counterparts [22,24]. Also, individuals who are 65 years and older are less likely to use EHRs, and while smartphone ownership has risen in this group in recent years, only 42% report owning a smartphone (compared to 77% of all adults over 18 years) [23,25]. African American, Asian, and Latino race, younger age (ie, <35 years), and low education level are other factors associated with low engagement with EHR systems [26]. However, there have been some positive trends in recent years. Nearly 60% of patients were offered an EHR patient portal by their health care provider—a 17% increase from 2014 to 2020 [27]. Additionally, the number of patients who downloaded their EHR data nearly doubled between 2017 and 2020, and roughly 20% of EHR users elected to link their health data to an outside caregiver, health service, or app [27]. Without improved access to the internet and connected digital health tools among UBR groups, clinical study outcomes will continue to lack diversity and thus have limited applicability.

The robustness of a patient's EHR is critically important to both providers and researchers to ensure reliable interpretation and analysis of health information. Patients who actively manage their EHR can help ensure their information is up-to-date and free of errors, but many EHR patient portals still do not offer patients editing permission to allow full control over their own health information. From a technical standpoint, as capabilities to aggregate and access data across different sources increase, so does the challenge to integrate data from multiple modalities, deal with missing data, and map discrepant terminology, including data in free-text form. The EHR itself must continue to evolve and expand its capability, for example, to enrich the clinical context for data such as images or lab results. Additionally, the persistent concern for privacy and data security must also not be overlooked as we seek to find new ways of verifying identity, securely transferring EHR data, and improving deidentification techniques. Lastly, if an EHR system lacks quality control and safeguards against erroneous information, including improper or fraudulent use of the system, serious problems can arise such as diminished quality of care and medical errors [28].

Moving Forward

As more health care organizations offer patients personalized tools to interact with and visualize their EHR data, patients will ultimately become more engaged with their provider and overall health management. Health IT professionals with expertise in user experience and interface design will serve increasingly important roles in optimizing patient engagement with their EHR and associated digital health tools. Equipping patients with permission to update information, correct errors, and connect external sources of health information is a critical step toward improving patient engagement with their EHR, which should become a universal feature across all EHR systems.

There are some promising technical solutions on the horizon. Increasingly, software services are available for cloud-based clinical data warehousing, entity extraction, terminology standardization, and record linkage, which leverage functionality developed by others at scale, obviating the need to solve these challenges for each application [29-32].

Accessing and sharing EHR data are not the only obstacles to recognizing the full potential of leveraging EHR data. Ideally, the EHR is more than a historical record of clinical outcomes but rather a dynamic asset in preventative interventions. However, for this to be realized, the EHR must continue toward a comprehensive capture of a patient's health information to meaningfully provide information back to the participant by way of at-risk assessments, prediction of outcomes, or personalized detection of disease.

Conclusion

EHR technology has made significant advances through improved compatibility across connected mobile devices, digital health products, and health IT software. The biomedical research community is beginning to harness the benefits of EHR connectedness by means of fully remote digital clinical trials, which help reduce burden on study participants and fosters diversity and inclusivity of study populations. As more patients become familiar with their EHR to manage their ever-growing sources of health information, engage with their provider, and partner with researchers, the health care community as a whole will be better equipped to optimize health and well-being for all.

Acknowledgments

Ting-Yang Hung, BS, assisted with data analysis in support of this manuscript.

Data Availability

All interested investigators will be allowed access to the analysis data set after approval of a proposal by a responsible authority at Scripps and with a data access agreement pledging to not reidentify individuals or share the data with a third party. All data inquiries should be initially addressed to the corresponding author.

Authors' Contributions

JP wrote the manuscript and performed data analysis; KB-M and JAP assisted with editing the manuscript; BBB assisted with editing the manuscript and performed data analysis; and ER assisted with editing the manuscript and supervised the project.

Conflicts of Interest

ER and BBB are current employees of CareEvolution, Inc, a health care technology company.

References

1. Miliard M. Patients want more from their EHRs. Healthcare IT News. 2014 Dec 11. URL: https://www.healthcareitnews.com/news/patients-want-more-their-ehrs [accessed 2022-06-24]
2. Adler-Milstein J, Jha AK. HITECH Act drove large gains in hospital electronic health record adoption. Health Aff (Millwood) 2017 Aug 01;36(8):1416-1422. [doi: 10.1377/hlthaff.2016.1651] [Medline: 28784734]
3. Adler-Milstein J, Holmgren A, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. J Am Med Inform Assoc 2017 Nov 01;24(6):1142-1148 [FREE Full text] [doi: 10.1093/jamia/ocx080] [Medline: 29016973]
4. Gold M, McLaughlin C. Assessing HITECH implementation and lessons: 5 years later. Milbank Q 2016 Sep 13;94(3):654-687 [FREE Full text] [doi: 10.1111/1468-0009.12214] [Medline: 27620687]

5. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021 Jul 30;9(7):e21929. [doi: 10.2196/21929]
6. Mandel J, Kreda D, Mandl K, Kohane I, Ramoni R. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
7. Saripalle R, Runyan C, Russell M. Using HL7 FHIR to achieve interoperability in patient health record. J Biomed Inform 2019 Jun;94:103188 [FREE Full text] [doi: 10.1016/j.jbi.2019.103188] [Medline: 31063828]
8. Karhade AV, Schwab JH, Del Fiol G, Kawamoto K. SMART on FHIR in spine: integrating clinical prediction models into electronic health records for precision medicine at the point of care. Spine J 2021 Oct;21(10):1649-1651. [doi: 10.1016/j.spinee.2020.06.014] [Medline: 32599144]
9. Scalia P, Ahmad F, Schubbe D, Forcino R, Durand M, Barr PJ, et al. Integrating option grid patient decision aids in the epic electronic health record: case study at 5 health systems. J Med Internet Res 2021 May 03;23(5):e22766 [FREE Full text] [doi: 10.2196/22766] [Medline: 33938806]
10. Jercich K. Apple's health data sharing feature now live. Healthcare IT News. 2021 Sep 21. URL: https://www.healthcareitnews.com/news/apples-health-data-sharing-feature-now-live [accessed 2022-06-24]
11. 21st Century Cures Act. US Food and Drug Administation. 2016. URL: https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/21st-century-cures-act [accessed 2022-03-01]
12. McGrail S. Final interoperability rule has implications for APIs, FHIR. HIT Infrastructure. 2020 Mar 12. URL: https://hitinfrastructure.com/news/final-interoperability-rule-has-implications-for-apis-fhir [accessed 2022-06-24]
13. Bokolo AJ. Use of telemedicine and virtual care for remote treatment in response to COVID-19 pandemic. J Med Syst 2020 Jun 15;44(7):132 [FREE Full text] [doi: 10.1007/s10916-020-01596-5] [Medline: 32542571]
14. Portz JD, Brungardt A, Shanbhag P, Staton EW, Bose-Brill S, Lin C, et al. Advance care planning among users of a patient portal during the COVID-19 pandemic: retrospective observational study. J Med Internet Res 2020 Aug 11;22(8):e21385 [FREE Full text] [doi: 10.2196/21385] [Medline: 32716900]
15. Osborne TF, Veigulis ZP, Arreola DM, Röösli E, Curtin CM. Automated EHR score to predict COVID-19 outcomes at US Department of Veterans Affairs. PLoS One 2020 Jul 27;15(7):e0236554 [FREE Full text] [doi: 10.1371/journal.pone.0236554] [Medline: 32716922]
16. Deeds SA, Hagan SL, Geyer JR, Vanderwarker C, Grandjean MW, Reddy A, et al. Leveraging an electronic health record note template to standardize screening and testing for COVID-19. Healthc (Amst) 2020 Sep;8(3):100454 [FREE Full text] [doi: 10.1016/j.hjdsi.2020.100454] [Medline: 32919584]
17. The Editors. Clinical trials have far too little racial and ethnic diversity. Scientific American. 2018 Sep 01. URL: https://tinyurl.com/bdfcrnnv [accessed 2022-06-24]
18. Covered entities and business associates. US Department of Health & Human Services. 2017. URL: https://www.hhs.gov/hipaa/for-professionals/covered-entities/index.html [accessed 2022-06-13]
19. Quer G, Radin JM, Gadaleta M, Baca-Motes K, Ariniello L, Ramos E, et al. Wearable sensor data and self-reported symptoms for COVID-19 detection. Nat Med 2021 Jan 29;27(1):73-77. [doi: 10.1038/s41591-020-1123-x] [Medline: 33122860]
20. Radin JM, Quer G, Ramos E, Baca-Motes K, Gadaleta M, Topol EJ, et al. Assessment of prolonged physiological and behavioral changes associated with COVID-19 infection. JAMA Netw Open 2021 Jul 01;4(7):e2115959 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.15959] [Medline: 34232306]
21. Mapes BM, Foster CS, Kusnoor SV, Epelbaum MI, AuYoung M, Jenkins G, All of Us Research Program. Diversity and inclusion for the All of Us research program: A scoping review. PLoS One 2020;15(7):1-10 [FREE Full text] [doi: 10.1371/journal.pone.0234962] [Medline: 32609747]
22. Perrin A. Mobile technology and home broadband 2021. Pew Research Center. 2021 Jun 03. URL: https://www.pewresearch.org/internet/2021/06/03/mobile-technology-and-home-broadband-2021/ [accessed 2022-06-24]
23. Perrin A. 10 facts about smartphones as the iPhone turns 10. Pew Research Center. 2017 Jun 28. URL: https://www.pewresearch.org/fact-tank/2017/06/28/10-facts-about-smartphones/ [accessed 2022-06-24]
24. Greenberg AJ, Haney D, Blake KD, Moser RP, Hesse BW. Differences in access to and use of electronic personal health information between rural and urban residents in the United States. J Rural Health 2018 Feb 11;34 Suppl 1:s30-s38 [FREE Full text] [doi: 10.1111/jrh.12228] [Medline: 28075508]
25. Calixte R, Islam S, Osakwe ZT, Rivera A, Camacho-Rivera M. Pattern of use of electronic health record (EHR) among the chronically ill: a Health Information National Trend Survey (HINTS) analysis. Int J Environ Res Public Health 2021 Jul 07;18(14):7254 [FREE Full text] [doi: 10.3390/ijerph18147254] [Medline: 34299705]
26. Irizarry T, DeVito Dabbs A, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015 Jun 23;17(6):e148 [FREE Full text] [doi: 10.2196/jmir.4255] [Medline: 26104044]
27. Johnson C, Richwine C, Patel V. Individuals' access and use of patient portals and smartphone health apps, 2020. The Office of the National Coordinator for Health Information Technology. 2021 Sep. URL: https://www.healthit.gov/data/data-briefs/individuals-access-and-use-patient-portals-and-smartphone-health-apps-2020 [accessed 2022-06-24]

28. Bowman S. Impact of electronic health record systems on information integrity: quality and safety implications. Perspect Health Inf Manag 2013;10:1c [FREE Full text] [Medline: 24159271]
29. Cloud healthcare API. Google. URL: https://cloud.google.com/healthcare-api/ [accessed 2022-06-13]
30. AWS for health. Amazon. URL: https://aws.amazon.com/health/ [accessed 2022-06-13]
31. Rosetta interface guide. CareEvolution. URL: https://rosetta-api.docs.careevolution.com/ [accessed 2022-06-13]
32. PPRL Linkage Honest Broker. Regenstrief Institute. URL: https://www.regenstrief.org/n3c-lhb/ [accessed 2022-06-13]

Abbreviations

DETECT: Digital Engagement & Tracking for Early Control & Treatment
DETECT-AHEAD: Digital Engagement & Tracking for Early Control & Treatment – At-Home Early Alert and Diagnosis
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HIPAA: Health Insurance Portability and Accountability Act
HITECH: Health Information Technology for Economic and Clinical Health
HL7: Health Level Seven International
SMART: Substitutable Medical Apps and Reusable Technology
UBR: underrepresented in biomedical research

Edited by C Lovis; submitted 02.05.22; peer-reviewed by X Xiong, E Kalenderian; comments to author 26.05.22; revised version received 16.06.22; accepted 21.06.22; published 08.07.22.

Please cite as:
Pawelek J, Baca-Motes K, Pandit JA, Berk BB, Ramos E
The Power of Patient Engagement With Electronic Health Records as Research Participants
JMIR Med Inform 2022;10(7):e39145
URL: https://medinform.jmir.org/2022/7/e39145
doi:10.2196/39145
PMID:35802410

Original Paper

The Effects of Information Continuity and Interpersonal Continuity on Physician Services Online: Cross-sectional Study

Yan Xuan[1*], MA; Chaojin Guo[1*], BA; Wei Lu[2], PhD

[1]Department of Nursing, Hainan Women and Children's Medical Center, Hainan, China

[2]School of Management, Hainan Medical University, Haikou, China

[*]these authors contributed equally

Corresponding Author:
Wei Lu, PhD
School of Management
Hainan Medical University
3 Xueyuan Road, Longhua District
Hainan Province
Haikou, 571199
China
Phone: 86 13546391040
Email: hy0205044@hainmc.edu.cn

Abstract

Background: Web-based medical services have become an effective supplement to traditional services in hospitals and an essential part of medical services. Studies have shown that web-based medical services are useful for shortening the delayed admission time and for enhancing the treatment effect from the service continuity perspective. However, the specific measures that patients and physicians should take to improve service continuity remain unknown.

Objective: Based on the information richness theory and continuity of care, this study investigates the dynamic effects of information continuity and interpersonal continuity on physician services online.

Methods: Data of 7200 patients with 360 physicians covering complete interaction records were collected from a professional web-based platform in China. Content analysis was performed to recognize matching patients and physicians, and least square regression analysis was performed to obtain all empirical results.

Results: Empirical results showed that in the short term, information continuity (including offline experience, medical records, and detailed information) influences physicians' web-based services, and their influences show heterogeneity. Moreover, if a patient's online physician is the same physician who he/she has visited offline, we find that interpersonal continuity is important for service. In the long term, information continuity and interpersonal continuity positively improve service continuity by facilitating repeat purchases.

Conclusions: Overall, our findings not only shed new light on patient behavior online and cross-channel behavior but also provide practical insights into improving the continuity of care in online health communities.

(JMIR Med Inform 2022;10(7):e35830) doi:10.2196/35830

KEYWORDS

continuity of care; web-based medical service; service quality; information continuity; interpersonal continuity

Introduction

Background

High continuity of care is the key to improving medical service quality and decreasing irrational use [1], which is an important theme of digital transformation that is receiving increasing attention. Currently, there is no universal definition of the concept and characteristics of continuity of care. However, experienced continuity, information continuity, coherence of medical records, cross-boundary and team continuity, longitudinal continuity, and interpersonal continuity are widely recognized as important elements of continuity of care [2]. As some medical services can be done using information technology, such as appointments and treatments, the use of information technology in health care could realize the mutual recognition of inspection results and sharing of medical records, thus improving the continuity of care [3].

Online health communities provide a channel for patient and physician contact conveniently by overcoming space-time limits and enriching information provision [4,5]. Web-based medical services have become an effective supplement to traditional services in hospitals and an essential part of medical services [6]. Patients can communicate with physicians via various types of services, including individual service (written consultation, phone consultation, video consultation) and team service. No matter which service patients use, they should post their questions mandatorily and provide offline treatment materials selectively if they have them. This offline information helps improve continuity of care and is useful for physicians to make an accurate diagnosis.

From the continuity of care perspective, “internet plus medical service” (a new application of the medical industry, which includes internet as the carrier and the technical method of health education, medical information query, electronic health records, disease risk evaluation, online consulting, electronic prescription, remote consultation, remote treatment and rehabilitation, and other forms of health care services) is believed to integrate medical treatment, health care, and rehabilitation, with extending medical services outside the hospital. The form of “internet plus medical service” is changing from “split” to “holistic medical treatment,” and this treatment plays a significant role in interpersonal continuity, information continuity, and geographical continuity [7]. However, the above benefits are only theoretical judgments and there are no empirical studies to examine the role of online health communities. To the best of our knowledge, this study is among the first to investigate the effects of information providing from the continuity of care perspective. Although the literature on online health communities is abundant [8-11], they rarely focus on the influence of offline experiences on online behaviors. In addition, prior studies have revealed that web-based medical services are useful for shortening the delayed admission time and for enhancing the treatment effect from the service continuity perspective [1]. However, the specific measures that patients and physicians should take to improve service continuity remain unknown. Based on the information richness theory and continuity of care, this study aims to investigate the dynamic effects of information continuity and interpersonal continuity on physicians’ services online. To fill the above research gap, we follow patient’s online information-providing behavior to examine the following research questions.

Research question 1: How does information continuity (offline treatment experience, medical records, and detailed information provision) influence physicians’ web-based services?

Research question 2: How does interpersonal continuity influence physicians’ web-based services?

Theoretical Foundation and Hypothesis Development

Information Richness Theory

Information richness theory, also called as media richness theory, takes the communication channel as an objective feature to determine the ability of information transmission [12]. It describes the ability to change people’s understanding within a time interval and consists of 4 standard features: the ability to give immediate feedback, the ability to communicate multiple social cues, linguistic diversity, and personalization [13]. The amount of information affects the communication outcomes by reducing uncertainty [14]. The appropriate amount is determined by the purpose of the communication and the content. The rich information can provide practical help for communication, coordination, collaboration, and information sharing. With the development of media, the standards for evaluating information richness have changed and a large number of important research results have been gained. Users’ perceived information richness would affect their satisfaction [15] and continued willingness to use [16]. Moreover, interactivity is an important factor in assessing the perceived richness of information [15,17,18] and could determine the platform development [19]. High interactivity would increase the willingness of users to use media or services [20,21]. High richness could decrease consumers’ uncertainty in online retail and increase their loyalty [22]. In the health field, the essential difference between web-based medical services and traditional medical services (ie, face-to-face) is information richness. However, with the development of web-based services, studies find that web-based psychological interventions are as effective as face-to-face psychotherapy [23]. For sensitive diseases, patients prefer a high information richness channel such as face-to-face therapy [24]. High information richness improves users’ perception of knowledge quality, source credibility, and knowledge consensus, especially under high health threats [25].

Continuity of Care

Service continuity was first proposed in the Folsom Report, Millis Report, and Willard Report in 1966, and then its concept has been developed and enriched. Subsequently, scholars have elaborated on various dimensions of continuity of care [2,26]. Continuity of care has also been defined in related studies as repeated contact between patients and physicians [27]. For the service provider, continuity of care can be divided into information continuity, multi-department continuity, time continuity, interpersonal continuity, and management continuity [2]. For the service receiver, experience continuity and geography continuity are important dimensions of continuity of care. The most widely used dimensions are information continuity, time continuity, and interpersonal continuity [28,29].

Information continuity means that different medical institutions have complete, timely, shareable, mutually recognized, and cohesive information in the aspects of disease prevention, examination, diagnosis, treatment, and rehabilitation of patients [2]. The health care provider uses information on past events to deliver care that is appropriate to the patient’s current circumstance [26]. Interpersonal continuity means providers develop an ongoing relationship with patients and the provider has knowledge of the patient as a person [26]. Interpersonal continuity is built on repeated (but not necessarily exclusive) contacts and is important for building trust and respect. The central skill fostered by interpersonal continuity over time is the ability to make and value a multidimensional diagnosis based on the biopsychosocial model within the patient’s context [2,30]. As many patients nowadays have more than one preferred health care provider, when transitions in care occur, communication and collaboration between health care providers (ie, information continuity) are more important than interpersonal continuity [31]. Continuity of care is associated with patient satisfaction, adherence to medical advice, and the use of hospital services [1]. Medical care is a special service for maintaining health; the continuity of life determines that medical care must be continuous. In the context of population aging, disease spectrum change, rapidly rising medical costs, and patients' increasing emphasis on self-worth, continuous medical care has become the focus of the establishment and improvement of health service systems in various countries.

Information Provision and Medical Service

The specialty of medical service leads to high information asymmetry between physicians and patients. It is difficult for both physicians and patients to fully explain the health condition within a limited time. Medical service is directly related to the health or safety of patients; thus, they often visit several physicians for rich information. Rich information helps improve physician-patient interaction and patient experience, thereby enhancing the information service capability and user satisfaction [32]. Quantitative information on the quality of health services can be more useful to patients by combining digital information with graphics [33]. Physicians' information has an important impact on the patient's decision [34].

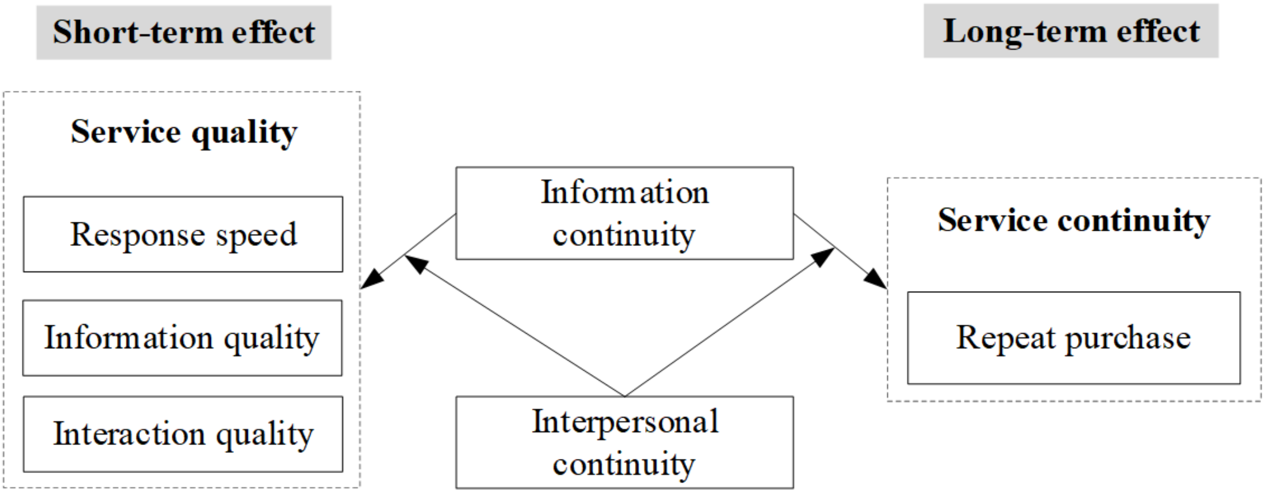
Since 1998, the government and private sectors have recognized the importance of using technology for improving care delivery and have made progress in setting the stage for transforming health care delivery through vastly improved use of health information technology [35]. There have been many government eHealth initiatives aiming to improve continuity and coordination through information, such as Personally Controlled Electronic Health Record [36], electronic health records [37], and telemedicine [38]. Although the use of online health communities is thought to help improve the continuity of care [3], only few empirical studies have been conducted to explore these influence mechanisms.

Online health communities serve as a bridge to help patients and physicians solve the problem of information asymmetry and improve the physician-patient relationship [39]. There are mainly 2 types of patients in online health communities. One type is those who have not seen a physician in hospitals and hope to receive advice on care through the web-based platform. The other category is the patients who have already seen a physician in hospitals and hope to receive more advice for disease treatment, rehabilitation, prognosis, and interpretation of the test report after receiving diagnosis and treatment offline. For the second type, as patients have received medical service in the hospital, they have more information, which they can provide to physicians in online health communities to improve continuity of care. Higher continuity is associated with higher quality of health care [40]. Based on the dimensions of continuity of care, we propose the following hypotheses:

1. Hypothesis 1: High information continuity helps improve a physician's web-based service. Previous studies have indicated that trust could change in different periods dynamically. In the case of medical service, the roles of information continuity and interpersonal continuity may change as the physician contacts patients over time [41]. Therefore, we examined the effects of information continuity in the short term (for the current purchase) and in the long term (for the future purchases). In short term, response speed, information quality, and interaction quality have been widely used in prior studies [8,9]. Repeat purchase is often used to measure the long-term effects [42]. Therefore, we included them and developed the following hypotheses. Hypothesis 1a: (short-term) high information continuity would improve the response speed of a physician's reply. Hypothesis 1b: (short-term) high information continuity would improve the information quality of a physician's reply. Hypothesis 1c: (short-term) high information continuity would improve the interaction quality of a physician's reply. Hypothesis 1d: (long-term) high information continuity would increase a patient's repeat purchase. Patients with a close continuous relationship with a specific physician are more likely to receive the recommended care [43]. Service content and service quality of health care can vary substantially across channels. Therefore, patients engaging in multiple visits with the same physician could help obtain a continuous and satisfactory outcome [44]. Based on the above arguments, we hypothesize that if it is the same physician online and offline, the effects of information continuity on the physician's service would be enhanced.
2. Hypothesis 2: High interpersonal continuity would enhance the relationships between information continuity and a physician's service. Based on the richness of information, we recognize whether a patient has offline treatment experience and has told the online physician, and then, we recognize whether a patient has provided his offline medical records to the online physician, and we calculate the degree of information provision.

The conceptual model for the abovementioned hypotheses is shown in Figure 1.

Figure 1. Conceptual model.



Methods

Ethics Approval

This study was approved by the institutional review board of Hainan Women and Children's Medical Center (HNWCMC202262).

Research Context and Data Collection

We collect data from one of the most professional and popular online health communities in China: Haodf.com [45]. Haodf.com was founded in 2006 and is one of China's leading online health care platforms. Haodf.com provides services such as hospital/physician information query, written consultation, phone consultation, video consultation, outpatient appointment, postdiagnosis disease management, family physician, disease knowledge, and popularization, and is widely trusted by physicians and patients. Haodf.com has a large number of high-quality physicians. By July 2021, Haodf.com had collected the information of more than 790,000 physicians in nearly 10,000 regular hospitals across the country. Among them, more than 240,000 physicians had registered on the platform, and those from AAA hospitals accounted for 73% of these active physicians. The hospitals in China are divided into 10 levels, and AAA is the best level. As of July 2021, Haodf.com has served more than 72 million patients. This online health community provides a physician-patient interaction platform for various diseases. Both individual services (eg, written consultation, phone consultation, video consultation) and team services are provided. Based on the aims of this study, we chose written consultation service and focused on physician-patient interaction content on diabetes for the following 2 reasons. First, chronic diseases have a long treatment period, and the patient often needs repeated communication with physicians. On Haodf.com, there is a large diabetic population, which was beneficial for the conduct of this research. Second, different from phone and video consultations, all interaction contents between physicians and patients based on written consultation are recorded on Haodf.com and shown publicly. We can obtain all the information that a patient has provided to his physician. By developing a web crawler, we firstly collected physician data from physician lists on Haodf.com, and 360 physicians were included. Then, for each physician, 20 complete physician-patient interaction contents were collected, including symptom description, offline experience, purchase times, medical records, or other material provision (shown in Figures 2 and 3). Finally, data of 7200 patients with 360 physicians were included in the empirical study.

Figure 2. Examples of patients' questions.

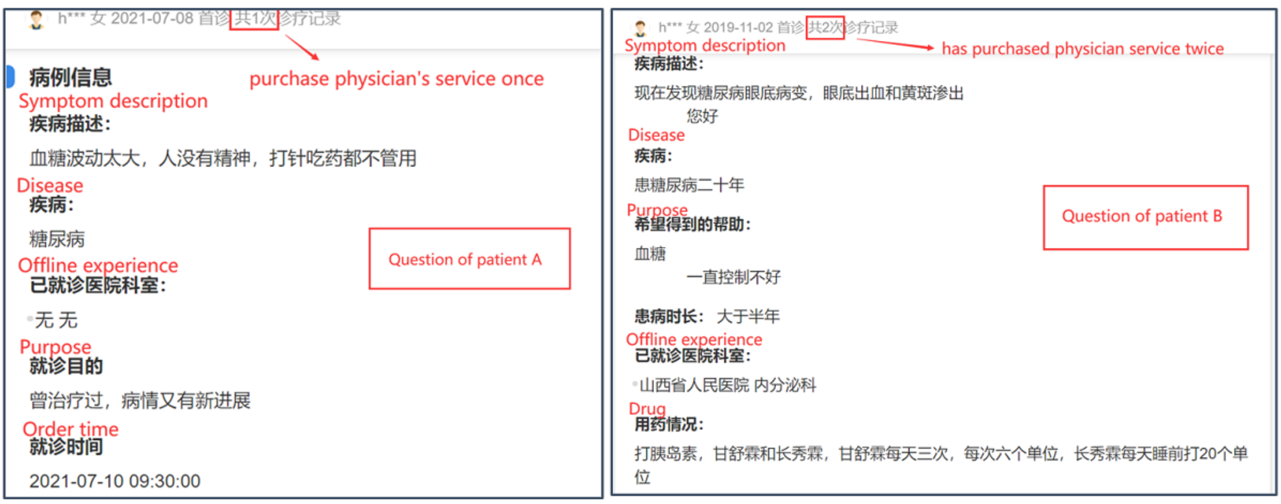


Figure 3. An example of interaction content.

Variables and Models

Dependent Variables

Four dependent variables were used to measure physicians' services: response speed (RS_{ij}), information quality ($InfQ_{ij}$), interaction quality ($IntQ_{ij}$), and repeat purchase (RP_{ij}). Response speed, information quality, and interaction quality, which were often used to measure the quality of physicians' online services in prior studies [34], were used to measure the short-term influence. The repeat purchase was used to measure the long-term influence.

Independent Variables

Information, including medical history, laboratory results, radiographs, and current diagnoses, as well as the history of medications and treatments, should be available to clinicians at the point of care whenever and wherever they need them, no matter where they were originally obtained [35]. Therefore, considering the information provision in online health communities, 3 independent variables were included to measure patient *i*'s offline information provision. Based on the degree of offline information provision, we measured whether patient *i* had offline experience (OE_{ij}) and mentioned it during the online consultation with physician *j*. If that was so, we measured the number of offline medical records or other material (OMR_{ij}) that patient *i* had provided to support the online service of physician *j*, and the number of words (ODI_{ij}) that patient *i* has described his offline experience to online physician *j*. These 3 variables describe the information continuity.

Moderating Variable: Interpersonal Continuity

Based on the interaction content, we recognized whether the physician in the patient's offline experience is the same as the physician who patient *i* had consulted online (SP_{ij}), and used a dummy variable in empirical models. This variable describes interpersonal continuity.

Control Variables

Other important information about physicians that may influence physician service was also included to control: physician medical title ($MTitle1_j$ and $MTitle2_j$), physician education title ($ETitle_j$), physician online reputation (POR_j), and hospital level

(*Level*$_j$). More details can be found in Table 1. Table 1 shows the definitions of the variables in the empirical analysis and their measurements. The unit of analysis is the individual online health community patient-physician interaction.

Accordingly, our empirical models are shown in , showing short-term effects and , showing the long-term effects, where i=1,…, N represents the patients, j=1,…, M represents the physicians, β_1 to β_7 are the focus parameters to be estimated. C represents control variables. ε is the error term associated with observation i and j.

Table 1. Description of the variables.

Variables	Description	Measures
Dependent variables		
Response speed (RS_{ij})	The response time that physician *j* could reply to a patient's question in 24 hours.	Use response time directly. The value is in days.
Information quality ($InfQ_{ij}$)	The level of detail in the physician *j*'s reply for patient *i*.	The number of words replied.
Interaction quality ($IntQ_{ij}$)	The frequency of physician-patient interaction.	The number of interactions between patient *i* and physician *j* is used.
Repeat purchase (RP_{ij})	Patient *i* may have purchased physician *j*'s service many times.	A dummy variable that describes whether patient *i* has repurchased physician *j*'s service.
Independent variables		
Offline experience (OE_{ij})	Patient *i* may have gone to a hospital for treatment before consulting online.	A dummy variable that describes whether patient *i* has provided his offline experience to the online physician. "1" refers to yes, and "0" refers to no.
Offline medical records (OMR_{ij})	Patient *i* may have gone to a hospital for treatment before consulting online, and undergo some tests.	The number of results of tests that patient *i* has provided to the online physician.
Offline detailed information (ODI_{ij})	Patient *i* may have gone to a hospital for treatment before consulting online.	The number of words that patient *i* has described his offline experience to the online physician.
Moderating variable		
Same physician (SP_{ij})	Whether the physician in patient's offline experience is same as the physician who patient *i* has consulted online.	A dummy variable that describes whether it is the same physician online and offline. "1" refers to yes, and "0" refers to no.
Control variables		
Physician medical titles ($MTitle1_j$ and $MTitle2_j$)	Physicians have medical titles, which are evaluated by the medical government based on their medical skills in China, including chief physician, associate chief physician, attending physician, and resident physician.	A dummy variable that describes whether physician *j* is a chief physician or associate chief physician. "1" refers to physician *j* as a chief physician or associate chief physician, and "0" refers to other medical titles.
Physician education title ($ETitle_j$)	Whether the physician *j* has worked at a university.	A dummy variable that describes whether physician *j* is a professor or associate professor at a university. "1" refers to physician *j* as a professor or associate professor, and "0" refers to other educational titles.
Physician online reputation (POR_j)	The reputation is based on physician *j*'s online work.	An indicator (ranges from 0 to 5) that is calculated by the website based on patients' feedbacks is used directly.
Hospital level ($Level_j$)	Hospitals have levels that are evaluated by the medical government based on their comprehensive health care quality in China.	A dummy variable indicating if the hospital where physician *j* works is AAA hospital. "1" refers to physician *j* works in an AAA-level hospital, and "0" refers to other level hospitals.

Results

Descriptive Statistics

Table S1 of Multimedia Appendix 1 shows the descriptive statistics and the correlations of the variables. On average, 46% (3312/7200) of the patients mentioned their offline experience. Each patient provided 6.11 offline medical records or other material and 38.5 words about the offline experience; 77% (5544/7200) of the patients chose the same physician online and offline. The response rate in 24 hours was 67.3% (242/360). The average numbers of information words and interactions were 12.61 and 17.97, respectively; 29% (2088/7200) of the patients purchased the physician service repeatedly. Multicollinearity is not an issue in our research as all variance inflation factors were less than 10.

Empirical Results: Short-term Effects

The ordinary least squares was used to obtain our short-term effect results, which are shown in Table 2, Table 3, and Table 4.

Table 2. Results for information continuity (offline experience provision): short-term effects.

Variables	Response speed				Information quality				Interaction quality			
	Model 1[a]		Model 2[b]		Model 1[c]		Model 2[d]		Model 1[e]		Model 2[f]	
	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value
Level	–.017 (.006)	.008	–.017 (.006)	.007	–.036 (.050)	.48	–.052 (.049)	.29	.022 (.003)	<.001	.022 (.003)	<.001
MTitle1[g]	–.021 (.008)	.006	–.019 (.008)	.01	.063 (.059)	.28	.152 (.058)	.009	.011 (.004)	.004	.010 (.004)	.009
MTitle2	–.002 (.007)	.80	–.002 (.007)	.83	.171 (.058)	.003	.188 (.057)	.001	.009 (.004)	.02	.009 (.004)	.02
ETitle[h]	–.017 (.007)	.01	–.017 (.007)	.01	–.120 (.052)	.02	–.117 (.051)	.02	.002 (.003)	.53	.002 (.003)	.53
POR[i]	.092 (.011)	<.001	.092 (.011)	<.001	1.930 (.082)	<.001	1.950 (.081)	<.001	–.105 (.005)	<.001	–.105 (.005)	<.001
OE[j]	N/A[k]	N/A	–.009 (.005)	.053	N/A	N/A	–.518 (.036)	<.001	N/A	N/A	.006 (.002)	.02

[a]Adjusted R^2=0.010; $F_{5,7720}$=16.808; P<.001.

[b]Adjusted R^2=0.010; $F_{1,7719}$=3.730; P=.053.

[c]Adjusted R^2=0.090; $F_{5,7720}$=153.226; P<.001.

[d]Adjusted R^2=0.113; $F_{1,7719}$=202.729; P<.001.

[e]Adjusted R^2=0.052; $F_{5,7720}$=85.391; P<.001.

[f]Adjusted R^2=0.052; $F_{1,7719}$=5.915; P=.02.

[g]MTitle1: physician medical title.

[h]ETitle: physician education title.

[i]POR: physician online reputation.

[j]OE: offline experience.

[k]N/A: not applicable.

Table 3. Results for information continuity (offline medical record provision): short-term effects.

Variables	Response speed				Information quality				Interaction quality			
	Model 1[a]		Model 2[b]		Model 1[c]		Model 2[d]		Model 1[e]		Model 2[f]	
	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value
Level	–.017 (.006)	.008	–.013 (.006)	.04	–.036 (.050)	.48	.052 (.049)	.28	.064 (.015)	<.001	.097 (.014)	<.001
MTitle1[g]	–.021 (.008)	.006	–.017 (.008)	.03	.063 (.059)	.28	.151 (.057)	.008	–.135 (.017)	<.001	–.102 (.016)	<.001
MTitle2	–.002 (.007)	.80	–.001 (.007)	.86	.171 (.058)	.003	.186 (.056)	.001	–.016 (.017)	.35	–.010 (.016)	.52
ETitle[h]	–.017 (.007)	.01	–.014 (.007)	.04	–.120 (.052)	.02	–.054 (.050)	.29	.003 (.015)	.86	.028 (.014)	.05
POR[i]	.092 (.011)	<.001	.057 (.011)	<.001	1.930 (.082)	<.001	1.148 (.087)	<.001	–.208 (.024)	<.001	–.501 (.025)	<.001
OMR[j]	N/A[k]	N/A	–.001 (.000)	<.001	N/A	N/A	–.025 (.001)	<.001	N/A	N/A	–.009 (.000)	<.001

[a]Adjusted R^2=0.010; $F_{5,7720}$=16.808; *P*<.001.

[b]Adjusted R^2=0.017; $F_{1,7719}$=56.802; *P*<.001.

[c]Adjusted R^2=0.090; $F_{5,7720}$=153.226; *P*<.001.

[d]Adjusted R^2=0.146; $F_{1,7719}$=511.611; *P*<.001.

[e]Adjusted R^2=0.037; $F_{5,7720}$=59.550; *P*<.001.

[f]Adjusted R^2=0.135; $F_{1,7719}$=882.911; *P*<.001.

[g]MTitle1: physician medical title.

[h]ETitle: physician education title.

[i]POR: physician online reputation.

[j]OMR: offline medical record.

[k]N/A: not applicable.

Table 4. Results for information continuity (offline detailed information provision): short-term effects.

Variables	Response speed				Information quality				Interaction quality			
	Model 1[a]		Model 2[b]		Model 1[c]		Model 2[d]		Model 1[e]		Model 2[f]	
	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value
Level	–.017 (.006)	.008	–.015 (.006)	.02	–.036 (.050)	.48	.032 (.009)	<.001	.064 (.015)	<.001	.068 (.014)	<.001
MTitle1[g]	–.021 (.008)	.006	–.022 (.007)	.003	.063 (.059)	.28	.023 (.011)	.03	–.135 (.017)	<.001	–.137 (.017)	<.001
MTitle2	–.002 (.007)	.80	–.006 (.007)	.40	.171 (.058)	.003	.003 (.011)	.74	–.016 (.017)	.35	–.025 (.017)	.14
ETitle[h]	–.017 (.007)	.01	–.013 (.007)	.05	–.120 (.052)	.02	.036 (.009)	<.001	.003 (.015)	.86	.011 (.015)	.46
POR[i]	.092 (.011)	<.001	.039 (.011)	<.001	1.930 (.082)	<.001	–.149 (.016)	<.001	–.208 (.024)	<.001	–.319 (.024)	<.001
OMR[j]	N/A[k]	N/A	.020 (.001)	<.001	N/A	N/A	.797 (.002)	<.001	N/A	N/A	.042 (.003)	<.001

[a]Adjusted R^2=0.010; $F_{5,7720}$=16.808; P<.001.

[b]Adjusted R^2=0.047; $F_{1,7719}$=297.243; P<.001.

[c]Adjusted R^2=0.090; $F_{5,7720}$=153.226; P<.001.

[d]Adjusted R^2=0.370; $F_{1,7719}$=4896.067; P<.001.

[e]Adjusted R^2=0.037; $F_{5,7720}$=59.550; P<.001.

[f]Adjusted R^2=0.067; $F_{1,7719}$=256.155; P<.001.

[g]MTitle1: physician medical title.

[h]ETitle: physician education title.

[i]POR: physician online reputation.

[j]OMR: offline medical record.

[k]N/A: not applicable.

Results for Information Continuity

Table 2 results suggest that offline experience negatively affects physician response speed (β=–.009, *P*=.053) and information quality (β=–.518, *P*<.001). Offline experience positively influences interaction quality (β=.006, *P*=.02). For offline experience provision, hypotheses 1a and 1c are supported but hypothesis 1b is not supported. Table 3 results show that offline medical record provision negatively affects physician response speed (β=–.001, *P*<.001), information quality (β=–.025, *P*<.001), and interaction quality (β=–.009, *P*<.001). For offline medical records provision, hypothesis 1a is supported but hypotheses 1b and 1c are not supported. The results in Table 4 present that offline detailed information provision positively affects physician response speed (β=.020, *P*<.001), information quality (β=.797, *P*<.001), and interaction quality (β=.042, *P*<.001). For offline detailed information provision, hypothesis 1a is not supported but hypotheses 1b and 1c are supported. Thus, hypotheses 1a, 1b, and 1c are partly supported.

Results for Interpersonal Continuity

The influences of interpersonal continuity on physician service are shown in Table S2 of Multimedia Appendix 2. We find that interpersonal continuity negatively moderates the relationship between offline experience provision and response speed (β=–.034, *P*=.006) and the relationship between offline experience provision and information quality (β=–.555, *P*<.001). We also find that interpersonal continuity positively moderates the relationship between offline medical record provision and interaction quality (β=–.010, *P*<.001), the relationship between offline detailed information provision and information quality (β=.015, *P*<.001), and the relationship between offline detailed information provision and interaction quality (β=.016, *P*=.01). Thus, for interpersonal continuity, hypothesis 2 is partly supported.

Empirical Results: Long-term Effects

The Probit regression was used to obtain our long-term effect results, which are shown in Tables 5 and 6. The results for information continuity are shown in Table 5. Our results suggest that offline experience positively affects physician response speed (β=.006, *P*=.04), information quality (β=.001, *P*<.001), and interaction quality (β=.602, *P*<.001). For the long-term effects of information continuity, hypothesis 1d is supported. The results for interpersonal continuity are shown in Table 6. The results indicate that the interpersonal continuity only positively moderates the relationship between offline detailed information and repeat purchase (β=.143, *P*=.04). Thus, hypothesis 2 is partly supported.

Table 5. Results for information continuity: long-term effects.

Variables	Offline experience				Offline medical record				Offline detailed information			
	Model 1[a]		Model 2[b]		Model 1[c]		Model 2[d]		Model 1[e]		Model 2[f]	
	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value
Level	.014 (.003)	.001	.014 (.003)	.001	.014 (.004)	.001	.012 (.004)	.004	.351 (.157)	.03	.402 (.153)	<.001
MTitle1[g]	.001 (.004)	.78	<.001 (.004)	.95	.001 (.005)	.78	<.001 (.005)	.97	–1.701 (.185)	<.001	–1.732 (.179)	<.001
MTitle2	.003 (.004)	.49	.003 (.004)	.52	.003 (.005)	.49	.003 (.005)	.52	–.243 (.183)	.18	–.369 (.177)	.04
ETitle[h]	.007 (.003)	.14	.007 (.003)	.14	.007 (.004)	.14	.006 (.004)	.20	–.017 (.162)	.92	.101 (.158)	.52
POR[i]	–.068 (.005)	<.001	–.068 (.005)	<.001	–.068 (.007)	<.001	–.057 (.008)	<.001	–.418 (.257)	.10	–1.988 (.261)	<.001
OE[j]	N/A[k]	N/A	.006 (.002)	.04	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
OMR[l]	N/A	N/A	N/A	N/A	N/A	N/A	.001 (.000)	<.001	N/A	N/A	N/A	N/A
ODI[m]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	.602 (.028)	<.001

[a]Adjusted R^2=0.013; $F_{5,7720}$=21.064; *P*<.001.
[b]Adjusted R^2=0.013; $F_{1,7719}$=4.162; *P*=.04.
[c]Adjusted R^2=0.013; $F_{5,7720}$=21.064; *P*<.001.
[d]Adjusted R^2=0.014; $F_{1,7719}$=12.792; *P*<.001.
[e]Adjusted R^2=0.025; $F_{5,7720}$=40.078; *P*<.001.
[f]Adjusted R^2=0.079; $F_{1,7719}$=455.791; *P*<.001.
[g]MTitle1: physician medical title.
[h]ETitle: physician education title.
[i]POR: physician online reputation.
[j]OE: offline experience.
[k]N/A: not applicable.
[l]OMR: offline medical record.
[m]ODI: offline detailed information.

Table 6. Results for interpersonal continuity: long-term effects.

