Contents

Review

Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review (e35724)
Carina Vorisek, Moritz Lehne, Sophie Klopfenstein, Paula Mayer, Alexander Bartschke, Thomas Haese, Sylvia Thun. ........................................... 3

Viewpoint

The Power of Patient Engagement With Electronic Health Records as Research Participants (e39145)
Jeff Pawelek, Katie Baca-Motes, Jay Pandit, Benjamin Berk, Edward Ramos. .......................................................... 20

Original Papers

The Effects of Information Continuity and Interpersonal Continuity on Physician Services Online: Cross-sectional Study (e35830)
Yan Xuan, Chaojin Guo, Wei Lu. .......................................................... 27

Electronic Patient Portal Access, Retention in Care, and Viral Suppression Among People Living With HIV in Southeastern United States: Observational Study (e34712)
Cassandra Schember, Sarah Scott, Cathy Jenkins, Peter Rebeiro, Megan Turner, Sally Furukawa, Carmen Bofill, Zhou Yan, Gretchen Jackson, April Pettit. .......................................................... 42

The Impact of Telemedicine on Physicians' After-hours Electronic Health Record "Work Outside Work" During the COVID-19 Pandemic: Retrospective Cohort Study (e34826)
Katharine Lawrence, Oded Nov, Devin Mann, Soumik Mandal, Eduardo Iturrate, Batia Wiesenfeld. .......................................................... 52

Extraction of Explicit and Implicit Cause-Effect Relationships in Patient-Reported Diabetes-Related Tweets From 2017 to 2021: Deep Learning Approach (e37201)
Adrian Ahne, Vivek Khetan, Xavier Tannier, Md Rizvi, Thomas Czernichow, Francisco Orchard, Charline Bour, Andrew Fano, Guy Fagherazzi. 6

Accurate Forecasting of Emergency Department Arrivals With Internet Search Index and Machine Learning Models: Model Development and Performance Evaluation (e34504)
Bi Fan, Jiaxuan Peng, Hainan Guo, Haobin Gu, Kangkang Xu, Tingting Wu. .......................................................... 77

Classification of Twitter Vaping Discourse Using BERTweet: Comparative Deep Learning Study (e33678)
William Baker, Jason Colditz, Page Dobbs, Huy Mai, Shyam Visweswaran, Justin Zhan, Brian Primack. .......................................................... 94
Impact of a Clinical Text–Based Fall Prediction Model on Preventing Extended Hospital Stays for Elderly Inpatients: Model Development and Performance Evaluation (e37913)
Yoshimasa Kawazoe, Kiminori Shimamoto, Daisaku Shibata, Emiko Shinohara, Hideaki Kawaguchi, Tomotaka Yamamoto. .......................... 102
Review

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

Carina Nina Vorisek¹, MD, MSc; Moritz Lehne¹, PhD; Sophie Anne Ines Klopfenstein¹², MD; Paula Josephine Mayer¹; Alexander Bartschke¹, MA; Thomas Haese¹, BSc; Sylvia Thun¹, MD, PhD

¹Core Facility Digital Medicine and Interoperability, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany
²Institute 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.

https://medinform.jmir.org/2022/7/e35724

JMIR Med Inform 2022 | vol. 10 | iss. 7 | e35724 | p.3 (page number not for citation purposes)
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 Document Architecture) and combines their advantages with established modern web technologies such as a Representational Document Architecture. Existing reviews on FHIR investigate the general use cases, implementation, goals, and limitations 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.
**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].
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].
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.

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

\(^a\)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.