Variables	Offline experience[a]		Offline medical records[b]		Offline detailed information[c]	
	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value
Level	.012 (.005)	.02	.011 (.005)	.03	.467 (.156)	.003
MTitle1[d]	.000 (.005)	.97	.000 (.005)	.98	–1.738 (.179)	<.001
MTitle2	.003 (.005)	.54	.003 (.005)	.52	–.372 (.177)	.04
ETitle[e]	.007 (.004)	.14	.006 (.004)	.20	.108 (.158)	.49
POR[f]	–.069 (.007)	<.001	–.057 (.008)	<.001	–1.960 (.261)	<.001
OE[g]	–.001 (.008)	.86	N/A[h]	N/A	N/A	N/A
SP[i]×OE	.009 (.008)	.28	N/A	N/A	N/A	N/A
OMR[j]	N/A	N/A	.000 (.000)	.05	N/A	N/A
SP×OMR	N/A	N/A	6.457E-5 (.000)	.65	N/A	N/A
ODI[k]	N/A	N/A	N/A	N/A	.523 (.066)	<.001
SP×ODI	N/A	N/A	N/A	N/A	.143 (.070)	.04

[a]Adjusted R^2=0.013.
[b]Adjusted R^2=0.014.
[c]Adjusted R^2=0.079.
[d]MTitle1: physician medical title.
[e]ETitle: physician education title.
[f]POR: physician online reputation.
[g]OE: offline experience.
[h]N/A: not applicable.
[i]SP: same physician.
[j]OMR: offline medical record.
[k]ODI: offline detailed information.

Robustness Check

In the main analysis, we did not consider whether a physician has also provided team service. As we only focused on individual service, we only included those physicians who did not provide team service; 25 physicians who provided team service were deleted. We used the new data to obtain empirical results (given the limited space, the robustness check results are included), and consistent results were found. Our results appear to be robust.

Discussion

Overview

Based on the information richness theory and continuity of care, this study investigates both short-term and long-term effects of information continuity and interpersonal continuity on physician service online by collecting data of 7200 patients with 360 physicians covering complete interaction records from a professional online platform in China. Our findings have theoretical and practical support for web-based managers and service providers to improve medical service quality.

Results Analysis

By collecting a data set from Haodf.com, we confirm the effects of information continuity and interpersonal continuity on the changing physician service. The summary of the results is shown in Table S3 and Table S4 of Multimedia Appendix 2. Our empirical study generated several important results.

First, both short-term and long-term effects of information continuity and interpersonal continuity were found. Continuity of care is important for medical service [1]. There is little understanding of how to improve the continuity of care and the effects of continuity of care. We find that providing offline experience is useful for improving the continuity of care and is helpful for physicians for providing high-quality service.

Second, the effects of information continuity showed heterogeneity. Offline experience and medical record provision are helpful for a physician to improve the response speed. However, detailed information provision increases response time. Offline experience and medical records could help refresh a physician's memory of the patient and then reply quickly. However, detailed offline information is written and provided by patients; therefore, it may contain a patient's personalized feelings, experience, and other questions, which takes the physician time to understand and then give a detailed reply to the patient's need. The above reasons can be used to explain the effects of the 3 independent variables on information quality. For the interaction quality, offline experience and detailed information provision help improve the interaction frequency between physicians and patients; however, offline medical records provision negatively affects interaction frequency. The reasons are that (1) web-based medical records are also a type of interaction and influence the calculation of interaction and (2) medical records contain much information about a disease condition, which a physician often needs to judge the disease. Without these medical records, the physician has to interact with patients to obtain relevant information.

Third, the effects of offline experience, medical records, and detailed information provisions on repeat purchases are consistent. Information continuity is helpful for a physician's service in the future.

Fourth, the moderating effects of interpersonal continuity were also consistent. Most of the moderating effects were positive and consistent with our hypotheses, that is, high interpersonal continuity would enhance the relationships between information continuity and physician service.

Implications

Our study produces several insights, which have implications for continuity of care, cross-channel behavior, and online health community literature. More importantly, these insights as a whole contribute to the design of integrated medical services. For the practical implications, first, for those who design and manage online health communities, attention needs to be paid not only to facilitating the transaction but also to interaction quality. From the continuity of care perspective, we have found significant influences of offline experience provision on physician online service. Our results suggest that mechanisms that can guide patients to provide offline experience should be established. In particular, the offline detailed information provision should be emphasized. Moreover, based on the positive effects of interpersonal continuity, online health community platforms should encourage patients to choose online physicians according to their offline physicians to improve consistency and then improve interpersonal continuity. Second, for the physicians, not only the short-term effects of offline experience provision should be valued but also long-term effects have to be highly regarded. Physicians can guide patients to remember the offline experience and provide their offline information, which is helpful for the physician to provide high-quality service and increase the repurchase rate further. Third, for the patient, our results suggest that patients could go to the nearby hospital to obtain medical records or other material and then provide them to the online physician to receive a better service.

Our study contributes to the current knowledge in several ways. First, our work extends our knowledge of the effects of information technology artifacts on the health care field from the continuity of care perspective. Although relevant departments believe that the use of information technology could realize the mutual recognition of inspection results, sharing of medical records, and thus improving the continuity of care [3], there are no empirical studies to examine the true effects. Our study has investigated the role of online health community use in improving the continuity of care. Moreover, we investigated the specific measures the patients and physicians should take to improve the continuity of care.

Second, our study enriches the literature on the continuity of care. Information continuity, interpersonal continuity, and time continuity have been widely discussed in previous studies [28,29]. However, they failed to examine the effects of different continuity dimensions on physician service, especially in a web-based environment. Our results show that the different dimensions of continuity of care have different effects on physician service behavior. Moreover, there are interaction effects between information continuity and interpersonal continuity.

Third, our study provides evidence on the cross-channel context. Although many studies have examined the channel effects in health care [6,46], they mainly focus on behaviors switching from online to offline. This study focuses on the effects of offline experience on online behavior, that is, behaviors switching from offline to online. Our results show that a patient's offline experience provision has a positive influence on the physician's web-based service.

Limitations of This Study

Several limitations and prospects in this study must be considered. First, we studied only 1 context, which helps us improve the internal validity, but it may also reduce the generalizability of our findings. Future studies could validate our results in other contexts. Second, word count and interaction count are used for measuring physician service. Future studies could use more accurate methods to measure physician services, such as text mining and sentiment analysis. Third, the unit of analysis is the individual online health community patient-physician interaction, and we do not have individual characteristics about patients. Future studies could try to obtain patient information and control them. Fourth, characteristics of physicians that may influence the use of web-based services are age, experience with computers/technology, and preferences toward in-person versus web-based delivery of services. Future studies could try to obtain more physician information and control them. Fifth, we assume that in-person experience and skills of physicians are transferrable to the online context. Future studies could obtain this skill of different physicians and control it.

Conclusions

Although abundant studies have investigated online health community behaviors and cross-channel behaviors, this study is among the first to investigate the effects of information providing from the continuity of care perspective and the influence of offline experience on online behaviors. Our study offers a better understanding of online behaviors, enriches the knowledge of the effects of information technology artifacts in the health care field, and contributes to the continuity of care literature. We have reported both short-term and long-term effects of the offline medical service experience on the online medical service experience. We believe that this paper could provoke some new thoughts on online health communities.

Acknowledgments

This study was funded by the Hainan Natural Science Foundation of China (821RC578 and 821RC757) and the Research and Cultivation Fund of Hainan Medical University (HYPY2020025) and Hainan Province Clinical Medical Center (QWYH202175).

Authors' Contributions

All the authors contributed to this paper. YX, CG, and WL conceived and designed the study, developed the research model, conducted data collection and analysis, and drafted as well as modified the manuscript. All authors approved the final version of the manuscript for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Descriptive statistics and correlations of variables.
[DOCX File , 16 KB - medinform_v10i7e35830_app1.docx]

Multimedia Appendix 2
Influences of interpersonal continuity on physician service and summary data of this study.
[DOCX File , 21 KB - medinform_v10i7e35830_app2.docx]

References

1. Walraven C, Oake N, Jennings A, Forster A. The association between continuity of care and outcomes: a systematic and critical review. Journal of Evaluation in Clinical Practice 2010;16(5):956. [doi: 10.1111/j.1365-2753.2009.01235.x]
2. Freeman GK, Olesen F, Hjortdahl P. Continuity of care: an essential element of modern general practice? Fam Pract 2003 Dec;20(6):623-627. [doi: 10.1093/fampra/cmg601] [Medline: 14701883]
3. Promote the internet plus medical model to improve the continuity of care. Sohu. 2018. URL: https://www.sohu.com/a/www.sohu.com/a/250661838_100063517 [accessed 2022-06-10]
4. Goh J, Gao G, Agarwal R. The Creation of Social Value: Can an Online Health Community Reduce Rural-Urban Health Disparities? MISQ 2016 Jan 1;40(1):247-263. [doi: 10.25300/misq/2016/40.1.11]
5. van der Eijk M, Faber MJ, Aarts JW, Kremer JA, Munneke M, Bloem BR. Using online health communities to deliver patient-centered care to people with chronic conditions. J Med Internet Res 2013 Jun 25;15(6):e115 [FREE Full text] [doi: 10.2196/jmir.2476] [Medline: 23803284]
6. Wu H, Lu N. Online written consultation, telephone consultation and offline appointment: An examination of the channel effect in online health communities. Int J Med Inform 2017 Nov;107:107-119. [doi: 10.1016/j.ijmedinf.2017.08.009] [Medline: 29029686]
7. Xu X, Zhou G, Zhang Y, Teng L, Zhang X. The influence of internet on traditional medical service pattern. Sohu. URL: https://www.sohu.com/a/www.sohu.com/a/417802176_139908 [accessed 2022-06-10]
8. Peng L, Wang Y, Chen J. Consequences of Gift Giving in Online Health Communities on Physician Service Quality: Empirical Text Mining Study. J Med Internet Res 2020 Jul 30;22(7):e18569 [FREE Full text] [doi: 10.2196/18569] [Medline: 32729834]
9. Wu H, Deng Z, Wang B, Wu T. Online service qualities in the multistage process and patients' compliments: A transaction cycle perspective. Information & Management 2020 Jul;57(5):103230. [doi: 10.1016/j.im.2019.103230]
10. Wang X, Zhao K, Street N. Analyzing and Predicting User Participations in Online Health Communities: A Social Support Perspective. J Med Internet Res 2017 Apr 24;19(4):e130 [FREE Full text] [doi: 10.2196/jmir.6834] [Medline: 28438725]
11. Lu X, Zhang R. Impact of Physician-Patient Communication in Online Health Communities on Patient Compliance: Cross-Sectional Questionnaire Study. J Med Internet Res 2019 May 13;21(5):e12891 [FREE Full text] [doi: 10.2196/12891] [Medline: 31094342]
12. Daft R, Lengel R. Information richness: A new approach to managerial behavior and organizational design. Research in Organizational Behavior. URL: https://psycnet.apa.org/record/1984-30194-001 [accessed 2022-07-12]
13. Daft RL, Lengel RH. Organizational Information Requirements, Media Richness and Structural Design. Management Science 1986 May;32(5):554-571. [doi: 10.1287/mnsc.32.5.554]
14. Perloff R. The Dynamics of Persuasion: Communication and Attitudes in the 21st Century. 2nd Edition. New York: Routledge; 2007.
15. Ogara SO, Koh CE, Prybutok VR. Investigating factors affecting social presence and user satisfaction with Mobile Instant Messaging. Computers in Human Behavior 2014 Jul;36:453-459. [doi: 10.1016/j.chb.2014.03.064]
16. Fernandez V, Simo P, Sallan JM, Enache M. Evolution of online discussion forum richness according to channel expansion theory: A longitudinal panel data analysis. Computers & Education 2013 Mar;62:32-40. [doi: 10.1016/j.compedu.2012.10.020]

17. Liu S, Liao H, Pratt JA. Impact of media richness and flow on e-learning technology acceptance. Computers & Education 2009 Apr;52(3):599-607. [doi: 10.1016/j.compedu.2008.11.002]
18. Otondo RF, Van Scotter JR, Allen DG, Palvia P. The complexity of richness: Media, message, and communication outcomes. Information & Management 2008 Jan;45(1):21-30. [doi: 10.1016/j.im.2007.09.003]
19. McMillan SJ, Hwang J. Measures of Perceived Interactivity: An Exploration of the Role of Direction of Communication, User Control, and Time in Shaping Perceptions of Interactivity. Journal of Advertising 2013 May 31;31(3):29-42. [doi: 10.1080/00913367.2002.10673674]
20. Gao Q, Rau P, Salvendy G. Perception of Interactivity: Affects of Four Key Variables in Mobile Advertising. International Journal of Human-Computer Interaction 2009 Aug 12;25(6):479-505. [doi: 10.1080/10447310902963936]
21. Gao Q, Rau PP, Salvendy G. Measuring perceived interactivity of mobile advertisements. Behaviour & Information Technology 2010 Jan;29(1):35-44. [doi: 10.1080/01449290802666770]
22. Li Y, Liu H, Lee M, Huang Q. Information privacy concern and deception in online retailing. INTR 2019 Jul 10;30(2):511-537. [doi: 10.1108/intr-02-2018-0066]
23. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlöf E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. Cogn Behav Ther 2018 Jan;47(1):1-18. [doi: 10.1080/16506073.2017.1401115] [Medline: 29215315]
24. Gieselmann A, Podleschka C, Rozental A, Pietrowsky R. Communication Formats and Their Impact on Patient Perception and Working Mechanisms: A Mixed-Methods Study of Chat-Based vs. Face-to-Face Psychotherapy for Insomnia. Behav Ther 2021 Mar;52(2):430-441. [doi: 10.1016/j.beth.2020.06.001] [Medline: 33622511]
25. Huo C, Zhang M, Ma F. Factors influencing people's health knowledge adoption in social media. LHT 2018 Jan 09;36(1):129-151. [doi: 10.1108/lht-04-2017-0074]
26. Haggerty JL, Reid RJ, Freeman GK, Starfield BH, Adair CE, McKendry R. Continuity of care: a multidisciplinary review. BMJ 2003 Nov 22;327(7425):1219-1221 [FREE Full text] [doi: 10.1136/bmj.327.7425.1219] [Medline: 14630762]
27. Gray DP, Evans P, Sweeney K, Lings P, Seamark D, Seamark C, et al. Towards a theory of continuity of care. J R Soc Med 2003 Apr;96(4):160-166 [FREE Full text] [doi: 10.1258/jrsm.96.4.160] [Medline: 12668701]
28. Curtis P, Rogers J. Continuity of care in a family practice residency program. J Fam Pract 1979 May;8(5):975-980. [Medline: 438757]
29. Shortell S. Continuity of medical care: conceptualization and measurement. Med Care 1976 May;14(5):377-391. [doi: 10.1097/00005650-197605000-00001] [Medline: 1271879]
30. Engel GL. The need for a new medical model: a challenge for biomedicine. Science 1977 Apr 08;196(4286):129-136. [doi: 10.1126/science.847460] [Medline: 847460]
31. Uijen AA, Schers HJ, van Weel C. Continuity of care preferably measured from the patients' perspective. J Clin Epidemiol 2010 Sep;63(9):998-999. [doi: 10.1016/j.jclinepi.2010.03.015] [Medline: 20656193]
32. Wu H, Deng Z, Wang B, Wang H. How online health community participation affects physicians' performance in hospitals: Empirical evidence from China. Information & Management 2021 Sep;58(6):103443. [doi: 10.1016/j.im.2021.103443]
33. Kenny P, Goodall S, Street DJ, Greene J. Choosing a Doctor: Does Presentation Format Affect the Way Consumers Use Health Care Performance Information? Patient 2017 Dec;10(6):739-751. [doi: 10.1007/s40271-017-0245-9] [Medline: 28477080]
34. Lu N, Wu H. Exploring the impact of word-of-mouth about Physicians' service quality on patient choice based on online health communities. BMC Med Inform Decis Mak 2016 Nov 26;16(1):151 [FREE Full text] [doi: 10.1186/s12911-016-0386-0] [Medline: 27888834]
35. Thompson T, Brailer D, Honorable T. The decade of health information technology: delivering consumer-centric and information-rich health care. US Department of Health and Human Services. 2004. URL: http://www.providersedge.com/ehdocs/ehr_articles/the_decade_of_hit-delivering_customer-centric_and_info-rich_hc.pdf [accessed 2022-07-12]
36. Pearce C, Bainbridge M. A personally controlled electronic health record for Australia. J Am Med Inform Assoc 2014;21(4):707-713 [FREE Full text] [doi: 10.1136/amiajnl-2013-002068] [Medline: 24650635]
37. Burton LC, Anderson GF, Kues IW. Using Electronic Health Records to Help Coordinate Care. The Milbank Quarterly 2004 Aug 25;82(3):457-481. [doi: 10.1111/j.0887-378x.2004.00318.x]
38. Mora F, Cone S, Rodas E, Merrell RC. Telemedicine and electronic health information for clinical continuity in a mobile surgery program. World J Surg 2006 Jun;30(6):1128-1134. [doi: 10.1007/s00268-005-0204-9] [Medline: 16736347]
39. Murray E, Lo B, Pollack L, Donelan K, Catania J, Lee K, et al. The impact of health information on the Internet on health care and the physician-patient relationship: national U.S. survey among 1.050 U.S. physicians. J Med Internet Res 2003;5(3):e17 [FREE Full text] [doi: 10.2196/jmir.5.3.e17] [Medline: 14517108]
40. Hussey PS, Schneider EC, Rudin RS, Fox DS, Lai J, Pollack CE. Continuity and the costs of care for chronic disease. JAMA Intern Med 2014 May;174(5):742-748 [FREE Full text] [doi: 10.1001/jamainternmed.2014.245] [Medline: 24638880]
41. Lewicki RJ, Tomlinson EC, Gillespie N. Models of Interpersonal Trust Development: Theoretical Approaches, Empirical Evidence, and Future Directions. Journal of Management 2016 Jul 01;32(6):991-1022. [doi: 10.1177/0149206306294405]
42. Li X, Hitt LM, Zhang ZJ. Product Reviews and Competition in Markets for Repeat Purchase Products. Journal of Management Information Systems 2014 Dec 08;27(4):9-42. [doi: 10.2753/mis0742-1222270401]

43. Peikes D, Chen A, Schore J, Brown R. Effects of Care Coordination on Hospitalization, Quality of Care, and Health Care Expenditures Among Medicare Beneficiaries. JAMA 2009 Feb 11;301(6):603. [doi: 10.1001/jama.2009.126] [Medline: 19211468]
44. Akçura MT, Ozdemir ZD. A Strategic Analysis of Multi-Channel Expert Services. Journal of Management Information Systems 2017 Apr 20;34(1):206-231. [doi: 10.1080/07421222.2017.1297637]
45. Online Health Community in China. URL: https://www.haodf.com/ [accessed 2022-07-12]
46. Li Y, Song Y, Zhao W, Guo X, Ju X, Vogel D. Exploring the Role of Online Health Community Information in Patients' Decisions to Switch from Online to Offline Medical Services. Int J Med Inform 2019 Oct;130:103951. [doi: 10.1016/j.ijmedinf.2019.08.011] [Medline: 31473534]

Edited by C Lovis; submitted 19.12.21; peer-reviewed by Y Chu, H Oh; comments to author 21.04.22; revised version received 04.05.22; accepted 03.07.22; published 21.07.22.

Please cite as:
Xuan Y, Guo C, Lu W
The Effects of Information Continuity and Interpersonal Continuity on Physician Services Online: Cross-sectional Study
JMIR Med Inform 2022;10(7):e35830
URL: https://medinform.jmir.org/2022/7/e35830
doi:10.2196/35830
PMID:35862183

Original Paper

Electronic Patient Portal Access, Retention in Care, and Viral Suppression Among People Living With HIV in Southeastern United States: Observational Study

Cassandra Oliver Schember[1], MPH, PhD; Sarah E Scott[2], MD; Cathy A Jenkins[3], MS; Peter F Rebeiro[1,3,4], MHS, PhD; Megan Turner[4], MA; Sally S Furukawa[4], BSc; Carmen Bofill[4], MPH; Zhou Yan[5], BSc; Gretchen P Jackson[6], MD, PhD; April C Pettit[1,4], MPH, MD

[1]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States
[2]Departments of Medicine and Pediatrics, Vanderbilt University Medical Center, Nashville, TN, United States
[3]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States
[4]Division of Infectious Diseases, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States
[5]Department of Health Information Technology Web Development, Vanderbilt University Medical Center, Nashville, TN, United States
[6]Departments of Surgery, Pediatrics and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Cassandra Oliver Schember, MPH, PhD
Division of Epidemiology
Department of Medicine
Vanderbilt University Medical Center
2525 West End Avenue, Suite 600
Nashville, TN, 37203
United States
Phone: 1 678 548 8656
Email: cassandra.schember@vanderbilt.edu

Abstract

Background: Approximately 1.1 million people living with HIV live in the United States, and the incidence is highest in Southeastern United States. Electronic patient portal prevalence is increasing and can improve engagement in primary medical care. Retention in care and viral suppression—measures of engagement in HIV care—are associated with decreased HIV transmission, morbidity, and mortality.

Objective: We aimed to determine if patient portal access among people living with HIV was associated with retention and viral suppression.

Methods: We conducted an observational cohort study among people living with HIV in care at the Vanderbilt Comprehensive Care Clinic (Nashville, Tennessee) from 2011-2016. Individual access was defined as patient portal account registration at any point in the year prior. Retention was defined as ≥2 kept appointments or HIV lab measurements ≥3 months apart within a 12-month period. Viral suppression was defined as the last viral load in the calendar year <200 copies/mL. We calculated adjusted prevalence ratios (aPRs) and 95% CIs using modified Poisson regression with generalized estimating equations to estimate the association of portal access with retention and viral suppression.

Results: We included 4237 people living with HIV contributing 16,951 person-years of follow-up (median 5, IQR 3-5 person-years). The median age was 43 (IQR 33-50) years. Of the 4237 people living with HIV, 78.1% (n=4237) were male, 40.8% (n=1727) were Black non-Hispanic, and 56.5% (n=2395) had access. Access was independently associated with retention (aPR 1.13, 95% CI 1.10-1.17) and viral suppression (aPR 1.18, 95% CI 1.14-1.22).

Conclusions: In this population, patient portal access was associated with retention and viral suppression. Future prospective studies should assess the impact of increasing portal access among people living with HIV on these HIV outcomes.

(JMIR Med Inform 2022;10(7):e34712) doi:10.2196/34712

KEYWORDS

HIV; viral suppression; retention in care; patient engagement; patient portal; observational study; United States; North America; eHealth; human immunodeficiency virus

Introduction

An estimated 1.1 million people living with HIV live in the United States, and the incidence is highest in Southeastern United States [1]. The US Department of Health and Human Services announced the Ending the HIV Epidemic plan in 2019 with the goals to reduce new HIV infections by 75% by 2025 and 90% by 2030 [2,3]. To achieve these goals, the Ending the HIV Epidemic plan identified the use of rapid and effective antiretroviral therapy to achieve viral suppression as a critical component [2,3]. The HIV Care Continuum outlines the sequential steps involved in sustained viral suppression, which include (1) HIV testing and diagnosis, (2) linkage to care, (3) retention in care, (4) receipt of antiretroviral therapy, and (5) viral suppression [4]. Despite advances in HIV treatment including lower pill burden and improved tolerability, US retention and viral suppression rates remain low at approximately 50% and 56%, respectively, suggesting that barriers to HIV treatment remain [5].

Electronic patient portals are web-based tools that allow patients and their families to interact with a health care system [6,7]. Portals promote patient-centered care, where all health care decisions and quality measurements are based on an individual's specific health needs and desired health outcomes. Electronic patient portal implementation and adoption has been rapidly increasing over the last decade [8,9]. These portals also assist health care facilities and providers in meeting the obligations of Meaningful Use within the Affordable Care Act, which requires that patients have web-based access to their health information [10]. Functionality varies across applications, but most portals allow patients to schedule appointments, access portions of their electronic health record, communicate with health care providers through secure messaging, and receive personalized health information [6,7,11].

Studies have demonstrated that electronic patient portals have increased patient engagement in care for various patient care populations and age groups [6,7,11,12]. Some studies have also assessed sociodemographic characteristics associated with patient portal use [8,13,14], such as one study that found that Black veterans living with HIV were less likely to register for and use a patient portal [15]. Few studies have assessed the impact of patient portals on HIV Care Continuum outcomes. Importantly, qualitative studies have demonstrated the acceptability of using patient portals to improve HIV care outcomes [16-18], and a study among US veterans found an association between electronic prescription refill through a patient portal and change from a detectable viral load to an undetectable viral load [19]. The objective of this study was to determine if patient portal access was independently associated with retention and viral suppression among people living with HIV engaged in care from 2011-2016 at the Vanderbilt Comprehensive Care Clinic (Nashville, Tennessee), a large HIV primary medical home in the Southeastern United States—a region disproportionately impacted by the HIV epidemic.

Methods

Study Population

We conducted a retrospective, observational cohort study among people living with HIV aged 18 years who had at least one HIV health care provider visit at the Vanderbilt Comprehensive Care Clinic from January 1, 2011, to December 31, 2015. The beginning of the study period was the first full year that clinic patients had access to the Vanderbilt electronic patient portal. Follow-up began on the date of the first HIV clinic visit during the study period and continued until the year prior to death or the end of the study period on December 31, 2016, allowing ≥1 year of follow-up for all people living with HIV included. We did not include data after 2016 due to a change in the Vanderbilt electronic patient portal application in 2017.

Data Sources and Study Definitions

The Vanderbilt University Medical Center deployed a robust electronic patient portal, *My Health at Vanderbilt*, in 2005. Within 10 years of this deployment, the portal had over 290,000 registered users and was accessed over 255,000 times per month [20]. *My Health at Vanderbilt* has similar features as other electronic patient portals, including secure messaging, appointment scheduling, bill management, and access to select laboratory results and electronic health record data, and all of these features were consistently available throughout the study period [21,22]. Further description of the *My Health at Vanderbilt* patient portal can be found in descriptions of the policies and procedures [21,22].

Clinical patient data were abstracted from the electronic health record which included information collected during routine clinical care. Our exposure of interest was electronic patient portal access, defined as whether a patient was registered for a *My Health at Vanderbilt* account at any point in the year prior. To register for *My Health at Vanderbilt*, patients are required to provide their name, social security number, birth date, and a valid email address [21]. This variable was lagged by 1 year, meaning that we assessed patient portal access in the year before our outcome. This ensured that the outcomes of interest were associated with *My Health at Vanderbilt* access in the year prior in an attempt to better establish temporality between patient portal access and HIV care outcomes.

The outcomes of interest were retention and viral suppression. Retention was defined as having ≥2 maintained in-person HIV clinic appointments, HIV-1 RNA viral load measurements, or CD4+ counts which occurred ≥3 months apart within a 12-month period based on the Health and Resources Services Administration HIV/AIDS Bureau definition of retention in care [4,23,24]. Viral suppression was defined as having ≥1 HIV-1 RNA viral load measurement within a given year with the last viral load measured in the year being <200 copies/mL [4,23]. Both outcomes were measured over each 12-month period after the first clinic visit during the study period.

Depending on the length of follow-up after the first visit, multiple outcomes per patient were possible. If an HIV-1 RNA viral load was missing during any 12-month interval, the patient was assumed to have a viral load of ≥200 copies/mL.

Covariates chosen based on a thorough review of the literature as well as in consultation with clinicians and epidemiologists who work directly with people living with HIV included birth sex, race/ethnicity, year of cohort entry, reported HIV transmission risk factor, insurance type, age, CD4+ cell count, and HIV-1 RNA viral load at the first clinic appointment attended. These covariates were chosen based on their connection to patient portal access and HIV care outcomes. Race/ethnicity was self-reported and categorized as White non-Hispanic, Black non-Hispanic, Hispanic, and other. Year of cohort entry was modeled continuously and defined as the year the patient entered the study. We categorized reported HIV transmission risk factors as male-male sexual contact (men who have sex with men; MSM), heterosexual contact, injection drug use (IDU), or other/unknown. If a patient had more than one type of transmission risk, IDU took precedence over MSM, which took precedence over heterosexual contact, in order of the risk of HIV transmission [25]. Insurance type was categorized as public (Medicare/Medicaid), private, or Ryan White. If an individual had more than one insurance type in a given year, Ryan White took precedence over public insurance, which took precedence over private insurance. Baseline CD4+ count was defined as the laboratory measurement closest to the first maintained appointment date (from 180 days prior to 30 days after); it was square-root transformed, modeled as a continuous covariate in the regression model, and displayed in our tables using the clinically salient CD4+ values of 100, 200, 350, and 500 cells/µL. Baseline HIV-1 RNA viral load was similarly defined as the laboratory measurement closest to the first maintained appointment date (from 180 days prior to 7 days after); it was $\log_{10}$-transformed and modeled continuously. Insurance status was time-updated during each 12-month period after enrollment. The remaining covariates were measured only at baseline.

Statistical Analysis

We reported demographic characteristics stratified by the existence of a patient portal account during follow-up, as we wanted to compare those who never accessed the patient portal to those who did. We reported categorical variables by frequency and proportion and used Pearson chi-squared test for comparisons. Continuous variables were reported as median and IQR, and Wilcoxon rank sum tests were used for comparisons [26,27]. Multiple imputation with 10 replications was used to account for missing CD4+ cell counts and HIV-1 RNA viral loads at baseline [28]. If missing, the reported HIV transmission risk factor was assumed to be other/unknown, and insurance type was handled by carrying forward the last observation. No patient was missing insurance type at baseline.

We estimated adjusted prevalence ratios (aPRs) and 95% CIs for retention and viral suppression using a modified Poisson regression [29]. Generalized estimating equations using an independence correlation structure accounted for multiple outcomes per individual [30,31]. A clustered sandwich estimator was used to estimate SEs [32-34]. In a sensitivity analysis, we excluded individuals with missing data to assess if a complete case analysis biased our results. All tests were 2-tailed and considered statistically significant if $P<.05$. All analyses were conducted using R statistical software (version 3.4; R Foundation for Statistical Computing).

Ethics Approval

Analyses were approved by the Vanderbilt University Institutional Review Board (approval number 170089) and conducted in accordance with the ethical standards set by the Declaration of Helsinki.

Results

Demographic Characteristics

The study population included 4237 people living with HIV followed for a total of 16,951 person-years. Of the 16,951 person-years, 74.8% (n=12,679) were categorized as retained in care and 71.4% (n=12,103) as virally suppressed. Median follow-up time per patient was 5 (IQR 3-5) person-years. The median age was 43 (IQR 33-50) years. Of the 4237 people living with HIV, 78.1% (n=3311) were male, 40.8% (n=1727) were Black non-Hispanic, and 41.2% (n=1747) reported MSM as an HIV transmission risk factor. The median baseline CD4+ count was 478 (IQR 288-692) cells/µL and median baseline HIV-1 RNA viral load was 100 (IQR 50-25,119) copies/mL (Table 1). Of the 4237 people living with HIV, reported HIV transmission risk factor, baseline CD4+ count, and HIV-1 RNA viral load were missing for 30.8% (n=1305), 34.2% (n=1449), and 44.8% (n=1898) of the participants, respectively. Insurance type varied over time; of the 16,951 person-years, 21% (n=3560) had private insurance, 40.1% (n=6797) had Ryan White, 27.7% (n=4695) had public insurance, and 11.2% (n=1899) were missing for which the last observation was carried forward.

Of the 4237 people living with HIV included, 56.5% (n=2395) had patient portal access at any point during follow-up. People living with HIV who had a *My Health at Vanderbilt* account were younger, with a median age of 42 (IQR 31-49) years, than those without an account, who had a median age of 44 (IQR 34-51) years. This difference was statistically significant, but a difference of 2 years is arguably not a clinically significant difference. A higher percentage (85.6%, 2050/2395) of those with an account were male, whereas only 68.5% (1261/1842) of those without an account were male. Fewer people living with HIV with patient portal access (30.2%, 724/2395) were Black non-Hispanic than people living with HIV without access (54.5%, 1003/1842). More people living with HIV with access (52.2%, 1250/2395) reported their HIV transmission risk factor as MSM than those without access (27%, 497/1842). Those with access also had a higher median baseline CD4+ count of 500 (IQR 309-702) cells/µL than those without access, who had a median baseline count of 444 (IQR 258-676) cells/µL. The baseline HIV-1 RNA viral load was similar between these 2 groups (Table 1).

Table 1. Baseline demographic characteristics of the study population stratified by patient portal account status.

Characteristic	No account during follow-up (n=1842)	Account existed during follow-up (n=2395)	All participants (N=4237)	*P* value[a]
Baseline age (years), median (IQR)	44 (34-51)	42 (31-49)	43 (33-50)	<.001
Sex, n (%)				<.001
Male	1261 (68.5)	2050 (85.6)	3311 (78.1)	
Female	581 (31.5)	345 (14.4)	926 (21.9)	
Race/ethnicity, n (%)				<.001
Black non-Hispanic	1003 (54.5)	724 (30.2)	1727 (40.8)	
Hispanic	138 (7.5)	102 (4.3)	240 (5.7)	
White non-Hispanic	407 (22.1)	1134 (47.3)	1541 (36.4)	
Other/unknown	294 (16)	435 (18.2)	729 (17.2)	
HIV risk factor, n (%)				<.001
MSM[b]	497 (27)	1250 (52.2)	1747 (41.2)	
Heterosexual	622 (33.8)	332 (13.9)	954 (22.5)	
IDU[c]	80 (4.3)	35 (1.5)	115 (2.7)	
Other/unknown	66 (3.6)	50 (2.1)	116 (2.7)	
Missing data	577 (31.3)	728 (30.4)	1305 (30.8)	
Baseline CD4+ count (cells/µL), median (IQR)	444 (258-676)	500 (309-702)	478 (288-692)	<.001
Baseline HIV-1 RNA viral load (copies/mL), median (IQR)	158.5 (50.1-19,952.6)	63.1 (50.1-25,118.9)	100.0 (50.1-25,118.9)	.30
Year of cohort entry, n (%)				.007
2011	1126 (61.1)	1452 (60.6)	2578 (60.8)	
2012	155 (8.4)	234 (9.8)	389 (9.2)	
2013	175 (9.5)	211 (8.8)	386 (9.1)	
2014	198 (10.7)	236 (9.9)	434 (10.2)	
2015	146 (7.9)	235 (9.8)	381 (9)	
2016	42 (2.3)	27 (1.1)	69 (1.6)	

[a]Wilcoxon rank sum test was used for continuous variables and Pearson chi-square test was used for categorical variables to compare those with an account to those without an account.

[b]MSM: men who have sex with men.

[c]IDU: injection drug use.

Retention in Care Outcome

In the multiple imputed, adjusted, and modified Poisson regression analysis, patient portal access was independently associated with better retention (aPR 1.13, 95% CI 1.10-1.17; Table 2). Other factors independently associated with better retention in this model included increased age at first visit (aPR 1.09, 95% CI 1.04-1.13) and MSM (aPR 1.13, 95% CI 1.03-1.23) and heterosexual contact (aPR 1.15, 95% CI 1.05-1.26) as reported HIV transmission risk factors compared to IDU (Table 2). A factor independently associated with worse retention was other/unknown race/ethnicity as compared to White non-Hispanic (aPR 0.93, 95% CI 0.90-0.97; Table 2).

Table 2. Adjusted prevalence ratios for the association of patient portal account existence and HIV outcomes of retention in care and viral suppression. All models adjusted for variables included in the table as well as the year of cohort entry.

Characteristic	Retention in care model, aPR[a] (95% CI)	Viral suppression model, aPR (95% CI)
Account status (variable lagged by 1 year)		
No account	REF[b]	REF
Account exists	1.13 (1.10-1.17)*	1.18 (1.14-1.22)*
Baseline age (per 10 years)	1.09 (1.04-1.13)*	1.09 (1.04-1.13)*
Sex		
Male	REF	REF
Female	1.04 (1.00-1.08)	0.99 (0.95-1.04)
Race/ethnicity		
Black non-Hispanic	0.99 (0.95-1.02)	0.95 (0.92-0.99)*
Hispanic	1.04 (0.98-1.11)	1.03 (0.96-1.10)
White non-Hispanic	REF	REF
Other/unknown	0.93 (0.90-0.97)*	0.94 (0.90-0.97)*
HIV risk factor		
MSM[c]	1.13 (1.03-1.23)*	1.11 (1.00-1.23)
Heterosexual	1.15 (1.05-1.26)*	1.15 (1.03-1.27)*
IDU[d]	REF	REF
Other/unknown	0.96 (0.88-1.05)	0.95 (0.86-1.06)
Insurance		
Private	REF	REF
Public	1.03 (0.99-1.07)	0.97 (0.94-1.01)
Ryan White	0.99 (0.95-1.02)	0.94 (0.90-0.98)*
Baseline CD4+ count (square-root transformed; cells/µL)		
100	0.99 (0.91-1.07)	0.99 (0.91-1.07)
200	0.99 (0.97-1.02)	0.99 (0.97-1.02)
350	REF	REF
500	1.01 (0.99-1.02)	1.00 (0.99-1.02)
Baseline HIV-1 RNA viral load ($\log_{10}$-transformed; copies/mL)	1.00 (0.98-1.01)	0.94 (0.92-0.96)*

[a]aPR: adjusted prevalence ratio.

[b]REF: reference.

[c]MSM: men who have sex with men.

[d]IDU: injection drug use.

**P*<.05.

Viral Suppression Outcome

In the multiple imputed, adjusted, and modified Poisson regression analysis, patient portal access was independently associated with improved viral suppression (aPR 1.18, 95% CI 1.14-1.22; Table 2). Other factors independently associated with better viral suppression included increased age at first visit (aPR 1.09. 95% CI 1.04-1.13) and heterosexual contact as a reported HIV transmission risk factor as compared to IDU (aPR 1.15, 95% CI 1.03-1.27). Factors independently associated with worse viral suppression included Black non-Hispanic (aPR 0.95, 95% CI 0.92-0.99) and other/unknown (aPR 0.94, 95% CI 0.90-0.97) race/ethnicity as compared to White non-Hispanic race/ethnicity; Ryan White coverage as compared to private insurance (aPR 0.94, 95% CI 0.90-0.98); and higher HIV-1 RNA viral load at first clinic visit (aPR 0.94, 95% CI 0.92-0.96; Table 2).

Sensitivity Analysis

We conducted a sensitivity analysis in which patients with missing data were excluded. This led to a complete case population of 1643 patients (38.8% of total cohort, N=4237) contributing 5589 person-years (33% of total person-years, N=16,951). The results were similar, but less precise, when the 2 full models from the primary analysis were used for retention and viral suppression (Table 3). Patient portal access remained associated with increased likelihood of retention (aPR 1.13, 95% CI 1.07-1.19) and viral suppression (aPR 1.16, 95% CI 1.10-1.23; Table 3).

Table 3. Adjusted prevalence ratios for the association of patient portal account existence and the HIV outcomes of retention in care and viral suppression—complete case analysis. All models adjusted for variables included in the table as well as the year of cohort entry.

Characteristic	Retention in care model, aPR[a] (95% CI)	Viral suppression model, aPR (95% CI)
Account status (variable lagged by 1 year)		
No account	REF[b]	REF
Account exists	1.13 (1.07-1.19)*	1.16 (1.10-1.23)*
Baseline age (per 10 years)	1.08 (1.06-1.10)*	1.08 (1.06-1.11)*
Sex		
Male	REF	REF
Female	1.05 (0.98-1.13)	1.01 (0.93-1.09)
Race/ethnicity		
Black non-Hispanic	0.96 (0.97-1.01)	0.92 (0.87-0.98)*
Hispanic	1.00 (0.91-1.10)	0.97 (0.87-1.08)
White non-Hispanic	REF	REF
Other/unknown	0.95 (0.88-1.01)	0.96 (0.90-1.03)
HIV risk factor		
MSM[c]	1.06 (0.93-1.21)	1.05 (0.92-1.20)
Heterosexual	1.12 (0.98-1.28)	1.12 (0.97-1.29)
IDU[d]	REF	REF
Other/unknown	1.01 (0.85-1.20)	1.01 (0.85-1.21)
Insurance		
Private	REF	REF
Public	0.97 (0.91-1.04)	0.94 (0.88-1.01)
Ryan White	1.00 (0.95-1.05)	0.96 (0.90-1.01)
Baseline CD4+ count (square-root transformed; cells/µL)		
100	1.00 (0.98-1.03)	0.99 (0.98-1.05)
200	1.00 (0.99-1.02)	0.99 (0.99-1.02)
350	REF	REF
500	1.00 (0.99-1.01)	1.01 (0.99-1.02)
Baseline HIV-1 RNA viral load (log_{10}-transformed; copies/mL)	1.00 (0.98-1.02)	0.96 (0.94-0.98)*

[a]aPR: adjusted prevalence ratio.
[b]REF: reference.
[c]MSM: men who have sex with men.
[d]IDU: injection drug use.
*P<.05.

Discussion

Principal Findings

Electronic patient portal access via Vanderbilt's *My Health at Vanderbilt* system was significantly associated with subsequent retention and viral suppression among people living with HIV in care at the Vanderbilt Comprehensive Care Clinic. This finding is consistent with previous findings from a Kaiser Permanente study that found patient portals increased patient membership retention for both people living with HIV and people not living with HIV [35]. There have been other studies of people living with HIV that found patient portals improve retention and viral suppression, but these were in less diverse or much smaller patient populations [16,19]. A small (n=22) prospective quality improvement project aimed to increase enrollment in a patient portal among women living with HIV to improve their retention in HIV care, given their increased risk of disengagement [16]. The authors found a significant association between enrollment in the patient portal and the number of scheduled visits but did not find a significant association with missed visits or viral suppression [16]. Another retrospective study among a population of 3374 veterans living with HIV found a significant association between messaging from a personal health record and viral suppression, but the authors did not assess retention or how patient portal access affected viral suppression [19]. The strengths of our study include having a large, demographically diverse cohort of people living with HIV living in Southeastern United States, a region of the country disproportionately affected by the HIV epidemic.

In our cohort, compared to patients without patient portal access, those with access were more likely to be younger, male, White non-Hispanic, and report MSM as their HIV transmission risk factor. They also had a higher CD4+ count at their first clinic visit compared to patients without patient portal access. Our results are consistent with previous studies in populations including people living with HIV and people without HIV, which showed that a higher proportion of those with access to patient portals tend to be younger and White, although the age difference in our study was only 2 years [8,11,13,14]. These differences could be due to increased technological literacy in using computers and smartphones [8,14]. Sex differences in patient portal access in other studies have varied, with some showing that women access patient portals more, and others showing that men preferred using patient portals than speaking in person with their health care providers [8,13]. In our cohort, men were more likely to have patient portal access. Our cohort had a higher proportion of men, but if there were no sex differences, we would expect the same proportion among those with and without access.

In addition to patient portal access, increasing age and reported HIV risk factor were independently associated with retention and viral suppression. People living with HIV in an older age group compared to those in a younger age group and people living with HIV who reported heterosexual activity or MSM compared to IDU as an HIV transmission risk factor were more likely to achieve retention and viral suppression. These findings are consistent with a systematic review of retention studies [36].

Factors that were independently associated with worse retention and viral suppression included race/ethnicity, insurance type, and HIV-1 RNA viral load at the first Vanderbilt Comprehensive Care Clinic visit. People living with HIV who are Black non-Hispanic (compared to those who are White non-Hispanic), have Ryan White coverage (compared to private insurance), or had a higher HIV-1 RNA viral load at their first clinic visit had worse retention and viral suppression. These findings are consistent with previous cohort studies assessing viral suppression trends over time, in which Black non-Hispanic race/ethnicity was associated with worse viral suppression and having Ryan White insurance was associated with worse HIV outcomes [37,38]. However, worse outcomes for people receiving care via the Ryan White HIV/AIDS Program is likely because it is a proxy for lower socioeconomic status. Our findings show that patient portal access follows similar trends to disparities in HIV care outcomes by age, race, HIV transmission risk factor, and insurance status, as groups with poor patient portal access also have poor HIV outcomes.

In the setting of the current COVID-19 pandemic, engaging care through electronic means such as patient portals and telehealth have increased [39,40]. This pandemic may have lasting effects on how individuals access and engage care, showing the importance of better understanding the effects of patient portal access on HIV care outcomes.

Our study is subject to several limitations. First, we had data on patient portal access but not on the frequency of or reasons for electronic patient portal use. It is possible to have patient portal access but never use the portal. However, regardless of use, patient portal access was associated with improved retention and viral suppression, demonstrating that providing access to patient portals is likely to improve HIV outcomes. Similarly, studies have stressed the importance of electronic health literacy in patient portal effectiveness and care outcomes. In our study, a patient may have had access to the electronic patient portal and used it but also had difficulty understanding the platform or information due to technological or health literacy barriers [41,42]. Both scenarios would have biased our results toward a null hypothesis; therefore, it is possible that the true relationship between patient portal access and retention and viral suppression may be stronger than what we described. Second, some people living with HIV in our cohort may have silently transferred to other clinics, which led them to be misclassified in our study as not retained in care. This may have led to an overestimation of those not retained in care, which could have biased our results in either direction depending on the population misclassified. Third, the reported HIV transmission risk factor, baseline CD4+ count, and baseline HIV-1 RNA viral load were missing for 31% to 45% of participants. The missing data for this risk factor and baseline measures of clinical variables were accounted for with multiple imputation. The results of the sensitivity analysis including only patients with complete records had similar results, suggesting that data were missing completely at random and therefore not a likely source of bias. Additionally, this was a single-site study and may not be generalizable to other settings, as electronic patient portal access may differ elsewhere. Lastly, these data are from 2011-2016. We were unable to provide more recent data because after 2016, Vanderbilt's patient portal changed. However, we were still able to establish a connection between an early patient portal and favorable HIV outcomes.