<table>
<thead>
<tr>
<th>Source, year</th>
<th>Country</th>
<th>Item mapped to FHIR&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Objective for FHIR use</th>
<th>FHIR resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banach et al [56], 2021</td>
<td>Germany</td>
<td>Medical and demographic data from free-text eligibility criteria</td>
<td>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</td>
<td>—&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Bauer et al [1], 2020</td>
<td>Australia</td>
<td>Questionnaire</td>
<td>Ontology-based standard questionnaire for linking genomic data with clinical outcomes</td>
<td>Questionnaire</td>
</tr>
<tr>
<td>Bialke et al [40], 2018</td>
<td>Germany</td>
<td>Modular consent templates</td>
<td>Support improvement for consent definition and consent documentation</td>
<td>Consent</td>
</tr>
<tr>
<td>Bild et al [28], 2020</td>
<td>Germany</td>
<td>Informed consent template</td>
<td>Cross-site interoperability layer for representing the validity of data use policies derived from signed informed consent templates and regulatory framework</td>
<td>Consent and Patient</td>
</tr>
<tr>
<td>Brandt et al [71], 2021</td>
<td>United States</td>
<td>Phenotype definitions from the Phenotype Knowledgebase repository</td>
<td>Repository of structured phenotype definitions for automation of cohort identification.</td>
<td>Patients, Encounter, Procedure, Medication orders, Condition, and Observation</td>
</tr>
<tr>
<td>Cheng et al [44], 2021</td>
<td>United States</td>
<td>EHR&lt;sup&gt;c&lt;/sup&gt; Data</td>
<td>Seamless data exchange between the REDCap&lt;sup&gt;d&lt;/sup&gt; research electronic data capture and any EHR system with a FHIR API&lt;sup&gt;e&lt;/sup&gt;</td>
<td>Patient, Observation, AllergyIntolerance,MedicationOrder, and Condition</td>
</tr>
<tr>
<td>Deppenwies et al [57], 2021</td>
<td>Germany</td>
<td>Oncology data</td>
<td>Provide a transformation tool from oncology data XML files to FHIR for oncological data to enable clinical research</td>
<td>Medication, MedicationStatement, and Procedure</td>
</tr>
<tr>
<td>Eapen et al [24], 2019</td>
<td>Canada</td>
<td>Electronic form components</td>
<td>Management, editing, and rendering of electronic forms in the form of an open-source framework</td>
<td>Questionnaire and QuestionnaireResponse</td>
</tr>
<tr>
<td>Fischer et al [35], 2020</td>
<td>Germany</td>
<td>Common data set from a German pulmonary hypertension registry</td>
<td>Feasibility of HL7&lt;sup&gt;f&lt;/sup&gt; FHIR Bundle and XSLT&lt;sup&gt;g&lt;/sup&gt; as a generic ETL&lt;sup&gt;h&lt;/sup&gt; process to populate an OMOP&lt;sup&gt;i&lt;/sup&gt; CDM&lt;sup&gt;j&lt;/sup&gt;</td>
<td>Patient, Encounter, and Observation</td>
</tr>
<tr>
<td>Garza et al [61], 2020</td>
<td>United States</td>
<td>Concomitant medications, demographics, eligibility, labs, medical history, therapeutic area–specific, procedure, encounters, vital signs, other, administrative, questionnaires, and study drug administration</td>
<td>Developing and implementing a systematic mapping approach for evaluating HL7 FHIR standard coverage in multicenter clinical trials.</td>
<td>Observation, Patient, Specimen, Encounter, Diagnostic Report, and Condition</td>
</tr>
<tr>
<td>González-Castro et al [55], 2021</td>
<td>Spain</td>
<td>Clinical patient data (from EHR) and patient-generated data</td>
<td>Collection and aggregation of survivorship data (use cases colon cancer and breast cancer)</td>
<td>Patient, Condition, Observation, MedicationStatement, Encounter, and Procedure</td>
</tr>
<tr>
<td>Gruendner et al [69], 2020</td>
<td>Germany</td>
<td>Clinical patient data</td>
<td>Analysis within and across institutions</td>
<td>—</td>
</tr>
<tr>
<td>Gruendner et al [42], 2021</td>
<td>Germany</td>
<td>Metadata</td>
<td>Developing a Metadata Schema based on FHIR to gather metadata on clinical, epidemiological, and public health studies; elevate data FAIRness&lt;sup&gt;k&lt;/sup&gt;; and widen analysis possibilities across health research domains</td>
<td>ResearchStudy, Questionnaire, and DocumentReference</td>
</tr>
<tr>
<td>Guérin et al [51], 2021</td>
<td>France</td>
<td>Clinical and omics data in oncology</td>
<td>Improve and accelerate retrospective and prospective clinical and genomic data sharing in oncology</td>
<td>MolecularSequence and Observation</td>
</tr>
<tr>
<td>Gulden et al [31], 2018</td>
<td>Germany</td>