We examined the association of an under-studied exposure with HIV care outcomes and found that electronic patient poral access was independently associated with retention and viral suppression in our cohort of people living with HIV. Studies have demonstrated that electronic patient portals offer a unique opportunity to improve outcomes that are a part of the HIV Care Continuum, such as retention and viral suppression [17,18]. Our study supports prior findings and fills a gap in previous literature by examining this association in a large cohort of people living with HIV in an area disproportionately affected by HIV with a median longitudinal follow-up of 5 years.

Conclusions

Retention and viral suppression are necessary for reducing HIV transmission and mortality, as well as increasing the quality of life for people living with HIV. We found that electronic patient portal access was associated with improved retention and viral suppression. This suggests that increased access to electronic patient portals among people living with HIV may be an effective method to promote better HIV Care Continuum outcomes. Large prospective studies assessing the impact of patient portal access on retention and viral suppression are needed to confirm these findings.

Acknowledgments

For their contributions to this study, we thank the patients, physicians, and staff of the Vanderbilt Comprehensive Care Clinic and the Tennessee Center for AIDS Research Epidemiology and Outcomes Working Group.

This work was supported by the National Institutes of Health (R01 MH113438 to PI-Pettit) and the Tennessee Center for AIDS Research (P30 AI110527). This work was also supported by the Clinical Translational Science Award (UL1TR000445) and Clinical Translational Science Award (TL1TR002244) from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent official views of the National Center for Advancing Translational Sciences or the National Institutes of Health.

Conflicts of Interest

PFR received funding from Gilead and Johnson & Johnson (paid to the individual) and from the National Institutes of Health/National Institute of Allergy and Infectious Diseases (paid to institution).

References

1. HIV Surveillance Report, 2018 (Preliminary); vol. 30. Centers for Disease Control and Prevention. 2019 Nov. URL: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-report-2018-preliminary-vol-30.pdf [accessed 2022-06-22]
2. Ending the HIV epidemic: a plan for America. United States Department of Health and Human Services. 2019 Feb. URL: https://www.hhs.gov/sites/default/files/ending-the-hiv-epidemic-fact-sheet.pdf [accessed 2022-06-22]
3. Fauci AS, Redfield RR, Sigounas G, Weahkee MD, Giroir BP. Ending the HIV epidemic: a plan for the United States. JAMA 2019 Mar 05;321(9):844-845. [doi: 10.1001/jama.2019.1343] [Medline: 30730529]
4. Hogg RS. Understanding the HIV care continuum. Lancet HIV 2018 Jun;5(6):e269-e270 [FREE Full text] [doi: 10.1016/S2352-3018(18)30102-4] [Medline: 29893238]
5. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data: United States and 6 dependent areas, 2018. HIV Surveillance Supplemental Report 2020;25(No. 2). Centers for Disease Control and Prevention. 2020 May. URL: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-25-2.pdf [accessed 2022-06-22]
6. Ammenwerth E, Schnell-Inderst P, Hoerbst A. The impact of electronic patient portals on patient care: a systematic review of controlled trials. J Med Internet Res 2012 Nov 26;14(6):e162 [FREE Full text] [doi: 10.2196/jmir.2238] [Medline: 23183044]
7. Goldzweig CL, Orshansky G, Paige NM, Towfigh AA, Haggstrom DA, Miake-Lye I, et al. Electronic patient portals: evidence on health outcomes, satisfaction, efficiency, and attitudes: a systematic review. Ann Intern Med 2013 Nov 19;159(10):677-687. [doi: 10.7326/0003-4819-159-10-201311190-00006] [Medline: 24247673]
8. Turner K, Clary A, Hong Y, Alishahi Tabriz A, Shea CM. Patient portal barriers and group differences: cross-sectional national survey study. J Med Internet Res 2020 Sep 17;22(9):e18870 [FREE Full text] [doi: 10.2196/18870] [Medline: 32940620]
9. Turner K, Hong YR, Yadav S, Huo J, Mainous AG. Patient portal utilization: before and after stage 2 electronic health record meaningful use. J Am Med Inform Assoc 2019 Oct 01;26(10):960-967 [FREE Full text] [doi: 10.1093/jamia/ocz030] [Medline: 30947331]
10. Blumenthal D, Abrams M, Nuzum R. The Affordable Care Act at 5 Years. N Engl J Med 2015 Oct 15;373(16):1580. [doi: 10.1056/NEJMc1510015] [Medline: 26466007]
11. Gordon NP, Hornbrook MC. Differences in access to and preferences for using patient portals and other eHealth technologies based on race, ethnicity, and age: a database and survey study of seniors in a large health plan. J Med Internet Res 2016 Mar 04;18(3):e50 [FREE Full text] [doi: 10.2196/jmir.5105] [Medline: 26944212]
12. Lyles CR, Nelson EC, Frampton S, Dykes PC, Cemballi AG, Sarkar U. Using electronic health record portals to improve patient engagement: research priorities and best practices. Ann Intern Med 2020 Jun 02;172(11 Suppl):S123-S129 [FREE Full text] [doi: 10.7326/M19-0876] [Medline: 32479176]

13. Ancker JS, Barrón Y, Rockoff ML, Hauser D, Pichardo M, Szerencsy A, et al. Use of an electronic patient portal among disadvantaged populations. J Gen Intern Med 2011 Oct 7;26(10):1117-1123 [FREE Full text] [doi: 10.1007/s11606-011-1749-y] [Medline: 21647748]
14. Elston Lafata J, Miller CA, Shires DA, Dyer K, Ratliff SM, Schreiber M. Patients' adoption of and feature access within electronic patient portals. Am J Manag Care 2018 Nov 01;24(11):e352-e357 [FREE Full text] [Medline: 30452203]
15. Javier S, Troszak L, Shimada S, McInnes D, Ohl M, Avoundjian T, et al. Racial and ethnic disparities in use of a personal health record by veterans living with HIV. J Am Med Inform Assoc 2019 Aug 01;26(8-9):696-702 [FREE Full text] [doi: 10.1093/jamia/ocz024] [Medline: 30924875]
16. Plimpton E. A quality improvement project to increase patient portal enrollment and utilization in women living with HIV at risk for disengagement in care. J Assoc Nurses AIDS Care 2020;31(1):60-65. [doi: 10.1097/JNC.0000000000000153] [Medline: 31834101]
17. Turner K, Klaman SL, Shea CM. Personal health records for people living with HIV: a review. AIDS Care 2016 Sep;28(9):1181-1187. [doi: 10.1080/09540121.2016.1153594] [Medline: 26917113]
18. Jackman K, Latkin CA, Maksut JL, Trent ME, Sanchez TH, Baral SD. Patient portals as highly acceptable tools to support HIV preventative behaviors among adolescent and young sexual minority men. J Adolesc Health 2020 Aug;67(2):278-281 [FREE Full text] [doi: 10.1016/j.jadohealth.2020.03.029] [Medline: 32387095]
19. McInnes DK, Shimada SL, Midboe AM, Nazi KM, Zhao S, Wu J, et al. Patient use of electronic prescription refill and secure messaging and its association with undetectable HIV viral load: a retrospective cohort study. J Med Internet Res 2017 Feb 15;19(2):e34 [FREE Full text] [doi: 10.2196/jmir.6932] [Medline: 28202428]
20. Cronin RM, Davis SE, Shenson JA, Chen Q, Rosenbloom ST, Jackson GP. Growth of secure messaging through a patient portal as a form of outpatient interaction across clinical specialties. Appl Clin Inform 2015;6(2):288-304 [FREE Full text] [doi: 10.4338/ACI-2014-12-RA-0117] [Medline: 26171076]
21. Osborn CY, Rosenbloom ST, Stenner SP, Anders S, Muse S, Johnson KB, et al. MyHealthAtVanderbilt: policies and procedures governing patient portal functionality. J Am Med Inform Assoc 2011 Dec 01;18 Suppl 1(Supplement 1):i18-i23 [FREE Full text] [doi: 10.1136/amiajnl-2011-000184] [Medline: 21807648]
22. Steitz B, Wong J, Cobb J, Carlson B, Smith G, Rosenbloom S. Policies and procedures governing patient portal use at an Academic Medical Center. JAMIA Open 2019 Dec;2(4):479-488 [FREE Full text] [doi: 10.1093/jamiaopen/ooz039] [Medline: 32025645]
23. Committee on Review Data Systems for Monitoring HIV Care, Institute of Medicine. In: Ford MA, Spicer CM, editors. Monitoring HIV care in the United States: Indicators and Data Systems. Washington, DC: National Academies Press; 2012.
24. Mugavero MJ, Westfall AO, Zinski A, Davila J, Drainoni M, Gardner LI, Retention in Care (RIC) Study Group. Measuring retention in HIV care: the elusive gold standard. J Acquir Immune Defic Syndr 2012 Dec 15;61(5):574-580. [doi: 10.1097/QAI.0b013e318273762f] [Medline: 23011397]
25. Patel P, Borkowf CB, Brooks JT, Lasry A, Lansky A, Mermin J. Estimating per-act HIV transmission risk: a systematic review. AIDS 2014 Jun 19;28(10):1509-1519. [doi: 10.1097/QAD.0000000000000298] [Medline: 24809629]
26. Camilli G, Hopkins KD. Applicability of chi-square to 2 × 2 contingency tables with small expected cell frequencies. Psychol Bull 1978;85(1):163-167. [doi: 10.1037/0033-2909.85.1.163]
27. Wilcoxon F, Katti SK, Wilcox RA. Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. Vol 1. Pearl River, NY: American Cyanamid; 1963:171-259.
28. Rubin DB. Multiple Imputation after 18+ Years. J Am Stat Assoc 1996 Jun;91(434):473-489. [doi: 10.1080/01621459.1996.10476908]
29. Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol 2004 Apr 01;159(7):702-706. [doi: 10.1093/aje/kwh090] [Medline: 15033648]
30. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986 Mar;42(1):121. [doi: 10.2307/2531248]
31. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. Int Stat Rev 1991 Apr;59(1):25. [doi: 10.2307/1403572]
32. Zeileis A. Econometric computing with HC and HAC covariance matrix estimators. J Stat Soft 2004;11(10):1-17. [doi: 10.18637/jss.v011.i10]
33. Zeileis A. Object-oriented computation of sandwich estimators. J Stat Soft 2006;16(9):1-16. [doi: 10.18637/jss.v016.i09]
34. Berger S, Graham N, Zeileis A. Various versatile variances: an object-oriented implementation of clustered covariances in R. Working Papers in Economics and Statistics 2017;2017(12) [FREE Full text]
35. Turley M, Garrido T, Lowenthal A, Zhou YY. Association between personal health record enrollment and patient loyalty. Am J Manag Care 2012 Jul 01;18(7):e248-e253 [FREE Full text] [Medline: 22823553]
36. Anderson AN, Higgins CM, Haardörfer R, Holstad MM, Nguyen MLT, Waldrop-Valverde D. Disparities in retention in care among adults living with HIV/AIDS: a systematic review. AIDS Behav 2020 Apr;24(4):985-997. [doi: 10.1007/s10461-019-02679-2] [Medline: 31555931]

37. Nance RM, Delaney JAC, Simoni JM, Wilson IB, Mayer KH, Whitney BM, et al. HIV viral suppression trends over time among HIV-infected patients receiving care in the United States, 1997 to 2015: a cohort study. Ann Intern Med 2018 Sep 18;169(6):376-384 [FREE Full text] [doi: 10.7326/M17-2242] [Medline: 30140916]
38. Ghiam MK, Rebeiro PF, Turner M, Rogers WB, Bebawy SS, Raffanti SP, et al. Trends in HIV continuum of care outcomes over ten years of follow-up at a large HIV primary medical home in the Southeastern United States. AIDS Res Hum Retroviruses 2017 Oct;33(10):1027-1034 [FREE Full text] [doi: 10.1089/AID.2017.0016] [Medline: 28462622]
39. Guaraldi G, Milic J, Martinez E, Kamarulzaman A, Mussini C, Waters L, et al. Human immunodeficiency virus (HIV) care models during the coronavirus disease 2019 (COVID-19) era. Clin Infect Dis 2021 Sep 07;73(5):e1222-e1227. [doi: 10.1093/cid/ciaa1864] [Medline: 34492689]
40. Jiang H, Zhou Y, Tang W. Maintaining HIV care during the COVID-19 pandemic. Lancet HIV 2020 May;7(5):e308-e309 [FREE Full text] [doi: 10.1016/S2352-3018(20)30105-3] [Medline: 32272084]
41. Reynolds R, Smoller S, Allen A, Nicholas PK. Health literacy and health outcomes in persons living with HIV disease: a systematic review. AIDS Behav 2019 Nov 19;23(11):3024-3043. [doi: 10.1007/s10461-019-02432-9] [Medline: 30783871]
42. Cheng C, Beauchamp A, Elsworth GR, Osborne RH. Applying the electronic health literacy lens: systematic review of electronic health interventions targeted at socially disadvantaged groups. J Med Internet Res 2020 Aug 13;22(8):e18476 [FREE Full text] [doi: 10.2196/18476] [Medline: 32788144]

Abbreviations

aPR: adjusted prevalence ratio
IDU: injection drug use
MSM: men who have sex with men

Edited by C Lovis; submitted 04.11.21; peer-reviewed by I Mircheva, A Midboe, P Jha; comments to author 21.12.21; revised version received 14.03.22; accepted 11.04.22; published 25.07.22.

Please cite as:
Schember CO, Scott SE, Jenkins CA, Rebeiro PF, Turner M, Furukawa SS, Bofill C, Yan Z, Jackson GP, Pettit AC
Electronic Patient Portal Access, Retention in Care, and Viral Suppression Among People Living With HIV in Southeastern United States: Observational Study
JMIR Med Inform 2022;10(7):e34712
URL: https://medinform.jmir.org/2022/7/e34712
doi:10.2196/34712
PMID:35877160

Original Paper

The Impact of Telemedicine on Physicians' After-hours Electronic Health Record "Work Outside Work" During the COVID-19 Pandemic: Retrospective Cohort Study

Katharine Lawrence[1,2,3*], MPH, MD; Oded Nov[1,4*], PhD; Devin Mann[1,2,3*], MD; Soumik Mandal[1,4*], PhD; Eduardo Iturrate[2*], MD; Batia Wiesenfeld[5*], PhD

[1]Department of Population Health, New York University Grossman School of Medicine, New York, NY, United States
[2]Department of Medicine, New York University Grossman School of Medicine, New York, NY, United States
[3]Medical Center Information Technology, New York University Langone Health, New York, NY, United States
[4]Department of Technology Management and Innovation, New York University Tandon School of Engineering, New York, NY, United States
[5]Management Department, New York University Stern School of Business, New York, NY, United States
[*]all authors contributed equally

Corresponding Author:
Katharine Lawrence, MPH, MD
Department of Population Health
New York University Grossman School of Medicine
227 E 30th St 6th Fl
New York, NY, 10016
United States
Phone: 1 646 501 3400
Email: katharine.lawrence@nyulangone.org

Abstract

Background: Telemedicine as a mode of health care work has grown dramatically during the COVID-19 pandemic; the impact of this transition on clinicians' after-hours electronic health record (EHR)–based clinical and administrative work is unclear.

Objective: This study assesses the impact of the transition to telemedicine during the COVID-19 pandemic on physicians' EHR-based after-hours workload (ie, "work outside work") at a large academic medical center in New York City.

Methods: We conducted an EHR-based retrospective cohort study of ambulatory care physicians providing telemedicine services before the pandemic, during the acute pandemic, and after the acute pandemic, relating EHR-based after-hours work to telemedicine intensity (ie, percentage of care provided via telemedicine) and clinical load (ie, patient load per provider).

Results: A total of 2129 physicians were included in this study. During the acute pandemic, the volume of care provided via telemedicine significantly increased for all physicians, whereas patient volume decreased. When normalized by clinical load (ie, average appointments per day by average clinical days per week), telemedicine intensity was positively associated with work outside work across time periods. This association was strongest after the acute pandemic.

Conclusions: Taking physicians' clinical load into account, physicians who devoted a higher proportion of their clinical time to telemedicine throughout various stages of the pandemic engaged in higher levels of EHR-based after-hours work compared to those who used telemedicine less intensively. This suggests that telemedicine, as currently delivered, may be less efficient than in-person–based care and may increase the after-hours work burden of physicians.

(JMIR Med Inform 2022;10(7):e34826) doi:10.2196/34826

KEYWORDS

telemedicine; telehealth; eHealth; COVID-19; EHR; electronic health record; clinician workload; impact; transition; workload; cohort; retrospective; physician; efficient; doctor; health care professional; pandemic

Introduction

The COVID-19 pandemic precipitated the rise of telemedicine—defined as the synchronous provision of health care services via telecommunications, either video or audio, to patients at remote sites—as a powerful disrupter of health care delivery [1-3]. Although not a new mode of work, the adoption and scaling of telemedicine prior to the pandemic was limited due to individual-, practice-, and system-level barriers that included technical and usability constraints, clinician practice patterns and preferences, security concerns, as well as payor and regulatory environments [4,5]. The significant disruptions to health care delivery caused by COVID-19 necessitated the rapid implementation of telemedicine in a variety of forms across practices and hospital systems in the United States and globally.

Prior to the pandemic, studies of the provision of clinical care through the medium of telemedicine identified potential benefits such as improved access to care in underserved regions or communities, better coordination of care, greater convenience, and lower costs [6,7]. Telemedicine may also have the potential to improve clinicians' well-being and reduce burnout by improving associated risk factors such as on-call burden, communication, and job satisfaction [8-10]. At the same time, however, the introduction of novel technologies that impact the provision and experience of health care work can also be detrimental; in particular, there is concern about the impact of electronic health records (EHRs) on clinicians' experience of work and its role in increasing both clinical and nonclinical administrative burden for physicians, including time spent on work-related tasks "outside" of clinical hours, often referred to as "work outside work" (WOW) or "pajama time" (PT) [11-14]. Shifting clinical and administrative work into personal time, particularly when physicians are at home, is a source of concern within the medical community, and it is unclear whether the proliferation of telemedicine as a form of health care work will exacerbate or ameliorate these conditions.

In this paper we focus on ambulatory physicians' WOW during a time of rapid telework transition spurred by the COVID-19 pandemic. Our goal is to evaluate the impact of telemedicine practice on ambulatory physicians' EHR-based WOW during the large-scale rollout of telemedicine in an urban academic hospital system during the COVID-19 pandemic.

Methods

Study Setting

New York University Langone Health (NYULH) is a large academic health care system in New York City, with over 8000 health care providers across 4 hospitals and over 500 ambulatory faculty group practices. The system is connected via a single EHR system, Epic, with over 7.5 million active patient accounts. Prior to the COVID-19 pandemic, NYULH offered limited telemedicine services only through pilot programs such as "virtual urgent care" (in emergency medicine), postoperative wound checks (in orthopedics), and some mental health services. Telemedicine for primary care and other routine health services was not available. During the pandemic, NYULH rapidly scaled its telemedicine offerings to include primary care, ambulatory specialty practice, and urgent care. NYULH "virtual health" was comprised of a single, enterprise-wide instance of synchronous, video-based telecommunications encounters between physicians and patients in remote locations accessed through a standardized EHR-based patient portal system and a third-party videoconferencing vendor. This platform provided a unified patient and provider experience between clinical practice sites and across specialties. At the height of the pandemic, this system saw an 8595% increase in monthly telemedicine visits between February (n=1699) and April (n=147,736), with over 2000 unique physicians engaging in video visits [15].

Study Design

This is an EHR-based retrospective cohort study including all ambulatory care physicians continuously practicing (defined as at least 5 appointments scheduled per week in the reporting period) at any New York-based NYULH faculty group practice site between January 1, 2020, and August 31, 2020. Nonphysician practitioners (eg, advanced-practice providers) and residents were not included in the study cohort, as with few exceptions, they did not provide telemedicine-based care during this period.

Ethical Considerations

This study was deemed part of a quality improvement and met the criteria for exemption from institutional review board's review according to NYULH institutional policy. All data were collected as part of routine clinical care and administrative management.

Study Measures

Definitions of key variables associated with study measures and analysis are provided in Table 1 and Table 2.

Table 1. Epic metric key terms and variables associated with study measures.

Epic metric	Description	Calculation
Reporting period	For a month, it starts on the Sunday on or immediately before the 1st and ends on the last Saturday of the month.	= (End date-start date)
Days with appointments	Percentage of days with at least one appointment within the reporting period.	For a reporting period:
Appointments per day	Average minutes a provider spent in the system outside of scheduled hours.	For a reporting period:
Time spent outside scheduled hours	Average minutes a provider spent in the system outside of scheduled hours.	For a reporting period:
Time spent on unscheduled days	Average minutes a provider spent in the system on days with no scheduled patients.	For a reporting period:

Table 2. Derived metric key terms and variables associated with study measures.

Derived metric	Calculation
Scheduled days	For a reporting period:
Unscheduled days	For a reporting period:
Time outside scheduled hours per month	For a reporting period:
Time on unscheduled days per month	For a reporting period:
Clinical load	For a reporting period:
“Work outside work” measure	For a reporting period:

Pandemic Time Period

To evaluate whether the effects of telemedicine intensity were influenced by the evolving stages of the COVID-19 pandemic, we aggregated monthly physician data into the following 3 successive time periods: (1) the prepandemic period of January 1-February 29, 2020; (2) the acute pandemic period of March 1-May 31 (with March 15th representing the date when most NYULH ambulatory practices were closed for in-person visits); and (3) after the acute pandemic period of June 1-August 31, representing the gradual resumption of in-person care.

Telemedicine Intensity

To create a measure of the relative volume of clinical care physicians provided via telemedicine, we calculated the proportion of total visits per month that were telemedicine-based for each physician (number of video visits per month divided by the total number of all patient visits per month per provider) with values that could range from 0 to 1.

Clinical Load

Prior research has found clinical load to be an important predictor of WOW burden [11,14] and recommended normalizing WOW by load [11]. To account for the reduction and gradual resumption of in-person care during the pandemic, we created a measure of clinical load reflecting the total number of patient appointments for each physician each month. This was calculated by multiplying Epic-reported values of average number of appointments per clinical day (in-person or via telemedicine) by average number of clinical days per week, for each physician each month.

WOW

Derived from EHR user activity logs from Epic, WOW was calculated by adding time outside scheduled hours (ie, the average minutes per day spent in the system outside of scheduled hours on scheduled days, where scheduled hours are determined using Epic Cadence scheduling data plus two 30-minute “buffer” periods added before the start of first appointment and after the end of last appointment) and time on unscheduled days (ie, the average number of minutes per day spent in the system on days with no scheduled patients). WOW was normalized for physicians’ patient load by dividing WOW by clinical load to create a measure reflecting WOW per appointment.

An alternative measure of WOW uses the Epic EHR’s own variable-generated data—PT. PT represents the average number of minutes per day spent in charting activities on weekdays outside a standard (local) 7 AM to 5:30 PM workday and any time on weekends. PT does not include time spent personalizing EHR tools (eg, documentation templates or preferences lists) or time using reporting tools such as SlicerDicer and Reporting Workbench during unscheduled days. Although PT can be used as a marker of after-hours clinical work, recent studies have called into question its accuracy and usefulness for this purpose [15,16]. These concerns are likely exacerbated during the pandemic due to the significant disruptions in clinical care hours and work schedules for practices and physicians (eg, the closure of clinics, physician illness and exposure, and the variable outpatient work hours of physicians who were asked to provide emergency inpatient care), and therefore, this value was not included in this study.

Statistical Analysis

We first computed telemedicine intensity, clinical load, WOW, and WOW per appointment for all physicians in the EHR that met our inclusion criteria. To evaluate whether WOW significantly varied across time periods, we ran one-way ANOVAs on both WOW and WOW per appointment. To evaluate the effect of telemedicine intensity and time period on after-hours work burden, as well as whether the relationship between telemedicine intensity and after-hours work varied across time periods, we conducted a hierarchical linear regression analysis in which the dependent variable was WOW per appointment. We first entered the main effects of telemedicine intensity and pandemic time period, followed by the interaction of telemedicine intensity and pandemic time period. To understand the nature of the interaction of telemedicine intensity and pandemic time period, we partitioned the data by time period and regressed WOW per appointment on telemedicine intensity in each time period. All analyses were conducted using SPSS (version 28; IBM Corp).

Results

We analyzed data on 2129 physicians from January to August 2020. The majority of physicians were from internal medicine subspecialties (eg, cardiology, pulmonology, and geriatrics), followed by ambulatory surgery (including general surgery and surgical subspecialists) and general medicine practice (eg, internal medicine and family medicine; Table 3).

One-way ANOVAs evaluating whether the average WOW per day and WOW per appointment varied by pandemic time period were significant across physicians (average WOW per day: $F_2(2,12822)=33.09$; $P<.001$; WOW per appointment: $F_2(2,12784)=42.68$; $P<.001$). Average WOW per day declined during the acute pandemic relative to the prepandemic period and then reverted back to prepandemic levels after the acute pandemic. However, WOW per appointment increased during the acute pandemic period across all physicians, before subsequently declining (approaching but not reaching prepandemic levels) after the acute pandemic (Table 4).

Across time periods (before the pandemic, during acute pandemic, and after acute pandemic) telemedicine intensity was positively associated with WOW per appointment (step 1 in Table 5), with physicians who spent a larger proportion of their time providing care via telemedicine devoting significantly more time to after-hours EHR work. Although the pandemic time period did not significantly affect WOW per appointment after controlling for telemedicine intensity, it significantly moderated the effect of telemedicine intensity on WOW per appointment (step 2 in Table 5). Regressions of WOW per appointment by telemedicine intensity for each time period showed that the positive relationship between telemedicine intensity and WOW per appointment was amplified over time, with the strongest positive relationship in the period after acute pandemic (Figure 1).

Table 3. Specialty of included study physicians (N=2129).

Clinical specialty	Values, n (%)
Internal medicine subspecialty	671 (31.5)
Surgery	377 (17.7)
General practice (eg, internal medicine and family doctors)	326 (15.3)
Pediatrics	175 (8.2)
Neurology	141 (6.6)
Obstetrician and gynecologist	134 (6.3)
Other	91 (4.3)
Psychiatry	72 (3.4)
Emergency medicine	68 (3.2)
Dermatology	36 (1.7)
Rehab	32 (1.5)
Pain medicine	6 (0.3)

Table 4. Work outside work (WOW) per day and per appointment, by time period.

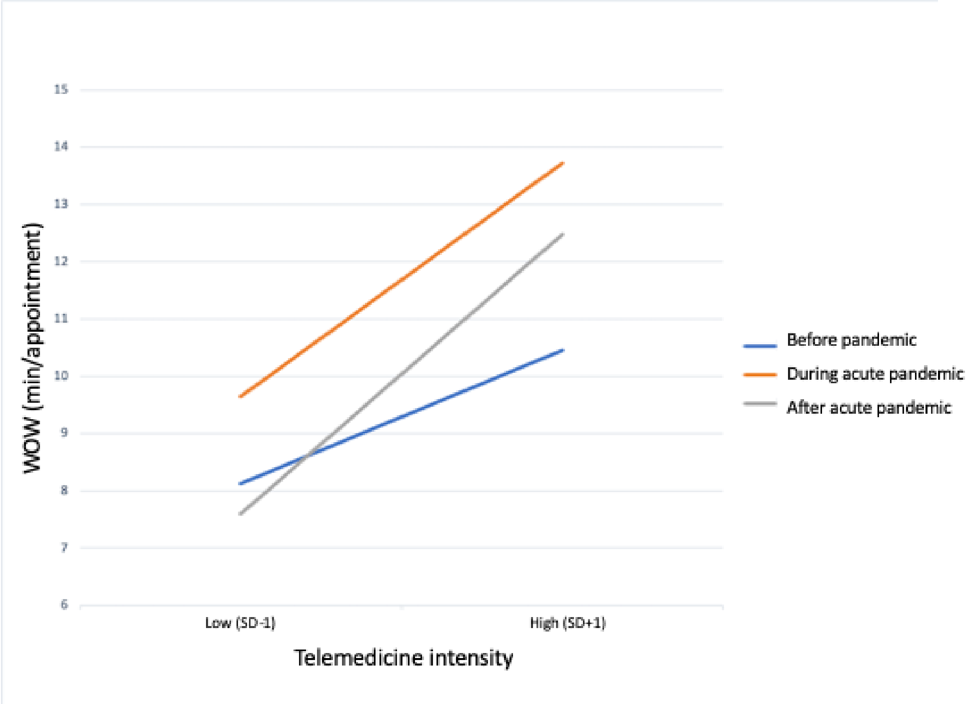
WOW	Time period								
	Before pandemic			During acute pandemic			After acute pandemic		
	Median	Mean	95% CI	Median	Mean	95% CI	Median	Mean	95% CI
WOW per day	27.19	34.50	33.52-35.47	23.96	30.20	29.50-30.91	26.94	34.11	33.31-34.91
WOW per appointment	5.73	9.29	8.92-9.65	7.52	11.68	11.31-12.05	6.04	10.03	9.70-10.37

Table 5. Hierarchical regression of work outside work (WOW) per appointment.

Study variables	Normalized WOW					
	Step 1			Step 2		
	Unstandardized coefficient	Standard error	*P* value	Unstandardized coefficient	Standard error	*P* value
COVID-19 time period	–0.27	0.13	0.05	–0.52	0.15	<.001
Telemedicine intensity	6.67	0.32	<.001	1.37	1.41	
Telemedicine intensity×time period	N/A[a]	N/A	N/A	2.48	0.64	<.001

[a]N/A: not applicable.

Figure 1. Work outside work (WOW) per appointment by telemedicine intensity and time period.



Discussion

Principal Results

Our study found that telemedicine was less efficient than in-person–based care and increased physicians' WOW burden. The overall EHR-based WOW declined for physicians in the context of the COVID-19 pandemic and the rapid transition to telemedicine; however, when controlling for changes in patient volume and clinical hours of care, physicians who devoted a higher proportion of their clinical time to telemedicine had higher levels of EHR-based WOW than those who used telemedicine less intensively. This relationship was present during all phases of the study (before the pandemic, during acute pandemic, and after acute pandemic) and was amplified over time, including in the after acute pandemic phase. These findings suggest that the observed decrease in the average WOW during the pandemic was the result of the overall decrease in clinical load for physicians rather than any benefits or efficiencies of telemedicine itself. Further, the amplification of the relationship between WOW per appointment and telemedicine intensity in the time period beyond the acute pandemic suggests that the WOW increasing effect of telemedicine was exacerbated over time, and therefore, the unique circumstances of the early COVID-19 pandemic alone are insufficient to explain the behavior patterns of physicians.

Limitations

There are several limitations to this study that future research could address. First, limitations in our Epic-based data set preclude the ability to review and analyze physician EHR activity with sufficient granularity beyond certain time periods; for example, time periods more specific than a calendar month or physician activity log data at smaller than 15-minute increments. Specifically, Epic does not count WOW in its time outside of scheduled hours if that work occurs within the 30 minutes before or after patient scheduled hours (a "shoulder period"), which our analysis is unable to reliably differentiate as WOW time and therefore excludes, resulting in a systematic underestimation of the true WOW. Moreover, because shoulder time is added for each clinical day regardless of length, this underestimation bias is greater for physicians who spread their patient time over more scheduled days relative to those who see the same number of patients on fewer days [11]. Similarly, we are unable to target more specific times of pandemic disruption (eg, March 15, which is the exact date when most of our institution's ambulatory clinics closed for in-person care). Second, we are limited in our ability to analyze activity at the level of physician or patient demographics; therefore, we are unable to comment on whether factors such as gender, age, or years in practice may have affected clinical load, telemedicine intensity, or WOW, and whether patient features such as patient complexity or acuity contributed to these outcomes. It is possible, for example, that telemedicine-based visits are overall less clinically intense compared to in-person visits due to differences in patient case mix, in which case our analysis would underestimate the time costs associated with telemedicine-based visits. Third, EHR-based data and work represent only part of the overall nonclinical burden of physicians. Additional time spent reviewing non-EHR based records, discussing care plans, working with interdisciplinary teams (eg, nurses and care managers), or advocating with insurers is not captured in this study; this work may have been increased during the pandemic due to disruptions in traditional office practices and workflows. Additionally, as our data are behavioral, we are unable to directly associate our measures with important factors such as physicians' attitudes (eg, stress and burnout). Finally, our findings represent only the experience of physicians at a single health care system during the unusual period of the COVID-19 pandemic and the rapid transition to telemedicine, which may limit generalizability across provider type, practice environment, or geographic location.

Interpretation of Findings in Clinical Context

To our knowledge, this is the first study to systematically evaluate the impact of the transition to telemedicine during the COVID-19 pandemic on physicians' after-hours workload and one of a few studies that used EHR-based data to objectively evaluate after-hours work burden [15-17]. Although research documenting the experience of health systems undergoing the transition to telemedicine in response to the pandemic has increased [18-20], there is limited research exploring the effects of telemedicine on health care delivery areas such as clinical workflows, administrative load, or practice efficiencies, either during the pandemic or prior to it; the most robust of these works are almost a decade old and reflect a dated telemedicine environment that may no longer be relevant to the current context of health care delivery [21,22]. Similarly, literature exploring the impact of telemedicine on important aspects of physician work experience such as burnout and quality of life are limited, with the majority of work prior to the pandemic coming out of the field of telepsychiatry as an "early adopter" of the technology [23,24]. This study contributes to the literature on telemedicine in health care by exploring both the novel context of its expansion during the COVID-19 pandemic and its relationship to EHR-based work burden for clinicians.

A number of factors may be responsible for our findings that telemedicine increased the after-hours work burden of physicians. First, it is possible that organizational and technological inefficiencies in the early design, deployment, and scaling of telemedicine may have resulted in increased after-hours EHR work burden for physicians using telemedicine more intensively. These include early and ongoing technological issues relating to the computer hardware, software functionality and integrations, and user experience of the "virtual health" platform deployed by our system. These issues have been highlighted elsewhere in EHR and digital health technology implementation research, particularly regarding usability and user experience barriers [25-27] exacerbated by the scale and abruptness of the transition to telemedicine due to the pandemic [28]. However, technological inefficiencies should be at least partially ameliorated over time as physicians learn to navigate and optimize their setup and systems (the "learning curve"), an assumption that is not supported by our after acute pandemic period findings of a continued "amplified" relationship between telemedicine intensity and after-hours EHR work. Similarly, telemedicine training for physicians during this period of rapid expansion was often ad hoc and likely suboptimal for the development of effective telemedicine competencies (eg, efficient platform navigation, technical troubleshooting, "virtual health" EHR documentation), and thereby, potentially worsening WOW; however, this would be expected to improve with time as physicians adapted their workflows and learned new skills, rather than, as our results found, establishing a pattern of increasing work burden in the later periods of telemedicine of use, even as access to quality telemedicine trainings and best practice knowledge sharing improved among institutions. This suggests that "virtual health" training as it existed during the early phases of the pandemic was not sufficient to improve after-hours work burden for physicians. Further exploration of the relationship between telemedicine training and "virtual health" practice patterns (including EHR-based activities) is warranted as training becomes more regularly integrated into medical education and professional development.

The second factor that might have impacted our findings is that it is likely that significant disruptions to the work norms of clinical practices during the pandemic affected after-hours work patterns. In clinics, individual- and practice-level adjustments to the demands of care provision during the pandemic likely resulted in a number of unique work structures and arrangements that could have likely affected physicians' work schedules, including time spent doing after-hours work. In particular, the shift to a telemedicine-based platform—particularly one with limited multiparty functionality—may have inhibited effective team-based care between physicians and clinical support staff (eg, medical assistants) and shifted both clinical and administrative tasks that had prior been completed by other staff members onto physicians. This "doctor does it all" phenomenon has been recently described as an unintended effect of the rapid transition to telemedicine during the pandemic [29]; within our own system, much of the current WOW involves responding to patient messages, phone calls, refill requests, and completing various EHR documentation requirements often left for the end of the day after direct patient care responsibilities are ended. NYULH is actively engaged in reducing this burden on providers by redistributing relevant work to support staff, as well as using novel technologies including machine learning to facilitate message triage and management, for example by suppressing messages that are not actionable by providers. More work is needed to fully understand the impact of the new virtual-first models of care delivery on interdisciplinary teams and team-based practice.

Learning from Other Fields and Implications for Health Care Practice

Overall, our results suggest that telemedicine is not panacea for the work challenges facing clinicians. In fact, our evidence during the acute pandemic and after the acute pandemic suggests that rather than reducing administrative burden, telemedicine intensity may increase it, shifting the work temporally and spatially to after-hours work and home. This suggests that a more thorough understanding of the implications of telemedicine in clinical practice is necessary prior to its indiscriminate expansion to ensure policies and practices that increase efficiency and work-life quality and counteract inefficiencies, waste, and work-related stress and burnout are implemented. Given the limited data available on the impact of telemedicine on important aspects of physicians' experience of work, it may be instructive to look to fields outside of medicine, where the study of "telework" (defined as a work arrangement that allows employees to perform work at approved alternative or remote worksites) [29] is more robust. Research in the industries of engineering, consulting, and software development has demonstrated varying effects of remote work on key elements of employees' work experience. Positive effects of telework in these fields include increased job satisfaction, performance, and work-life balance, as well as reduced employee turnover, real estate costs, commute time, and environmental impact [30-32].

Conversely, negative effects include reduced career development and feelings of reduced energy, confidence, and engagement due to the loss of high-quality interaction with colleagues and clients [33,34]. Significantly, telework has been associated with workers' inability to disconnect from their work and increased stress-inducing work intensification [35,36]. This relationship may apply to telemedicine and help explain the findings in this study. Although more investigation is needed to understand the full scope and implications of medical telework beyond the direct care provided by telemedicine (including tasks such as remote teaching, non-EHR–based clinical work, administrative work, and research), general learning from these fields may help identify and guide key areas of future telemedicine and telework research.

Conclusions

In this study, we evaluated the impact of the transition to telemedicine during the COVID-19 pandemic on physicians' EHR-based after-hours workload; we found that when controlling for the clinical load of patient visits, physicians who devoted a higher proportion of their clinical time to telemedicine engaged in higher levels of EHR-based after-hours work compared to those who used telemedicine less intensively; this relationship persisted and was amplified over time, even after the acute pandemic period. This suggests that telemedicine, as currently delivered, may be less efficient than in-person–based care and may contribute to after-hours work burden of physicians. Further study is needed on the detailed impacts of telemedicine on physician work practices, particularly in contexts beyond the COVID-19 pandemic and relating to administrative burden, after-hours clinical responsibilities (particularly the EHR-related in-basket and patient portal messaging responsibilities), and experience of work. Learning from other industries where telework is more established can help identify areas of need and opportunity in future telemedicine care delivery.

Acknowledgments

We would like to acknowledge the work of our colleagues in the Department of Population Health. KL, ON, DM, and BW's research is supported by the National Science Foundation (grants 1928614 and 2129076). The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the supporting agencies.

Data Availability

The data underlying this article will be shared upon reasonable request to the corresponding author.

Authors' Contributions

KL, DM, ON, and BW significantly contributed to the conception and design of the study. EI and SM acquired the data. KL, BW, EI, and SM analyzed the data. KL, BW, ON, and DM drafted the initial manuscript. All authors were involved in data interpretation and manuscript revision and approved the final version submitted for publication.

Conflicts of Interest

None declared.

References

1. Using telehealth to expand access to essential health services during the COVID-19 pandemic. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/coronavirus/2019-ncov/hcp/telehealth.html [accessed 2020-12-01]
2. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: evidence from the field. J Am Med Inform Assoc 2020 Apr 23:1132-1135 [FREE Full text] [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]
3. Bokolo Anthony Jnr. Use of telemedicine and virtual care for remote treatment in response to COVID-19 pandemic. J Med Syst 2020 Jun 15;44(7):132 [FREE Full text] [doi: 10.1007/s10916-020-01596-5] [Medline: 32542571]
4. Shigekawa E, Fix M, Corbett G, Roby DH, Coffman J. The current state of telehealth evidence: a rapid review. Health Aff (Millwood) 2018 Dec;37(12):1975-1982. [doi: 10.1377/hlthaff.2018.05132] [Medline: 30633674]
5. Scott KC, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018 Jan;24(1):4-12 [FREE Full text] [doi: 10.1177/1357633X16674087] [Medline: 29320966]
6. Hailey D, Roine R, Ohinmaa A. Systematic review of evidence for the benefits of telemedicine. J Telemed Telecare 2002 Dec 02;8 Suppl 1(1_suppl):1-30. [doi: 10.1258/1357633021937604] [Medline: 12020415]
7. Roine R, Ohinmaa A, Hailey D. Assessing telemedicine: a systematic review of the literature. CMAJ 2001 Sep 18;165(6):765-771 [FREE Full text] [Medline: 11584564]
8. Doshi A, Platt Y, Dressen JR, Mathews BK, Siy JC. Keep calm and log on: telemedicine for COVID-19 pandemic response. J Hosp Med 2020 May;15(5):302-304. [doi: 10.12788/jhm.3419] [Medline: 32379036]

9. Reliford A, Adebanjo B. Use of telepsychiatry in pediatric emergency room to decrease length of stay for psychiatric patients, improve resident on-call burden, and reduce factors related to physician burnout. Telemed J E Health 2019 Sep 01;25(9):828-832. [doi: 10.1089/tmj.2018.0124] [Medline: 30379635]
10. Hjelm NM. Benefits and drawbacks of telemedicine. J Telemed Telecare 2005 Jun 24;11(2):60-70. [doi: 10.1258/1357633053499886] [Medline: 15829049]
11. Sinsky CA, Rule A, Cohen G, Arndt B, Shanafelt TD, Sharp CD, et al. Metrics for assessing physician activity using electronic health record log data. J Am Med Inform Assoc 2020 Apr 01;27(4):639-643 [FREE Full text] [doi: 10.1093/jamia/ocz223] [Medline: 32027360]
12. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan W, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426 [FREE Full text] [doi: 10.1370/afm.2121] [Medline: 28893811]
13. Marmor R, Clay B, Millen M, Savides T, Longhurst C. The impact of physician ehr usage on patient satisfaction. Appl Clin Inform 2018 Jan 03;9(1):11-14 [FREE Full text] [doi: 10.1055/s-0037-1620263] [Medline: 29298451]
14. Saag HS, Shah K, Jones SA, Testa PA, Horwitz LI. Pajama time: working after work in the electronic health record. J Gen Intern Med 2019 Sep 9;34(9):1695-1696 [FREE Full text] [doi: 10.1007/s11606-019-05055-x] [Medline: 31073856]
15. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan W, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426 [FREE Full text] [doi: 10.1370/afm.2121] [Medline: 28893811]
16. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: Time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020 Apr 01;27(4):531-538 [FREE Full text] [doi: 10.1093/jamia/ocz220] [Medline: 32016375]
17. Melnick ER, Ong SY, Fong A, Socrates V, Ratwani RM, Nath B, et al. Characterizing physician EHR use with vendor derived data: a feasibility study and cross-sectional analysis. J Am Med Inform Assoc 2021 Jul 14;28(7):1383-1392 [FREE Full text] [doi: 10.1093/jamia/ocab011] [Medline: 33822970]
18. Haider Z, Aweid B, Subramanian P, Iranpour F. Telemedicine in orthopaedics during COVID-19 and beyond: a systematic review. J Telemed Telecare 2022 Jul 06;28(6):391-403. [doi: 10.1177/1357633X20938241] [Medline: 32762270]
19. Novara G, Checcucci E, Crestani A, Abrate A, Esperto F, Pavan N, Research Urology Network (RUN). Telehealth in urology: a systematic review of the literature. How much can telemedicine be useful during and after the COVID-19 pandemic? Eur Urol 2020 Dec;78(6):786-811 [FREE Full text] [doi: 10.1016/j.eururo.2020.06.025] [Medline: 32616405]
20. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 01;20(1):1193 [FREE Full text] [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]
21. Ekeland AG, Bowes A, Flottorp S. Effectiveness of telemedicine: a systematic review of reviews. Int J Med Inform 2010 Nov;79(11):736-771. [doi: 10.1016/j.ijmedinf.2010.08.006] [Medline: 20884286]
22. Mistry H. Systematic review of studies of the cost-effectiveness of telemedicine and telecare. Changes in the economic evidence over twenty years. J Telemed Telecare 2012 Jan;18(1):1-6. [doi: 10.1258/jtt.2011.110505] [Medline: 22101609]
23. Gardner JS, Plaven BE, Yellowlees P, Shore JH. Remote telepsychiatry workforce: a solution to psychiatry's workforce issues. Curr Psychiatry Rep 2020 Jan 27;22(2):8. [doi: 10.1007/s11920-020-1128-7] [Medline: 31989417]
24. Vogt EL, Mahmoud H, Elhaj O. Telepsychiatry: implications for psychiatrist burnout and well-being. Psychiatr Serv 2019 May 01;70(5):422-424. [doi: 10.1176/appi.ps.201800465] [Medline: 30813863]
25. Goodison R, Borycki EM, Kushniruk AW. Use of agile project methodology in health care it implementations: a scoping review. Stud Health Technol Inform 2019;257:140-145. [Medline: 30741186]
26. Shah S, Diwan S, Kohan L, Rosenblum D, Gharibo C, Soin A, et al. The technological impact of COVID-19 on the future of education and health care delivery. Pain Physician 2020 Aug;23(4S):S367-S380 [FREE Full text] [Medline: 32942794]
27. What is telework? United States Office of Personnel Management. URL: https://www.opm.gov/faqs/QA.aspx?fid=b48bf83b-440c-4f1e-a88c-3cdc9d802ac8&pid=75346675-3b92-4aec-831d-58cf5b0e86d2&result=1 [accessed 2021-01-04]
28. Sherwin J, Lawrence K, Gragnano V, Testa PA. Scaling virtual health at the epicentre of coronavirus disease 2019: A case study from NYU Langone Health. J Telemed Telecare 2020 Jul 19;28(3):224-229. [doi: 10.1177/1357633x20941395]
29. Sinsky CA, Jerzak JT, Hopkins KD. Telemedicine and team-based care: the perils and the promise. Mayo Clin Proc 2021 Feb;96(2):429-437 [FREE Full text] [doi: 10.1016/j.mayocp.2020.11.020] [Medline: 33549262]
30. Gajendran RS, Harrison DA. The good, the bad, and the unknown about telecommuting: meta-analysis of psychological mediators and individual consequences. J Appl Psychol 2007 Nov;92(6):1524-1541. [doi: 10.1037/0021-9010.92.6.1524] [Medline: 18020794]
31. Bloom N, Liang J, Roberts J, Ying ZJ. Does working from home work? Evidence from a Chinese experiment. Q J Econ 2015 Feb;130(1):165-218 [FREE Full text] [doi: 10.1093/qje/qju032]
32. Guimaraes T, Dallow P. Empirically testing the benefits, problems, and success factors for telecommuting programmes. Eur J Inf Syst 2017 Dec 19;8(1):40-54. [doi: 10.1057/palgrave.ejis.3000317]

33. Leonardi PM, Treem JW, Jackson MH. The connectivity paradox: using technology to both decrease and increase perceptions of distance in distributed work arrangements. JACR 2010 Feb;38(1):85-105. [doi: 10.1080/00909880903483599]
34. Felstead A, Henseke G. Assessing the growth of remote working and its consequences for effort, well-being and work-life balance. New Technol Work Employ 2017 Oct 04;32(3):195-212. [doi: 10.1111/ntwe.12097]
35. Wiesenfeld BM, Raghuram S, Garud R. Communication patterns as determinants of organizational identification in a virtual organization. Organ Sci 1999 Dec;10(6):777-790. [doi: 10.1287/orsc.10.6.777]
36. Wiesenfeld BM, Raghuram S, Garud R. Organizational identification among virtual workers: the role of need for affiliation and perceived work-based social support. J Manag 2016 Jun 30;27(2):213-229. [doi: 10.1177/014920630102700205]

Abbreviations

EHR: electronic health record
NYULH: New York University Langone Health
PT: pajama time
WOW: work outside work

Edited by C Lovis; submitted 09.11.21; peer-reviewed by C Sinsky, A Behmanesh, P Yellowlees; comments to author 24.01.22; revised version received 03.05.22; accepted 26.05.22; published 28.07.22.