<td>Eligibility criteria of clinical trials</td>
<td>Recruitment of patients for clinical trials using eligibility criteria</td>
<td>Condition and Patient</td>
</tr>
<tr>
<td>Gulden et al [30], 2021</td>
<td>Germany</td>
<td>Clinical trial data</td>
<td>Multisite clinical trial registry</td>
<td>ResearchStudy</td>
</tr>
<tr>
<td>Hong et al [25], 2017</td>
<td>United States</td>
<td>Ovarian cancer data</td>
<td>Support of clinical statistics and analysis leveraging standardized data exchange and access services based on FHIR</td>
<td>Patient, Observation, Condition, and Procedure</td>
</tr>
<tr>
<td>Source, year</td>
<td>Country</td>
<td>Item mapped to FHIR</td>
<td>Objective for FHIR use</td>
<td>FHIR resources</td>
</tr>
<tr>
<td>-------------</td>
<td>---------</td>
<td>---------------------</td>
<td>------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Hund et al [53], 2021</td>
<td>Germany</td>
<td>Process data</td>
<td>Developing a framework to enable standardized, shared processes using Business Process Model and Notation and FHIR for arbitrary biomedical research</td>
<td>ActivityDefinition, Binary, Bundle, CodeSystem, Endpoint, Group, NamingSystem, Organization, Practitioner, PractitionerRole, ResearchStudy, StructureDefinition, Subscription, and Task</td>
</tr>
<tr>
<td>Jiang et al [70], 2017</td>
<td>United States</td>
<td>Clinical research data</td>
<td>Development and assessment of a consensus-based approach for harmonizing the OHDSI CDM with HL7 FHIR</td>
<td>Observation</td>
</tr>
<tr>
<td>Kilintzis et al [59], 2022</td>
<td>Greece</td>
<td>Clinical information from ICU COVID-19 patients</td>
<td>Fusion of clinical information with chest sounds and imaging of COVID-19 ICU patients</td>
<td>Media</td>
</tr>
<tr>
<td>Klopfenstein et al [41], 2021</td>
<td>Germany</td>
<td>Metadata of clinical, epidemiological and public health studies</td>
<td>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</td>
<td>ResearchStudy, Questionnaire, and DocumentReference</td>
</tr>
<tr>
<td>Khalique and Khan [38], 2017</td>
<td>Pakistan</td>
<td>EHR</td>
<td>Analysis or mining of EHR data and contextual information to assess the population’s health</td>
<td>—</td>
</tr>
<tr>
<td>Khvastova et al [12], 2020</td>
<td>Germany</td>
<td>Open-source research platform (XNAT®)</td>
<td>Feasibility study for the full integration of FHIR into XNAT</td>
<td>Patient</td>
</tr>
<tr>
<td>Lackerbauer et al [32], 2019</td>
<td>Austria</td>
<td>Informed consent or questionnaires</td>
<td>Automated verification of answers</td>
<td>Questionnaire and QuestionnaireResponse</td>
</tr>
<tr>
<td>Lambarki et al [58], 2021</td>
<td>Germany</td>
<td>Oncology data</td>
<td>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</td>
<td>Condition, Observation, Procedure, MedicationStatement, Patient, Organization, Specimen, ClinicalImpression, Encounter, and ServiceRequest</td>
</tr>
<tr>
<td>Lee et al [23], 2020</td>
<td>Taiwan</td>
<td>IPS®</td>
<td>FHIR-based global infectious disease surveillance and case-tracking model</td>
<td>MedicationStatement, Medication, AllergyIntolerance, Condition, Immunization, Procedure, Organization, Observation, CarePlan, and Location</td>
</tr>
<tr>
<td>Lenert et al [50], 2021</td>
<td>United States</td>
<td>Clinical data</td>
<td>Availability of data for research</td>
<td>Patient, Encounter, Condition, Procedure, Observation, MedicationRequest, and MedicationAdministration</td>
</tr>
<tr>
<td>Leroux et al [67], 2017</td>
<td>Australia</td>
<td>Data model</td>
<td>Mapping CDISC ODM to FHIR</td>
<td>Patient, Observation, EpisodeOfCare, Encounter, QuestionnaireResponse, Questionnaire, and CarePlan</td>
</tr>
<tr>
<td>Majeed et al [60], 2021</td>
<td>Germany</td>
<td>General patient information, encounter, or visit related information; individual data points; observations; measurements; and surveys</td>
<td>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</td>
<td>Patient, Observation, and Encounter</td>
</tr>
<tr>
<td>Metke-Jimenez et al [43], 2019</td>
<td>Australia</td>
<td>REDCap forms</td>
<td>Data export from REDCap into FHIR resources</td>
<td>Encounter, Observation, Condition, and Patient</td>
</tr>
<tr>
<td>Peng et al [52], 2021</td>
<td>Germany</td>
<td>Genomic Variant Cell Format data</td>
<td>Coverage of Variant Cell Format data in OMOP CDM with and without using FHIR as intermediate layer</td>
<td>MolecularSequence, Patient, and Condition</td>
</tr>
<tr>
<td>Source, year</td>
<td>Country</td>
<td>Item mapped to FHIR(^a)</td>
<td>Objective for FHIR use</td>
<td>FHIR resources</td>
</tr>
<tr>
<td>-------------</td>
<td>---------</td>
<td>----------------------------</td>
<td>------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Pfiffner et al [22], 2016</td>
<td>United States</td>
<td>ResearchKit data</td>
<td>Patient-reported outcomes</td>
<td>Contract, Questionnaire, QuestionnaireResponse, Patient, and Observation</td>
</tr>
<tr>
<td>Reinecke et al [29], 2020</td>
<td>Germany</td>
<td>Patient ID lists</td>
<td>Data-driven recruitment of patients for clinical trials, storage of patient lists, and generation of notifications</td>
<td>List</td>
</tr>
<tr>
<td>Rinaldi et al [45], 2021</td>
<td>Germany</td>
<td>Microbiology data</td>
<td>Standardization of clinical data from patient care and medical research in the field of infection control</td>
<td>DiagnosticReport, Observation, Specimen, and ServiceRequest</td>
</tr>
<tr>
<td>Rinaldi et al [47], 2021</td>
<td>Germany</td>
<td>OpenEHR Template</td>
<td>Mapping infection control related data across 3 different standards—OpenEHR, FHIR, and OMOP CDM—to maximize analysis capabilities</td>
<td>DiagnosticReport, Observation, Specimen, ServiceRequest, and Encounter</td>
</tr>
<tr>
<td>Sass et al [26], 2020</td>
<td>Germany</td>
<td>COVID-19 data</td>
<td>Standardized data model</td>
<td>Patient, Consent, Observation, Condition, Procedure, Encounter, Medication, and MedicationStatement</td>
</tr>
<tr>
<td>Sass et al [46], 2021</td>
<td>Germany</td>
<td>Medication chapter of the German Procedure Classification and Identification of Medicinal Products–compliant medication terminology</td>
<td>Representation of structured medication data</td>
<td>Patient, Procedure, MedicationStatement, and Medication</td>
</tr>
<tr>
<td>Tanaka et al [27], 2020</td>
<td>Japan</td>
<td>SS-MIX2(^2)</td>
<td>Mapping electronic medical record items between SS-MIX2 and HL7 FHIR</td>
<td>Patient, Encounter, Condition, AllergyIntolerance, Observation, Specimen, ServiceRequest, MedicationRequest, and MedicationDispense</td>
</tr>
<tr>
<td>Ulrich et al [34], 2016</td>
<td>Germany</td>
<td>Metadata or CRF(^3)</td>
<td>Metadata repository</td>
<td>Questionnaire</td>
</tr>
<tr>
<td>Wagholikar et al [36], 2017</td>
<td>United States</td>
<td>Common data model demographics, laboratory results, and diagnoses</td>
<td>Clinical apps sharing via a platform</td>
<td>—</td>
</tr>
<tr>
<td>Wang et al [48], 2021</td>
<td>United States</td>
<td>FDA’s Adverse Event Reporting System data</td>
<td>Potential use of FHIR for postmarket safety surveillance for drug products</td>
<td>AdverseEvent</td>
</tr>
<tr>
<td>Weber et al [39], 2020</td>
<td>Switzerland</td>
<td>Electronic consent form</td>
<td>Designing of a FHIR-based eConsent app for Android and evaluation of acceptance</td>
<td>Contract</td>
</tr>
<tr>
<td>Wettstein et al [62], 2021</td>
<td>Germany</td>
<td>Clinical data</td>
<td>Using FHIR for automated and distributed feasibility queries to find available cohort sizes across institutions</td>
<td>Group, ResearchStudy, and Task</td>
</tr>
<tr>
<td>Wettstein et al [63], 2021</td>
<td>Germany</td>
<td>Medical routine data</td>
<td>HL7 FHIR version R4 is used to define the necessary communication messages as well as process input and output variables.</td>
<td>Group, ResearchStudy, and Task</td>
</tr>
<tr>
<td>Wu et al [37], 2018</td>
<td>United Kingdom</td>
<td>EHR data and unstructured documents</td>
<td>Semantic search system for obtaining clinical insights from unstructured clinical notes</td>
<td>Patient and DocumentReference</td>
</tr>
<tr>
<td>Xu et al [64], 2020</td>
<td>United States</td>
<td>Data set of patients with “asthma-like” conditions</td>
<td>Impact of airborne pollutant exposures on asthma (research question)</td>
<td>—</td>
</tr>
<tr>
<td>Zong et al [65], 2020</td>
<td>United States</td>
<td>Colorectal cancer report</td>
<td>Automatic population of eCRFs in colorectal clinical cancer trials</td>
<td>Questionnaire and QuestionnaireResponse</td>
</tr>
<tr>
<td>Zong et al [66], 2021</td>
<td>United States</td>
<td>Colorectal cancer data model</td>
<td>Framework for capturing common data elements from CRFs and FHIR resources to identify clinical information needs</td>
<td>DiagnosticReport and Observation</td>
</tr>
</tbody>
</table>
**Study Objectives**