Please cite as:
Lawrence K, Nov O, Mann D, Mandal S, Iturrate E, Wiesenfeld B
The Impact of Telemedicine on Physicians' After-hours Electronic Health Record "Work Outside Work" During the COVID-19 Pandemic: Retrospective Cohort Study
JMIR Med Inform 2022;10(7):e34826
URL: https://medinform.jmir.org/2022/7/e34826
doi:10.2196/34826
PMID:35749661

Original Paper

Extraction of Explicit and Implicit Cause-Effect Relationships in Patient-Reported Diabetes-Related Tweets From 2017 to 2021: Deep Learning Approach

Adrian Ahne[1,2], PhD; Vivek Khetan[3], MSc; Xavier Tannier[4], Prof Dr; Md Imbesat Hassan Rizvi[5], MSc; Thomas Czernichow[2], PhD; Francisco Orchard[2], MSc; Charline Bour[6], MSc; Andrew Fano[3], PhD; Guy Fagherazzi[6], PhD

[1]Center of Epidemiology and Population Health, Inserm, Hospital Gustave Roussy, Paris-Saclay University, Villejuif, France

[2]Epiconcept Company, Paris, France

[3]Accenture Labs, San Francisco, CA, United States

[4]Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, Inserm, University Sorbonne Paris Nord, Sorbonne University, Paris, France

[5]Indian Institute of Science, Bengaluru, India

[6]Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, Strassen, Luxembourg

Corresponding Author:
Adrian Ahne, PhD
Center of Epidemiology and Population Health
Inserm, Hospital Gustave Roussy
Paris-Saclay University
20 Rue du Dr Pinel
Villejuif, 94800
France
Phone: 33 142115386
Email: adrian.ahne@protonmail.com

Abstract

Background: Intervening in and preventing diabetes distress requires an understanding of its causes and, in particular, from a patient's perspective. Social media data provide direct access to how patients see and understand their disease and consequently show the causes of diabetes distress.

Objective: Leveraging machine learning methods, we aim to extract both explicit and implicit cause-effect relationships in patient-reported diabetes-related tweets and provide a methodology to better understand the opinions, feelings, and observations shared within the diabetes online community from a causality perspective.

Methods: More than 30 million diabetes-related tweets in English were collected between April 2017 and January 2021. Deep learning and natural language processing methods were applied to focus on tweets with personal and emotional content. A cause-effect tweet data set was manually labeled and used to train (1) a fine-tuned BERTweet model to detect causal sentences containing a causal relation and (2) a conditional random field model with Bidirectional Encoder Representations from Transformers (BERT)-based features to extract possible cause-effect associations. Causes and effects were clustered in a semisupervised approach and visualized in an interactive cause-effect network.

Results: Causal sentences were detected with a recall of 68% in an imbalanced data set. A conditional random field model with BERT-based features outperformed a fine-tuned BERT model for cause-effect detection with a macro recall of 68%. This led to 96,676 sentences with cause-effect relationships. "Diabetes" was identified as the central cluster followed by "death" and "insulin." Insulin pricing–related causes were frequently associated with death.

Conclusions: A novel methodology was developed to detect causal sentences and identify both explicit and implicit, single and multiword cause, and the corresponding effect, as expressed in diabetes-related tweets leveraging BERT-based architectures and visualized as cause-effect network. Extracting causal associations in real life, patient-reported outcomes in social media data provide a useful complementary source of information in diabetes research.

(JMIR Med Inform 2022;10(7):e37201) doi:10.2196/37201

KEYWORDS

causality; deep learning; natural language processing; diabetes; social media; causal relation extraction; social media data; machine learning

Introduction

Diabetes distress refers to psychological factors such as emotional burden, worries, frustration, or stress in the day-to-day management of all types of diabetes [1-3]. Diabetes distress is associated with poor quality of life [4], high hemoglobin A_{1C} levels [5,6], and low medication adherence [7]. Reducing diabetes distress may improve hemoglobin A_{1c} levels and reduce the burden of disease among people with diabetes [8]. Social media is a useful observatory resource for patient-reported diabetes issues and could help to contribute directly to public and clinical decision-making from a patient's perspective, given the active online diabetes community [9,10]. Identifying causal relations in expressed text data in social media platforms might help to discover unknown etiological results, specifically, causes of health problems, concerns, and symptoms.

To intervene and potentially prevent diabetes distress, it is necessary to understand the causes of diabetes distress from a patient's perspective to understand how patients see their disease. Causal relation extraction in natural language text has gained popularity in clinical decision-making, biomedical knowledge discovery, or emergency management [11]. In particular, causal relations on Twitter have been examined for diverse factors causing stress and relaxation [12], adverse drug reactions [13], or causal associations related to insomnia or headache [14]. Most approaches examine *explicit* causality in text [14-16], when cause and effect are explicitly stated, for instance, by connective words (eg, so, hence, because, lead to, since, if-then) [11,17]. An example for an *explicit* cause-effect pair is "diabetes causes hypoglycemia." However, *implicit* causality is more complicated to detect such as in "I reversed diabetes with lifestyle changes" with cause "lifestyle changes" and effect "reversed diabetes."

Natural language processing methods explore among other things how computers can be used to extract useful information from natural language documents. In combination with machine learning and deep learning models, which are artificial intelligence algorithms designed to learn from experience, they have also been applied to extract causal relations [18,19]. Machine learning methods are able to explore implicit relations and provide better generalization contrary to rule-based approaches [11,20-22]. An interesting approach leveraging the transfer learning paradigm and addressing both explicit and implicit cause-effect extraction is provided by Khetan et al [23]. They fine-tuned pretrained transformer-based Bidirectional Encoder Representations from Transformers (BERT) language models [24,25] to detect "cause-effect" relationships by using publicly available data sets such as the adverse drug effect data set [26]. More generally, the idea of transfer learning is to leverage the knowledge of a model that has been trained on an auxiliary domain [27].

In this study, we aimed to extract spans of text as 2 distinct events from diabetes and diabetes-related tweets such that one event directly (explicit) or indirectly (implicit) impacts another event. We categorized these events as cause-event and effect-event depending upon the expressed context of each tweet. The identified cause and effect will then be aggregated into clusters and ultimately visualized in an interactive cause-effect network.

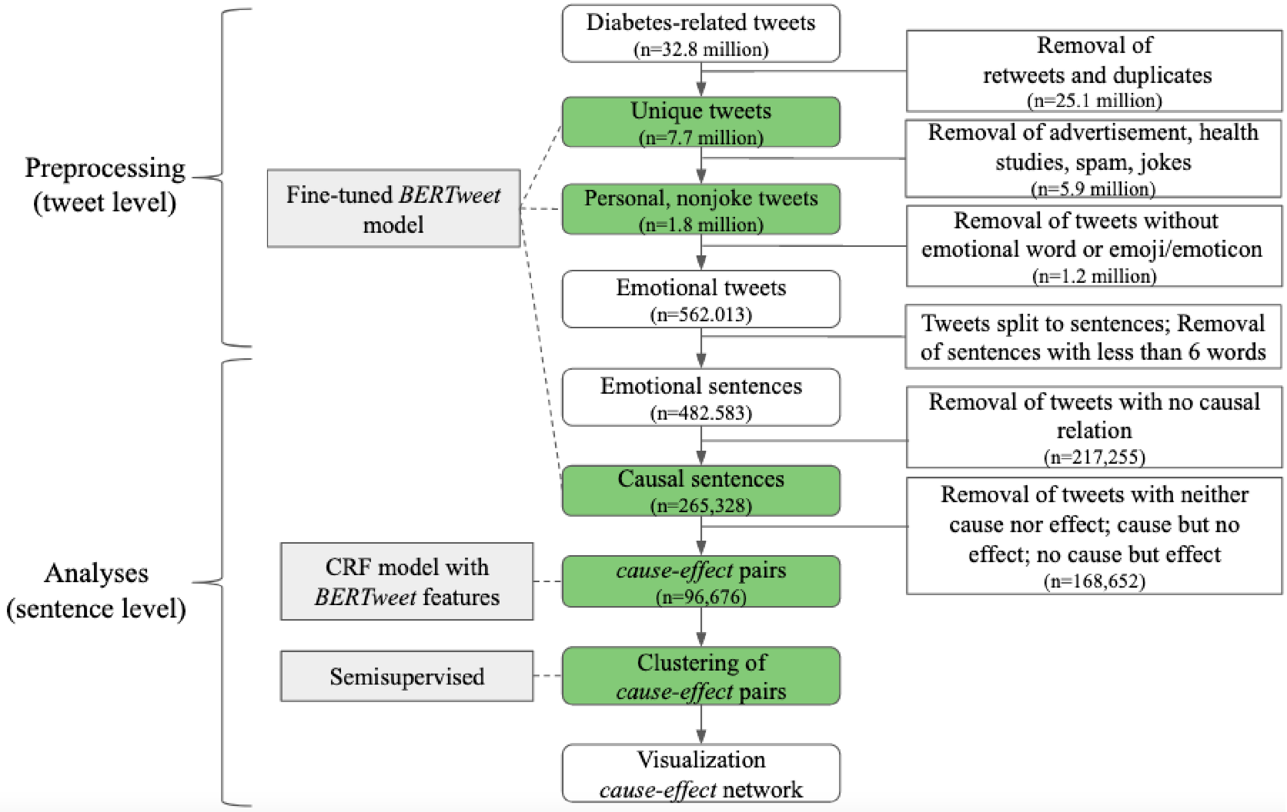
This work is realized in the frame of the World Diabetes Distress Study, which aims to analyze what is shared on social media worldwide to better understand what people with diabetes and diabetes distress are experiencing [28,29]. The social network "Twitter" is a popular data resource among diabetes researchers owing to its public character and its active online diabetes community compared to other social media [30,31]. Recent studies suggest an overrepresentation of people with type 1 diabetes compared to those with type 2 diabetes who are active on Twitter [9,31].

Methods

Overview

On the basis of diabetes-related tweets, we first preprocessed tweets to only focus on personal, nonjoke, and emotional content. Second, after this preprocessing step, we split tweets into sentences for our analyses, as we aimed to identify the cause-effect relationships between events within a sentence (sentence level) and not across multiple sentences (tweet level). This also simplifies model training and helps with easier learning. Third, we identified sentences in which causal information (opinion, observation, etc) is communicated. In the fourth step, causes and their corresponding effects were extracted. Lastly, those cause-effect pairs were aggregated, described, and visualized. The entire workflow is illustrated in Figure 1.

Figure 1. Workflow. The steps shown in green include machine learning methods. CRF: conditional random field.



Data Collection and Ethical Considerations

Via Twitter's streaming application programming interface, 32 million diabetes-related tweets in English were collected between April 2017 and January 2021 based on a list of diabetes-related keywords such as *diabetes*, *hypoglycemia*, *hyperglycemia*, and *insulin* from all over the world (see Multimedia Appendix 1 for the full list of keywords used). This is an extended data set of the one used in earlier works [9]. All data collected in this study were publicly posted on Twitter. Therefore, according to the privacy policy of Twitter, users agree to have this information available to the general public [30].

Data Preprocessing

Tweets are noisy and unstructured. They contain many misspelled or nonstandard English words. To reduce noise in the data set, we applied a preprocessing pipeline similar to that in earlier works, the details of which are summarized in Figure 1 [9]. First, retweets and duplicates were removed to obtain a database with 7.7 million unique tweets. Second, we determined only tweets with *personal* content where feelings, emotions, and opinions could be shared by people with or talking about diabetes and excluded *institutional* tweets referring to commercial, news, or health information. To identify *personal* content in tweets, we leveraged the transfer learning paradigm and fine-tuned the already pretrained transformer-based language model *BERTweet*, which was pretrained on 850 million English tweets (16 billion word tokens ~ 80 GB) [25,32]. To use the model and fine-tune it for a binary sentence classification, a linear layer was added on top of the last transformer layer of the *BERTweet* model by using the *transformers* package of HuggingFace [33]. The model was then fine-tuned with an extended data set of one used in earlier works, leading to a total of 4303 tweets (1539 *personal* and 2764 *institutional*) to account for a possible temporal divergence of the way people tweet [9]. The model performance to identify tweets with personal content had accuracy of 91.2%, precision of 86.2%, recall of 90.9%, and F1 score of 88.5%. The trained model was then applied to all unique tweets, resulting in a total of 2.5 million tweets with personal content. Moreover, jokes around diabetes are common on Twitter and were considered out of scope for this study as well. Similar to the *personal* content classifier, *BERTweet* was fine-tuned to detect if a tweet is a joke. For this purpose, a joke tweet data set from earlier works was extended to 1648 tweets (486 jokes, 1162 nonjokes) [9]. The performance to identify if a tweet is a joke had accuracy of 90.4%, precision of 78.5%, recall of 90.8%, and F1 score of 84.2%. Applying the joke classifier on all tweets with personal content led to a data set of 1.8 million personal nonjoke tweets.

A particular focus of this study was on studying diabetes distress and thus, the psychological factors and emotions. To capture these factors in tweets, only tweets containing an emotional element such as emojis/emoticons or emotional words were kept. Emotional words were identified based on a combination of the psychologue Parrot's hierarchical classification of emotions with the 6 primary emotions (*joy*, *love*, *surprise*, *sadness*, *anger*, *fear*) and emotional words present in common questionnaires to study diabetes distress such as the Problem Areas in Diabetes scale and Diabetes Distress Scale [34-36]. This led to 562,013 tweets containing personal, nonjoke, and emotional content. More details on the preprocessing pipeline are summarized in Multimedia Appendix 2 [9,25,32-40].

Data Annotation

In order to identify causal sentences and *cause-effect* association, 5000 randomly chosen diabetes-related tweets were selected, preprocessed, split into sentences, and then manually labeled. We did not restrict ourselves to a specific area of diabetes-related causal relationships, and we included potentially all types. Table 1 illustrates some example sentences. Only causal relationships related to diabetes were labeled as positive samples, whereas non–diabetes-related or unclear cause-effect relationships were labeled as negative samples. For a more detailed explanation on the annotation, please refer to our annotation guidelines in Multimedia Appendix 3.

Table 1. Sample sentences in different label scenarios. The examples are fictive to ensure privacy.

Sentences	Cause	Effect	Causal association	Explanation
Diabetes causes me to have mood swings	Diabetes	mood swings	1	Possible causal association
I just want to eat, I hate #diabetes	#diabetes	hate	1	Possible causal association related to diabetes distress
Scary, have a diabetic daughter but I read thousands of people a year die in the United Kingdom just from flu so why panic over corona.	—[a]	—	0	Nondiabetes or diabetes distress–related relationship. “Flu” is not diabetes-related
Had two strokes and recover now and also have high blood pressure and diabetes.	—	—	0	Unclear cause-effect relationship. Not clear if “high blood pressure” or “diabetes” caused the stroke
Not sure if I've been up since 3:30 to watch Titanic or because of my anxiety over my glucose test is what keeps me up	glucose test	anxiety	1	Chaining cause-effect relationship (A->B->C) Event A: glucose test Event B: anxiety Event C: been up since 3:30 => label the relationship which is closest to our study objective: diabetes and diabetes distress
My 14-year-old daughter is type 1 = malfunctioning pancreas, meaning not enough insulin being made to regulate	type 1	malfunctioning pancreas; not enough insulin	1	Negation in a cause/effect is considered being part of the cause/effect as it does not alter the meaning
It is not true to think that insulin makes you feel so bad	insulin	feel so bad	0	Negation is not part of cause/effect and alters the meaning

[a]Not available.

Labeling cause-effect pairs is a complex task. To verify the reliability of the labeling, 2 authors labeled 500 sentences independently and we calculated Cohen κ score, a statistical measure expressing the level of agreement between 2 annotators [41]. We obtained a score of 0.83, which is interpreted as an *almost perfect* agreement according to Altman [42] and Landis and Koch [43]. Disagreements were discussed between 2 authors, and 1 author labelled the other samples, resulting in 8235 labelled sentences (7218 noncausal sentences and 1017 causal sentences) from 5000 tweets.

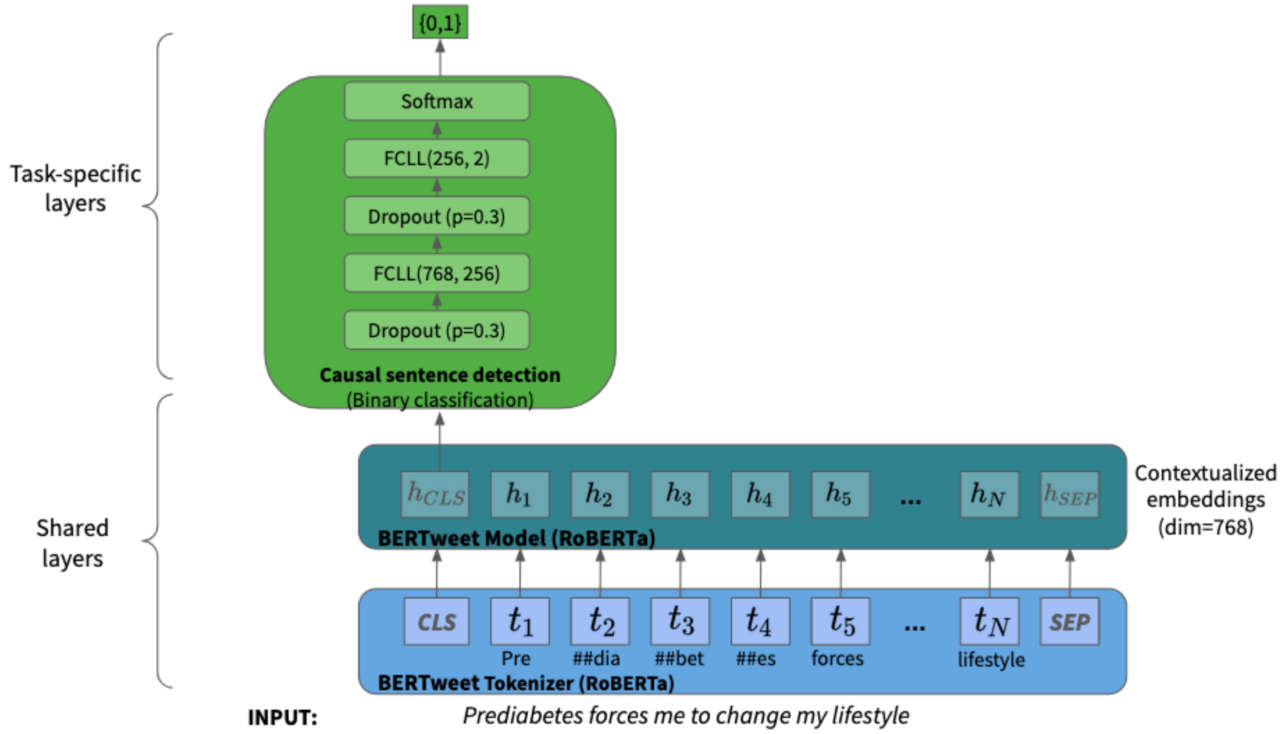
Models

The first model was trained to predict if a sentence contains a potential cause-effect association (causal sentence), and the second model extracted the specific cause and the associated effect from the causal sentence. Thus, the first model acts like a barrier and filters noncausal sentences out. These sentences may have either a cause, an effect, none of them, but not both. To simplify the model training, we hypothesized that cause-effect pairs only occur in the same sentence and we removed all sentences with less than 6 words owing to a lack of context. For this reason, we operated on a sentence level and not at the tweet level. Additional challenges in our setting were that *causes* and *effects* could be multiword entities and the language used on Twitter is nonstandard with frequent slang and misspelled words.

Causal Sentence Detection

The identification of causal sentences is a binary classification task. The pretrained language model *BERTweet* served as a foundation for the model architecture capable of handling the nonstandard nature of Twitter data [32]. A feed-forward network is built on top of the *BERTweet* [32] architecture consisting of 2 fully connected layers with dropout layers with a probability of 0.3, finalized by a softmax layer, which translates the model predictions into probabilities (Figure 2). To adjust for the class imbalance in the labeled data, class weights were included as parameters in the categorical cross-entropy loss function to penalize mispredictions for causal sentences strongly. Initially, labelled data were stratified, and 10% of it was kept as test set. The remaining 90% of the samples were further separated into training and validation sets with 80:20 split.

Figure 2. Model architecture for causal sentence detection. FCLL: fully connected linear layer; p: probability of an element to be zeroed.

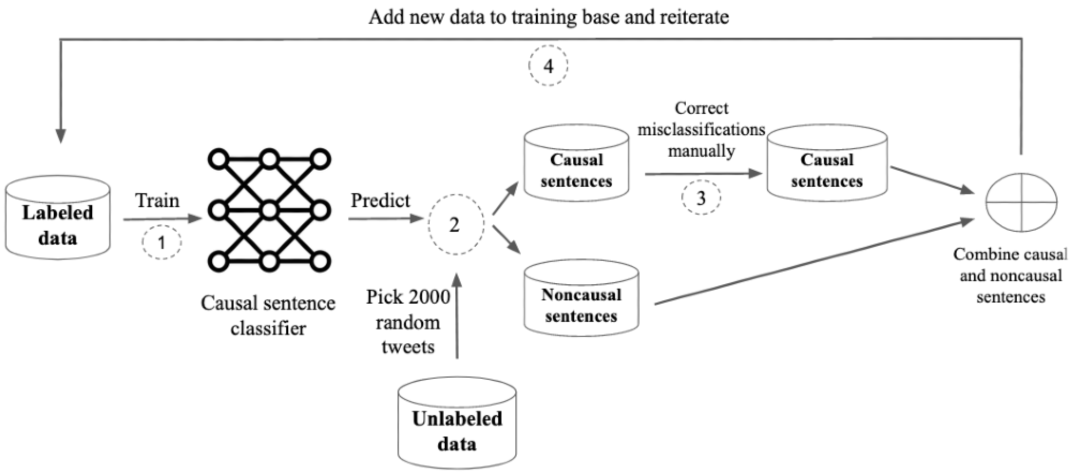


Data Augmentation Through Active Learning

Data imbalance on the one hand and the limited number of positive training examples for each cause-effect pair on the other hand (as causes and effects could potentially be related to any concept in the diabetes domain) drove us to adopt an active learning approach to increase the training data. Active learning is a sample selection approach aiming to minimize the annotation cost while maximizing the performance of machine learning–based models [44]. It has been widely applied on textual data [45,46]. The training data were increased in several iterations, as illustrated in Figure 3.

The first iteration started by training the causal sentence classifier on sentences from the 5000 tweets. The trained classifier was then applied on 2000 randomly selected unlabeled tweets, which were preprocessed and split into sentences, resulting in a set of causal sentences and a set of noncausal sentences. The sentences predicted as causal sentences were examined manually, and possible misclassifications were corrected to ensure clean positive training samples. The noncausal sentence set remained untouched. As a consequence, potential misclassifications remained in the noncausal sentence set, which should then be considered noisy. Both the causal and noncausal sentence set were then combined and added as new training data to the already labeled data, leading to an updated training set of 7000 tweets. This process was iterated 4 times and allowed us to augment the labelled data much faster and more efficiently than that without active learning, as it enables us to focus on the few positive samples. The final training set was used to train the classification model and the cause-effect extraction model.

Figure 3. Active learning loop to augment the training set in a time-efficient fashion.



Cause-Effect Pairs

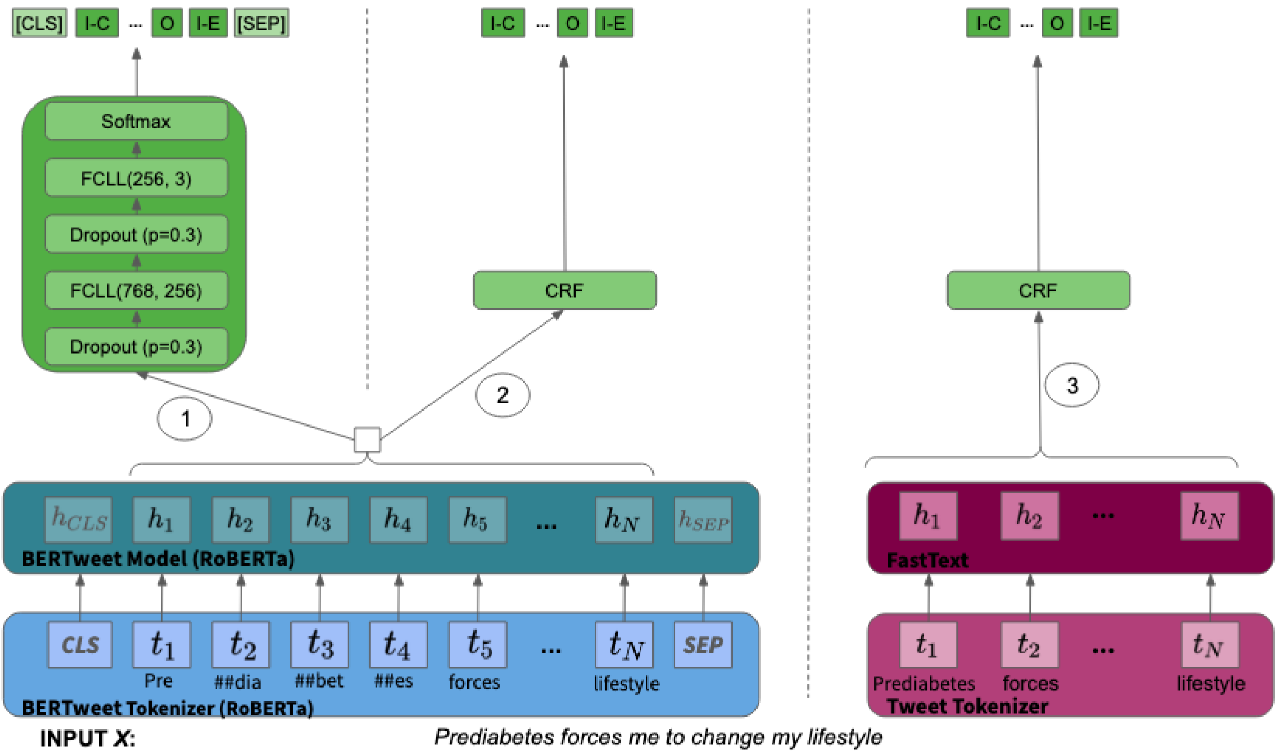
After having trained the causal sentence classifier to detect sentences with causal information, we identified the specific cause-effect pairs in the causal sentences. The identification of cause-effect pairs was casted as an event extraction or named-entity recognition task, that is, assigning a label cause or effect to a sequence of words. The manually labeled causes and effects were encoded in an IO tagging format based on the common tagging format BIO (Beginning, Inside, Outside), introduced by Ramshaw and Marcus [47]. Here, "I-C" denotes inside the cause and "I-E" inside the effect. Those 2 tags were completed by the outside tag "O," symbolizing that the word is neither cause nor effect. The IO tagging scheme for the example sentence with cause "prediabetes" and effect "change my lifestyle" is summarized:

Sentence: Prediabetes, forces, me, to, change, my, lifestyle

IO tags: I-C, O, O, O, I-E, I-E, I-E

Note that a word can be both cause or effect depending on the context. For instance "prediabetes" in "Prediabetes forces me to change my lifestyle" takes the role of a cause, whereas in "Limited exercising may lead to prediabetes," it is a possible effect. IO tagging was preferred over BIO tagging to simplify the model learning by reducing the number of class from 5 to 3. Moreover, the task is complex and considered open domain, as causes and effects are not restricted to 1 specific topic but can be related to any concept in our target domain (diabetes). As a consequence, the creation of a representative training set is challenging, as most cause-effect pairs occur rarely. This complexity drove us to test several model architectures; refer to Figure 4 for an overview.

Figure 4. Model architectures of cause-effect identification. CRF: conditional random field; FCLL: fully connected linear layer; p: probability of an element to be zeroed.



1. BERT_FFL: Pretrained BERTweet language model and on top, 2 feed forward layers with a dropout of 0.3, followed by a softmax layer. For the model training, the cross-entropy loss function is selected and weighted by the class weights to penalize mispredictions for causes and effects stronger.
2. WE_BERT_CRF: Single conditional random field (CRF) layer with BERTweet embeddings as features augmented by discrete features such as if the word is lowercase, digit, or the word length. CRFs are a standard statistical sequential classification method to identify entities in a text [48]. The CRF function is implemented with the python package sklearn-crfsuite [49] based on CRFsuite [50]. As parameters for the CRF function, the default algorithm "Gradient descent using the Limited Memory Broyden-Fletcher-GoldfarbShanno method" was chosen, and the coefficient for L1 and L2 regularization was 0.1.
3. FastText_CRF: Similar to WE_BERT_CRF, with the difference that BERTweet embeddings were replaced by FastText embeddings in the feature vector for each word. FastText vectors trained on similar diabetes-related tweets, which were well adapted to our use case [9].

Clustering of Causes and Effects

A large part of *causes* and *effects* can be regrouped into similar concepts (clusters) to facilitate analyses and allow effective network analyses. We chose a semisupervised, time-efficient approach in which 1000 *causes* and 1000 *effects* were randomly chosen and 2 researchers manually grouped these into clusters such as "diabetes," "death," "family," and "fear," hereinafter referred to as "parent clusters" to simplify understanding. The remaining *causes* and *effects* were then automatically compared to each element of all the clusters based on *FastText* vectors and cosine similarity and associated with the cluster containing the most similar element. Experimentally, a similarity threshold of 0.55 was determined; if a cause/effect had a similarity smaller than this threshold for all elements, a new cluster was created for this cause/effect. These clusters were also visualized in an interactive cause-effect network, developed in D3, to enable further exploration of the cause-effect association about diabetes distress communication in social media. Python (version 3.8.8) and the deep learning framework PyTorch (version 1.8.1) were used to implement the abovementioned methods. The algorithms are open sourced under [51].

Results

The following results were obtained from 482,583 sentences, which were obtained from splitting the 562,013 personal, emotional, and nonjoke tweets into sentences, excluding questions and including only sentences with more than 5 words.

Model Training and Performance

Causal Sentences

Hyperparameters for the model training were optimized, and the best model was trained with an Adam optimizer with a learning rate of 1e-3 among [1e-2, 1e-3, 1e-4] and a scheduler with linearly decreasing learning rate with 0 warmup steps. The optimal batch size was obtained for 16 among [8,16,32], and we trained for 35 epochs with early stopping. The performances to detect causal sentences for the imbalanced data set are illustrated in Table 2 for each round of the active learning loop, with each round having been trained on more data. The highest accuracy was reached in round 4 with 71%. We applied the model of round 4 on all the remaining tweets, as it was trained on the largest training data set, including difficult causal examples missed by earlier models and is thus better at identifying complex causal sentences. The active learning strategy led us to increase the training data much quicker than that without active learning and without loss in performance. This led to a clean database of 265,328 causal sentences with the most noisy sentences removed.

Table 2. Performance measures (macro) for each round of more training data.

Round	Sentences in training set (n)	Sentences in test set (n)	Accuracy (%)	Precision (%)	Recall (%)
0	6024	837	64.5	58.0	67.4
1	7536	1047	67.7	61.2	71.6
2	8804	1223	67.7	60.3	66.3
3	10,284	1429	65.4	60.0	68.8
4	11,861	1648	71.0	61.0	67.8

Cause and Effect Detection

After having identified the causal sentences, the cause-effect models were trained to extract the specific cause-effect pairs. The active learning strategy led to an extended data set of 2118 causal sentences, that is, containing both cause and effect, of which 10% were used as a test set while the remaining 90% were further used to create a training and validation set with an 80:20 split. The performances of the different cause-effect models are listed in Table 3. The best performing model was the CRF model with BERT-embedding features (WE_BERT_CRF) with a precision, recall, and F1 score of 0.68. Surprisingly, it outperforms fine-tuning a BERT model, which is considered the gold standard of current named-entity recognition tasks. A potential explanation for this is that BERT-based models make local decisions at every point of the sequence taking the neighboring words into account before its decision. In a situation like ours, with strong uncertainty on all elements, owing to the complexity of the task, a single CRF layer model leveraging BERT features, making global decisions using the local context of each word, maximizes the probability of the whole sequence of the decision better. Moreover, the CRF model with simpler FastText models achieved strong results as well with one reason being probably that the word embeddings were specifically trained on this diabetes corpus.

Consequently, the WE_BERT_CRF model was applied on all causal sentences leading to a data set of 96,676 sentences with the *cause* and associated *effect* predicted.

Table 3. Performance measures for each of the 4 architectures.

Models	Precision	Recall	F1 score
BERT_FFL			
I-C	0.48	0.46	0.47
I-E	0.20	0.48	0.29
O	0.91	0.77	0.83
macro	0.53	0.57	0.53
WE_BERT_CRF			
I-C	0.63	0.61	0.62
I-E	0.49	0.49	0.49
O	0.93	0.93	0.93
macro	0.68	0.68	0.68
FastText_CRF			
I-C	0.59	0.57	0.58
I-E	0.45	0.38	0.41
O	0.92	0.94	0.93
macro	0.65	0.63	0.64

Cause-Effect Description

The semisupervised clustering led to 1751 clusters. To remove noisy clusters through potential misclassifications, only clusters with a minimal number of 10 cause/effect occurrences were considered for the following analyses, resulting in 763 clusters. Note that the order of documents might affect the results, as different clusters might have been created. Please refer to Multimedia Appendix 4 for an overview over the 100 largest clusters (automatically added clusters have "other" as "parent cluster").

Table 4 provides an overview over the largest clusters, containing either cause or effect. Table 5 provides the most frequent cause-effect associations, excluding the largest cluster "diabetes," as it will be studied separately. The cluster "diabetes" is the largest one with 66,775 occurrences of "diabetes" as either cause or effect (eg, diabetes, #diabetes, diabetes mellitus) followed by "death" with 16,989 (eg, passed away, killed, died, suicide) and "insulin" (eg, insulin, insulin hormone) with 14,148 occurrences. From the 30 largest clusters, 6 refer to nutrition, 4 to diabetes, and 3 to each of insulin, emotions, and the health care system. The most frequent cause-effect is "unable to afford insulin," which causes "death" expressed in 1246 cases, followed by "insulin" causing "death" with 1156 cases and "type 1 diabetes" causing "fear" with 1054 cases.

The largest cluster "diabetes" mainly occurs as a cause and its 10 most frequent effects are death (n=7446), fear (n=4836), sick (n=2799), neuropathy (n=2477), hypoglycemia (n=2062), anger (n=1908), suffer (n=1808), insulin (n=1605), overweight (n=1506), and reduce weight (n=1487). From the 30 most numerous effects for "diabetes," 6 were related to "nutrition" and 5 to "complications and comorbidities" and 3 to each of "diabetes distress," "emotions," and "health care system."

The interactive visualization in D3 with filter options is published in [52]. Figure 5 provides an example graph of this visualization showing only cause-effect relationships with at least 250 occurrences to ensure readability. It is striking that "death" seems to play such a central role as *effect* with various causes (unable to afford insulin, rationing insulin, finance, insulin, type 1 diabetes, overweight) pointing at it. Other central nodes are type 1 diabetes acting as cause for insulin pump, insulin, hypoglycemia (hypo), sickness, finance, and anger, and fear emotions, where the latter has the strongest association, or the node "insulin" mostly relating as cause for sickness, medication, finance, death, or hypoglycemia and fear and anger.

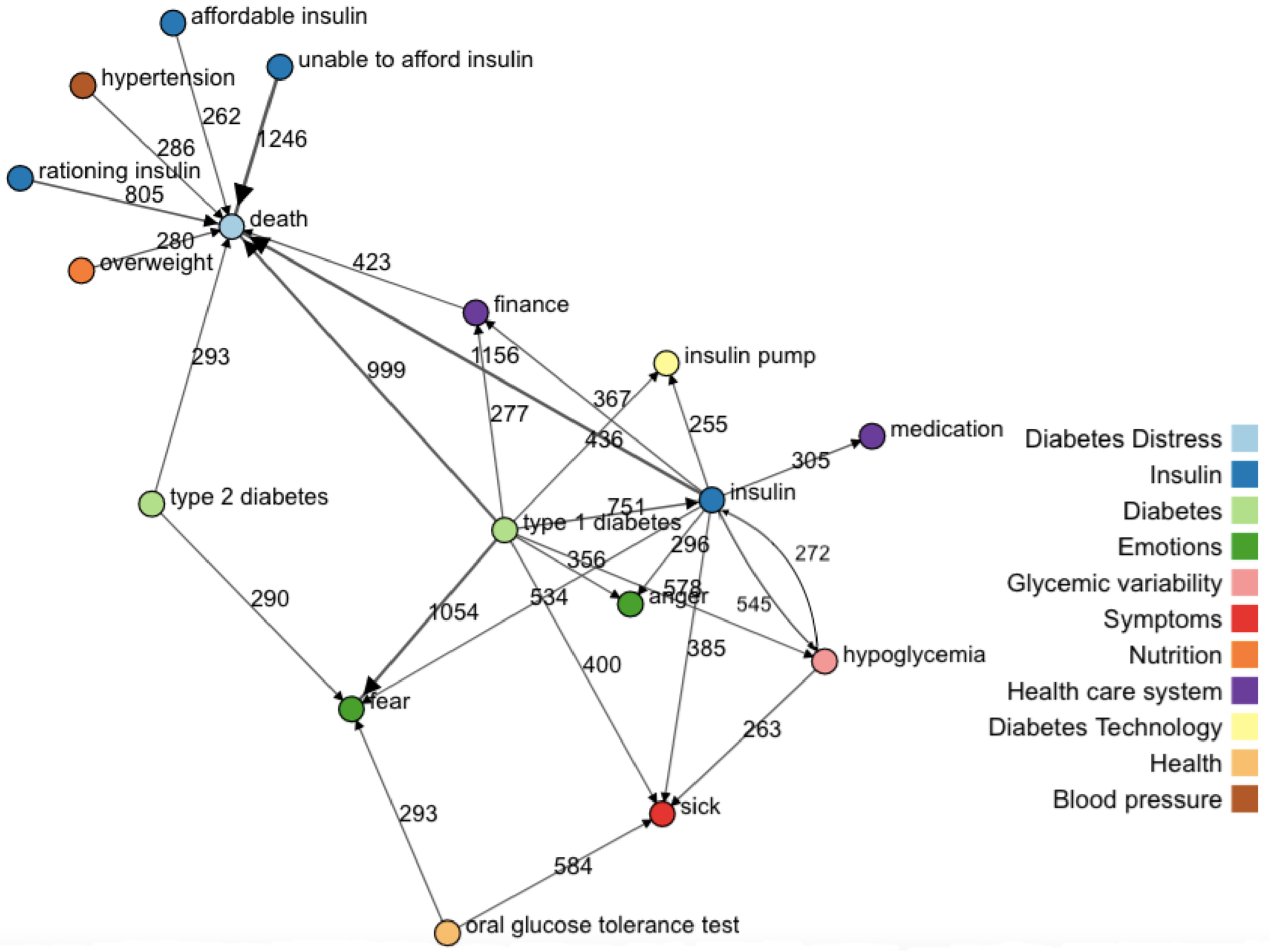
Table 4. The most frequent clusters (causes and effects) with the number of occurrences.

Parent cluster	Cluster	Value (n)
Diabetes	diabetes	66,775
Death	death	16,989
Insulin	insulin	14,148
Diabetes	type 1 diabetes	11,693
Emotions	fear	10,160
Glycemic variability	hypoglycemia	9547
Symptoms	sick	6549
Nutrition	overweight	5186
Diabetes	type 2 diabetes	4909
Complications and comorbidities	neuropathy	4481
Health care system	medication	4389
Diabetes Technology	insulin pump	4307
Nutrition	nutrition	4230
Emotions	anger	4149
Health	oral glucose tolerance test	4053
Blood pressure	hypertension	3782
Health care system	finance	3767
Nutrition	reduce weight	3589
Insulin	unable to afford insulin	3381
Nutrition	diet	3325
Emotions	sadness	3153
Glycemic variability	hyperglycemia	3144
Diabetes	suffer	3132
Diabetes Distress	depression	2810
Health care system	hospital	2721
Diabetes Distress	stress	2681
Nutrition	sugar	2369
Nutrition	fasting	2363
Insulin	rationing insulin	2244
Health	gestational diabetes	2076

Table 5. The most frequent cause-effect relationships excluding the cluster “diabetes” with the number of occurrences.

Cause	Effect	Value (n)
unable to afford insulin	death	1246
insulin	death	1156
type 1 diabetes	fear	1054
type 1 diabetes	death	999
rationing insulin	death	805
type 1 diabetes	insulin	751
oral glucose tolerance test	sick	584
type 1 diabetes	hypoglycemia	578
insulin	hypo	545
insulin	fear	534
type 1 diabetes	insulin pump	436
finance	death	423
type 1 diabetes	sick	400
insulin	sick	385
insulin	finance	367
type 1 diabetes	anger	356
insulin	medication	305
insulin	anger	296
oral glucose tolerance test	fear	293
type 2 diabetes	death	293
type 2 diabetes	fear	290
hypertension	death	286
overweight	death	280
type 1 diabetes	finance	277
hypoglycemia	insulin	272
hypoglycemia	sick	263
affordable insulin	death	262
insulin	insulin pump	255
complications	death	248
insulin	sadness	240

Figure 5. Cause-effect network with a minimum number of associations (edges) of 250. Accessible in [52].