In terms of medical specialty, most (55%, 27/49) of the studies [24,27-32,34,36-42,44,46,48,49,53,56,60-63,70] 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 [1,12,22,23,25,35,50,52,58,64,69] 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
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]
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]


71. Vorisek et al. JMIR MEDICAL INFORMATICS 2022 | vol. 10 | iss. 7 | e35724 | p.18 https://medinform.jmir.org/2022/7/e35724


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

©Carina Nina Vorisek, Moritz Lehne, Sophie Anne Ines Klopfenstein, Paula Josephine Mayer, Alexander Bartschke, Thomas Haese, Sylvia Thun. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.07.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
The Power of Patient Engagement With Electronic Health Records as Research Participants

Jeff Pawelek¹, BSc; Katie Baca-Motes¹, MBA; Jay A Pandit¹, MD; Benjamin B Berk², MD; Edward Ramos¹², PhD

¹Digital Trials Center, Scripps Research Translational Institute, La Jolla, CA, United States
²CareEvolution, 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

https://medinform.jmir.org/2022/7/e39145

JMIR Med Inform 2022 | vol. 10 | iss. 7 | e39145 | p. 20
(page number not for citation purposes)
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 choose 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 (e.g., 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 (e.g., 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 (ClinicalTrials.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-AHEADa enrollment numbers in the underrepresented in biomedical research category (self-reported).