Discussion

Principal Findings

Our findings suggest that it is feasible to extract both explicit and implicit causes and associated effects from diabetes-related Twitter data. We demonstrated that by adopting the transfer learning paradigm and fine-tuning a pretrained language model, we were able to detect causal sentences. Moreover, we have shown that simply fine-tuning a BERT-based model does not always outperform more traditional methods such as relying on CRFs in the case of the cause-effect pair detection. The precision, recall, and F1 scores, given the challenging task and the imbalanced data set, were satisfying. The semisupervised clustering and interactive visualization enabled us to identify "diabetes" as the largest cluster acting mainly as the cause for "death" and "fear." Besides, a central cluster was detected in "death" acting as an effect for various causes related to insulin pricing—a link that was already detected in earlier works [9]. From a patient's perspective, we were able to show that their main fear is insulin pricing, which is expressed in the most frequent cause-effect relationship "unable to afford insulin" causing "death" or "rationing insulin" causing "death." As the main diabetes distress–related causes, we identified fear of hypoglycemia, insulin, hypertension, or the oral glucose tolerance test.

Comparison With Previous Works

Several former works have addressed causality on Twitter data. Doan et al [14] focused on 3 health-related concepts, namely, stress, insomnia, and headache as effects and identified causes by using manually crafted patterns and rules. However, they only focused on explicit causality and excluded causes and effects encoded in hashtags and synonymous expressions [14]. On the contrary, we tackled both explicit and implicit causality, including causes and effects in hashtags and exploiting synonymous expressions through the use of word embeddings. Kayesh et al [16] proposed an innovative approach, a novel technique based on neural networks, which uses common sense background knowledge to enhance the feature set, but they focused on the simplified version of explicit causality in tweets. Bollegala et al [53] developed a causality-sensitive approach for detecting adverse drug reactions from social media by using lexical patterns and thereby aiming at explicit causality. Dasgupta et al [54] proposed one of the few deep learning approaches due to the unavailability of appropriate training data, leveraging a recursive neural network architecture to detect cause-effect relations from text, but they also only targeted explicit causality. A BERT-based approach tackling both explicit and implicit causality is provided by Khetan et al [23] who used already existing labeled corpora not based on social media data. Recently, they further extended their work of explicit and implicit causality understanding in single and multiple sentences but in clinical notes [55]. To the best of our knowledge, this is the first paper investigating both explicit and implicit cause-effect relationships on diabetes-related Twitter data.

Strengths and Limitations

This study demonstrates various strengths. First, by leveraging powerful language models, we were able to identify a large number of tweets containing *cause-effect* relationships, which enabled us to the detect cause-effect associations in 20% (96,676/482,583) of the sentences, contrary to other approaches that were able to identify causality in less than 2% of tweets [14]. Second, contrary to most previous work, we tackled both explicit and implicit *causal relationships*, an additional explanation for the higher number of *cause-effect* associations we obtained, compared to other studies focusing only on explicit associations [14]. Third, relying fully on automatic machine learning algorithms avoided us from defining manually crafted patterns to detect causal associations. Fourth, operating on social media data that are expressed spontaneously and in real time offers the opportunity to gain knowledge from an alternative data source and, in particular, from a patient’s perspective, which might complement traditional epidemiological data sources. Lastly, the data-driven approach to identify cause-effect relationships, as reported from Twitter users, can be used in the next step to generate new hypotheses that can be tested in a more clinical setting, for example, in a clinical trial.

A strong limitation is that *cause-effect* relations are expressed in tweets and this cannot be used for causal inference as the Twitter data source is uncertain and the information shared can be an opinion or an observation. Another shortcoming is that the performance of our algorithms to detect *cause-effect* pairs is not perfect. However, the overall process and the vast amount of data minimize this issue. The lack of recall is counterbalanced by the sheer amount of data, and the lack of precision is counterbalanced by the clustering approach in which nonfrequent causes or effects are discarded [56]. Labeling causes and effects in a data set is a highly complicated task, and we would like to emphasize that mislabeling in the data set may occur. Here, the actual prevalence of causal sentences is lower, as we wanted to catch as many causal sentences as possible, which led to also having captured some noncausal sentences. Enhancing data quality certainly is a strong point to address to further improve performance. The causal association structures learnt by the model from the training set might not generalize completely when applied on the large amount of Twitter data. Besides, the active learning strategy certainly added noise to the model, as only positive samples were corrected, which could be improved in future investigations. Moreover, we would like to highlight that the diabetes-related information shared on Twitter may not be representative for all people with diabetes. For instance, we observed a bigger cluster of causes/effects related to type 1 diabetes compared to that related to type 2 diabetes, which is contrary to that in the real world [57]. A potential explanation for that is the age distribution of Twitter users [58]. However, owing to the large number of tweets analyzed, a significant variability in the tweets could be observed.

Conclusion

In this work, we developed an innovative methodology to identify possible cause-effect relationships among diabetes-related tweets. This task was challenging owing to addressing both explicit and implicit causality, multiword entities, the fact that a word could be both cause or effect, the open domain of causes and effects, the biases occurring during labeling of causality, and the relatively small data set for this complex task. We overcame these challenges by augmenting the small data set via an active learning loop. The feasibility of our approach was demonstrated using modern BERT-based architectures in the preprocessing and causal sentence detection. A combination of BERT features and CRF layer were leveraged to extract causes and effects in diabetes-related tweets, which were then aggregated to clusters in a semisupervised approach. The visualization of the cause-effect network based on Twitter data can deepen our understanding of diabetes, in a way of directly capturing patient-reported outcomes from a causal perspective. The fear of death owing to the inability to afford insulin was the main concern expressed.

Acknowledgments

This work was supported by the MSDAVENIR Foundation, the French Speaking Diabetes Society, and the Luxembourg Institute of Health. These study sponsors had no role in the design or the interpretation of the results of this study. AA, FO, and TC are supported by Epiconcept Company. Epiconcept was involved in the data collection and writing of the report. No study sponsor influenced the decision to submit the paper for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of diabetes-related keywords for the Twitter application programming interface tweet extraction.
[PDF File (Adobe PDF File), 47 KB - medinform_v10i7e37201_app1.pdf]

Multimedia Appendix 2
Preprocessing pipeline.
[PDF File (Adobe PDF File), 64 KB - medinform_v10i7e37201_app2.pdf]

Multimedia Appendix 3
Annotation guidelines.
[PDF File (Adobe PDF File), 120 KB - medinform_v10i7e37201_app3.pdf]

Multimedia Appendix 4
Most frequent clusters.
[PDF File (Adobe PDF File), 75 KB - medinform_v10i7e37201_app4.pdf]

References

1. Fisher L, Hessler DM, Polonsky WH, Mullan J. When is diabetes distress clinically meaningful?: establishing cut points for the Diabetes Distress Scale. Diabetes Care 2012 Feb;35(2):259-264. [doi: 10.2337/dc11-1572] [Medline: 22228744]
2. Fisher L, Polonsky WH, Hessler DM, Masharani U, Blumer I, Peters AL, et al. Understanding the sources of diabetes distress in adults with type 1 diabetes. J Diabetes Complications 2015;29(4):572-577. [doi: 10.1016/j.jdiacomp.2015.01.012] [Medline: 25765489]
3. Coccaro E, Lazarus S, Joseph J, Wyne K, Drossos T, Phillipson L, et al. Emotional Regulation and Diabetes Distress in Adults With Type 1 and Type 2 Diabetes. Diabetes Care 2021 Jan;44(1):20-25. [doi: 10.2337/dc20-1059] [Medline: 33444157]
4. Carper MM, Traeger L, Gonzalez JS, Wexler DJ, Psaros C, Safren SA. The differential associations of depression and diabetes distress with quality of life domains in type 2 diabetes. J Behav Med 2014 Jun;37(3):501-510. [doi: 10.1007/s10865-013-9505-x] [Medline: 23515932]
5. Cummings DM, Lutes L, Littlewood K, DiNatale E, Hambidge B, Schulman K, et al. Regimen-Related Distress, Medication Adherence, and Glycemic Control in Rural African American Women With Type 2 Diabetes Mellitus. Ann Pharmacother 2014 Aug;48(8):970-977. [doi: 10.1177/1060028014536532] [Medline: 24904183]
6. Fisher L, Mullan J, Skaff M, Glasgow R, Arean P, Hessler D. Predicting diabetes distress in patients with Type 2 diabetes: a longitudinal study. Diabet Med 2009 Jun;26(6):622-627. [doi: 10.1111/j.1464-5491.2009.02730.x] [Medline: 19538238]
7. Pandit AU, Bailey SC, Curtis LM, Seligman HK, Davis TC, Parker RM, et al. Disease-related distress, self-care and clinical outcomes among low-income patients with diabetes. J Epidemiol Community Health 2014 Jun;68(6):557-564. [doi: 10.1136/jech-2013-203063] [Medline: 24489044]
8. Schmidt CB, van Loon BJP, Vergouwen ACM, Snoek FJ, Honig A. Systematic review and meta-analysis of psychological interventions in people with diabetes and elevated diabetes-distress. Diabet. Med 2018 Jun 30;35(9):1157-1172. [doi: 10.1111/dme.13709] [Medline: 29896760]
9. Ahne A, Orchard F, Tannier X, Perchoux C, Balkau B, Pagoto S, et al. Insulin pricing and other major diabetes-related concerns in the USA: a study of 46 407 tweets between 2017 and 2019. BMJ Open Diabetes Res Care 2020 Jun;8(1):e001190 [FREE Full text] [doi: 10.1136/bmjdrc-2020-001190] [Medline: 32503810]
10. Balkhi AM, Reid AM, McNamara JP, Geffken GR. The diabetes online community: the importance of forum use in parents of children with type 1 diabetes. Pediatr Diabetes 2014 Sep 25;15(6):408-415. [doi: 10.1111/pedi.12110] [Medline: 24372986]
11. Yang J, Han S, Poon J. A survey on extraction of causal relations from natural language text. ArXiv. 2021 Nov 01. URL: http://arxiv.org/abs/2101.06426 [accessed 2022-05-29]
12. Doan S, Ritchart A, Perry N, Chaparro JD, Conway M. How Do You #relax When You're #stressed? A Content Analysis and Infodemiology Study of Stress-Related Tweets. JMIR Public Health Surveill 2017 Jun 13;3(2):e35 [FREE Full text] [doi: 10.2196/publichealth.5939] [Medline: 28611016]
13. 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]
14. Doan S, Yang EW, Tilak SS, Li PW, Zisook DS, Torii M. Extracting health-related causality from twitter messages using natural language processing. BMC Med Inform Decis Mak 2019 Apr 04;19(Suppl 3):79 [FREE Full text] [doi: 10.1186/s12911-019-0785-0] [Medline: 30943954]
15. Khoo C, Chan S, Niu Y. Extracting causal knowledge from a medical database using graphical patterns. 2000 Presented at: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics; October; Hong Kong p. 336-343. [doi: 10.3115/1075218.1075261]
16. Kayesh H, Islam M, Wang J. On event causality detection in tweets. ArXiv. 2019. URL: http://arxiv.org/abs/1901.03526 [accessed 2022-01-31]
17. Khoo C, Chan S, Niu Y. The many facets of the cause-effect relation. In: The Semantics of Relationships. Dordrecht: Springer; 2002.
18. Chowdhury GG. Natural language processing. Ann Rev Info Sci Tech 2005 Jan 31;37(1):51-89. [doi: 10.1002/aris.1440370103]
19. El Naqa I, Murphy M. What is machine learning? In: Machine Learning in Radiation Oncology. Cham: Springer International Publishing; 2015.

20. Xu Y, Mou L, Li G, Chen Y, Peng H, Jin Z. Classifying relations via long short term memory networks along shortest dependency paths. 2015 Presented at: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; September; Lisbon p. 1785-1794. [doi: 10.18653/v1/d15-1206]
21. Wang L, Cao Z, Melo GD, Liu Z. Relation classification via multi-level attention CNNs. 2016 Presented at: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; August; Berlin p. 1298-1307. [doi: 10.18653/v1/P16-1123]
22. Ponti E, Korhonen A. Event-related features in feedforward neural networks contribute to identifying causal relations in discourse. In; 2017 Presented at: LSDSem 2017 - 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-Level Semantics, Proceedings of the Workshop; April; Valencia, Spain p. 25-30. [doi: https://doi.org/10.17863/CAM.9725]
23. Khetan V, Ramnani R, Anand M, Sengupta S, Fano A. Causal BERT: language models for causality detection between events expressed in text. In: Arai K, editor. Intelligent Computing. Cham: Springer International Publishing; Jul 2021:965-980.
24. Devlin J, Chang M, Lee K, Toutanova K. BERT: pretraining of deep bidirectional transformers for language understanding. ArXiv. 2018. URL: http://arxiv.org/abs/1810.04805 [accessed 2021-05-31]
25. Vaswani A, Shazeer N, Parmar N. Attention is all you need. ArXiv. 2017. URL: http://arxiv.org/abs/1706.03762 [accessed 2021-09-09]
26. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Inform 2012 Oct;45(5):885-892 [FREE Full text] [doi: 10.1016/j.jbi.2012.04.008] [Medline: 22554702]
27. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data 2016 May 28;3(1):1-40. [doi: 10.1186/s40537-016-0043-6]
28. Fagherazzi G. Challenges and perspectives for the future of diabetes epidemiology in the era of digital health and artificial intelligence. Diabetes Epidemiology and Management 2021 Jan;1:100004. [doi: 10.1016/j.deman.2021.100004]
29. Fagherazzi G, Ahne A, Guillot C, Riveline J, Bonnet F, Mebarki A, et al. Étude mondiale de la détresse liée au diabète : le potentiel du réseau social Twitter pour la recherche médicale. Revue d'Épidémiologie et de Santé Publique 2018 Jun;66:S197-S198. [doi: 10.1016/j.respe.2018.04.002]
30. Twitter privacy policy. Twitter. 2021. URL: https://twitter.com/en/privacy [accessed 2021-07-06]
31. Liu Y, Mei Q, Hanauer DA, Zheng K, Lee JM. Use of Social Media in the Diabetes Community: An Exploratory Analysis of Diabetes-Related Tweets. JMIR Diabetes 2016 Nov 07;1(2):e4 [FREE Full text] [doi: 10.2196/diabetes.6256] [Medline: 30291053]
32. Nguyen D, Vu T, Nguyen A. BERTweet: A pretrained language model for English tweets. ArXiv. 2020. URL: http://arxiv.org/abs/2005.10200 [accessed 2021-12-12]
33. Wolf T, Debut L, Sanh V. HuggingFace's transformers: state-of-the-art natural language processing. ArXiv. 2020. URL: http://arxiv.org/abs/1910.03771 [accessed 2021-12-12]
34. Parrott W. Emotions in Social Psychology: Essential Readings. Hove, East Sussex, United Kingdom: Psychology Press; 2001.
35. Polonsky W, Anderson B, Lohrer P, Welch G, Jacobson AM, Aponte JE, et al. Assessment of diabetes-related distress. Diabetes Care 1995 Jun;18(6):754-760. [doi: 10.2337/diacare.18.6.754] [Medline: 7555499]
36. Polonsky WH, Fisher L, Earles J, Dudl RJ, Lees J, Mullan J, et al. Assessing psychosocial distress in diabetes: development of the diabetes distress scale. Diabetes Care 2005 Mar;28(3):626-631. [doi: 10.2337/diacare.28.3.626] [Medline: 15735199]
37. Beguerisse-Díaz M, McLennan AK, Garduño-Hernández G, Barahona M, Ulijaszek SJ. The 'who' and 'what' of #diabetes on Twitter. Digit Health 2017;3:2055207616688841 [FREE Full text] [doi: 10.1177/2055207616688841] [Medline: 29942579]
38. Johnsen JAK, Eggesvik TB, Rørvik TH, Hanssen MW, Wynn R, Kummervold PE. Differences in Emotional and Pain-Related Language in Tweets About Dentists and Medical Doctors: Text Analysis of Twitter Content. JMIR Public Health Surveill 2019 Feb 06;5(1):e10432 [FREE Full text] [doi: 10.2196/10432] [Medline: 30724738]
39. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. Arxiv. 2019. URL: https://arxiv.org/abs/1907.11692 [accessed 2022-06-27] [doi: 10.48550/ARXIV.1907.11692]
40. Princeton University. About WordNet. 2010. URL: https://wordnet.princeton.edu/ [accessed 2019-04-10]
41. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 2016 Jul 02;20(1):37-46. [doi: 10.1177/001316446002000104]
42. Altman D. Practical Statistics for Medical Research. United Kingdom: Chapman & Hall; 1990.
43. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics 1977 Mar;33(1):159. [doi: 10.2307/2529310]
44. Settles B. Active Learning Literature Survey. 2010. URL: http://burrsettles.com/pub/settles.activelearning.pdf?source=post_page [accessed 2021-08-08]
45. Zhang Y, Lease M, Wallace B. Active discriminative text representation learning. ArXiv. 2016. URL: http://arxiv.org/abs/1606.04212 [accessed 2021-09-12]

46. Tong S, Koller D. Support vector machine active learning with applications to text classification. J Mach Learn Res 2001;2:45-66 [FREE Full text] [doi: 10.1162/153244302760185243]
47. Ramshaw L, Marcus M. Text chunking using transformation-based learning. ArXiv. 1999. URL: http://arxiv.org/abs/cmp-lg/9505040 [accessed 2021-09-07]
48. Lafferty J, Mccallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. University of Pennsylvania. 2001. URL: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers [accessed 2022-06-27]
49. Korobov M. sklearn-crfsuite. URL: https://sklearn-crfsuite.readthedocs.io/en/latest/index.html [accessed 2021-09-14]
50. Okazaki N. A fast implementation of conditional random fields (CRFs). CRFsuite. 2007. URL: http://www.chokkan.org/software/crfsuite/ [accessed 2021-10-11]
51. WDDS/Causal-associations-diabetes-twitter. GitHub. URL: https://github.com/WDDS/Causal-associations-diabetes-twitter/ [accessed 2022-06-27]
52. Cause and effect associations in diabetes-related tweets. Adahne. URL: https://observablehq.com/@adahne/cause-and-effect-associations-in-diabetes-related-tweets [accessed 2022-06-27]
53. Bollegala D, Maskell S, Sloane R, Hajne J, Pirmohamed M. Causality Patterns for Detecting Adverse Drug Reactions From Social Media: Text Mining Approach. JMIR Public Health Surveill 2018 May 09;4(2):e51 [FREE Full text] [doi: 10.2196/publichealth.8214] [Medline: 29743155]
54. Dasgupta T, Saha R, Dey L, Naskar A. Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks. 2018 Presented at: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue; April; Melbourne p. 306-316 URL: https://www.superlectures.com/sigdial2018/automatic-extraction-of-causal-relations-from-text-using-linguistically-informed-deep-neural-networks [doi: 10.18653/v1/w18-5035]
55. Khetan V, Rizvi M, Huber J, Bartusiak P, Sacaleanu B, Fano A. MIMICause: Defining, identifying and predicting types of causal relationships between biomedical concepts from clinical notes. ArXiv. URL: http://arxiv.org/abs/2110.07090 [accessed 2021-12-12]
56. Tannier X. NLP-driven Data Journalism: Time-Aware Mining and Visualization of International Alliances. 2016. URL: https://hal.archives-ouvertes.fr/hal-02407145/document [accessed 2022-06-27]
57. International Diabetes Federation Diabetes Atlas, 9th edn. 2019. URL: https://www.diabetesatlas.org [accessed 2021-12-09]
58. Percentage of US adults who use Twitter as of February 2021, by age group. Statista. 2021. URL: https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/ [accessed 2021-10-23]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BIO: Beginning, Inside, Outside
CRF: conditional random field

Edited by C Lovis; submitted 10.02.22; peer-reviewed by A Hulman, C Pollack; comments to author 21.04.22; revised version received 17.05.22; accepted 04.06.22; published 19.07.22.

Please cite as:
Ahne A, Khetan V, Tannier X, Rizvi MIH, Czernichow T, Orchard F, Bour C, Fano A, Fagherazzi G
Extraction of Explicit and Implicit Cause-Effect Relationships in Patient-Reported Diabetes-Related Tweets From 2017 to 2021: Deep Learning Approach
JMIR Med Inform 2022;10(7):e37201
URL: https://medinform.jmir.org/2022/7/e37201
doi:10.2196/37201
PMID:35852829

Original Paper

Accurate Forecasting of Emergency Department Arrivals With Internet Search Index and Machine Learning Models: Model Development and Performance Evaluation

Bi Fan[1], PhD; Jiaxuan Peng[2], MSc; Hainan Guo[1], PhD; Haobin Gu[3], MSc; Kangkang Xu[4], PhD; Tingting Wu[1], MSc

[1]College of Management, Institute of Business Analysis and Supply Chain Management, Shenzhen University, Shenzhen, China

[2]Faculty of Science, University of St Andrews, St Andrews, United Kingdom

[3]School of Management Science and Engineering, Dongbei University of Finance and Economics, Dalian, China

[4]School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou, China

Corresponding Author:
Jiaxuan Peng, MSc
Faculty of Science
University of St Andrews
St Andrews, Fife
Scotland
St Andrews, KY16 0US
United Kingdom
Phone: 44 07421733982
Email: jiaxuan_peng@163.com

Abstract

Background: Emergency department (ED) overcrowding is a concerning global health care issue, which is mainly caused by the uncertainty of patient arrivals, especially during the pandemic. Accurate forecasting of patient arrivals can allow health resource allocation in advance to reduce overcrowding. Currently, traditional data, such as historical patient visits, weather, holiday, and calendar, are primarily used to create forecasting models. However, data from an internet search engine (eg, Google) is less studied, although they can provide pivotal real-time surveillance information. The internet data can be employed to improve forecasting performance and provide early warning, especially during the epidemic. Moreover, possible nonlinearities between patient arrivals and these variables are often ignored.

Objective: This study aims to develop an intelligent forecasting system with machine learning models and internet search index to provide an accurate prediction of ED patient arrivals, to verify the effectiveness of the internet search index, and to explore whether nonlinear models can improve the forecasting accuracy.

Methods: Data on ED patient arrivals were collected from July 12, 2009, to June 27, 2010, the period of the 2009 H1N1 pandemic. These included 139,910 ED visits in our collaborative hospital, which is one of the biggest public hospitals in Hong Kong. Traditional data were also collected during the same period. The internet search index was generated from 268 search queries on Google to comprehensively capture the information about potential patients. The relationship between the index and patient arrivals was verified by Pearson correlation coefficient, Johansen cointegration, and Granger causality. Linear and nonlinear models were then developed with the internet search index to predict patient arrivals. The accuracy and robustness were also examined.

Results: All models could accurately predict patient arrivals. The causality test indicated internet search index as a strong predictor of ED patient arrivals. With the internet search index, the mean absolute percentage error (MAPE) and the root mean square error (RMSE) of the linear model reduced from 5.3% to 5.0% and from 24.44 to 23.18, respectively, whereas the MAPE and RMSE of the nonlinear model decreased even more, from 3.5% to 3% and from 16.72 to 14.55, respectively. Compared with each other, the experimental results revealed that the forecasting system with extreme learning machine, as well as the internet search index, had the best performance in both forecasting accuracy and robustness analysis.

Conclusions: The proposed forecasting system can make accurate, real-time prediction of ED patient arrivals. Compared with the static traditional variables, the internet search index significantly improves forecasting as a reliable predictor monitoring continuous behavior trend and sudden changes during the epidemic (P=.002). The nonlinear model performs better than the linear

counterparts by capturing the dynamic relationship between the index and patient arrivals. Thus, the system can facilitate staff planning and workflow monitoring.

(JMIR Med Inform 2022;10(7):e34504) doi:10.2196/34504

KEYWORDS

emergency department; internet search index; machine learning; nonlinear model; patient arrival forecasting

Introduction

Background

The emergency department (ED), which provides instant and efficient medical services for patients all day, is one of the most important parts in the health care system [1]. Unfortunately, the ED as the main entrance in modern hospitals is now under the threat of overcrowding, which can lead to serious negative consequences, such as treatment delays, increased patient mortality, and financial losses [2]. The common causes of overcrowding are inadequate resource allocation and increased demand for ED services, particularly during the epidemic period [3].

The management of patient flow is a challenge faced by many EDs. The ability to accurately forecast the demand for medical service in EDs has considerable implications for hospitals, as it can improve staff and equipment resource allocation. Considering the high cost of purchasing new medical resources in a short time, it is more reasonable to develop an accurate forecasting model of ED patient arrivals, which could enable better matching of current resources and ED visits. By forecasting the level of demand for ED care in advance, medical staff have the opportunity to prepare for this demand, which can improve the ED service throughput, avoid overcrowding, and ensure the safety of patients [4].

Prior Work

Previous studies mainly focused on the relationship between patient arrivals and the traditional variables, including the historical data of patient arrivals, calendar, weather, and holidays [4-12]. There have been many successful applications. However, the sudden and transient changes in people's behavior cannot be captured by the traditional variables. This information should be applied to predict ED visits before such changes are noticed in the ED [4,11,13]. Recently, there has been an increasing interest to apply internet data to predict the behaviors and intentions of people in many areas, such as tourist arrivals, product sales, stock returns, and unemployment rate [14-16]. In health care, the weekly information report from Google Trends can be used for weekly influenza epidemic detection [17,18]. Moreover, internet data have been shown to be useful for predicting disease trends [17-19]. However, in some scenarios, the reliability of Google Trend is of concern as it is vulnerable to the mass media and statistical anomalies [20,21]. Dugas et al [22] studied the association between influenza rates and crowding metrics using the Google Flu Trends. However, only few studies have been published regarding the potential of internet data to improve ED visit forecasting. Ekström et al [4] monitored the visits to a special, regional medical website to predict the daily ED attendance with linear regression. Combining calendar, weather, and autoregressive (AR) terms, the least absolute shrinkage and selection operator (LASSO) regression was applied to forecast ED patient arrivals [11]. Ho et al [13] predicted ED patient volume in the Singapore General Hospital using multiple regression and publicly available Google data [13]. Moreover, they even developed a software suite to enable data visualization and prediction of patient arrivals, which is convenient for hospital management. Although these methods work well in their scenarios, there remains room for further improvement of the ED forecasting, neither limited to a specific region nor relying on expert experience to collect internet information. Moreover, the aforementioned studies are mainly based on linear model, and the possible nonlinearity may be ignored. In our paper, the nonlinearity is among patient arrivals and all the independent variables (eg, calendar, holiday, weather, and internet search index) are considered. A general method is, however, needed to overcome the aforesaid limitations.

Objective

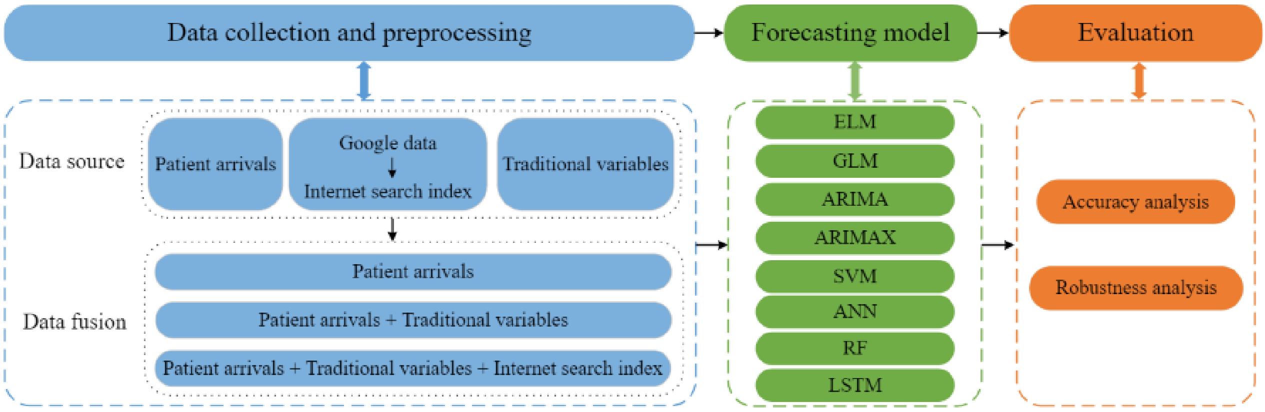
The objective of this study is to develop an intelligent forecasting system with a machine learning model and internet search index to provide accurate prediction of ED patient arrivals, to verify the effectiveness of the internet search index, and to explore whether nonlinear models can improve the forecasting accuracy. First, the internet search index was constructed from 266 search queries and verified as a novel variable by a systematic method. The data were generated from Google search queries, covering disease names, causes, symptoms, treatments, and others. The different types of information required by the patient, as reflected by the search query, might capture population-level interaction with events, such as infectious diseases, that traditional data sources alone may miss. The relationship between the internet search index and the ED visits was examined by Pearson correlation coefficient, Johansen cointegration, and Granger causality. Second, linear and nonlinear models were applied to predict ED patient arrivals with or without internet search index, respectively [4-12]. In addition, a nonlinear model, the extreme learning machine (ELM), was introduced because of its good generalization abilities and high prediction performance in flow prediction [23].

Methods

Overview

This study aimed to establish an intelligent system for predicting patient arrivals accurately and timely. The system consisted of 3 parts: data collection and processing, the establishment of forecasting model, and performance evaluation. In addition to the ED patient arrivals and traditional variables (weather, holidays, calendar), the internet search index, which extracted and integrated ED-related human behavior information scattered in Google search engines, was generated as a new variable. The correlation between the internet search index and patient arrivals was verified by Pearson correlation coefficient, Johansen cointegration, and Granger causality analysis. We then applied 8 forecasting models to predict ED patient arrivals, including ELM, generalized linear model (GLM), autoregressive integrated moving average model (ARIMA), ARIMA with explanatory variables (ARIMAX), support vector machine (SVM), artificial neural network (ANN), random forest (RF), and long short-term memory (LSTM) [24-33]. After that, their performances were evaluated in terms of accuracy and robustness analysis. The details of the intelligent system are shown in Figure 1.

Figure 1. A framework of the intelligent forecasting system with the internet search index. ANN: artificial neural network; ARIMA: autoregressive integrated moving average model; ARIMAX: ARIMA with explanatory variables; ELM: extreme learning machine; GLM: generalized linear model; LSTM: long short-term memory; RF: random forest; SVM: support vector machine.

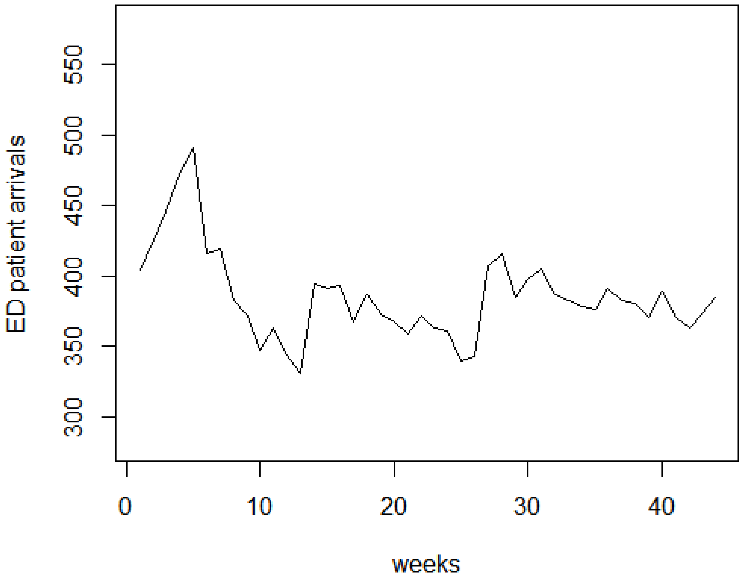


Data Collection and Processing

Data on ED Patient Arrivals

In Hong Kong, patients can directly visit public ED without an appointment and be reimbursed for most medical expenses, so the ED is usually overcrowded. About 65% of patients walk-in and half of them are semiurgent or nonurgent [34]. The Cooperative Hospital is one of the largest public hospitals in Hong Kong, which is also the first one to receive patients with COVID-19. It has more than 4000 medical staff members and 1700 beds, and the hospital provides services for all residents living in Hong Kong, especially those in Kowloon. The number of ED patient arrivals from July 12, 2009, to June 27, 2010, had an annual flow of 139,910 ED visits, with an average of about 380 visits every day. The H1N1 pandemic broke out in Hong Kong during this period, which was a global epidemic before the outbreak of COVID-19. This ED provides 24/7 service for patients. As weekly scheduling arrangements have many applicable scenarios, we focused on the dynamic characteristics of weekly patient visits in this work. The hospital administrators use the ED weekly visits forecasting to optimize their human and material resources, as well as to enhance their preparedness for a crisis [10]. For the same purpose, some scholars forecasted ED weekly visits by considering the week of the year seasonality [3]. Every week is from Sunday to the next Saturday. For our analysis, all numerical data variables were converted to their corresponding weekly data by averages per week, which can represent the difference among weeks. In this way, the total number of holidays within a week was used to present the impact of this factor on patient arrivals. For temperature variables, we applied the data from the previous week to forecast patient arrivals in the current week. The data set has been examined and there were no heavy outliers. All variables were transformed with the minimum-maximum normalization technique before modeling. Therefore, the data of 51 weeks are the total data set. The data of the first 27 weeks were treated as the training data set and the rest as the testing data set. In the analysis, we divided the data set into 2 parts (60:40). We validated the model's forecasting power with more testing data and the convenience of setting the split point at the beginning of the month. For example, if we were to predict ED patient arrivals for week t+1, then the data we applied included the number of ED patients at week t, the month of t+1 week, the highest and lowest temperatures at week t, and the total number of holidays for school and public holidays in the t+1 week. The search queries were from week t-6 to t from Google Trends. The normal and the outbreak conditions were considered in the training data. The weekly patient arrivals to the ED are shown in Figure 2.

Figure 2. Weekly patient arrivals to ED. ED: emergency department.



Traditional Variables

According to previous studies [4-12], 5 exogenous variables were chosen as input and divided into 3 groups: calendar data (ie, months), weather data (ie, the daily highest and lowest temperatures), and holiday data (ie, school and public holidays). The daily highest and lowest temperatures were collected from Hong Kong Observatory. The public holidays included the following: The first day of January, the day preceding Lunar New Year's Day, the first to third day of the Lunar New Year, Good Friday, the day following Good Friday, Easter Monday, the day following the Ching Ming Festival, Labour Day, the Buddha's Birthday, the Tuen Ng Festival, Hong Kong Special Administrative Region Establishment Day, National Day, Chinese Mid-Autumn Festival, the Chung Yeung Festival, Christmas Day, and the first weekday after Christmas Day. The school summer holiday was from July 11, 2009, to August 31, 2009, and 2 school winter holidays were included: one from December 19, 2009, to January 3, 2010, and the other from February 10, 2010, to February 21, 2010. The boxplot of ED arrivals by month (Figure 3) shows that patient arrivals were much higher in July and January. Although the contribution of this variable is limited in our analysis, it was still considered because of its importance and the generalization of the model [4,7]. The boxplot of holidays shows that more patients visit the ED during public holidays (Figure 4).

Figure 3. Boxplot of patient volumes of ED visits per month. ED: emergency department.

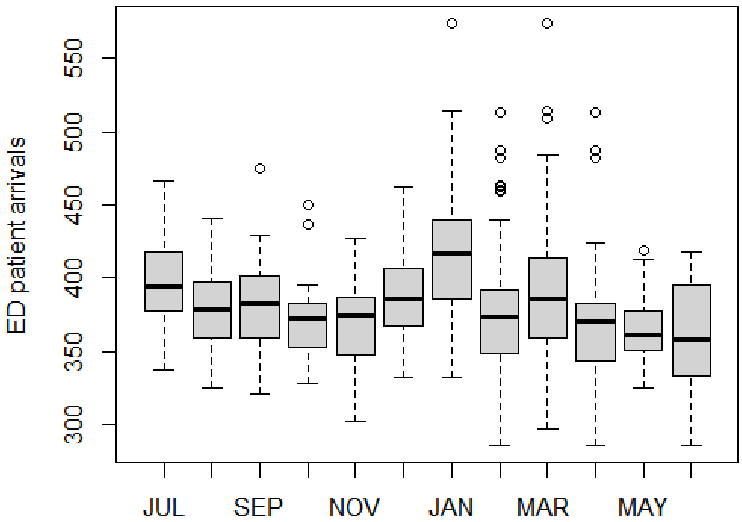
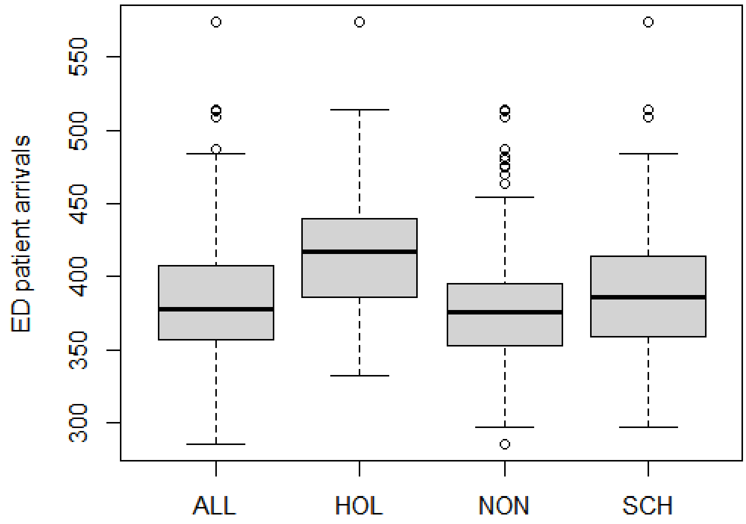


Figure 4. Boxplot of patient volumes in ED of holidays. ED: emergency department; HOL: holidays; NON: nonholidays; SCH: school closure.



Internet Search Index

Overview

In the Information Age, many patients prefer to search the internet to seek information about their problem before attending the ED. This was particularly the case during the flu outbreak in 2009-2010 [35]. Internet data can be monitored in near real time, showing the weekly dynamics of patient flow. Thus, sudden and transient changes in people’s behavior can be measured and used for prediction before such changes are actually noticed by the ED [4]. Internet data may also be a feasible surveillance tool for ED to prevent overcrowding. According to the data from Statcounter [36], the Google search engine has now become mainstream in Hong Kong, and thus using data from this search engine is conducive to the consistency of future forecasts. The search queries from Google Hong Kong were collected as internet data through Statistical Analysis Tools (Google LLC). In addition, the normalization of Google data only slightly affects the experiments because the data are renormalized in every iteration. However, because the normalization of Google Trends data is within a specific period, it is necessary to monitor the current values of the queries.

Google Trends is a common data-aggregating tool for measuring and analyzing Google search data, which can timely reflect the changes and trends in a society based on the popularity of specific Google search queries. The internet data collected had the same geographic area and period as ED patient arrivals. Data were collected by selecting “All categories” on Google Trends and Google Web Search between July 12, 2009, and June 27, 2010, in Hong Kong on the Google Trends official website. An internet search index was constructed by combining the ED-related search queries. The fusion method is a 4-stage process [12].

Step 1: Queries Generation

The selection of the initial queries is important to comprehensively collect internet information. As the current methods are mostly based on empirical intuition, the initial selection in this work was designed to expand the related search scope as much as possible. We defined and organized the information (Table 1) into 5 specific categories inspired by expert knowledge and well-studied papers: names of diseases, causes, symptoms, treatments, and others [13,22,37]. The initial queries were selected based on expert knowledge and information from the Hong Kong Department of Health. This includes, in particular, the experience of ED staff, the most common search queries in health-related references as well as the information on infectious diseases and virus surveillance from the Department of Health [13,22,37,38]. Results of some queries indicated that potential patients with specific conditions may visit the ED. For example, poor weather contributes to the development of numerous ailments, such as asthma. Claritin is a common antiallergic medication among patients with allergy. Massages can help with lumbar muscle strain, muscle atrophy, and migraine headaches. In Hong Kong, honey is one of the most popular health care remedies for curing sore throat. Ultimately, the initial 20 queries were selected. Hong Kong is a multicultural city, and both Chinese and English are often used in search engines. For a better understanding, the English translations of Chinese search queries are shown in parentheses.

Table 1. Initial search queries related to emergency department patients

Aspects	Index
Names	癌症 (*cancer*), 流感 (*influenza*), abortion, flu, h1n1 symptoms
Causes	天氣 (*weather*), 病毒 (*virus*), pregnancy, skin problem, tobacco
Symptoms	喉嚨痛 (*sore throat*), 發燒 (*fever*), 出汗 (*sweat*), infections
Treatments	克拉汀 (*claritin*), 按摩 (*massage*)
Others	蜂蜜 (*honey*), 醫生 (*doctor*), 冬季 (*winter*), depression

Step 2: Queries Expansion

A total of 20 basic queries were used as seed words. The related queries were recommended by Google Trends. These queries were then applied in the second-round search. This process was repeated until the queries became unavailable. A total of 268 search queries were collected by this process. As multiple comparisons are involved, the *P* value modified by false discovery rate was applied to make it hard to reject the null hypothesis.

Step 3: Queries Selection

The Pearson correlation coefficients were calculated between ED patient arrivals and the search queries. As the actual distributions of queries and patient arrivals are unknown, they were assumed to be normally distributed by convention [11]. Pearson correlation can help find some interpretable queries to ensure their information is useful for prediction. Considering that the actual visit is later than the online search behavior, it is necessary to test search queries with different lags. For every query, 7 Pearson correlation coefficients were generated from the data of 7 weeks before the forecast week, denoted as lag1 to lag7. Among them, we selected queries with the largest correlation coefficient no less than 0.30, which is calculated between ED patient arrivals and the search queries in the training data set. Finally, 9 queries were selected as shown in Table 2. Taking into account their lags, they were shifted (ie, previous queries moved to the corresponding rows of the current week) and summed to build the index.

Table 2. Maximum correlation coefficient of search queries from Google Trends.

Number	Index	Aspects	Lag[a]	Correlation coefficient	*P* value[b]
1	ginger	Treatment	1	–0.33	.02
2	swine flu[c] symptoms	Disease	1	0.50	<.001
3	Infect	Symptom	1	0.36	.01
4	衛生署 (Department of Health)	Others	1	0.43	.00
5	fever	Symptom	2	0.31	.04
6	豬流感 診所 (swine flu clinic)	Disease	2	0.49	<.001
7	牙醫 (dentist)	Others	2	–0.32	.04
8	肠病毒 (enterovirus)	Disease	6	0.38	<.001
9	cough	Symptom	7	–0.42	<.001

[a]The unit of lag is week(s).

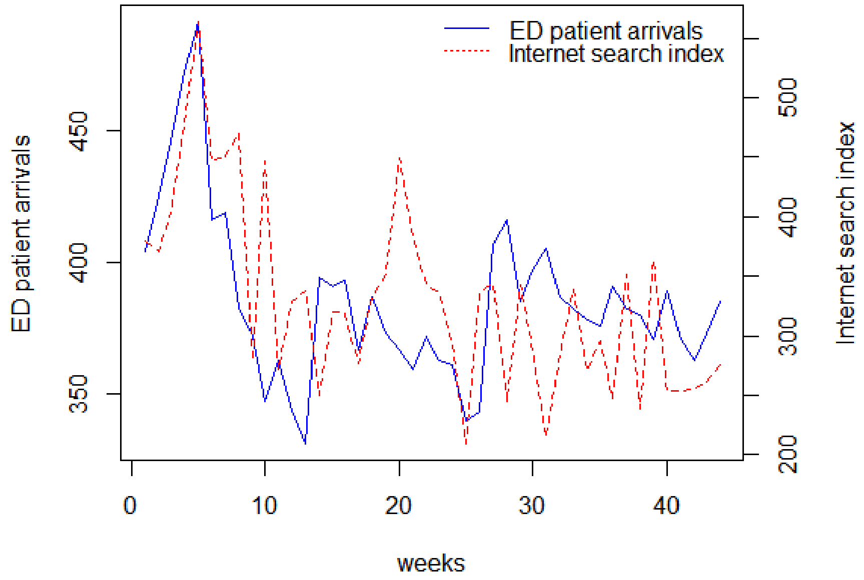
[b]The *P* value is modified by false discovery rate (significance level=.05).

[c]Swine flu is the nickname of H1N1 influenza in Hong Kong.

Step 4: Internet Search Index Construction

To illustrate the contribution of the overall related internet information, the internet search index was employed by shifting and summing. In addition, the internet search index can effectively reduce the dimension of the data compared with the queries. According to the lag term, the 9 queries selected above were shifted in such a way that the previous queries moved to the corresponding rows of the current week. That is the reason why we applied search queries from at least one week before. All of the shifted search queries were summed to form the internet search index as a new time series. Although the fluctuation of the internet search index is greater than ED patient arrivals, it presents a similar trend to ED patient arrivals (Figure 5).

Figure 5. Trend of ED patient arrivals and internet search index. ED: emergency department.



According to the aforementioned steps, 9 valuable queries were selected from a total of 268 queries, of which 4 were Chinese queries: 衛生署 (Department of Health), 豬流感 診所 (swine flu clinic), 牙醫 (dentist), and 肠病毒 (enterovirus). The remaining 5 were English queries (ginger, swine flu symptoms, infect, fever, and cough).