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

aDETECT-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
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

©Jeff Pawelek, Katie Baca-Motes, Jay A Pandit, Benjamin B Berk, Edward Ramos. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 08.07.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
The Effects of Information Continuity and Interpersonal Continuity on Physician Services Online: Cross-sectional Study

Yan Xuan1*, MA; Chaojin Guo1*, BA; Wei Lu2, PhD

1Department of Nursing, Hainan Women and Children’s Medical Center, Hainan, China
2School 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 interaction 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.
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.
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 \((MTitle_{1j})\) and \((MTitle_{2j})\), 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 Eq. 1, showing short-term effects and Eq. 2, showing the long-term effects, where \(i=1,\ldots N\) represents the patients, \(j=1,\ldots M\) represents the physicians, \(\beta_1\) to \(\beta_7\) are the focus parameters to be estimated. \(C\) represents control variables. \(\varepsilon\) is the error term associated with observation \(i\) and \(j\).

Table 1. Description of the variables.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response speed (RS(_{ij}))</td>
<td>The response time that physician (j) could reply to a patient’s question in 24 hours.</td>
<td>Use response time directly. The value is in days.</td>
</tr>
<tr>
<td>Information quality (InfQ(_{ij}))</td>
<td>The level of detail in the physician (j)’s reply for patient (i).</td>
<td>The number of words replied.</td>
</tr>
<tr>
<td>Interaction quality (IntQ(_{ij}))</td>
<td>The frequency of physician-patient interaction.</td>
<td>The number of interactions between patient (i) and physician (j) is used.</td>
</tr>
<tr>
<td>Repeat purchase (RP(_{ij}))</td>
<td>Patient (i) may have purchased physician (j)’s service many times.</td>
<td>A dummy variable that describes whether patient (i) has repurchased physician (j)’s service.</td>
</tr>
<tr>
<td><strong>Independent variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Offline experience (OE(_{ij}))</td>
<td>Patient (i) may have gone to a hospital for treatment before consulting online.</td>
<td>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.</td>
</tr>
<tr>
<td>Offline medical records (OMR(_{ij}))</td>
<td>Patient (i) may have gone to a hospital for treatment before consulting online, and undergo some tests.</td>
<td>The number of results of tests that patient (i) has provided to the online physician.</td>
</tr>
<tr>
<td>Offline detailed information (ODI(_{ij}))</td>
<td>Patient (i) may have gone to a hospital for treatment before consulting online.</td>
<td>The number of words that patient (i) has described his offline experience to the online physician.</td>
</tr>
<tr>
<td><strong>Moderating variable</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same physician (SP(_{ij}))</td>
<td>Whether the physician in patient’s offline experience is same as the physician who patient (i) has consulted online.</td>
<td>A dummy variable that describes whether it is the same physician online and offline. “1” refers to yes, and “0” refers to no.</td>
</tr>
<tr>
<td><strong>Control variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician medical titles (MTitle(<em>{1j}) and MTitle(</em>{2j}))</td>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td>Physician education title (ETitle(_{j}))</td>
<td>Whether the physician (j) has worked at a university.</td>
<td>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.</td>
</tr>
<tr>
<td>Physician online reputation (POR(_{j}))</td>
<td>The reputation is based on physician (j)’s online work.</td>
<td>An indicator (ranges from 0 to 5) that is calculated by the website based on patients’ feedbacks is used directly.</td>
</tr>
<tr>
<td>Hospital level (Level(_j))</td>
<td>Hospitals have levels that are evaluated by the medical government based on their comprehensive health care quality in China.</td>
<td>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.</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Response speed</th>
<th>Information quality</th>
<th>Interaction quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1a β (SD)</td>
<td>P value</td>
<td>Model 2b β (SD)</td>
</tr>
<tr>
<td>Level</td>
<td>-0.17 (.006)</td>
<td>.008</td>
<td>-0.17 (.006)</td>
</tr>
<tr>
<td>MTitle1f</td>
<td>-0.21 (.008)</td>
<td>.006</td>
<td>-0.19 (.008)</td>
</tr>
<tr>
<td>MTitle2</td>
<td>-0.02 (.007)</td>
<td>.80</td>
<td>-0.02 (.007)</td>
</tr>
<tr>
<td>ETitleh</td>
<td>-0.017 (.007)</td>
<td>.01</td>
<td>-0.17 (.007)</td>
</tr>
<tr>
<td>PORi</td>
<td>.092 (.111)</td>
<td>&lt;.001</td>
<td>.092 (.111)</td>
</tr>
<tr>
<td>OEj</td>
<td>N/A k</td>
<td>-.009</td>
<td>N/A</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>Variables</th>
<th>Response speed</th>
<th>Information quality</th>
<th>Interaction quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Model 2&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Model 1&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>β (SD)</td>
<td>P value</td>
<td>β (SD)</td>
</tr>
<tr>
<td>Level</td>
<td>-.017 (.006)</td>
<td>.008</td>
<td>-.013 (.006)</td>
</tr>
<tr>
<td>MTitle1&lt;sup&gt;g&lt;/sup&gt;</td>
<td>-.021 (.008)</td>
<td>.006</td>
<td>-.017 (.008)</td>
</tr>
<tr>
<td>MTitle2</td>
<td>-.002 (.007)</td>
<td>.80</td>
<td>-.001 (.007)</td>
</tr>
<tr>
<td>ETitle&lt;sup&gt;h&lt;/sup&gt;</td>
<td>-.017 (.007)</td>
<td>.01</td>
<td>-.014 (.007)</td>
</tr>
<tr>
<td>POR&lt;sup&gt;i&lt;/sup&gt;</td>
<td>.092 (.011)</td>
<td>&lt;.001</td>
<td>.057 (.011)</td>
</tr>
<tr>
<td>OMR&lt;sup&gt;j&lt;/sup&gt;</td>
<td>N/A</td>
<td>N/A</td>
<td>-.001 (.000)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Adjusted $R^2$=0.010; $F_{5,7720}=16.808$; $P<.001$.
<sup>b</sup>Adjusted $R^2$=0.017; $F_{1,7719}=56.802$; $P<.001$.
<sup>c</sup>Adjusted $R^2$=0.090; $F_{5,7720}=153.226$; $P<.001$.
<sup>d</sup>Adjusted $R^2$=0.146; $F_{1,7719}=511.611$; $P<.001$.
<sup>e</sup>Adjusted $R^2$=0.037; $F_{5,7720}=59.550$; $P<.001$.
<sup>f</sup>Adjusted $R^2$=0.135; $F_{1,7719}=882.911$; $P<.001$.
<sup>g</sup>MTitle1: physician medical title.
<sup>h</sup>ETitle: physician education title.
<sup>i</sup>POR: physician online reputation.
<sup>j</sup>OMR: offline medical record.
<sup>k</sup>N/A: not applicable.
Thus, hypotheses 1a, 1b, and 1c are partly supported. Hypothesis 1a is not supported but hypotheses 1b and 1c are supported. The results in Table 4 show that offline experience negatively affects physician response speed (β = −.025, P < .001), and interaction quality (β = −.025, P < .001). For offline medical records provision, hypothesis 1a is 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 = −.343, 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 = .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 = .025, 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 provision and repeat purchase (β = .143, P = .04). Thus, hypothesis 2 is partly supported.
Table 5. Results for information continuity: long-term effects.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Offline experience</th>
<th>Offline medical record</th>
<th>Offline detailed information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Model 2&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Model 1&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Level</td>
<td>0.014 (.003)</td>
<td>0.014 (.003)</td>
<td>0.014 (.004)</td>
</tr>
<tr>
<td>MTitle 1&lt;sup&gt;e&lt;/sup&gt;</td>
<td>0.001 (.004)</td>
<td>&lt;.001 (.004)</td>
<td>0.78 (.157)</td>
</tr>
<tr>
<td>MTitle 2</td>
<td>0.003 (.004)</td>
<td>0.003 (.004)</td>
<td>0.003 (.005)</td>
</tr>
<tr>
<td>ETitle&lt;sup&gt;h&lt;/sup&gt;</td>
<td>0.007 (.003)</td>
<td>0.007 (.003)</td>
<td>0.007 (.004)</td>
</tr>
<tr>
<td>POR&lt;sup&gt;i&lt;/sup&gt;</td>
<td>-.068 (.005)</td>
<td>&lt;.001 (.005)</td>
<td>-.068 (.007)</td>
</tr>
<tr>
<td>OE&lt;sup&gt;j&lt;/sup&gt;</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>OMR&lt;sup&gt;k&lt;/sup&gt;</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>ODI&lt;sup&gt;m&lt;/sup&gt;</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