It has been observed that pandemic outbreaks can be captured by query terms [11,39]. The 3 queries here, swine flu symptoms, 豬流感 診所 (swine flu clinic), 肠病毒 (enterovirus), can be associated with the swine flu outbreak in 2009 and enteroviruses outbreak in 2010 in Hong Kong. In addition, the query 衛生署 (Department of Health) is related to ED visits. The Department of Health is the official hospital management agency in Hong Kong [38]. Its website provides reliable and comprehensive medical-related information and regularly issues outbreak alerts. The remaining queries (cold, infect, fever) belong to the common emergency services. Moreover, in Hong Kong, ginger is an effective medicinal spice used in daily life and widely used for the prevention and early treatment of cold. Therefore, “ginger” as a search query might suggest that the user is a potential ED patient. These queries capture local population–level health information and were translated into an internet search index. It provides information rarely found in traditional data sources, and thus, the forecasting of ED visits can be improved.

Extreme Learning Machine

For the forecasting model, we employed 8 different methods: ELM, GLM, ARIMA, ARIMAX, SVM, ANN, RF, and LSTM. Besides ELM, the others are well known. To the best of our knowledge, this is the first time that ELM has been applied for the prediction of ED visits.

ELM is a single hidden layer neural network algorithm [40-43]. It has been widely used in many fields because of simple mathematical description, lower computational burden, and faster learning speed [40]. The main feature of the ELM is that the algorithm can randomly generate the input weights and node biases. The least-square method is used to determine the output weight by simple matrix computations. These made it computationally attractive.

We used the sample data set , where x_i is the input and y_i is the output; n and m are the dimension numbers of input and output, respectively; and N is the number of samples. The forecasting model can be established using the ELM algorithm with L hidden neurons as follows:

where a_j and b_j denote the input weight and the bias of the hidden layer, respectively; $g(\cdot)$ represents the activation function of hidden neurons; β*j* is the output weight representing the connected output neuron and the *j*th hidden neuron.

The following objective function is constructed to find the output weight β.

Equations (2) and (3) can be rewritten as:

$$H\beta = Y \quad (4)$$

where H is the hidden layer output matrix.

Through the least-squares method, the output weight β can be obtained as follows:

where H^+ denotes the Moore-Penrose generalized inverse of matrix H. Using equations 1 and 7, the resulting ELM model can be estimated.

Evaluation Metrics

Two evaluation metrics were used to qualify the forecasting performance of different models: the root mean square error (RMSE) and mean absolute percentage error (MAPE). These are written as follows:

where N is the number of observations; y_i is the real value; and indicates the forecast value.

In addition, the Diebold-Mariano (DM) test was used to compare the forecast accuracy of forecast models [44]. The null hypothesis is that the reference model *re* is more accurate than the test model *te*. The DM statistic can be written as:

where

; y_i is the actual value; and $y_{re,i}$ and $y_{te,i}$ are forecasting values of the reference and test models, respectively; N is the number of observations; h(>1) is h-step-ahead forecasts; is the autocovariance of the loss differential at lag k. The loss differential time series d_i is confirmed as stationary with the augmented Dickey-Fuller test.

Results

Relationship Between ED Visits and Internet Search Index

We first analyzed the association between the internet search index and ED patient arrivals with the Pearson correlation coefficient. We then also applied the Johansen cointegration and Granger causality to verify their relationship [45,46]. All 3 analyses were based on the training data set.

Initially, the Pearson correlation coefficients indicated that ED patient arrivals were significantly correlated with the internet search index (r=0.46, P=.002).

We next report the results of cointegration (Multimedia Appendix 1). Given the logarithmic form of the 2 variables to reduce the impact of outliers, a stability test was performed. These 2 time series were stably validated by the augmented Dickey-Fuller test [47]. The cointegration results indicated that ED arrivals and internet search index were cointegrated. The first hypothesis, r=0, tests for the presence of cointegration. As the test statistic exceeds the 1% level (66.22>23.52), we have strong evidence to reject the null hypothesis of *no cointegration*. The second test for r<1 against the alternative hypothesis of r>1 also provides evidence to reject r<1 because the test statistic again exceeds the 1% level (20.41>11.65). Thus, the cointegration results demonstrate that ED patient arrivals and the internet search index were cointegrated.

Meanwhile, the Granger causality test was applied to verify whether the internet search index is a predictor of ED arrivals. According to Multimedia Appendix 1, log(internet search index) is the Granger cause of log(patient arrivals). It indicates a causal relationship between internet search index and patient arrivals. We examined the relationship between internet search index and patient arrivals by Pearson correlation coefficient, Johansen cointegration, and Granger causality test. As the internet search index was correlated with ED patient arrivals, it can be included as a novel variable in the forecasting model.

Forecasting Performance Evaluation

The forecasting models are linear and nonlinear models, including ELM, GLM, ARIMA, ARIMAX, SVM, ANN, RF, and LSTM. To test the predictive power of adding different data sources, the data can be classified into 3 types: "patient arrivals," "patient arrivals + traditional variables," "patient arrivals + traditional variables + internet search index."

The prediction accuracy of the models is evaluated by MAPE and RMSE. As shown in Table 3, the results are promising because the models perform with a fairly high level of accuracy overall. It is obvious that combining the internet search index provides a higher prediction accuracy than both "patient arrivals + traditional variables" and "patient arrivals." As stated earlier, internet search data can improve prediction.

Table 3. Prediction performance of weekly emergency department patient arrivals.

Models	Training		Testing	
	MAPE[a] (%)	RMSE[b]	MAPE (%)	RMSE
Patient arrivals				
ARIMA[c]	3.6	17.02	5.2	24.28
ANN[d]	3.5	19.19	3.6	19.20
SVM[e]	2.2	18.52	4.1	18.59
RF[f]	2.5	19.36	3.0	19.67
LSTM[g]	2.9	16.93	4.2	20.79
ELM[h]	2.8	16.52	3.2	16.99
Patient arrivals + traditional variables				
GLM[i]	3.2	16.79	5.3	24.44
ARIMAX[j]	3.5	17.85	5.1	23.16
ANN	3.4	16.10	4.0	18.48
SVM	2.8	16.24	3.9	17.45
RF	2.9	17.05	3.7	18.36
LSTM	3.2	16.43	4.4	19.94
ELM	2.7	13.17	3.5	16.72
Patient arrivals + traditional variables + internet search index				
GLM	3.2	16.24	5.0	23.18
ARIMAX	3.4	17.84	5.1	22.00
ANN	3.0	14.51	3.3	15.45
SVM	2.6	14.84	3.1	15.09
RF	2.9	15.92	3.3	16.32
LSTM	3.0	15.15	3.4	16.69
ELM	2.6	13.10	3.0	14.55

[a]MAPE: average mean absolute percentage error.

[b]RMSE: root mean square error.

[c]ARIMA: autoregressive integrated moving average model.

[d]ANN: artificial neural network.

[e]SVM: support vector machine.

[f]RF: random forest.

[g]LSTM: long short-term memory.

[h]ELM: extreme learning machine.

[i]GLM: generalized linear model.

[j]ARIMAX: ARIMA with explanatory variables.

In particular, the performance of the models varies based on the value of the hyperparameters. The process of tuning the hyperparameters is performed to balance the relationship between optimal solution and regularization in the training data set, and thus to achieve the best generalization ability in the testing data set. As 2 commonly used parameter selection methods, trial-and-error and grid search guarantee good performance. Using these methods, we applied different models in this study. Both methods utilized different combinations of parameters and then built the best performance model with the selected parameters. With the trial-and-error method, the final selected GLM was fitted with Gaussian distribution rather than with other error distributions. The performance of ARIMA(X) is determined by the AR order (p), the degree of difference (d), and the moving average (MA) order (q). Autocorrelation function and partial autocorrelation were used to identify the value of AR and MA after verifying stationary by differencing the time series. The sigmoid activation function is applied in ANN. The specific values of hidden layer and hidden neuron are chosen from grid search. Similarly, the radial is used in

SVM. A grid search is employed to select the number of costs, gamma, and epsilon. As for RF, the number of trees grown and the number of variables sampled at each split are decided through a grid search. Moreover, it is applied to tune the batch size, hidden units, and epochs. For ELM, the number of hidden nodes is set to 20, 120, and 150 for the "patient arrivals," "patient arrivals + traditional variables," and "patient arrivals + traditional variables + internet search index" data set, respectively. The kernel function is set to "satlins" for all data sets. Furthermore, as the initial weights were generated randomly, the parameter was decided by the average performance of the experiments (n<10) to ensure reliability. The optimal forms of ARIMA and ARIMAX were estimated by minimizing Akaike information criteria and Bayesian information criterion.

The ED experts informed that they had to increase additional medical staff members when configuration was mismatched by more than 18% [11]. Therefore, the aforesaid results indicate that there are 7 and 5 mismatch days for GLM without and with internet search index, respectively. ELM with internet search index had 2 mismatching days, which is the least value among all the forecasting models. Compared with "patient arrivals + traditional variables," it can prevent 1 mismatching day theoretically.

Moreover, the best performance is achieved by ELM with independent variables of "patient arrivals + traditional variables + internet search index" in the training and testing data sets. It achieved an MAPE of 3%, with RMSE of 14.55. SVM also performed well, followed by ANN, RF, LSTM, and ARIMAX; GLM ranked last. The dynamic characteristic of the patient arrivals can be well represented by the ELM model.

The DM test was used to compare accuracy of forecasting models from a statistical point of view. The DM statistic results are shown in Table 4. With internet search index, when the ELM is applied as a test model with medium significance (P<.001), the model was superior to other forecasting models. By contrast, GLM had the lowest prediction performance among the 7 forecasting models. In addition, the performance of ELM, ANN, RF, SVM, and LSTM was better than that of ARIMAX and GLM. Therefore, nonlinear models may be more suitable for predicting the arrival of ED patients with the internet search index.

We next measured the DM test results between the reference model without internet data and the test model with internet data (Table 5). One of the critical findings is that the same models with internet data are better than those without internet data. Among all the models, neither GLM nor ARIMAX had a good performance, even with the internet data. All nonlinear models with internet data had higher accuracy than those without.

We assessed the robustness of the 7 forecasting models with or without the internet search index. All forecasting models were run 20 times using data set with different lengths. The robustness was evaluated by the SD of MAPE and RMSE. ELM was the most stable model with minimum SD of MAPE and RMSE (Table 6). By contrast, GLM was the most unstable forecasting model because it had maximum SD of MAPE and RMSE. The results also indicated that the forecasting models with the internet search index are more stable. Moreover, the robustness of nonlinear models is better than that of linear models. Compared with linear models, the rate of decline for nonlinear models is faster.

Table 4. DM[a] test results of testing data set for same data set.

Test model	Reference model[b,c]					
	GLM[d]	ARIMAX[e]	ANN[f]	SVM[g]	RF[h]	LSTM[i]
Patient arrivals + traditional variables						
ELM	2.8297 (<.001)	3.0624 (<.001)	2.012 (<.001)	1.8178 (<.001)	2.8481 (<.001)	2.1002 (<.001)
GLM		0.2935 (.31)	0.86595 (.11)	1.0707 (.06)	0.7643 (.09)	0.1663 (.54)
ARIMAX			0.64435 (.17)	1.0691 (.06)	0.3957 (.23)	0.4876 (.68)
ANN				0.13244 (.38)	0.2746 (.45)	1.9512 (.01)
SVM					0.5823 (.27)	0.8714 (.12)
RF						1.0045 (.08)
Patient arrivals + traditional variables + internet search index						
ELM	2.5062 (<.001)	3.79 (<.001)	2.0047 (<.001)	2.0325 (<.001)	2.0476 (<.001)	1.6659 (.02)
GLM		0.32675 (.30)	1.1462 (.12)	1.6064 (.07)	1.0467 (.09)	0.3647 (.64)
ARIMAX			1.7314 (.06)	2.2885 (.05)	1.5946 (.11)	1.2671 (.07)
ANN				0.14419 (.40)	0.2104 (.49)	1.2304 (.08)
SVM					0.2593 (.36)	1.4391 (.04)
RF						1.2992 (.06)

[a]DM: Diebold-Mariano.

[b]The *P* value modified by false discovery rate is given in brackets. The significance level is .05.

[c]Values are presented as the Diebold-Mariano statistic (*P* value modified by false discovery rate).

[d]GLM: generalized linear model.

[e]ARIMAX: ARIMA with explanatory variables.

[f]ANN: artificial neural network.

[g]SVM: support vector machine.

[h]RF: random forest.

[i]LSTM: long short-term memory.

Table 5. DM test results of testing data set for different data sets.

Test model (with internet data)	Reference model (without internet data)[a]						
	GLM[b]	ARIMAX[c]	ANN[d]	SVM[e]	RF[f]	LSTM[g]	ELM[h]
GLM	2.4848 (<.001)	0.4041 (.31)	0.2797 (.37)	0.2314 (.40)	1.8806 (.22)	2.1002 (.31)	0.8635 (.16)
ARIMAX	1.5701 (.04)	2.5818 (<.001)	0.4968 (.28)	1.7698 (.12)	0.7756 (.20)	1.5748 (.24)	0.0337 (.51)
ANN	2.2546 (<.001)	1.7547 (.02)	4.1945 (<.001)	2.8276 (<.001)	3.4393 (<.001)	1.2432 (.09)	4.4291 (<.001)
SVM	2.244 (<.001)	1.7374 (.02)	6.0597 (<.001)	2.3394 (<.001)	1.6767 (.02)	0.8602 (.07)	3.2791 (<.001)
RF	1.7886 (.02)	1.9591 (.02)	2.7599 (<.001)	1.8785 (.02)	2.3097 (<.001)	0.5800 (.05)	2.3075 (<.001)
LSTM	0.8556 (.01)	3.3685 (<.001)	1.3620 (.07)	1.6441 (.04)	1.0560 (.12)	2.4263 (<.001)	1.9995 (.02)
ELM	2.2546 (<.001)	1.7547 (.03)	4.1946 (<.001)	2.8276 (<.001)	3.4394 (<.001)	2.175 (.01)	4.4291 (<.001)

[a]Values are presented as the Diebold-Mariano statistic. The *P* value modified by false discovery rate is in brackets. The significance level is .05.
[b]GLM: generalized linear model.
[c]ARIMAX: ARIMA with explanatory variables.
[d]ANN: artificial neural network.
[e]SVM: support vector machine.
[f]RF: random forest.
[g]LSTM: long short-term memory.
[h]ELM: extreme learning machine.

Table 6. Robustness analysis.

SD	Forecasting model						
	GLM[a]	ANN[b]	SVM[c]	ARIMAX[d]	LSTM[e]	RF[f]	ELM[g]
Patient arrivals + traditional variables							
SD of MAPE[h] (%)	2.5	1.0	1.0	1.7	1.7	1.2	1.0
SD of RMSE[i]	15.638	4.385	5.158	5.843	7.409	5.371	4.099
Patient arrivals + traditional variables + internet search index							
SD of MAPE (%)	2.4	0.8	0.7	0.9	1.3	0.8	0.7
SD of RMSE	15.212	3.577	4.008	5.681	5.797	3.985	3.370

[a]GLM: generalized linear model.
[b]ANN: artificial neural network.
[c]SVM: support vector machine.
[d]ARIMAX: ARIMA with explanatory variables.
[e]LSTM: long short-term memory.
[f]RF: random forest.
[g]ELM: extreme learning machine.
[h]MAPE: average mean absolute percentage error.
[i]RMSE: root mean square error.

These analyses have revealed some interesting findings: (1) The forecasting performance is improved by the internet search index, which might reflect the behavioral trends of potential patients during the period of the H1N1 pandemic. (2) The accuracy of the ELM model was far superior than that of other forecasting models and the model captures the nonlinearities between the variables and ED patients. (3) Including internet search index results in more stable models and the proposed ELM was the most stable among the models.

Discussion

Principal Findings

As the number of patients increases continually, ED needs more information to make timely and target resource configuration strategies, thus preventing overcrowding and reducing social pressure. In the era of big data, internet data have been used in many areas and may help formulate new and appropriate measures to provide early warning signals to decision makers. In this study, we mainly focused on introducing internet data and nonlinear models to predict ED visits during the pandemic. The 3 contributions are summarized as follows. First, we compared the performance of linear and nonlinear models in the data set with or without internet search index to predict patient arrivals. The observed increase in forecasting accuracy could be attributed to internet data and kernel-based ELM. In addition, we investigated the performance metrics of previous studies. The visits to a special, regional medical website were monitored to predict the daily ED attendance with linear regression, with an MAPE of 4.8% [4]. Another linear model (LASSO) was employed, in combination with traditional variables, to reduce the MAPE and RMSE to 7.58% and 12.07 [11]. Recently, a multiple regression was applied with Google data to predict ED arrivals in the Singapore General Hospital. Its prediction curve indicated that MAPE was close to 8% [13]. Compared with the performance metrics, the minimum MAPE and RMSE obtained in our study were 3.0% and 14.55, respectively. The comparison reveals that our work is competitive. Although the compared studies have different scenarios, data, models, and environments, all found that the internet data can help in the prediction of ED patient arrivals. We further examined the accuracy and robustness from a statistics perspective. Second, a systematic method was applied to build the internet search index that reflects patient-related information as comprehensively as possible in search queries, including common diseases, possible causes, current symptoms, self-treatment, and others. Statistically, the effectiveness of the internet search index was also verified by Pearson correlation coefficient, Johansen cointegration, and Granger causality. Third, the characteristics of the ED visits during the outbreak of H1N1 pandemic were also modeled. The problem of ED overcrowding at such a serious time was more intense than at normal times, which could be a typical environment for the proposed method using the internet search index.

The proposed intelligent forecasting system predicts ED patient arrivals accurately and timely. The predictive power provided by the system stems from 2 parts. First, the internet search index that integrates with relevant internet search queries greatly contributes to the improvement of forecasting. According to the selected queries, the lag term of most queries was lag 1, indicating that patients are very likely to visit the ED within a week after identifying their symptoms. As we took internet data into account, if we extend the forecasting scale, it may miss some queries and fail to reflect the near–real-time trends and sudden changes. A trade-off between accuracy and time was thus necessary here. Ultimately, we made 1-week-ahead forecasting of ED patient arrivals. Whenever the new incoming data are higher than previously highest values, we first updated the internet data and then predicted the patient arrivals. More specifically, the new incoming higher value will be normalized as 1, meaning that we have obtained another extreme value and thus it is necessary to update the entire query data. The prediction model will also be retrained with the new query data. When the highest value remains the same after adding the new data, we simply add the new data to the existing ones. Second, kernel-based ELM can explore the nonlinearities in data to achieve better forecasting accuracy. It is a novel computational intelligence method based on single-hidden layer feedforward networks for regression and classification. The essence of ELM is that almost all nonlinear piecewise continuous functions can be used as the hidden-node output function, and thus, the feature mappings used in ELM can be very diversified to approximate arbitrary nonlinearity. Moreover, input weights and the hidden layer parameters are randomly generated independent of the training samples, and only the output weights are calculated through the least-square method. This characteristic leads to a significant improvement in the learning speed of ELM. Therefore, ELM can be applied in the identification of nonlinearity to forecast ED patient arrivals. The proposed system is utilized to forecast ED patient arrivals for a Hong Kong hospital. Our experimental results reveal that the forecasting system with ELM is significantly superior over the traditional linear models and some other nonlinear models. Meanwhile, the internet search index increases the forecasting power of all models. Therefore, this system will provide more information for the predicted values, and then well-matched resource allocation plans will be developed in real-time or near real-time per week.

Limitations

The limitations of this paper are as follows. First, some patients may not be able to access the internet. As a result, their behaviors will not be recorded by the internet data. However, with the development of the mobile internet, it will be more convenient for people to obtain the information through search engines. Second, the queries we selected only contained Chinese and English. Thus, some internet searches in other languages are likely ignored; however, languages other than the aforesaid are less popular in Hong Kong. Third, the search queries we chose may be limited. Although we had 266 queries, some queries may have been missed. We believe that the search queries could be updated in another comprehensive query selection method over time. Moreover, the Granger causality may result in spurious causality. Finally, to ensure the reliability of the data from Google Trends, the influencing factors, including the mass media interference and the statistical anomalies, need to be considered. Other advanced methods in selecting informative queries, such as Spearman, will be seriously considered to improve the forecasting power of our method.

Conclusions

This study supports the possibility of using internet data to predict ED visits during a pandemic and this is, to the best of our knowledge, the first study to use internet data and nonlinear models to predict ED visits. Compared with several related papers, we mainly focused on dynamic characteristics of patient arrivals during the H1N1 influenza, which was declared as a pandemic in Hong Kong in 2009 [4,11,13]. The problem of ED overcrowding at such that time was more serious than in normal times. Using the proposed framework, the ED-related human behavior information can be effectively extracted and introduced into the prediction model. In this study, an intelligent forecasting system was proposed with machine learning and internet search index to accurately predict weekly ED patient arrivals. Initially, we used a comprehensive and systematic method to build the internet search index with related search queries, which contained information about disease, causes, symptoms, and treatments. The relationship between the internet search index and ED patient arrivals was then verified by Pearson correlation coefficient, Johansen cointegration, and Granger causality. Finally, forecasting models were applied to different combinations of data with or without internet search index.

Our experimental results indicated that all of the forecasting models are more accurate when the internet search index is considered, as the internet data can timely reflect the changes and trends. Compared with other popular forecasting methods, the proposed kernel-based ELM model was more accurate and robust to present the nonlinearities between the variables and ED patients. In general, the performance of nonlinear models is better than linear models. This may imply that the dynamic relationship between variables and patient arrivals can be well represented by the nonlinear models. This intelligent forecasting system can be widely applied in other EDs, with the need to only update the internet search index according to regional or special requirements. It may also help ED managers to improve staff scheduling and allocate resources more effectively to prevent overcrowding by giving an early warning, especially during a pandemic like H1N1 or even during COVID-19 times.

Our future work will explore ED-related data from social media platforms, such as Twitter, Facebook, and Weibo, to investigate their impact on the ED patient arrivals. In addition, we plan to predict the ED patient flow with different severity levels. As there are 5 levels of ED patient arrival triages in Hong Kong, the relationship between these 5 levels of patient arrivals and internet information will be further studied. This will help ED managers develop a more flexible and targeted strategy to balance the need of different patients. Furthermore, the spatiotemporal changes in ED patient visits are worth studying in-depth. To further improve the accuracy of the forecasting model, deep learning algorithms will be of great interest in our future work, especially the ability to find efficient representations in large amounts of data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant numbers 71701136, 71701132, and 51905109) and by the Natural Science Foundation of Guangdong Province (grant numbers 2022A1515011009, 2021A1515010987, and 2021A1515011971).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Cointegration test and Granger causality test result.
[DOCX File , 14 KB - medinform_v10i7e34504_app1.docx]

References

1. Carvalho-Silva M, Monteiro MTT, Sá-Soares FD, Dória-Nóbrega S. Assessment of forecasting models for patients arrival at Emergency Department. Operations Research for Health Care 2018 Sep;18:112-118. [doi: 10.1016/j.orhc.2017.05.001]
2. Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. Emergency department crowding: A systematic review of causes, consequences and solutions. PLoS One 2018;13(8):e0203316 [FREE Full text] [doi: 10.1371/journal.pone.0203316] [Medline: 30161242]
3. Afilal M, Yalaoui F, Dugardin F, Amodeo L, Laplanche D, Blua P. Forecasting the Emergency Department Patients Flow. J Med Syst 2016 Jul 7;40(7):175. [doi: 10.1007/s10916-016-0527-0] [Medline: 27272135]
4. Ekström A, Kurland L, Farrokhnia N, Castrén M, Nordberg M. Forecasting emergency department visits using internet data. Ann Emerg Med 2015 Apr;65(4):436-442.e1. [doi: 10.1016/j.annemergmed.2014.10.008] [Medline: 25487026]
5. Gul M, Celik E. An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. Health Syst (Basingstoke) 2018 Nov 19;9(4):263-284 [FREE Full text] [doi: 10.1080/20476965.2018.1547348] [Medline: 33354320]
6. Jiang S, Chin K, Wang L, Qu G, Tsui KL. Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient department. Expert Systems with Applications 2017 Oct;82:216-230. [doi: 10.1016/j.eswa.2017.04.017]
7. Xu Q, Tsui K, Jiang W, Guo H. A Hybrid Approach for Forecasting Patient Visits in Emergency Department. Qual. Reliab. Engng. Int 2016 Oct 06;32(8):2751-2759. [doi: 10.1002/qre.2095]
8. Qin Q, Xie K, He H, Li L, Chu X, Wei Y, et al. An effective and robust decomposition-ensemble energy price forecasting paradigm with local linear prediction. Energy Economics 2019 Sep;83:402-414. [doi: 10.1016/j.eneco.2019.07.026]

9. Ordu M, Demir E, Tofallis C. A comprehensive modelling framework to forecast the demand for all hospital services. Int J Health Plann Manage 2019 Apr 22;34(2):e1257-e1271. [doi: 10.1002/hpm.2771] [Medline: 30901132]
10. Khaldi R, Afia AE, Chiheb R. Forecasting of weekly patient visits to emergency department: real case study. Procedia Computer Science 2019;148:532-541. [doi: 10.1016/j.procs.2019.01.026]
11. Tideman S, Santillana M, Bickel J, Reis B. Internet search query data improve forecasts of daily emergency department volume. J Am Med Inform Assoc 2019 Dec 01;26(12):1574-1583 [FREE Full text] [doi: 10.1093/jamia/ocz154] [Medline: 31730701]
12. Zhang Y, Luo L, Yang J, Liu D, Kong R, Feng Y. A hybrid ARIMA-SVR approach for forecasting emergency patient flow. J Ambient Intell Human Comput 2018 Sep 26;10(8):3315-3323. [doi: 10.1007/s12652-018-1059-x]
13. Ho AFW, To BZYS, Koh JM, Cheong KH. Forecasting Hospital Emergency Department Patient Volume Using Internet Search Data. IEEE Access 2019;7:93387-93395. [doi: 10.1109/access.2019.2928122]
14. Sun S, Wei Y, Tsui K, Wang S. Forecasting tourist arrivals with machine learning and internet search index. Tourism Management 2019 Feb;70:1-10. [doi: 10.1016/j.tourman.2018.07.010]
15. Oliveira N, Cortez P, Areal N. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Systems with Applications 2017 May;73:125-144. [doi: 10.1016/j.eswa.2016.12.036]
16. Naccarato A, Falorsi S, Loriga S, Pierini A. Combining official and Google Trends data to forecast the Italian youth unemployment rate. Technological Forecasting and Social Change 2018 May;130:114-122. [doi: 10.1016/j.techfore.2017.11.022]
17. Deiner MS, Lietman TM, McLeod SD, Chodosh J, Porco TC. Surveillance Tools Emerging From Search Engines and Social Media Data for Determining Eye Disease Patterns. JAMA Ophthalmol 2016 Sep 01;134(9):1024-1030 [FREE Full text] [doi: 10.1001/jamaophthalmol.2016.2267] [Medline: 27416554]
18. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb 19;457(7232):1012-1014. [doi: 10.1038/nature07634] [Medline: 19020500]
19. Lampos V, Miller AC, Crossan S, Stefansen C. Advances in nowcasting influenza-like illness rates using search query logs. Sci Rep 2015 Aug 03;5:12760 [FREE Full text] [doi: 10.1038/srep12760] [Medline: 26234783]
20. Cervellin G, Comelli I, Lippi G. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. J Epidemiol Glob Health 2017 Dec;7(3):185-189 [FREE Full text] [doi: 10.1016/j.jegh.2017.06.001] [Medline: 28756828]
21. Sato K, Mano T, Iwata A, Toda T. Need of care in interpreting Google Trends-based COVID-19 infodemiological study results: potential risk of false-positivity. BMC Med Res Methodol 2021 Jul 18;21(1):147. [doi: 10.1186/s12874-021-01338-2]
22. Dugas AF, Hsieh Y, Levin SR, Pines JM, Mareiniss DP, Mohareb A, et al. Google Flu Trends: correlation with emergency department influenza rates and crowding metrics. Clin Infect Dis 2012 Feb 15;54(4):463-469 [FREE Full text] [doi: 10.1093/cid/cir883] [Medline: 22230244]
23. Xu K, Yang H, Zhu C. A novel extreme learning Machine-based Hammerstein-Wiener model for complex nonlinear industrial processes. Neurocomputing 2019 Sep;358:246-254. [doi: 10.1016/j.neucom.2019.05.049]
24. Jilani T, Housley G, Figueredo G, Tang P, Hatton J, Shaw D. Short and Long term predictions of Hospital emergency department attendances. Int J Med Inform 2019 Sep;129:167-174. [doi: 10.1016/j.ijmedinf.2019.05.011] [Medline: 31445251]
25. Whitt W, Zhang X. Forecasting arrivals and occupancy levels in an emergency department. Operations Research for Health Care 2019 Jun;21:1-18. [doi: 10.1016/j.orhc.2019.01.002]
26. Aboagye-Sarfo P, Mai Q, Sanfilippo FM, Preen DB, Stewart LM, Fatovich DM. A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in Western Australia. J Biomed Inform 2015 Oct;57:62-73 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.022] [Medline: 26151668]
27. Juang W, Huang S, Huang F, Cheng P, Wann S. Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. BMJ Open 2017 Dec 01;7(11):e018628. [doi: 10.1136/bmjopen-2017-018628] [Medline: 29196487]
28. Choudhury A, Urena E. Forecasting hourly emergency department arrival using time series analysis. British Journal of Healthcare Management 2020 Jan 02;26(1):34-43. [doi: 10.12968/bjhc.2019.0067]
29. Calegari R, Fogliatto FS, Lucini FR, Neyeloff J, Kuchenbecker RS, Schaan BD. Forecasting Daily Volume and Acuity of Patients in the Emergency Department. Comput Math Methods Med 2016;2016:3863268 [FREE Full text] [doi: 10.1155/2016/3863268] [Medline: 27725842]
30. Ram S, Zhang W, Williams M, Pengetnze Y. Predicting asthma-related emergency department visits using big data. IEEE J Biomed Health Inform 2015 Jul;19(4):1216-1223. [doi: 10.1109/JBHI.2015.2404829] [Medline: 25706935]
31. Chen Y, Chen Y, Long J, Shi D, Chen X, Hou M, et al. Achieving a sub-10 nm nanopore array in silicon by metal-assisted chemical etching and machine learning. Int. J. Extrem. Manuf 2021 May 25;3(3):035104. [doi: 10.1088/2631-7990/abff6a]
32. Fan B, Li H, Hu Y. An Intelligent Decision System for Intraoperative Somatosensory Evoked Potential Monitoring. IEEE Trans Neural Syst Rehabil Eng 2016 Feb;24(2):300-307. [doi: 10.1109/TNSRE.2015.2477557] [Medline: 26415181]

33. Fang X, Liu W, Ai J, He M, Wu Y, Shi Y, et al. Forecasting incidence of infectious diarrhea using random forest in Jiangsu Province, China. BMC Infect Dis 2020 Mar 14;20(1):222 [FREE Full text] [doi: 10.1186/s12879-020-4930-2] [Medline: 32171261]
34. Chen W, Guo H, Tsui K. A new medical staff allocation via simulation optimisation for an emergency department in Hong Kong. International Journal of Production Research 2019 Sep 16;58(19):6004-6023. [doi: 10.1080/00207543.2019.1665201]
35. Cocco AM, Zordan R, Taylor DM, Weiland TJ, Dilley SJ, Kant J, et al. Dr Google in the ED: searching for online health information by adult emergency department patients. Med J Aust 2018 Oct 15;209(8):342-347. [doi: 10.5694/mja17.00889] [Medline: 30107763]
36. Statcounter G. Browser Market Share Worldwide. Online. URL: https://gs.statcounter.com/ [accessed 2022-06-25]
37. Wang S, Paul MJ, Dredze M. Exploring health topics in Chinese social media: An analysis of Sina Weibo. 2014 Jul Presented at: AAAI Workshop on the World Wide Web and Public Health Intelligence, July; 23; 2014; Québec Convention Centre p. A.
38. Department OH, The GOTHKSAR. Online. Available. URL: https://www.dh.gov.hk/english/index.html [accessed 2022-06-25]
39. Chen Y, Zhang Y, Xu Z, Wang X, Lu J, Hu W. Avian Influenza A (H7N9) and related Internet search query data in China. Sci Rep 2019 Jul 18;9(1):10434. [doi: 10.1038/s41598-019-46898-y] [Medline: 31320681]
40. Adnan RM, Liang Z, Trajkovic S, Zounemat-Kermani M, Li B, Kisi O. Daily streamflow prediction using optimally pruned extreme learning machine. Journal of Hydrology 2019 Oct;577:123981. [doi: 10.1016/j.jhydrol.2019.123981]
41. Cao W, Wang X, Ming Z, Gao J. A review on neural networks with random weights. Neurocomputing 2018 Jan;275:278-287. [doi: 10.1016/j.neucom.2017.08.040]
42. Xu K, Fan B, Yang H, Hu L, Shen W. Locally Weighted Principal Component Analysis-Based Multimode Modeling for Complex Distributed Parameter Systems. IEEE Trans. Cybern 2021:1-11. [doi: 10.1109/tcyb.2021.3061741]
43. Liu F, Ma Z, Wang B, Lin W. A Virtual Machine Consolidation Algorithm Based on Ant Colony System and Extreme Learning Machine for Cloud Data Center. IEEE Access 2020;8:53-67. [doi: 10.1109/access.2019.2961786]
44. Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning. The Review of Financial Studies, 2020 Feb 2020;33(5):2223-2273. [doi: 10.1093/rfs/hhaa009]
45. Ur Rahman Z, Iqbal Khattak S, Ahmad M, Khan A. A disaggregated-level analysis of the relationship among energy production, energy consumption and economic growth: Evidence from China. Energy 2020 Mar;194:116836. [doi: 10.1016/j.energy.2019.116836]
46. Zyphur MJ, Allison PD, Tay L, Voelkle MC, Preacher KJ, Zhang Z, et al. From Data to Causes I: Building A General Cross-Lagged Panel Model (GCLM). Organizational Research Methods 2019 May 21;23(4):651-687. [doi: 10.1177/1094428119847278]
47. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in Brief 2020 Apr;29:105340. [doi: 10.1016/j.dib.2020.105340]

Abbreviations

ANN: artificial neural network
AR: autoregressive
ARIMA: autoregressive integrated moving average model
ARIMAX: ARIMA with explanatory variables
DM: Diebold-Mariano
ED: emergency department
ELM: extreme learning machine
GLM: generalized linear model
LASSO: the least absolute shrinkage and selection operator
LSTM: long short-term memory
MA: moving average
MAPE: average mean absolute percentage error
M-P: Moore-Penrose
RF: random forest
RMSE: root mean square error
SVM: support vector machine

Edited by C Lovis; submitted 27.10.21; peer-reviewed by B Wang, A Staffini, SD Boie, A Rovetta; comments to author 25.02.22; revised version received 22.04.22; accepted 25.05.22; published 20.07.22.

Please cite as:
Fan B, Peng J, Guo H, Gu H, Xu K, Wu T
Accurate Forecasting of Emergency Department Arrivals With Internet Search Index and Machine Learning Models: Model Development and Performance Evaluation
JMIR Med Inform 2022;10(7):e34504
URL: https://medinform.jmir.org/2022/7/e34504
doi:10.2196/34504
PMID:35857360

Original Paper

Classification of Twitter Vaping Discourse Using BERTweet: Comparative Deep Learning Study

William Baker[1], BSc; Jason B Colditz[2], MEd, PhD; Page D Dobbs[3], PhD; Huy Mai[1], BSc; Shyam Visweswaran[4], MD, PhD; Justin Zhan[1], PhD; Brian A Primack[5], MD, PhD

[1]Department of Computer Science and Computer Engineering, University of Arkansas, Fayetteville, AR, United States

[2]Division of General Internal Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States

[3]Health, Human Performance and Recreation Department, University of Arkansas, Fayetteville, AR, United States

[4]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States

[5]College of Public Health and Human Science, Oregon State University, Corvallis, OR, United States

Corresponding Author:
Page D Dobbs, PhD
Health, Human Performance and Recreation Department
University of Arkansas
308A HPER Building
Fayetteville, AR, 72701
United States
Phone: 1 476 575 8680
Email: pdobbs@uark.edu

Abstract

Background: Twitter provides a valuable platform for the surveillance and monitoring of public health topics; however, manually categorizing large quantities of Twitter data is labor intensive and presents barriers to identify major trends and sentiments. Additionally, while machine and deep learning approaches have been proposed with high accuracy, they require large, annotated data sets. Public pretrained deep learning classification models, such as BERTweet, produce higher-quality models while using smaller annotated training sets.

Objective: This study aims to derive and evaluate a pretrained deep learning model based on BERTweet that can identify tweets relevant to vaping, tweets (related to vaping) of commercial nature, and tweets with provape sentiment. Additionally, the performance of the BERTweet classifier will be compared against a long short-term memory (LSTM) model to show the improvements a pretrained model has over traditional deep learning approaches.

Methods: Twitter data were collected from August to October 2019 using vaping-related search terms. From this set, a random subsample of 2401 English tweets was manually annotated for relevance (vaping related or not), commercial nature (commercial or not), and sentiment (positive, negative, or neutral). Using the annotated data, 3 separate classifiers were built using BERTweet with the default parameters defined by the Simple Transformer application programming interface (API). Each model was trained for 20 iterations and evaluated with a random split of the annotated tweets, reserving 10% (n=165) of tweets for evaluations.

Results: The relevance, commercial, and sentiment classifiers achieved an area under the receiver operating characteristic curve (AUROC) of 94.5%, 99.3%, and 81.7%, respectively. Additionally, the weighted F1 scores of each were 97.6%, 99.0%, and 86.1%, respectively. We found that BERTweet outperformed the LSTM model in the classification of all categories.

Conclusions: Large, open-source deep learning classifiers, such as BERTweet, can provide researchers the ability to reliably determine if tweets are relevant to vaping; include commercial content; and include positive, negative, or neutral content about vaping with a higher accuracy than traditional natural language processing deep learning models. Such enhancement to the utilization of Twitter data can allow for faster exploration and dissemination of time-sensitive data than traditional methodologies (eg, surveys, polling research).

(JMIR Med Inform 2022;10(7):e33678) doi:10.2196/33678

KEYWORDS

vaping; social media; deep learning; transformer models; infoveillance

Introduction

Background

Since its launch in 2006, Twitter has exploded in popularity to become one of the top social media platforms. As of 2021, the site hosts 192 million daily active users worldwide [1]. The 280-character constraint on a Twitter text post, called a tweet, lends itself well to spontaneous and organic interactions. The candid nature of the tweets provides invaluable data for the public health realm. Patients spend relatively little time with health care professionals, with some only seeing their primary care physician once every other year, and therefore it can be difficult for health care workers to accurately address needs or feelings that patients often find uncomfortable disclosing to others [2].

While Twitter provides a valuable platform for the surveillance and monitoring of public health topics, manually categorizing large quantities of Twitter data by hand presents challenges to identify major trends and sentiments in a timely manner. Machine and deep learning methods have previously been proposed to provide a framework for systematic and automated processing and analysis of Twitter data to develop surveillance systems with applications to public health [3]. While these models achieve high accuracy, they require large sets of annotated data to be trained. By contrast, public pretrained deep learning classification models, such as BERTweet, produce higher-quality models while using smaller annotated training sets [4]. In this study, we derive and evaluate a pretrained deep learning model based on BERTweet that can identify tweets relevant to vaping, tweets of commercial nature, and tweets with provape sentiment. We compare the results of the BERTweet-based classifier with a long short-term memory model (LSTM) to show the improvements a pretrained model has over traditional deep learning approaches.

Traditional Deep Learning

Deep learning is a class of machine learning algorithms that uses multiple layers to progressively extract higher-level features from raw input [4]. Several types of deep learning architectures exist, such as deep neural networks, recurrent neural networks (RNNs), and convolutional neural networks (CNNs). Applications of deep learning include computer vision, speech recognition, natural language processing, and drug design.

In their work, Visweswaran et al [3] found that LSTM models performed particularly well on tweet classification for relevance, sentiment, and commercial nature [3]. An LSTM network is a special kind of RNN capable of learning long-term dependencies [5]. Unlike standard feedforward networks, such as CNNs, LSTMs have a feedback connection. This feedback connection allows the network to not only process a single data point (ie, a word), but also entire sequences of data (ie, sentence or phrase), which make them extremely powerful in classifying sentiment of a message.

Pretrained Transformer Models

Over the last few years, transformer models have been very effective for a large variety of natural language processing tasks. First proposed by Colditz et al [6], transformers use a self-attention mechanism to capture what aspects of a sequence are important in a series of tokens. In simple terms, self-attention mechanisms aim to create real natural language understanding in machines.

In 2018, Google AI Language released the Bidirectional Encoder Representations from Transformers (BERT) model, which improves upon the original transformer model by learning token representations in both directions [7]. In normal transformers, a sequence is analyzed either left to right or right to left, but not in both directions. To achieve this, BERT uses a revamped pretraining procedure that includes masked language model and next sentence prediction objectives [2]. Several BERT models pretrained on a variety of texts, languages, and topics are available freely to the public. This creates a ready-made approach for researchers trying to create models for a number of language tasks, including text classification. Researchers can use BERT in its default settings, or they can apply fine-tuning on a data set closely applicable to the task at hand. For instance, in this study, the created model is fine-tuned on a set of hand-annotated tweets before testing the classification accuracy of the system.

After BERT was introduced, the "Robustly optimized BERT pre-training approach" (RoBERTa) was published [8]. RoBERTa was created out of the authors' experimentation with the default hyperparameters of BERT. They found that BERT was significantly undertrained, and that with some minor changes, the modified BERT model was able to outperform newer and even larger transformer models. Pretraining optimizations in RoBERTa include dynamic masking, large mini-batches, larger byte-pair encodings, and using full sentences across documents. We refer to Liu et al [8] for a more detailed discussion of the optimizations performed in RoBERTa. Like BERT, many pretrained variations of RoBERTa are available online.

BERTweet is a public BERT-based model trained using the RoBERTa pretraining procedure [9]. Released in 2020, it was the first large-scale pretrained language model for English tweets to be released to other researchers for further improvements and novel applications. BERTweet was trained on 850 million English tweets collected from 2012 to 2019, which prepares it well for novel downstream classification tasks on a set of tweets. This pipeline of pretraining on a large text corpus and then fine-tuning the model for classification tasks is called transfer learning [2]. It has been shown that pretraining is integral to model performance on downstream tasks, and it follows that pretraining a model on material that is similar to the texts in the downstream task will yield improved performance. Therefore, having access to a model trained on a large corpus of tweets is invaluable for the creation of a Twitter-based public health surveillance system. We refer to Nguyen et al [9] for a more detailed explanation of how the BERTweet model functions.

Objective

It is our goal to produce an accurate BERTweet-based deep learning classifier that can improve upon existing Twitter surveillance systems that are focused on vaping-related tweets. Additionally, we aim to produce a classifier that is reliable and accurate in assessing a tweet for relevance (relevant or not), sentiment (positive, negative, or neutral), and commercial nature (commercial or not). Leveraging Twitter as a complement to traditional surveillance will allow for real-time identification of changes that can be used by public health practitioners. For example, when positive sentiment toward vaping rises, practitioners may be able to determine the exact reasons why and respond accordingly. Similarly, when there is a notable spike in misinformation about vaping and its effects on health, health experts will be able to act immediately to correct this information [3].

Related Work

Several works have proposed classifiers to classify Twitter data in terms of sentiment. Further, the last few years have seen a surge in publications on creating classifiers to analyze public health trends as depicted on Twitter. Gohil et al [10] performed a review of current sentiment analysis tools available for researchers. They found that while multiple methods existed for analyzing the sentiment of tweets in the health care setting, there is still the need for an accurate and verified tool for sentiment analysis of tweets trained using a health care setting–specific tweet. Edara et al [11] developed an LSTM to classify cancer-related tweets based on the tone of the tweet and compared the results against several traditional machine learning approaches. They found that the LSTM model outperformed all of the other approaches. Ji et al [12] utilized the Twitter platform to monitor the spread of public concern about epidemics by separating personal tweets from new tweets and then further categorizing the personal tweets into those that are negative and nonnegative using a naïve Bayes classifier.

For a general approach to performing a sentiment analysis on Twitter data, Agarwal et al [13] introduced unigram, feature-based, and tree-based models to classify tweets as either a binary task (positive or negative) or a 3-way task (positive, negative, and neutral). Harjule et al [14] proposed another general approach to classifying the sentiment of tweets. The authors analyzed several lexicon and machine learning–based tweet sentiment classifiers on a large group of data sets and found that the machine learning models were more accurate at classifying sentiment. Kharde and Sonawane [15] performed a similar comparative analysis and verified the claim from Harjule et al [14] that machine learning classifiers yield higher accuracy, with the caveat that lexicon-based methods can be more affective in some cases.

Beyond general sentiment and public health monitoring, several studies have looked at using Twitter to monitor trends toward vaping and e-cigarettes [16,17]. Han and Kavuluru [18] implemented several machine learning models, such as support vector machines, logistic regression, and CNNs, to identify marketing and nonmarketing e-cigarette tweets. Further, Myslín et al [19] and Cole-Lewis et al [20] annotated tobacco-related tweets and derived several machine learning classifiers to predict sentiment. Huang et al [21] analyzed tweets using machine learning classifiers to find trend in the commercial nature of tweets relating to vaping. They found that tweets related to e-cigarettes were about 90% commercial and about 10% mentioned smoking cessation. Resende and Culotta [22] derived a sentiment classifier for e-cigarette–related tweets that identified positive and negative tweets with 96% and 70% precision, respectively. Visweswaran et al [3] performed an in-depth comparison of traditional machine learning classifiers (regression, random forest, linear support vector machine, and multinomial naïve Bayes) with deep learning classifiers (CNN, LSTM, LSTM-CNN, and bidirectional LSTM), and found that among all the tested networks, LSTM achieved the highest classification accuracy.

Methods

Data Collection

Tweets were collected continuously from August to October 2019 using the Real-Time Infoveillance of Twitter Health Messages (RITHM) framework [6]. The RITHM framework is an open-source software for collecting and formatting Twitter data. It additionally provides procedures for maximizing the efficiency and effectiveness of subsequent human data coding. The keywords that we used for data collection include *vape*, *vapes*, *vaper*, *vapers*, *vaping*, *juul*, *juuls*, and *juuling*. The vaping-related keywords are based on previous Twitter research [6,10] and, in particular, we included keywords to identify the highly popular e-cigarette brand, JUUL, which had the highest market share at the time from which data were collected [23]. We identified and collected all tweets that matched 1 or more keywords from the list above.

Annotation

After data collection, a random subsample of 2401 English tweets was annotated for relevance (vaping related or not), commercial nature (commercial or not), and sentiment (positive, negative, or neutral). This annotation was done in accordance with the 3-level hierarchical annotation schema, as depicted in Table 1. A tweet was first annotated for relevance. Then, only if the tweet was relevant, was it annotated for commercial nature and sentiment.

A team of 2 trained annotators independently annotated batches of 400 tweets at a time. Adjudicated annotation disagreements were carried out under the presence of the supervising investigator. All annotates codes have a Cohen κ value over 0.70, indicating strong internal agreement among annotators. The full set of 2401 adjudicated annotations and tweet content were used in the training of the classifier models. A detailed description of the annotations can be found in Table 2.

Table 1. Descriptions of labels used for annotating vaping-related tweets.

Labels	Descriptions	Example quotes
Relevant	• Is the tweet in English and related to the vaping topic at hand (eg, vape use or users, vaping devices, or products)?	
Not relevant	• Typically, non-English tweets or tweets that referenced vaping cannabis products specifically.	
Commercial	• Is the tweet an advertisement/marketing for vaping products?	• *Today only! Buy one JUUL get the second half price with our online coupon code #JUUL4LIFE*
Noncommercial	• Includes tweets that demonstrate favorability toward a product but do not directly advocate for purchasing it.	

Positive	• The tweet is associated with positive emotions or contexts regarding vaping.	• The tweeter is currently, or has recently used, or is going to vape:
		  • *Currently juuling in the bathroom at school!*
		• The tweeter shows positivity or neutral acceptance from others' usage or others' positive comments about vaping:
		  • *Just got Hannah to try vaping for the first time! She loved it.*
		• The tweeter mentions a vape pen in association with other positive aspects of society or popular culture.
		  • *We need a Disney princess that rips her JUUL in the middle of a serious conversation.*
		• The tweeter asks a question using first-person pronouns:
		  • *Where can I buy a JUUL?*

Negative	• The tweet is associated with negative emotions or contexts regarding vaping.	• The tweeter believes smoking a vape is disgusting, uncool, or unattractive:
		  • *Cannot believe everyone is smoking JUULs these days. I think it's disgusting.*
		• The tweeter criticizes/ridicules others for using a vape:
		  • *ur mcm says 'cigarettes are gross' yet is addicted to nicotine through cool cucumber flavored JUUL pods.*
		• The tweeter prefers to use a different substance, such as cigarettes or marijuana:
		  • *Tried a JUUL today for the first time but I still prefer cigarettes over it.*

Neutral		• The tweet is factual but not opinionated or is a question about unbiased facts/information about vaping:
		  • *They are selling JUUL pens at my local tobacco shop for anyone interested.*
		  • *What is a JUUL?*
		  • *Is a JUUL better than tobacco?*

Table 2. Description of annotated training and test data sets (N=2401).[a]

Targets	Number of tweets with a positive target, n (%)	Number of tweets with a negative target, n (%)	Number of tweets with a neutral target, n (%)
Relevance	Relevant: • Total: 1802 (75.05) • Training: 1637 (90.84) • Test: 165 (9.16)	Nonrelevant: • Total: 599 (24.95) • Training: 524 (87.48) • Test: 75 (12.52)	N/A[b]
Commercial	Commercial: • Total: 117 (4.87) • Training: 106 (90.60) • Test: 11(9.40)	Noncommercial: • Total: 1685 (70.18) • Training: 1516 (89.97) • Test: 169 (10.03)	N/A
Sentiment	Positive: • Total: 172 (7.16) • Training: 158 (91.86) • Test: 14 (8.14)	Negative: • Total: 130 (5.41) • Training: 119 (91.54) • Test: 11 (8.46)	Neutral: • Total: 1372 (57.14) • Training: 1229 (89.58) • Test: 143 (10.42)

[a]Percentages may not add up to 100% as classification was made for sentiment only if the tweet was relevant.

[b]Sentiment-only code with neutral target.

LSTM Model

We will briefly recount the process explained in Visweswaran et al [3] to train and evaluate an LSTM model to classify a tweet related to vaping as relevant; commercial; and if it was positive, negative, or neutral in sentiment. Our LSTM model was developed using the built-in functionality of the TensorFlow machine learning library. We utilized rectified linear unit (ReLU) as the activation function of the hidden layers and the sigmoid activation function for the output layer. Additionally, we utilized binary cross entropy as the loss function with Adam as the optimizer. In accordance with Visweswaran et al's study [3], we used nondomain-specific GloVe word vectors.

After first testing a 70/30 split to create the relevance classifier and testing random splits to prevent over fitting, we found optimal results with a 90/10 split of the entire annotated data set, as all tweets were coded as either relevant or nonrelevant. We used the 90% split (n=1637) to train the LSTM relevance classifier, and then tested on the remaining 10% (n=165). We trained the model for 5 epochs using a batch size of 64. Both the commercial and sentiment classifiers followed the same training and testing procedures as the relevance classifier. The one difference being that only tweets labeled as relevant were used in the commercial and sentiment data sets. All nonrelevant tweets were filtered out and discarded.

BERTweet

To create a classifier for relevance, 90% of the tweets labeled as either relevant (n=1637) or nonrelevant (n=524) were used to fine-tune the BERTweet model, and the remaining 10% were used to test the final model (relevant n=165; nonrelevant n=75). This splitting, training, and testing process was repeated multiple times with random splits, and the accuracy results are the averages of each individual run. BERTweet was trained for 20 epochs with a batch size of 32 and a learning rate of 5×10^{-5}. All other hyperparameters were left to the default values according to Simple Transformers API, which was used to accelerate the fine-tuning process for BERTweet and decrease the amount of proprietary code needed to be written. Tokenization of input tweet text was handled by Simple Transformers API, which automatically uses the BERTweet tokenizer defined by the creators of the model.

To create the commercial and sentiment classifiers, annotated tweets were first filtered by relevance; nonrelevant tweets were discarded for these classifiers, and tweets marked relevant were then split into training and testing sets, and models were fine-tuned using the same process as the relevance classifier.

Results

Overview

We compared the performance of the LSTM and BERTweet classifiers in terms of F1 and AUROC scores. Additionally, each score is the average of 3 different testing iterations of the respective models. F1 is a function of precision and recall:

F1 = 2×(Precision × Recall)/(Precision + Recall) (1)

Precision = True positive/(True positive + False positive) (2)

Recall = True positive/(True positive + False negative) (3)

For F1, values closer to 1 on a scale of 0 to 1 indicate good balance between precision and recall.

AUROC is the measure of the discrimination of the models, that is, for example, how well a classifier differentiates between positive, negative, and neutral tweets. The larger the AUROC score is, the better the model performs.

Relevance

With regard to classifying a tweet as relevant or nonrelevant, the BERTweet classifier obtained an F1 score of 0.976 and an AUROC score of 0.945. The LSTM classifier achieved an F1 score of 0.924 and an AUROC score of 0.924. All runs of the

BERTweet classifier achieved higher F1 and AUROC scores than the LSTM model.

Commercial

In classifying commercial tweets (commercial or noncommercial) the BERTweet classifier performed well with an F1 score of 0.990 and an AUROC of 0.993. Of all classes, the BERTweet performed best in commercial classification. The LSTM model produced a lower F1 score of 0.727 and a lower AUROC score of 0.903 in comparison to the BERTweet model (Table 3).

Table 3. Comparison of BERTweet and LSTM[a] F1 and AUROC[b] scores.

Classifier/metric	Relevance	Commercial	Sentiment
BERTweet			
F1	0.976	0.990	0.861
AUROC	0.945	0.993	0.817
LSTM			
F1	0.924	0.727	0.250
AUROC	0.924	0.903	0.776

[a]LSTM: long short-term memory.

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

Sentiment

Both the BERTweet and LSTM models performed the worst in the classification of sentiment (positive, negative, or neutral). BERTweet obtained an F1 of 0.861 with an AUROC of 0.817. The LSTM model had an F1 of 0.250 with an AUROC of 0.776.

Discussion

Principal Findings

This is the first study to use BERTweet to classify vaping-related tweets. Based on the analyses, we found that pretrained deep learning classifiers such as BERTweet perform exceptionally well at classifying a tweet as being relevant to vaping, being a commercial-natured tweet about vaping, as well as the sentiment of a tweet toward vaping. Compared with the LSTM classifier, the BERTweet classifier had AUROC values of 0.945, 0.993, and 0.817 for relevance, commercial nature, and sentiment, respectively. In general, these results show that pretrained classifiers can be utilized to monitor social medial platforms such as Twitter for public health trends. Such enhancement to the utilization of Twitter data can allow for faster exploration and dissemination of time-sensitive data than traditional methodologies such as surveys and polling research.

Practically, our work also serves to provide public health practitioners with vaping-related information on Twitter. For example, if there is an increase in positive sentiments of tweets, public health practitioners may find that a particular area is ready for policy change. Using the classification results, practitioners can also understand how many tweets are related to marketing of vaping and the relationship between sentiment of people and number of commercial tweets.

Limitations

This study was performed with several limitations. First, a relatively small set of 2401 tweets was annotated by hand. Compared with another study [3], this was just over half the size of the data set they annotated. While the set was small, it was enough to produce accurate results when using BERTweet, which is another testament to the power that pretrained transformer models have. However, this limitation does make it difficult to compare results directly with Visweswaran et al [3]. Second, while we matched keywords with Visweswaran et al's study [3], due to the evolving nature of language on Twitter, our collection methods could have overlooked new products or trends that have become prevalent on the Twitter platform. Third, we analyzed tweets that were written in English only. This limits the populations from which this classifier can accurately classify tweets. For instance, other countries may have different sentiments toward vaping that were not supported in this study. Finally, the date range of the tweets was limited to a 2-month time span, which limits the generalizability of the classifier over time, and therefore, more analysis would need to be performed to discover the longevity of the classifier.

Future Research

Several different research endeavors relating to utilizing pretrained deep learning models to classifying tweets could be explored. First, we could expand from analyzing only English tweets to diversify this work for global regions and languages. Additionally, analysis on the number of annotated tweets needed to create an equivalent LSTM model could be performed to give substantial evidence that pretrained models provide evidence just beyond higher classification accuracy. Finally, the BERTweet model developed in this paper could be extended to create a real-time analysis platform for sentiment toward vaping to better inform public health officials, allowing them to understand the impacts of current and future policy interventions.

Conclusion

In this study, we produced a deep learning classification model based on BERTweet that was able to classify a vaping-related tweet along several viewpoints such as relevance (relevant or not), commercial nature (commercial or not), and sentiment (positive, negative, or neutral). We then compared the classification performance of the BERTweet model with that of an LSTM model for the classification of 2401 hand-coded tweets. We found that in all classification cases BERTweet achieved higher levels of accuracy. The strong performance of BERTweet shows that it can increase the ability to accurately monitor social platforms such as Twitter with regard to public health trends such as vaping.

Acknowledgments

The authors thank Eric Schisler, Caroline Stokenbury, and Emily Abby Norton for data annotation. This work was supported by awards from the National Cancer Institute of the National Institutes of Health (R01-CA225773), the National Library of Medicine of the National Institutes of Health (R01- LM012095), and the National Science Foundation (ACI-1548562 and ACI-1445606 to the Pittsburgh Supercomputing Center). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

Conflicts of Interest

None declared.

References

1. Ying L. 10 Twitter Statistics Every Marketer Should Know in 2021. Infographic. URL: https://www.oberlo.com/blog/twitter-statistics [accessed 2021-04-16]
2. Baker W. Using Large Pre-Trained Language Models to Track Emotions of Cancer Patients on Twitter. Computer Science and Computer Engineering Undergraduate Honors Theses. URL: https://scholarworks.uark.edu/csceuht/92/ [accessed 2022-05-24]
3. Visweswaran S, Colditz JB, O'Halloran P, Han N, Taneja SB, Welling J, et al. Machine learning classifiers for Twitter surveillance of vaping: comparative machine learning study. J Med Internet Res 2020 Aug 12;22(8):e17478 [FREE Full text] [doi: 10.2196/17478] [Medline: 32784184]
4. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444 [FREE Full text] [doi: 10.1038/nature14539] [Medline: 26017442]
5. Zhao R, Yan R, Wang J, Mao K. Learning to monitor machine health with convolutional bi-directional LSTM networks. Sensors 2017 Jan 30;17(2):273 [FREE Full text] [doi: 10.3390/s17020273] [Medline: 28146106]
6. Colditz JB, Chu K, Emery SL, Larkin CR, James AE, Welling J, et al. Toward real-time infoveillance of Twitter health messages. Am J Public Health 2018 Aug;108(8):1009-1014. [doi: 10.2105/ajph.2018.304497]
7. Devlin J, Chang M, 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. 2019 Presented at: Association for Computational Linguistics; 2019; Minneapolis, Minnesota p. 4171-4186 URL: https://www.aclweb.org/anthology/N19-1423 [doi: 10.18653/v1/n18-2]
8. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pre-training approach. ArXiv 0219:abs/1907.1 2019:1692. [doi: doi.org/10.48550/arXiv.1907.11692]
9. Nguyen D, Vu T, Nguyen A. BERTweet: A Pre-trained Language Model for English Tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. BERTweet: A Pre-trained Language Model for English Tweets. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processingystem Demonstrations. Association for Computational Linguistics 2020 Oct; 2020 Presented at: Association for Computational Linguistics; July 5-10, 2020; Virtual p. A URL: https://www.aclweb.org/anthology/2020.emnlp-demos.2/ [doi: 10.18653/v1/2020.emnlp-demos.2]
10. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care tweets: review of the methods used. JMIR Public Health Surveill 2018 Apr 23;4(2):e43 [FREE Full text] [doi: 10.2196/publichealth.5789] [Medline: 29685871]
11. Edara DC, Vanukuri LP, Sistla V, Kolli VKK. Sentiment analysis and text categorization of cancer medical records with LSTM. J Ambient Intell Human Comput 2019 Jul 16:1. [doi: 10.1007/s12652-019-01399-8]
12. Ji X, Chun SA, Wei Z, Geller J. Twitter sentiment classification for measuring public health concerns. Soc Netw Anal Min 2015 May 12;5(1):13 [FREE Full text] [doi: 10.1007/s13278-015-0253-5] [Medline: 32226558]
13. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R. Sentiment Analysis of Twitter Data. In: Department of Computer Science Columbia University. 2011 Presented at: In Proceedings of the workshop on language in social media; June 2011; New York, NY p. 30-38 URL: http://www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf
14. Harjule P, Gurjar A, Seth H, Thakur P. Text Classification on Twitter Data. 2020 Presented at: 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things; February 7-8, 2020; Jaipur, India p. 160-164. [doi: 10.1109/ICETCE48199.2020.9091774]
15. Kharde V, Sonawane S. Sentiment analysis of Twitter data: a survey of techniques. IJCA 2016 Apr 15;139(11):5-15. [doi: 10.5120/ijca2016908625]

16. Colditz JB, Welling J, Smith NA, James AE, Primack BA. World vaping day: contextualizing vaping culture in online social media using a mixed methods approach. Journal of Mixed Methods Research 2017 Apr 09;13(2):196-215. [doi: 10.1177/1558689817702753]
17. Sidani JE, Colditz JB, Barrett EL, Shensa A, Chu K, James AE, et al. I wake up and hit the JUUL: Analyzing Twitter for JUUL nicotine effects and dependence. Drug and Alcohol Dependence 2019 Nov;204:107500. [doi: 10.1016/j.drugalcdep.2019.06.005]
18. Han S, Kavuluru R. Exploratory Analysis of Marketing and Non-Marketing E-cigarette Themes on Twitter. 2016 Nov 11 Presented at: International Conference on Social Informatics; 2016; Bellevue. [doi: 10.1007/978-3-319-47874-6_22]
19. Myslín M, Zhu S, Chapman W, Conway M. Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res 2013 Aug 29;15(8):e174 [FREE Full text] [doi: 10.2196/jmir.2534] [Medline: 23989137]
20. Cole-Lewis H, Varghese A, Sanders A, Schwarz M, Pugatch J, Augustson E. Assessing electronic cigarette-related tweets for sentiment and content using supervised machine learning. J Med Internet Res 2015 Aug 25;17(8):e208 [FREE Full text] [doi: 10.2196/jmir.4392] [Medline: 26307512]
21. Huang J, Kornfield R, Szczypka G, Emery SL. A cross-sectional examination of marketing of electronic cigarettes on Twitter. Tob Control 2014 Jul 16;23 Suppl 3(suppl 3):iii26-iii30 [FREE Full text] [doi: 10.1136/tobaccocontrol-2014-051551] [Medline: 24935894]
22. Resende E, Culotta A. A Demographic and Sentiment Analysis of E-cigarette Messages on Twitter. In: Computer Science Department, Illinois Institute of Technology. 2015 Presented at: 6th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics; September 9-12, 2015; Atlanta URL: http://cs.tulane.edu/~aculotta/pubs/resende15demographic.pdf
23. Huang J, Duan Z, Kwok J, Binns S, Vera LE, Kim Y, et al. Vaping versus JUULing: how the extraordinary growth and marketing of JUUL transformed the US retail e-cigarette market. Tob Control 2019 Mar;28(2):146-151 [FREE Full text] [doi: 10.1136/tobaccocontrol-2018-054382] [Medline: 29853561]

Abbreviations

API: application programming interface
AUROC: area under the receiver operating characteristic curve
BERT: bidirectional encoder representations from transformer
CNN: convolutional neural network
LSTM: long short-term memory
ReLU: rectified linear unit
RITHM: Real-time Time Infoveillance of Twitter Health Messages
RNN: recurrent neural network
RoBERTa: robustly optimized BERT pre-training approach

Edited by A Mavragani, C Lovis; submitted 18.09.21; peer-reviewed by S Chen, S Doan, X Zhou, S Kim; comments to author 30.12.21; revised version received 21.03.22; accepted 08.05.22; published 21.07.22.

Please cite as:
Baker W, Colditz JB, Dobbs PD, Mai H, Visweswaran S, Zhan J, Primack BA
Classification of Twitter Vaping Discourse Using BERTweet: Comparative Deep Learning Study
JMIR Med Inform 2022;10(7):e33678
URL: https://medinform.jmir.org/2022/7/e33678
doi:10.2196/33678
PMID:35862172

Original Paper

Impact of a Clinical Text–Based Fall Prediction Model on Preventing Extended Hospital Stays for Elderly Inpatients: Model Development and Performance Evaluation

Yoshimasa Kawazoe[1], MD, PhD; Kiminori Shimamoto[1], MEng; Daisaku Shibata[1], PhD; Emiko Shinohara[1], PhD; Hideaki Kawaguchi[2,3], MD, PhD; Tomotaka Yamamoto[4], MD, PhD

[1]Artificial Intelligence in Healthcare, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[2]Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[3]Quantum Computing Center, Keio University, Tokyo, Japan

[4]Department of Performance Monitoring and Risk Management, The University of Tokyo Hospital, Tokyo, Japan

Corresponding Author:
Yoshimasa Kawazoe, MD, PhD
Artificial Intelligence in Healthcare
Graduate School of Medicine
The University of Tokyo
7-3-1 Hongo
Bunkyo-ku
Tokyo, 113-0033
Japan
Phone: 81 03 3815 5411 ext 34432
Email: kawazoe@m.u-tokyo.ac.jp

Abstract

Background: Falls may cause elderly people to be bedridden, requiring professional intervention; thus, fall prevention is crucial. The use of electronic health records (EHRs) is expected to provide highly accurate risk assessment and length-of-stay data related to falls, which may be used to estimate the costs and benefits of prevention. However, no studies to date have investigated the extent to which hospital stays could be shortened through fall avoidance resulting from the use of prediction tools.

Objective: We first estimated the extended length of hospital stay caused by falls among elderly inpatients. Next, we developed a model that predicts falls using clinical text as input and evaluated its accuracy. Finally, we estimated the potentially shortened hospital stay that would be made possible by appropriate interventions based on the prediction model.

Methods: Patients aged 65 years or older were selected as subjects, and the EHRs of 1728 falls and 70,586 nonfalls were subjected to analysis. The extended-stay lengths were estimated using propensity score matching of 49 associated variables. Bidirectional encoder representations from transformers and bidirectional long short-term memory methods were used to predict falls from clinical text. The estimated length of stay and the outputs of the prediction model were used to determine stay reductions.

Results: The extended length of hospital stay due to falls was estimated to be 17.8 days (95% CI 16.6-19.0), which dropped to 8.6 days when there were unobserved covariates at an odds ratio of 2.0. The accuracy of the prediction model was as follows: area under the receiver operating characteristic curve, 0.851; F-value, 0.165; recall, 0.737; precision, 0.093; and specificity, 0.839. When assuming interventions with 25% or 100% effectiveness against cases where the model predicted a fall, the stay reduction was estimated at 0.022 and 0.099 days/day, respectively.

Conclusions: The accuracy of the prediction model using clinical text is considered to be higher than the prediction accuracy of conventional assessments. However, our model's precision remained low at 9.3%. This may be due, in part, to the inclusion of cases in which falls did not occur because of preventative interventions during hospitalization. Nonetheless, it is estimated that interventions for cases when falls were predicted will reduce medical costs by 886 Yen/day (~US $6.50/day) of intervention, even if the preventative effect is 25%. Limitations include the fact that these results cannot be extrapolated to short- or long-term hospitalization cases, and that this was a single-center study.

(JMIR Med Inform 2022;10(7):e37913) doi:10.2196/37913

KEYWORDS

accidental falls; accident prevention; inpatients; machine learning; natural language processing; propensity score; hospital; elderly; prediction model; patient; risk assessment

Introduction

Falls in older adults represent a serious social issue, as they can cause grave injuries that may result in the victim becoming bedridden and in need of professional care. These risks also exist within medical institutions, where falls among elderly patients considerably contribute toward extended hospital stays, increased costs, and decreased quality of life. The incidence of falls in Japanese hospitals is reported to be 4.40/1000 patient days, and the incidence of falls accompanying disabilities is reported to be 0.29/1000 days [1], which is comparable to the respective values of 3.56/1000 and 0.93/1000 patient days reported in the United States [2].

Risk factors of falls include intrinsic variables such as muscle weakness, history of falls, gait deficit, balance deficit, utilization of assistive devices, visual deficit, arthritis, impaired activities of daily living, depression, and cognitive impairment. Extrinsic risk factors include specific medications, polypharmacy, dark lighting, loose carpets, and a lack of bathroom safety devices [3]. Risk assessment tools are often used by medical institutions to assess the susceptibility to falling based on these risk factors. Morita et al [4] investigated the predictive performance of risk factors using a multivariate logistic regression model on 19 fall-related explanatory variables: (1) age of 70 years or older, (2) previous history of falls, (3) decreased lower-limb muscle strength, (4) use of a cane or walker, (5) wobbling, (6) disturbed behaviors, (7) strong independence, (8) decreased comprehension, (9) overestimation of self, (10) need for someone else to stand by during excretion, (11) need for assistance during excretion, (12) nocturia, (13) narcotics, (14) antidepressants, (15) laxatives, (16) sleep stabilizers, (17) antihypertensive agents, (18) clinical department or room transfers, and (19) oxygen inhalation drip. The results showed that the prediction accuracy reflected an area under the receiver operating characteristic curve (AUC) value of 0.822, a recall of 74.5%, and a specificity of 79.6%. Tools for assessing fall risk factors are commonplace, such as the renowned Morse Fall Scale [5], St. Thomas Risk Assessment Tool [6], Resident Assessment Instrument [7], and Hendrich Fall Risk Model [8]. Their use requires manual responses by health care professionals. Hence, the tendency is for the number of actions to be small, which improves clinician interpretability but may negatively affect the results. Furthermore, there remain significant differences in the input terms applied by medical professionals to electronic health records (EHRs). However, there are expectations that computers will be able to help predict falls with high accuracy and thus improve patient safety.

Among EHR types, clinician-input text data (ie, clinical text) contain information relating to falls, including patient condition. Previous research has applied natural language processing (NLP) techniques to EHR text to classify entries related to falls and to predict whether patients would fall during hospitalization. Toyabe [9] investigated the frequency of true fall event entries from progress notes, discharge summaries, image orders, and incident reports via text mining using dependency parsing. Bjarnadottir et al [10] reported that information on true fall events was most frequently recorded in progress notes (100%), incident reports (65.0%), and image orders (12.5%). They further analyzed intensive care unit nursing records from the Medical Information Mart for Intensive Care database, finding meaningful information related to the risk and prevention of falls [10]. Nakatani et al [11] extracted the nursing records of 335 fallen and 408 unfallen individuals from the EHR system of an acute care hospital, and reported the accuracy of fall prediction by morphological analysis and machine learning methods. The average AUC value from five independent experiments was 0.834 (SD 0.005), and the prediction model contained many words closely related to known risk factors [11]. These studies showed that entries related to patient falls can be extracted from EHRs using NLP, but only with a certain level of accuracy. Nevertheless, fall probability can be predicted during hospitalization, and the results suggest that it may be a useful risk management tool.

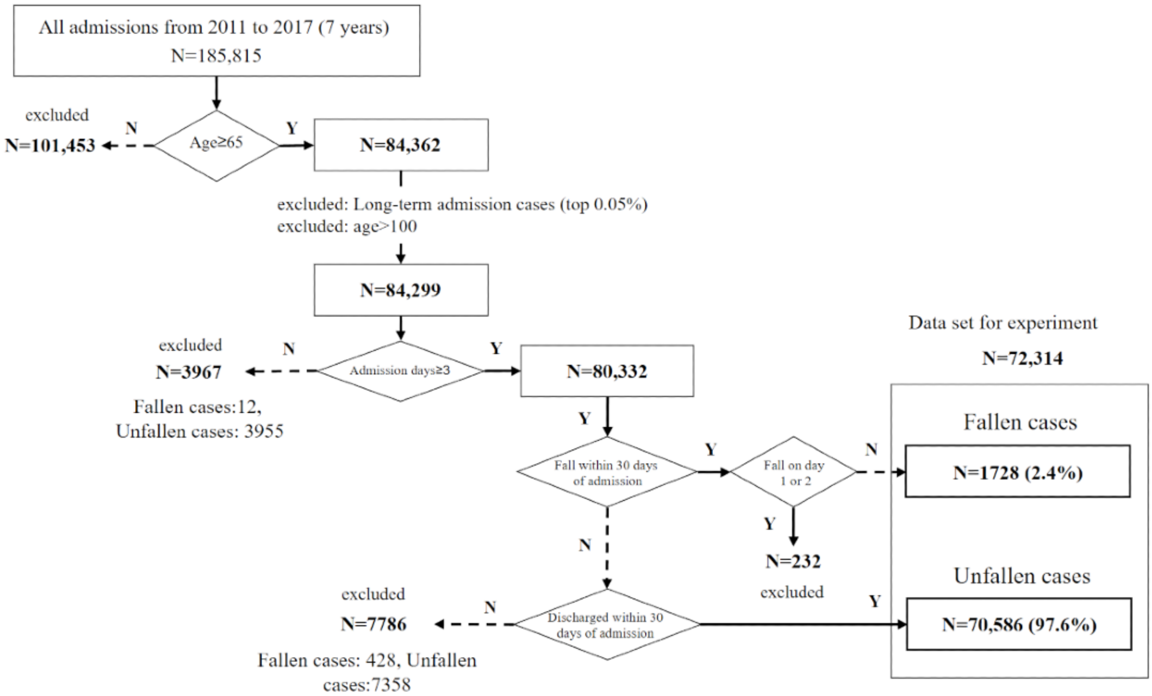
To the best of our knowledge, no studies have investigated the extent to which hospital stays could be shortened through fall avoidance resulting from the use of prediction tools. If the extended hospital stay by a fall can be quantitatively classified, then the costs of developing predictive accuracy and preventative measures can be estimated based on the performance of these aspects. Therefore, in this study, the subject demographic was narrowed down to elderly inpatients, and we estimated the extended length of hospital stay caused by falls using the propensity score matching method. In the United States, it has been reported that patients injured by falls during hospitalization had an average stay extension of 6-12 days, incurring additional hospitalization costs of US $13,316 [2,12-14]. However, differences in medical systems and patient demographics compared with those in Japan prohibit the generalization of these figures. Thus, we conducted this investigation anew for Japan. Specifically, we compared the length of hospital stay in fallen and unfallen groups with adjustments made for patient demographics, which were obtained by propensity score matching using 49 covariates that are considered to influence both falls and length of hospital stay to ultimately estimate the average treatment effect on treatment (ATET). Additionally, the effect of unobserved covariates on ATET was investigated using sensitivity analysis. Next, we used clinical entries made at the time of hospitalization of an elderly inpatient with annotations of the presence or absence of a fall to create a data set. The proposed method was built upon bidirectional encoder representations from transformers (BERT) [15], a general-purpose NLP model. Predictions were made by inputting the clinical text up to the second day of hospitalization and setting the objective of prediction as whether the patient would fall within the next 30 days of hospitalization. Finally, the results were used as a basis to estimate the shortened length of hospital stay and reduced medical costs as a result of fall prevention measures. We then investigated the potential costs incurred in implementing the model and the associated precautionary measures.

Methods

Data Set

Among all hospitalizations that overlapped in the 7-year period from January 1, 2011, to December 31, 2017, patients aged 65 years or older at the time of hospitalization were included. However, those with a hospitalization period that was extensive (top 0.05% number of days) and those aged 100 years or older were excluded as outliers. As a result, a total of 84,299 hospitalizations were obtained from the EHR system of the University of Tokyo Hospital. Results of comparing these hospitalizations with the occurrence of falls that were reported in incident reports indicated that 2402 falls were reported and 82,089 were not. In the second half of this study, we used clinical text from the first 2 days of hospitalization to predict the occurrence of falls from subsequent days. However, it was considered that predicting future falls from 2 days' worth of clinical text would be difficult. Therefore, the prediction period was limited to the period from day 3 to day 30 of hospitalization, during which falls resulted in the hospitalization being classified as "fallen hospitalization" and the nonoccurrence of a fall resulted in the hospitalization being classified as "unfallen hospitalization." Experimental subjects included 72,314 cases (1727 fallen and 70,586 unfallen) after excluding those among the previously mentioned 84,299 that did not meet all criteria. Figure 1 shows the extraction flow of the experiment subjects.

Figure 1. Flowchart of data collection and selection procedure.



Ethics Approval

All experiments and data collection were approved by the institutional review board at the University of Tokyo Hospital (approval number 201919NI). All experiments described below were carried out under relevant ethical guidelines and regulations.

Variables

Occurrence of Falls

We used falls that were reported in incident reports, which have a high degree of completeness, as such reporting is mandatory. These reports distinguish between falls during walking and falls from bed, including mild and severe classifications. However, these cases were not classified separately in this study.

Risk Factors for Falls

Factors other than falls influence the length of hospital stay; thus, determining the extended length of stay caused by falls requires the elimination of covariates that affect both falls and the length of hospital stay. A total of 49 covariates were identified by propensity score matching to adjust their effects on diagnosis procedure combinations (DPCs), incident report data sets, blood test results, and prescription drugs.

DPCs contain information entered by medical staff for all inpatients regarding diagnostic procedures. We used several factors influencing falls and length of stay, including age, gender, consciousness disorder at admission, emergency transport at admission, dementia at admission, purpose for chemotherapy at admission, and the disease that triggered hospitalization. The latter was coded using the 10th revision of the International Statistical Classification of Diseases and Related Health Problems, and 17 types of dummy variables were developed based on the major classification code (A to U). All variables, apart from age, were treated as binary variables.

Incident reports comprise systematic reviews showing that past falls are high predictors of subsequent falls [16-18]. Previous history includes cases of hospitalization where falls were reported in the respective incident report.

Blood test results were used to determine the presence or absence of anemia and poor nutritional status, which are known risk factors that affect falls. Seven variables were adapted as test results reflecting these risks, including decreased hemoglobin, decreased protein/albumin, increased urea nitrogen (suggesting chronic kidney disease), increased liver enzymes, decreased blood glucose, abnormal electrolytes, and elevated C-reactive protein. Each threshold value was set as a binary variable. Table A1 in Multimedia Appendix 1 provides the thresholds for each variable.

Prescription drugs in this case include hypnotics and antipsychotics, which have been identified as contributors to falls [3]. Binary variables were set for these drugs using the criteria of the drug corresponding to its three-digit drug efficacy classification code from the subcategory “87 drugs and related products” of the Japanese standard product classification. The following 12 drug groups were considered: hypnotics, antiepileptics, nonsteroidal anti-inflammatory drugs (NSAIDs), anti-Parkinson drugs, antipsychiatric drugs, other neuroactive drugs, muscle relaxants, diuretics, antihypertensive drugs, diabetes drugs, narcotics, and laxatives. Furthermore, polypharmacy is known to contribute to falls. This includes cases in which 10 or more of the above-mentioned drugs were prescribed simultaneously.

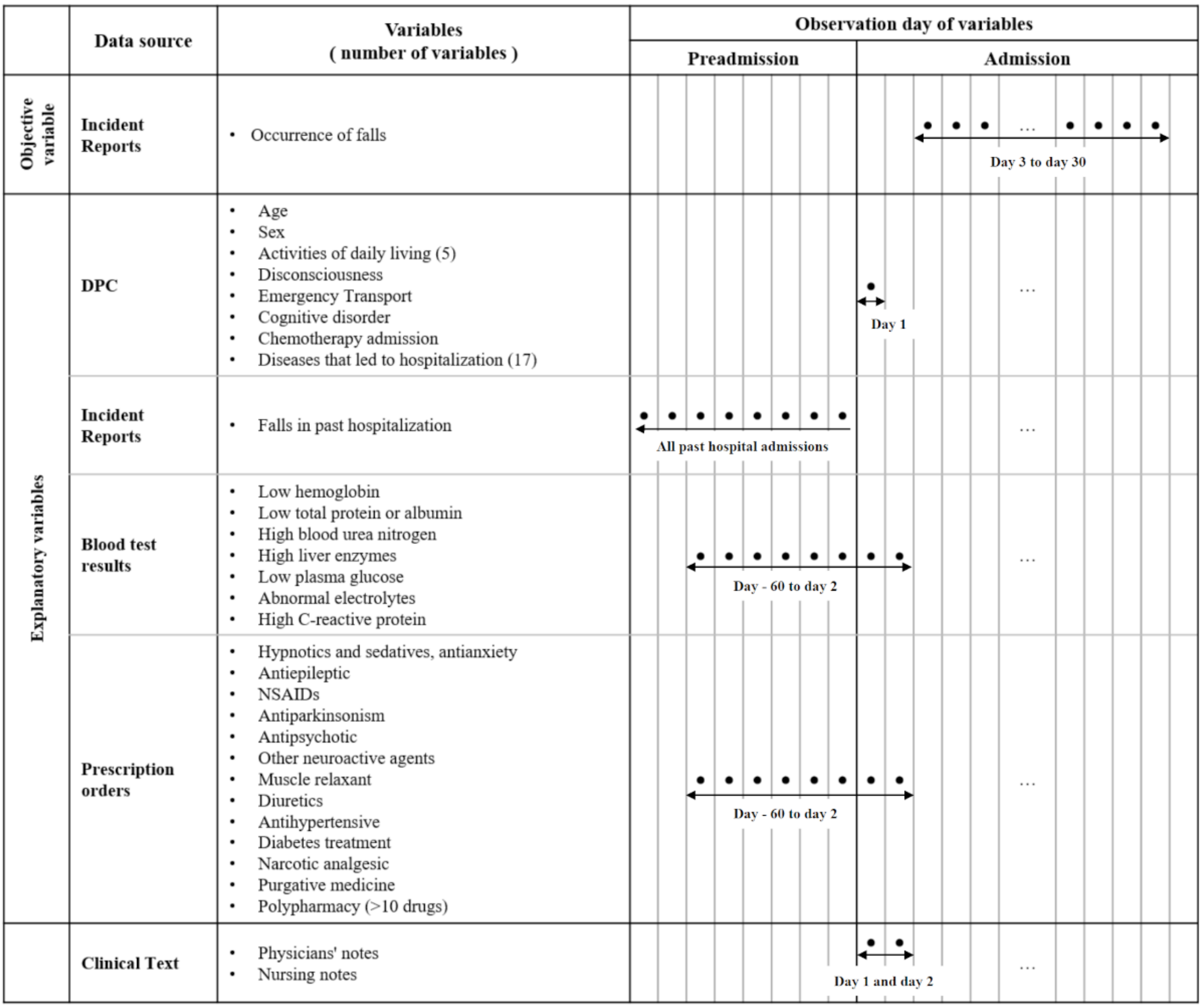
Clinical Notes

Clinical text was used as input to the fall prediction model without distinguishing the type of clinician.

Period of Data Extraction

It was desirable to obtain the above-mentioned 49 variables from the clinical text entered on the day of hospitalization. However, there was a concern that the number of missing values would increase if the target period for variable extraction was limited to that day. Therefore, variables relating to blood test results and prescribed drugs were taken from the 60 days before hospitalization to the second day of hospitalization. For the clinical text used as input, the subject period included the first and second days. Figure 2 shows the variables used and their target periods.

Figure 2. Period and variables of data extraction. DPC: diagnosis procedure combination; NSAID: nonsteroidal anti-inflammatory drug.



Missing Values

Figure 2 shows that there were no missing values found in the DPC data. However, the blood test results and prescription orders showed cases in which these entries did not exist during the target period. These missing values were estimated using the multiple imputation by chained equation (MICE) method 20 times [19].

ATET Estimation by Propensity Score Matching

The extended length of hospital stay caused by falling was estimated using propensity score matching [20]. Matching unfallen cases with tendencies similar to those of fallen cases and comparing the lengths of hospital stays between the two groups were achieved by repeating this method, resulting in an ATET estimation of the effect of falls on the length of stay. The propensity score was obtained using a multivariate logistic regression model with the 49 explanatory variables and the presence or absence of falls as the objective variable. Some variables had missing values, as described above. Thus, values estimated from 20 MICE calculations were used as inputs to the multivariate logistic regression model. The one-to-one nearest-neighbor matching with replacement method [21] was used to match the fallen and unfallen groups. Here, propensity score matching estimations strongly assumed that the fall allocation depended only on the explanatory variables used; however, not all variables were observed. Therefore, the effects of the unobserved ATET covariates were also investigated using sensitivity analysis to the maximum *P* value and minimum Hodges-Lehmann point estimate [22] according to Rosenbaum's [23,24] procedure. Here, the null hypothesis is *fall events do not influence the extended length of hospital stay*, and the *P* value is the value of the one-sided Wilcoxon signed-rank sum test.

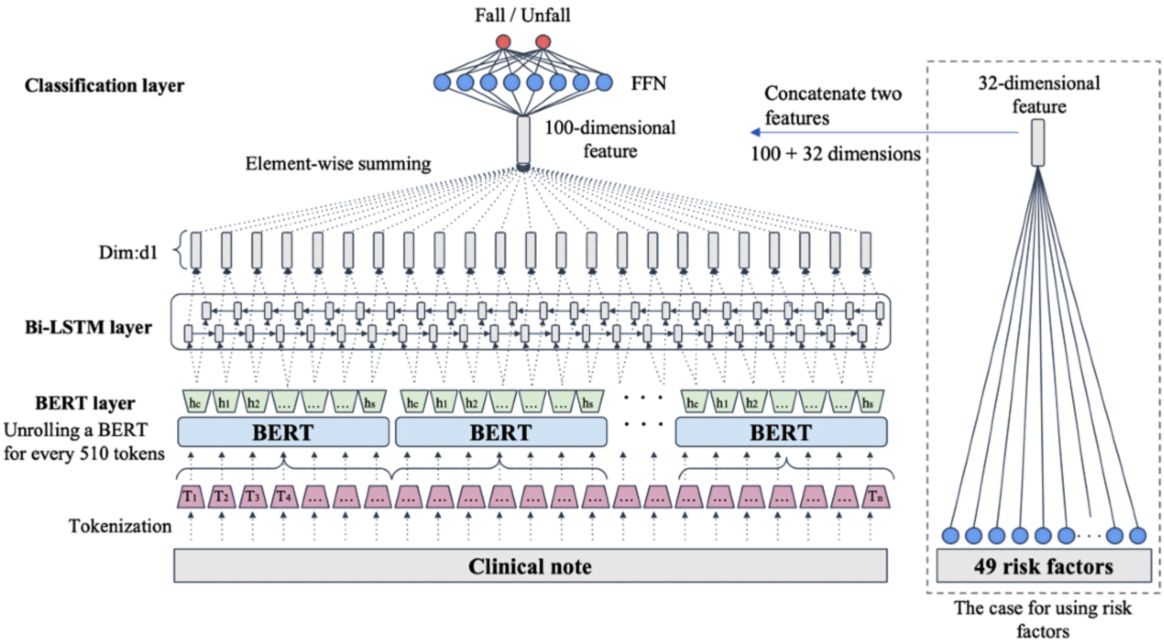
NLP Fall Prediction From Clinical Text

Fall prediction learning and evaluation were performed on 71,943 cases, excluding 371 cases with missing clinical text from the 72,314 experimental data, as shown in Figure 1. Cases in which hospitalization occurred between 2011 and 2016 (1500 fallen and 60,060 unfallen) were used as learning data; cases in which the day of hospitalization was in 2017 (228 fallen and 10,158 unfallen) were divided into two groups so that the number of fall cases was even. Subsequently, two-fold cross-validation was performed using alternating models for model selection and evaluation. The AUC, F-value, recall, precision, and specificity were used as evaluation indicators, and the 2-time average value was used for performance evaluation.

We adopted a model that leveraged UTH-BERT [25], which was pretrained on Japanese clinical text using bidirectional long short-term memory (Bi-LSTM) [26] to predict falls. The model-learning process involved dividing clinical text into vocabulary tokens unique to UTH-BERT, and adding the special tokens for classification ([CLS]) and separation ([SEP]) to the beginning and end of sequences. In BERT, a fixed-length sequence of up to 512 tokens is taken as input, and the embeddings of [CLS] and those corresponding to each input token are considered as output. [CLS] embeddings are used as input to the classifier, after which fine tuning is performed [15]. Owing to this limitation, it was proposed to divide the input by 512 so that the tokens could be input sequentially. In this way, [CLS] embeddings could be output sequentially to a classifier (eg, recurrent neural network) that can use the series to classify sentences consisting of longer sequences [27]. However, [CLS] embeddings do not always aggregate the contents of an entire sentence, and the likelihood of reduced performance was a concern [28]. Therefore, we instead adopted a model in which the output of BERT token embedding was input into a single-layer Bi-LSTM so that a 100-dimensional feature value output could be obtained. This was then used to perform the binary classification of fallen and unfallen cases. Furthermore, the structure provided that a 32-dimensional feature value would be obtained by linearly converting the 49 fall-related variables from the clinical text, followed by their binary classification. Figure 3 shows the structure of the BERT+Bi-LSTM network.

The median number of characters in the clinical text of fallen and unfallen cases was 4144 and 2105, respectively, and the amount of text used to describe fallen cases tended to be larger. Additionally, the median number of tokens obtained from tokenizing the UTH-BERT vocabulary was 2531 and 1288, respectively. The sequential input of long sequences to BERT required maintaining an error gradient; thus, GPU memory limitations resulted in the curtailment of the input token (text) length. In this study, we used eight Tesla-V-100 processors with 16 GB GPU memory. However, there was a limit of 13 BERT inputs (6630 tokens; 510 tokens×13). Therefore, text exceeding this limit had to be truncated. There were a total of 444 hospitalization cases that exceeded 6630 tokens, which comprised 0.6% of the entire data set. Ultimately, it was determined that this limitation would not have a large effect on model performance.

Figure 3. Overview of the bidirectional encoder representations from transformers (BERT) classification model. The input document was divided into 510 tokens; classification [CLS] and separation [SEP] tokens were added at each end, and the input was sequential. All token embeddings output sequentially were used as inputs to the bidirectional long short-term memory (Bi-LSTM) model, and the 50-dimensional vectors in the forward and reverse directions that were output for each were combined to form 100-dimensional vectors. The feature value obtained from the document was set as the sum of each dimension of the multiple 100-dimensional vectors, which were converted linearly and output as binary fallen or unfallen values using a sigmoid function. FFN: feedforward neural network.



Measures Against Imbalanced Data

Since the number of fallen and unfallen cases was uneven (see Figure 1), to reduce the impact of imbalanced data on learning, the inverse of the class frequency calculated from the training data set was weighted to the loss function. This is a simple heuristic method that is widely adopted in the presence of class imbalance [29].

Experimental Settings

We evaluated the performance of three prediction models: two-layer multilayer perceptron (inputs=49 fall-related variables), BERT+Bi-LSTM (inputs=only clinical text), and BERT+Bi-LSTM (inputs=clinical text+49 fall-related variables). For all prediction models, output binary values for each fallen and unfallen case were obtained using a sigmoid function to minimize the value with cross-entropy loss. It was determined that the learning stop condition would occur when the AUC value stopped improving for five epochs. Performance differences between the models were then investigated via net reclassification improvement (NRI) [30]. MeCab [31] was used as the morphological analyzer of the clinical text, and Mecab-ipadic-Neologd [32] and the Japanese disease name dictionary [33] were used as analyzer dictionaries. To develop the prediction models, we used Python v.3.8.5 (Python Software Foundation) and the PyTorch v.1.7.1 machine learning framework (Facebook's Artificial Intelligence Research Lab). All statistical analyses were conducted using the STATA v.16.1 integrated statistical software package.

Results

Fall-Related Variables

Table 1 lists the mean value, missing value rate, adjusted odds ratio, and standardized difference of the 49 fall-related variables. The average length of hospital stay was 30.3 days (SD 23.7) for fallen hospitalization and 10.6 days (SD 6.8) for unfallen hospitalization, with the difference being 19.7 days. No missing values were found in the basic patient and disease characteristics. The variable with the most missing values in the blood test results was plasma glucose at a missing rate of 19.7%. The missing value rate of variables related to prescription drugs was 8.3%. The 20-time AUC average was 0.73 (95% CI 0.72-0.74). Variables showing a significant difference at P<.05 for basic patient information included age, gender, assistance in bathing and movement in activities of daily living, impaired consciousness at admission, and previous history of falls in past admissions. Similarly, several diseases were significantly correlated with falls in terms of hospitalization triggers: diseases of the blood and blood-forming organs, mental and behavioral disorders, diseases of the eye and adnexa, diseases of the circulatory system, diseases of the digestive system, diseases of the skin and subcutaneous tissue, and diseases of the musculoskeletal system and connective tissue. For blood tests, low hemoglobin, low total protein or albumin, and abnormal electrolytes were significantly correlated with falls. For prescription drugs, NSAIDs, anti-Parkinson drugs, antipsychotics, other neuroactive agents, and diuretics were significantly correlated with falls. Among all fall-related variables, mental and behavioral disorders had the highest odds ratio and diseases of the eye and adnexa had the lowest odds ratio (Table 1).

Table 1. Statistics of fall-related variables.

Variables	Fallen cases (n=1728)	Unfallen cases (n=70,586)	Multivariate regression[a]		Standardized difference	
			Adjusted odds ratio (95% CI)	*P* value[b]	Before matching (n=1728 fallen cases, n=70,586 unfallen cases)	After matching (n=1728 fallen cases, n=1728 unfallen cases)
Hospital days, mean (SD)	30.3 (23.7)	10.6 (6.8)	N/A[c]	N/A	N/A	N/A
Demographics						
Age (years), mean (SD)	76.5 (6.8)	74.3 (SD 6.4)	1.03 (1.02-1.03)[d]	<.001	0.33	–0.04
Sex (male 0, female 1), positive rate (%)	40.6	43.8	0.71 (0.63-0.80)	<.001	–0.06	0.01
ADL[e] Eats, positive rate (%)	9.2	2.4	1.08 (0.83-1.40)	.57	0.29	0.04
ADL Bathe, positive rate (%)	19.2	5.5	1.37 (1.06-1.77)	.02	0.43	0.04
ADL Dressing[f], positive rate (%)	15.3	4.4	0.76 (0.57-1.02)	.07	0.37	0.04
ADL Transferring[g], positive rate (%)	26.2	8.6	1.79 (1.48-2.18)	<.001	0.48	0.02
ADL Continence[h], positive rate (%)	13.0	3.5	1.04 (0.80-1.37)	.75	0.35	0.05
Unconsciousness (JCS[i] 0,≠0), positive rate (%)	18.1	5.6	1.70 (1.44-2.00)	<.001	0.39	0.04
Emergency transport, positive rate (%)	8.6	3.9	0.96 (0.78-1.17)	.68	0.20	–0.01
Cognitive disorder, positive rate (%)	11.1	4.0	1.10 (0.92-1.32)	.28	0.01	0.01
Chemotherapy admission, positive rate (%)	11.7	11.4	1.08 (0.91-1.27)	.39	0.27	0.02
Past fallen, positive rate (%)	8.1	3.5	1.37 (1.13-1.65)	.001	0.20	0.01
Disease						
Certain infectious and parasitic diseases (A00-B99), positive rate (%)	8.6	6.8	0.98 (0.82-1.17)	.78	0.07	–0.02
Neoplasms (C00-D48), positive rate (%)	40.8	41.1	1.10 (0.97-1.25)	.12	–0.01	–0.01
Diseases of the blood and blood-forming organs (D50-D89), positive rate (%)	8.3	6.3	1.28 (1.07-1.53)	.01	0.08	–0.002
Endocrine, nutritional, and metabolic diseases (E00-E90), positive rate (%)	23.8	18.5	1.09 (0.96-1.24)	.19	0.13	–0.01
Mental and behavioral disorders (F00-F99), positive rate (%)	4.6	1.1	2.09 (1.61-2.71)	<.001	0.21	0.05
Diseases of the nervous system (G00-G99), positive rate (%)	8.4	4.7	1.14 (0.95-1.38)	.16	0.16	–0.001
Diseases of the eye and adnexa (H00-H59), positive rate (%)	3.8	13.3	0.47 (0.36-0.61)	<.001	–0.35	0.04
Diseases of the ear and mastoid process (H60-H95), positive rate (%)	0.3	0.8	0.47 (0.19-1.14)	.10	–0.07	0.01

Variables	Fallen cases (n=1728)	Unfallen cases (n=70,586)	Multivariate regression[a]		Standardized difference	
			Adjusted odds ratio (95% CI)	*P* value[b]	Before matching (n=1728 fallen cases, n=70,586 unfallen cases)	After matching (n=1728 fallen cases, n=1728 unfallen cases)
Diseases of the circulatory system (I00-I99), positive rate (%)	33.9	26.1	1.15 (1.02-1.29)	.02	0.17	0.00
Diseases of the respiratory system (J00-J99), positive rate (%)	9.5	6.2	1.01 (0.85-1.20)	.91	0.12	–0.01
Diseases of the digestive system (K00-K93), positive rate (%)	17.8	16.6	0.77 (0.67-0.87)	<.001	0.03	–0.01
Diseases of the skin and subcutaneous tissue (L00-L99), positive rate (%)	3.0	1.6	1.46 (1.09-1.95)	.01	0.09	–0.00
Diseases of the musculoskeletal system and connective tissue (M00-M99), positive rate (%)	11.9	8.4	1.22 (1.04-1.43)	.02	0.11	0.00
Diseases of the genitourinary system (N00-N99), positive rate (%)	10.0	7.3	0.94 (0.79-1.12)	.50	0.10	–0.003
Pregnancy, perinatal period, congenital malformations (O00-Q99), positive rate (%)	0.3	0.4	1.03 (0.42-2.52)	.94	–0.10	0.00
Symptoms, signs, and abnormal clinical and laboratory findings (R00-R99), positive rate (%)	5.8	3.3	1.03 (0.83-1.28)	.80	0.12	–0.01
Injury, poisoning and certain other consequences of external causes (S00-T98), positive rate (%)	5.1	3.1	1.11 (0.88-1.40)	.38	0.10	0.00
Blood tests						
Low hemoglobin (3.9% missing data), positive rate (%)	71.8	57.5	1.34 (1.19-1.53)	<.001	0.24	–0.04
Low total protein or albumin (5.0% missing data), positive rate (%)	48.7	33.8	1.20 (1.08-1.34)	.001	0.32	–0.04
High blood urea nitrogen (4.4% missing data), positive rate (%)	3.4	1.6	1.20 (0.90-1.61)	.22	0.12	–0.004
High liver enzymes (AST[j], ALT[k]; 4.0% missing data), positive rate (%)	6.0	3.6	1.22 (0.98-1.52)	.07	0.12	–0.01
Low plasma glucose (19.7% missing data), positive rate (%)	2.5	1.7	1.14 (0.80-1.62)	.48	0.05	–0.01
Abnormal electrolytes (Na, K, Cl; 12.1% missing data), positive rate (%)	35.1	21.6	1.40 (1.26-1.57)	<.001	0.32	–0.02
High C-reactive protein (6.8% missing data), positive rate (%)	10.9	5.0	1.12 (0.94-1.34)	.21	0.22	–0.005
Prescription						
Hypnotics and sedatives, antianxietics	37.4	30.7	1.09 (0.97-1.22)	.13	0.13	–0.001
Antiepileptic	4.4	1.8	1.30 (1.00-1.69)	.05	0.16	0.03
NSAIDs[l]	43.5	32.6	1.21 (1.08-1.36)	.001	0.22	–0.03

Variables	Fallen cases (n=1728)	Unfallen cases (n=70,586)	Multivariate regression[a]		Standardized difference	
			Adjusted odds ratio (95% CI)	*P* value[b]	Before matching (n=1728 fallen cases, n=70,586 unfallen cases)	After matching (n=1728 fallen cases, n=1728 unfallen cases)
Antiparkinsonism	3.2	1.0	1.61 (1.18-2.21)	.003	0.16	0.02
Antipsychotic	21.4	9.6	1.44 (1.25-1.66)	<.001	0.33	0.02
Other neuroactive agents	13.8	6.6	1.19 (1.01-1.39)	.03	0.23	0.01
Muscle relaxant	0.3	0.1	1.70 (0.60-4.85)	.32	0.03	0.01
Diuretic	23.4	13.7	1.33 (1.16-1.53)	<.001	0.24	–0.003
Antihypertensive	31.4	25.9	0.88 (0.77-1.00)	.05	0.11	–0.01
Diabetes treatment	15.9	12.7	1.08 (0.92-1.26)	.48	0.09	–0.01
Narcotic analgesic	3.3	1.4	1.11 (0.81-1.51)	.52	0.12	–0.001
Purgative medicine	38.3	32.6	1.09 (0.97-1.22)	.15	0.11	0.00
Polypharmacy (>10 drugs)	48.7	35.8	1.02 (0.89-1.17)	.77	0.26	–0.004

[a]Multivariate logistic regression on the results of missing value estimation by the multiple imputation method.

[b]Based on the two-tailed Z-test for a coefficient of zero.

[c]N/A: not applicable.

[d]The odds ratio for age was calculated by univariate logistic regression with the age range from 65 to 99 years equally transformed from 0.0 to 1.0.

[e]ADL: activities of daily living.

[f]Assistance is required for dressing or personal maintenance.

[g]Assistance is required for walking, going up and down stairs, getting into/out of bed or chair, or going to the toilet.

[h]Assistance is required for either defecation or urination.

[i]JCS: Japan Coma Scale, which has been widely used to assess patients' consciousness level in Japan.

[j]AST: aspartate aminotransferase.

[k]ALT: alanine aminotransferase.

[l]NSAID: nonsteroidal anti-inflammatory drug.

Impact of Falls on Hospital Stay

The AUC of the logistic regression model for which the propensity score was calculated was 0.73. Figure 4 shows the distribution of propensity scores before and after matching. The upper IQR was distributed at a low range of less than 0.2 both before and after matching. The results of matching the fallen and unfallen cases showed a sample size of 1728 for each, and the distribution of propensity scores in each group was similar. Furthermore, as shown in Table 1, the standardized differences [20] for all variables after matching were less than 0.1, and the differences between groups became sufficiently small for all variables [20]. The average length of hospital stays in the unfallen group, in which propensity score matching was performed, was 12.5 days (SD 7.0) and the ATET was 17.8 days (95% CI 16.6-19.0). Based on these results, it was estimated that the average length of hospital stay was extended by 17.8 to 30.3 days from 12.5 days, which was the estimated average length of hospital stay if the fallen cases had not fallen as a result of an elderly inpatient falling.

Table 2 summarizes the results of the Rosenbaum sensitivity analysis for the estimated ATET according to the upper limit of the extent of influence of the unobserved variables on the fall propensity score (Γ), which corresponds to the upper limit when the odds of allocation to a fallen case of the matched pair fluctuate in the range of (1/Γ,Γ) due to the unobserved variables. The maximum *P* value and minimum Hodges–Lehmann point estimate [22] reflect the maximum value of the null hypothesis' significance level and the minimum ATET value for each Γ value, respectively. Here, the null hypothesis is fall events do not influence the extended length of hospital stay, and the *P* value is the value of the one-sided Wilcoxon signed-rank sum test.

As shown in Table 2, when Γ was 7.5, the lower limit of the estimated value of ATET was 0.8 days, and the null hypothesis could not be rejected at the significance level of .05. By contrast, when Γ<7.5, a significant causal effect was observed. The bias of Γ=7.5 was huge [23], and the robustness of the hypothesis that falls cause an increased length of stay is demonstrated. Furthermore, as shown in Table 1, the highest odds ratio among the 49 covariates was 2.09 for mental and behavioral disorders. However, even with Γ=2.0, which assumes the presence of unobserved factors having the same degree of influence as the above variables, it was estimated that the length of hospital stays of fallen inpatients was extended by at least 8.6 days.

Figure 4. Box-and-whisker plots of the propensity scores (a) before matching and (b) after matching. Boxes show lower and upper IQR, and whiskers show the highest and lowest values, excluding outliers (>1.5 times IQR; rounds). Propensity score matching was performed using one-to-one nearest-neighbor matching with the replacement method on fallen cases.

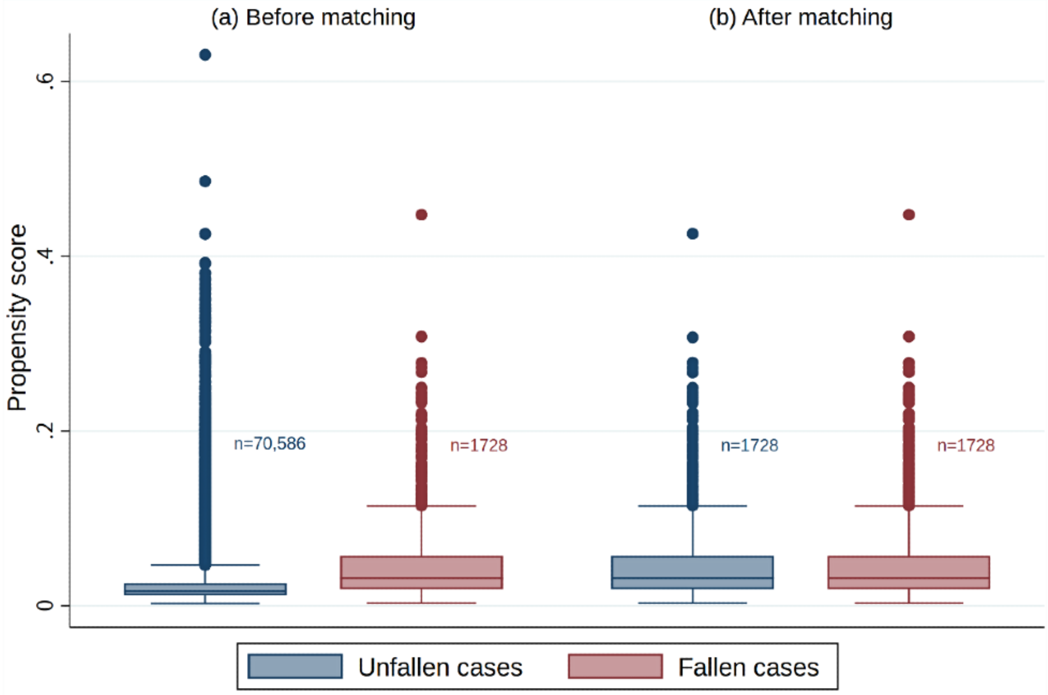


Table 2. Sensitivity analysis for *P* value and Rosenbaum bounds estimates (average values calculated over 20 imputed data sets) to unobserved biases.

Γ[a]	Maximum *P* value[b]	Minimum Hodges–Lehmann point estimate (days)
1.0	<.001	14.1
2.0	<.001	8.6
3.0	<.001	6.0
4.0	<.001	4.1
5.0	<.001	2.9
6.0	<.001	2.0
7.0	0.01	1.1
7.5	.05	0.8
8.0	.16	0.5

[a]Γ: odds of differential assignment due to unobserved factors.

[b]The *P* value is based on a one-tailed Wilcoxon signed-rank test for the null hypothesis of no extension of hospital stay caused by falls.

Performance of Fall Prediction Models

Table 3 summarizes the evaluation results of the machine learning models. Model 1, a multilayer perceptron with only the 49 fall-related factors as input, had the lowest AUC at 0.735. Model 2, the BERT+Bi-LSTM with only the clinical text as input, had the highest AUC at 0.851. Model 3, the BERT+Bi-LSTM using the clinical text and 49 fall-related factors as input, had an AUC of 0.850.

Tables A2 and A3 in Multimedia Appendix 1 list the NRIs for the reclassifications conducted to investigate the performance differences between models. Table A2 shows the result of comparing models 1 and 3; the NRIs of the fallen and unfallen cases were 0.123 (*P*<.001) and 0.068 (*P*<.001), respectively, and the integrated NRI was 0.191 (*P*<.001). This result showed that the performance of Model 3 was significantly improved over that of Model 1, suggesting that using clinical text improved predictive performance. Table A3 shows the result of comparing models 2 and 3, and the integrated NRI of the fallen and unfallen cases was –0.015 (*P*=.48), with no significant differences observed. This result indicates that there were no significant improvements to the performance of Model 3 over Model 2. Thus, adding the 49 fall-related factors to the clinical text did not improve the predictive performance.

Table 3. Performance comparison of machine learning models with input data categories.

Model	Input data		Evaluation accuracy[a]				
	49 fall-related factors	Clinical text	AUC[b]	F1[c]	Sensitivity	Specificity	Precision
Model 1: MLP[d]	✓		0.735	0.090	0.662	0.708	0.048
Model 2: BERT[e]+Bi-LSTM[f]		✓	0.851	0.165	0.737	0.839	0.093
Model 3: BERT+Bi-LSTM	✓	✓	0.850	0.138	0.794	0.776	0.076

[a]The accuracies are the average values of two cross-validation tests based on the cutoff determined by the Youden index.

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

[c]F1 is the harmonic mean of precision and recall.

[d]MLP: multilayer perceptron.

[e]BERT: bidirectional encoder representations from transformers.

[f]Bi-LSTM: bidirectional long short-term memory.

Impact of Prediction-Based Interventions

Table 4 shows a cross-table summary of the evaluation results of two Model 2 cross-validations based on the cutoff determined by the Youden index. It can be assumed from these results that some positive interventions were conducted on the 1806 hospitalization cases predicted to result in a fall and that some falls were completely prevented across 19,463 days (168 hospitalizations×12.5 days=average days of unfallen cases matched to fallen cases; 1638 hospitalizations×10.6 days=average days of unfallen cases). As a result, the hospitalized stay was shortened by a total of 2990 days (168 hospitalizations×17.8 days=ATET) among cases that were otherwise destined to experience a fall. This corresponds to 0.154 days per day of interventions (2990/19,463 days). Of the 8580 cases that were predicted to be unfallen, 60 cases actually experienced a fall (ie, false negatives). This indicates that 1068 (60 hospitalizations×17.8 days=ATET) shortened hospitalization stays were lost. Thus, the net reduced length of hospital stay was 1922 days (2990–1068 days). This corresponds to 0.099 days per day of interventions (1922/19,463 days). The average daily hospitalization cost in Japan is approximately 40,000 Yen (US $1=~136 Yen) [34]. Thus, the net reduced daily medical costs by active intervention were estimated to have been approximately 3950 Yen (1922 days×40,000 Yen per day/19,463 days) per day of interventions. This interpretation assumes that the preventive effect of aggressive intervention was 100%. However, Table 5 presents estimates when the presumed effect was adjusted to 25.5% and the ATET was set to 8.6 days. While the results up to this point were based on fixed cut-off values determined by the Youden index, Figure 5 shows how the net reduced daily medical costs for scenarios 1-4 in Table 5 change when the cutoff is changed. In Figure 5, the horizontal axis shows the sensitivity to changing the cutoff of Model 2 in the range of sensitivity≥0.5; the vertical axis shows the net reduced daily medical cost. For example, if the sensitivity is set to 0.95, the net reduced daily medical costs are 2249, 538, 1054, and 258 Yen, respectively.

Table 4. Cross-table summary of the results of the two Model 3 cross-validations. The cutoff was determined using the Youden index.

Predictions	Fallen cases	Unfallen cases	Sum
Predicted fallen cases	168	1638	1806
Predicted unfallen cases	60	8520	8580
Sum	228	10,158	10,386

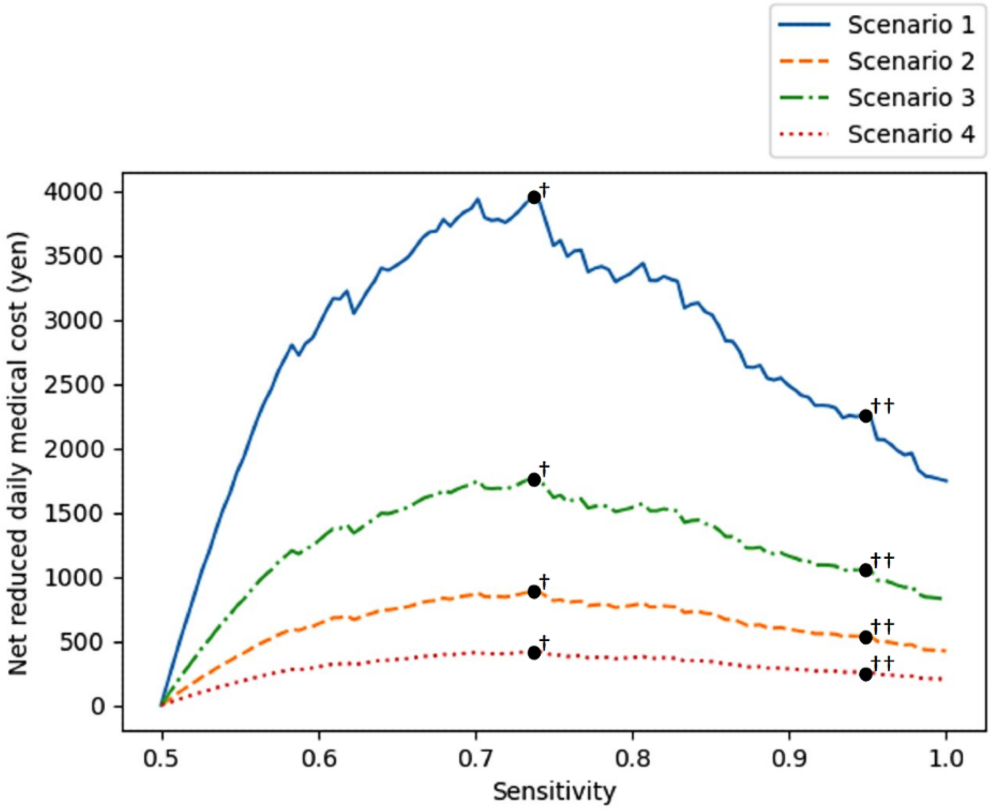
Table 5. Estimated hospital days reduced by interventions based on Model 2 predictions (sensitivity 73.7%, precision 9.3%).

Scenario	ATET[a] (days)	Fall prevention rate (%)	Reduced length of hospital stay (number of days per day of interventions)	Hospital stays that could not be reduced (number of days per day of interventions)	Net reduced length of hospital stay (number of days per day of interventions)	Net reduced daily medical costs (Yen per day)[b]
Scenario 1	17.8	100	0.154	0.055	0.099	3950
Scenario 2	17.8	25	0.035	0.012	0.022	886
Scenario 3	8.6	100	0.069	0.025	0.044	1769
Scenario 4	8.6	25	0.017	0.006	0.011	420

[a]ATET: average treatment effect on treatment.

[b]Medical costs were estimated at 40,000 Yen per day (US $1=~136 Yen).

Figure 5. Estimated net reduced daily medical costs by interventions based on Model 2 sensitivity. The maximum points in Scenarios 1-4 are indicated by a circle with † and their values are 3951, 886, 1768, and 420 (Yen; US $1=~136 Yen), respectively. These are taken with a sensitivity of 0.737; the sensitivity is the same as determined using the Youden index. The points with 0.95 sensitivity in Scenarios 1-4 are indicated by a circle with ††, and their values are 2249, 538, 1054, and 258 (Yen), respectively.



Discussion

Principal Results

In this study, we verified the performance of a fall prediction model using clinical EHR text pertaining to elderly patients, and we estimated the reduction in medical costs incurred if fall prevention interventions had been successfully conducted according to the prediction results.

Extended Hospital Stays Due to Fall

The extended length of hospital stay due to falls (ATET) was estimated at 17.8 days. This value was 1.9 days shorter than the simple difference (19.7 days) between the average days of hospitalized stay between fallen and unfallen groups. This is the result of a positive correlation between fall susceptibility and length of stay, with the exclusion of confounding background factors between groups. Falls include incidental falls, which intuitively lead to 17.8 days of extended stay. In these examples, the analysis subject was aged 65 years or older and was hospitalized for 3 days or more. It is also common for severe falls to result in extended hospitalized stays of 1 month or longer. Thus, it is further intuitive that this may be the effect of averaging incidents and accidents. Meanwhile, this ATET was obtained from 49 variables automatically extracted from the EHR system; thus, there may have been unobserved covariates. The verification of the *P* value of causal effect and robustness of the ATET by Rosenbaum sensitivity analysis showed that the causal effect of falls extending the length of hospitalized stay was significant at a level of $P<.05$, even when assuming unobserved covariates with large odds ratios such as Γ=7. As reported in previous studies [2,12-14], this supports the finding that falls extend the length of hospital stay. Moreover, when assuming a more realistic Γ, of the 49 variables shown in Table 1, if there were unobserved covariates with Γ=2.0 corresponding to mental and behavioral disorders (the largest odds ratio), then the extended length of stay caused by falls was estimated to be at least 8.6 days. This value falls within the 6-12 days reported in US studies [2,12-14]. However, comparisons between acute-care hospitalized stays in 2019 [35] showed an average length of hospital stay in the United States of 6.1 days. The average length of hospital stay in Japan was 16.0 days, which is 2.5 times longer. Therefore, it is intuitive that the extended length of hospital stay due to falls will be longer in Japan. Thus, the extension of 8.6 days is thought to be conservative.

Fall Prediction Model Performance

The accuracy of the proposed prediction model was investigated by comparing the prediction accuracy of the 19-item multivariate logistic regression model (AUC 0.82), including nurse observations, performed in a previous study [4]. The AUC of Model 1 (multilayer perceptron), which used only the 49 fall-related variables, was 0.735. This was lower than the AUC of 0.82 obtained in the previous study, which used items obtained only by nurse observations as explanatory variables for the multivariate logistic regression models, including decreased lower-limb muscle strength, use of a cane or walker, wobbling, disturbing behaviors, strong independence, and decreased comprehension. These variables are known to affect prediction accuracy. The fact that such items were not included in the 49 variables in this study is clearly the reason for the relatively low accuracy of Model 1. However, the AUC of Model 3, in which clinical text was added, was 0.850. Additionally, because this study evaluated generalization performance using past data for learning and future data for evaluation, this value is intuitively higher than the AUC of 0.82. As described below, clinical nurse risk assessments of falls and fall prevention interventions may have improved model performance.

The AUC of Model 3, which used clinical text, was more than 0.1 higher than that of Model 1, which did not use clinical text. A two-sided Z-test of the NRI between models showed that Model 3 was significantly more accurate. It is therefore rational to conclude that the prediction accuracy of a model that uses clinical text is high because, at the time of hospitalization, the nurse observes the patient, conducts a risk assessment, and records the evaluation results as necessary. Therefore, clinical text contained more information related to fall risk than the 49 fall-related variables, which likely contributed to the improvement in prediction accuracy. Meanwhile, no significant difference was found between the prediction accuracy of Models 2 and 3, suggesting that the clinical text also contained information corresponding to the 49 variables at the time of hospitalization.

It has often been reported that BERT exhibits high performance, even with clinical text [36-39]. This is also true for this study, in which a model combining BERT and Bi-LSTM using clinical text recorded in daily practice allowed for fall prediction with an accuracy equal to or higher than that of conventional risk assessment tools. Although not limited to BERT, prediction models that use neural networks also show high performance. However, they lack a means of explaining the prediction, as opposed to linear and tree models. Application of explanatory techniques such as SHapley Additive exPlanations [40] would lead to remarkable explanatory findings related to falls. Hence, this is a future study direction.

Regarding model precision, Model 2 had the best precision of 9.3%, which was higher than the value of 6.9% obtained in previous research [4]. However, this shows that many false positives were likely present. Predicting a patient's future is an inherently difficult task; however, the data set used in this study involved fall prevention measures based on risk-assessment results. Thus, it is thought that there were likely some cases in which falls were prevented when the risk was high. Fall prevention measures include a mat-type buzzer installed inside the bed and a mechanism that sounds like a buzzer when the patient leaves the bed. A limitation of this study is that the data set did not contain information about this and other prevention measures. Hence, future studies should not rule these out.

Impact of Fall Prevention Interventions Based on the Prediction Model

Table 5 shows four scenarios in which the length of hospital stay was shortened when assuming that active fall prevention was conducted for all cases in which Model 2 predicted falls. The net reduced length of hospital stay per day of interventions was 0.099 days/day when the preventive effect was set to 100% (Scenario 1) and 0.022 days/day when the effect was set to 25% (Scenario 2). Additionally, when assuming the presence of unobserved covariates with odds ratios equivalent to 2.0 times, the shortened number of days was 0.044 days/day (Scenario 3) and 0.011 days/day (Scenario 4). The results showed that in cases where medical expenses per day of hospitalization were 40,000 Yen/day, the break-even costs of 3950-420 Yen/day in Scenarios 1-4 were found based on the costs of introducing the prediction model and fall prevention measures. Figure 5, which shows the net reduced daily medical costs when the cutoff changed, reveals that the break-even cost of Scenarios 1-4 was 2249-258 Yen/day when the sensitivity was set to 0.95. Although not shown in Figure 5, as an extreme cut-off setting, the net reduced daily medical costs of applying fall prevention interventions to all cases without using the prediction model were 1469, 357, 696, and 172 Yen in Scenarios 1-4, respectively. There are sensitivity points at which the net reduced daily medical cost is higher using our prediction model than without the prediction model in all scenarios, which shows the advantage of using our prediction model over not using the prediction model. Medical expenses vary depending on the size of the hospital; thus, the break-even point is higher in larger hospitals. Hence, the incentive for prediction should be high. These results reflect the costs of introducing preventive measures in addition to those already taken. Thus, more effective preventive measures are needed. An ideal solution would be to include methods to further prevent falls by attaching a motion sensor to patients when a fall is predicted and using its data to predict near-future behaviors. These technologies are expected to be available in the near future. Furthermore, higher prediction performance and improved fall prevention intervention will further reduce hospitalized stays and medical costs.

Limitations and Future Work

One limitation of this study pertains to extant preventive measures that may have negated true positives. Another limitation pertains to the results of this study not being applicable to patients with short-term (1-2 days) or long-term (31 days or more) hospital stays. In this study, 232 cases of falls that occurred during the first or second day of hospitalization were excluded. However, these constituted 11.8% of the 1960 total cases (Figure 1). Additionally, although our data set was relatively large, it was limited in that it was obtained from a single facility; thus, it is not generalizable to all of Japan. Future studies should obtain more robust data using multicenter information and analyze the prediction results using techniques that visualize the basis of prediction.

Conclusions

In this study, it was estimated that the general length of hospital stay in Japan was extended by 17.8 days due to falls among elderly inpatients. The predictive performance of the proposed model, which predicts falls up to the 30th day of hospitalization using clinical text from the second day of hospitalization, showed an AUC of 0.85. Thus, it was suggested that this may be more accurate than traditional risk assessment tools. However, its precision was still low, at 9.3%. A possible reason for this discrepancy may be the inclusion of cases where falls did not occur because of successful fall prevention interventions during hospitalization, which were not accounted for. Fall prevention interventions for cases predicted by this model were shown to reduce medical costs by up to 886 Yen per day, even if the preventive effect was as low as 25%. Limitations include the fact that short- and long-term patients were not included, and only a single-center demographic was applied.

Acknowledgments

The authors thank Shinsuke Ogue for technical support and expertise. This research was also supported by the Japan Science and Technology Agency, Promoting Individual Research to Nurture the Seeds of Future Innovation and Organizing Unique, Innovative Networks (JPMJPR1654).

Authors' Contributions

YK, TY, and ES conceived and designed the research. TY provided the incident reports of the hospital. YK, KS, and HK designed and conducted the statistical analysis. KS and DS developed and evaluated the machine learning models. YK and KS wrote the manuscript and prepared the figures and tables. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

YK, KS, DS, and ES belong to the Artificial Intelligence in Healthcare, Graduate School of Medicine, The University of Tokyo, which is an endowment department supported by an unrestricted grant from I&H Co, Ltd and EM SYSTEMS Co. However, these sponsors had no influence over the interpretation, writing, or publication of this work.

Multimedia Appendix 1
Tables A1-A3.
[DOCX File , 18 KB - medinform_v10i7e37913_app1.docx]

References

1. Rate of fall incident; report of quality improvement of medical care project. Japan Federation of Democratic Medical Institutions. 2016. URL: https://www.min-iren.gr.jp/hokoku/hokoku_h28.html [accessed 2022-07-05]
2. Bouldin ELD, Andresen EM, Dunton NE, Simon M, Waters TM, Liu M, et al. Falls among adult patients hospitalized in the United States: prevalence and trends. J Patient Saf 2013 Mar;9(1):13-17 [FREE Full text] [doi: 10.1097/PTS.0b013e3182699b64] [Medline: 23143749]
3. American Geriatrics Society, British Geriatrics Society, and American Academy of Orthopaedic Surgeons Panel on Falls Prevention. Guideline for the prevention of falls in older persons. J Am Geriatr Soc 2001 May;49(5):664-672. [Medline: 11380764]
4. Morita E, Iijima S, Hirai S, Kazawa M, Anzai Y. Revision of an assessment score sheet for fall prediction and examination on inter-rater agreement among the nurse evaluators. J Japan Acad Nurs Admin Policies 2010;14(1):51-58. [doi: 10.19012/janap.14.1_51]
5. Morse JM, Morse RM, Tylko SJ. Development of a scale to identify the fall-prone patient. Can J Aging 2010 Nov 29;8(4):366-377. [doi: 10.1017/s0714980800008576]
6. Oliver D, Britton M, Seed P, Martin FC, Hopper AH. Development and evaluation of evidence based risk assessment tool (STRATIFY) to predict which elderly inpatients will fall: case-control and cohort studies. BMJ 1997 Oct 25;315(7115):1049-1053 [FREE Full text] [doi: 10.1136/bmj.315.7115.1049] [Medline: 9366729]
7. Morris JN, Nonemaker S, Murphy K, Hawes C, Fries BE, Mor V, et al. A commitment to change: revision of HCFA's RAI. J Am Geriatr Soc 1997 Aug;45(8):1011-1016. [doi: 10.1111/j.1532-5415.1997.tb02974.x] [Medline: 9256856]
8. Hendrich A, Nyhuis A, Kippenbrock T, Soja ME. Hospital falls: development of a predictive model for clinical practice. Appl Nurs Res 1995 Aug;8(3):129-139. [doi: 10.1016/s0897-1897(95)80592-3] [Medline: 7668855]
9. Toyabe S. Detecting inpatient falls by using natural language processing of electronic medical records. BMC Health Serv Res 2012 Dec 05;12:448 [FREE Full text] [doi: 10.1186/1472-6963-12-448] [Medline: 23217016]
10. Bjarnadottir RI, Lucero RJ. What can we learn about fall risk factors from EHR nursing notes? A text mining study. eGEMS 2018 Sep 20;6(1):21 [FREE Full text] [doi: 10.5334/egems.237] [Medline: 30263902]
11. Nakatani H, Nakao M, Uchiyama H, Toyoshiba H, Ochiai C. Predicting inpatient falls using natural language processing of nursing records obtained from Japanese electronic medical records: case-control study. JMIR Med Inform 2020 Apr 22;8(4):e16970 [FREE Full text] [doi: 10.2196/16970] [Medline: 32319959]
12. Bates DW, Pruess K, Souney P, Platt R. Serious falls in hospitalized patients: correlates and resource utilization. Am J Med 1995 Aug;99(2):137-143. [doi: 10.1016/s0002-9343(99)80133-8] [Medline: 7625418]
13. Brand CA, Sundararajan V. A 10-year cohort study of the burden and risk of in-hospital falls and fractures using routinely collected hospital data. Qual Saf Health Care 2010 Dec;19(6):e51. [doi: 10.1136/qshc.2009.038273] [Medline: 20558479]

14. Wong CA, Recktenwald AJ, Jones ML, Waterman BM, Bollini ML, Dunagan WC. The cost of serious fall-related injuries at three Midwestern hospitals. Jt Comm J Qual Patient Saf 2011 Feb;37(2):81-87. [doi: 10.1016/s1553-7250(11)37010-9] [Medline: 21939135]
15. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/N19-1423]
16. Oliver D, Daly F, Martin FC, McMurdo MET. Risk factors and risk assessment tools for falls in hospital in-patients: a systematic review. Age Ageing 2004 Mar;33(2):122-130. [doi: 10.1093/ageing/afh017] [Medline: 14960426]
17. Matarese M, Ivziku D, Bartolozzi F, Piredda M, De Marinis MG. Systematic review of fall risk screening tools for older patients in acute hospitals. J Adv Nurs 2015 Jun;71(6):1198-1209. [doi: 10.1111/jan.12542] [Medline: 25287867]
18. Aranda-Gallardo M, Morales-Asencio JM, Canca-Sanchez JC, Barrero-Sojo S, Perez-Jimenez C, Morales-Fernandez A, et al. Instruments for assessing the risk of falls in acute hospitalized patients: a systematic review and meta-analysis. BMC Health Serv Res 2013 Apr 02;13:122 [FREE Full text] [doi: 10.1186/1472-6963-13-122] [Medline: 23547708]
19. van Buuren S, Groothuis-Oudshoorn CG. MICE: Multivariate Imputation by Chained Equations in R. J Stat Soft 2011;45(3):2000-6743. [doi: 10.18637/jss.v045.i03]
20. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011 May;46(3):399-424 [FREE Full text] [doi: 10.1080/00273171.2011.568786] [Medline: 21818162]
21. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica 2006 Jan;74(1):235-267. [doi: 10.1111/j.1468-0262.2006.00655.x]
22. Rosenbaum PR. Hodges-Lehmann point estimates of treatment effect in observational studies. J Am Stat Assoc 1993 Dec;88(424):1250-1253. [doi: 10.1080/01621459.1993.10476405]
23. Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika 1987;74(1):13-26. [doi: 10.1093/biomet/74.1.13]
24. Rosenbaum PR. Various practical issues in matching. In: Design of observational studies. New York, NY: Springer; 2010:187-195.
25. Kawazoe Y, Shibata D, Shinohara E, Aramaki E, Ohe K. A clinical specific BERT developed using a huge Japanese clinical text corpus. PLoS One 2021 Nov 9;16(11):e0259763 [FREE Full text] [doi: 10.1371/journal.pone.0259763] [Medline: 34752490]
26. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv. 2015 Aug 09. URL: https://arxiv.org/pdf/1508.01991.pdf [accessed 2022-07-05]
27. Mulyar A, Schumacher E, Rouhizadeh M, Dredze M. Phenotyping of clinical notes with improved document classification models using contextualized neural language models. arXiv. URL: https://arxiv.org/abs/1910.13664 [accessed 2022-07-05]
28. Choi H, Kim J, Joe S, Gwon Y. Evaluation of BERT and ALBERT sentence embedding performance on downstream NLP tasks. 2021 Presented at: 25th International Conference on Pattern Recognition, ICPR2020; January 10-15, 2021; Milano-Virtual. [doi: 10.1109/icpr48806.2021.9412102]
29. Fernando KRM, Tsokos CP. Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. IEEE Trans Neural Netw Learning Syst 2021:15247-15260. [doi: 10.1109/tnnls.2020.3047335]
30. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008 Jan 30;27(2):157-72; discussion 207. [doi: 10.1002/sim.2929] [Medline: 17569110]
31. MeCab: yet another part-of-speech and morphological analyzer. GitHub. 2005. URL: https://github.com/taku910/mecab/blob/master/index.html [accessed 2021-11-02]
32. Sato T. Mecab-ipadic-NEologd: Neologism dictionary for MeCab. GitHub. URL: https://github.com/neologd/mecab-ipadic-neologd [accessed 2021-11-02]
33. Ito, Kaoru. J-MeDic: A Japanese disease name dictionary based on real clinical usage. 2018 Presented at: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC); May 7-12, 2018; Miyazaki, Japan URL: https://sociocom.naist.jp/manbyo-dic-en/
34. Publication of the Trends of Medical Care Expenditures in FY2020 (in Japanese). Ministry of Health, Labor and Welfare, Japan. URL: https://www.mhlw.go.jp/topics/medias/year/20/dl/iryouhi_data.pdf [accessed 2022-07-05]
35. OECD Health Statistics 2021. Organization for Economic Cooperation and Development. URL: https://www.oecd.org/health/health-data.htm [accessed 2022-07-05]
36. Kades K, Sellner J, Koehler G, Full PM, Lai TYE, Kleesiek J, et al. Adapting bidirectional encoder representations from transformers (BERT) to assess clinical semantic textual similarity: algorithm development and validation study. JMIR Med Inform 2021 Feb 03;9(2):e22795 [FREE Full text] [doi: 10.2196/22795] [Medline: 33533728]
37. Nakamura Y, Hanaoka S, Nomura Y, Nakao T, Miki S, Watadani T, et al. Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers. BMC Med Inform Decis Mak 2021 Sep 11;21(1):262 [FREE Full text] [doi: 10.1186/s12911-021-01623-6] [Medline: 34511100]

38. Xiong Y, Chen S, Chen Q, Yan J, Tang B. Using character-level and entity-level representations to enhance bidirectional encoder representation from transformers-based clinical semantic textual similarity model: ClinicalSTS modeling study. JMIR Med Inform 2020 Dec 29;8(12):e23357 [FREE Full text] [doi: 10.2196/23357] [Medline: 33372664]
39. Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-tuning bidirectional encoder representations from transformers (BERT)-based models on large-scale electronic health record notes: an empirical study. JMIR Med Inform 2019 Sep 12;7(3):e14830 [FREE Full text] [doi: 10.2196/14830] [Medline: 31516126]
40. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. arXiv. 2017. URL: https://arxiv.org/abs/1705.07874 [accessed 2022-07-05]

Abbreviations

ATET: average treatment effect on treatment
AUC: area under the receiver operating characteristic curve
BERT: bidirectional encoder representations from transformers
Bi-LSTM: bidirectional long short-term memory
CLS: classification
DPC: diagnosis procedure combination
EHR: electronic health record
MICE: multiple imputation by chained equation
NLP: natural language processing
NRI: net reclassification improvement
NSAID: nonsteroidal anti-inflammatory drug
SEP: separation

Edited by T Hao; submitted 13.03.22; peer-reviewed by H Uchiyama, M Rodrigues; comments to author 26.04.22; revised version received 13.06.22; accepted 27.06.22; published 27.07.22.

Please cite as:
Kawazoe Y, Shimamoto K, Shibata D, Shinohara E, Kawaguchi H, Yamamoto T
Impact of a Clinical Text–Based Fall Prediction Model on Preventing Extended Hospital Stays for Elderly Inpatients: Model Development and Performance Evaluation
JMIR Med Inform 2022;10(7):e37913
URL: https://medinform.jmir.org/2022/7/e37913
doi:10.2196/37913
PMID:35896017

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org