<sup>a</sup>Adjusted $R^2=0.013$; $F_{5,7720}=21.064$; $P<.001$.
<sup>b</sup>Adjusted $R^2=0.013$; $F_{1,7719}=4.162$; $P=.04$.
<sup>c</sup>Adjusted $R^2=0.013$; $F_{5,7720}=21.064$; $P<.001$.
<sup>d</sup>Adjusted $R^2=0.014$; $F_{1,7719}=12.792$; $P<.001$.
<sup>e</sup>Adjusted $R^2=0.025$; $F_{5,7720}=40.078$; $P<.001$.
<sup>f</sup>Adjusted $R^2=0.079$; $F_{1,7719}=455.791$; $P<.001$.
<sup>g</sup>MTitle1: physician medical title.
<sup>h</sup>ETitle: physician education title.
<sup>i</sup>POR: physician online reputation.
<sup>j</sup>OE: offline experience.
<sup>k</sup>N/A: not applicable.
<sup>l</sup>OMR: offline medical record.
<sup>m</sup>ODI: offline detailed information.
### Table 6. Results for interpersonal continuity: long-term effects.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Offline experience&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Offline medical records&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Offline detailed information&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β (SD)</td>
<td>P value</td>
<td>β (SD)</td>
</tr>
<tr>
<td>Level</td>
<td>.012 (.005)</td>
<td>.02</td>
<td>.011 (.005)</td>
</tr>
<tr>
<td>MTitle¹</td>
<td>.000 (.005)</td>
<td>.97</td>
<td>.000 (.005)</td>
</tr>
<tr>
<td>MTitle²</td>
<td>.003 (.005)</td>
<td>.54</td>
<td>.003 (.005)</td>
</tr>
<tr>
<td>ETitle²</td>
<td>.007 (.004)</td>
<td>.14</td>
<td>.006 (.004)</td>
</tr>
<tr>
<td>POR⁵</td>
<td>−.069 (.007)</td>
<td>&lt;.001</td>
<td>−.057 (.008)</td>
</tr>
<tr>
<td>OE⁶</td>
<td>−.001 (.008)</td>
<td>.86</td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>SP×OE</td>
<td>.009 (.008)</td>
<td>.28</td>
<td>N/A</td>
</tr>
<tr>
<td>OMR¹</td>
<td>N/A</td>
<td>N/A</td>
<td>.000 (.000)</td>
</tr>
<tr>
<td>SP×OMR</td>
<td>N/A</td>
<td>N/A</td>
<td>6.457E−5 (.000)</td>
</tr>
<tr>
<td>ODI⁶</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>SP×ODI</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

<sup>a</sup>Adjusted $R^2=0.013$.  
<sup>b</sup>Adjusted $R^2=0.014$.  
<sup>c</sup>Adjusted $R^2=0.079$.  
<sup>d</sup>MTitle¹: physician medical title.  
<sup>e</sup>ETitle²: physician education title.  
<sup>f</sup>POR: physician online reputation.  
<sup>g</sup>OE: offline experience.  
<sup>h</sup>N/A: not applicable.  
<sup>i</sup>SP: same physician.  
<sup>j</sup>OMR: offline medical record.  
<sup>k</sup>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, we 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


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

Cassandra Oliver Schember¹, MPH, PhD; Sarah E Scott², MD; Cathy A Jenkins³, MS; Peter F Rebeiro¹,³,⁴, MHS, PhD; Megan Turner⁴, MA; Sally S Furukawa⁴, BSc; Carmen Bofill³, MPH; Zhou Yan⁵, BSc; Gretchen P Jackson⁶, MD, PhD; April C Pettit¹,³, MP...
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.

https://medinform.jmir.org/2022/7/e34712 JMIR Med Inform 2022 | vol. 10 | iss. 7 | e34712 | p.43 (page number not for citation purposes)
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 proportion and used Pearson chi-squared test 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.

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

\(^a\)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 1.09-1.17; 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.

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

<sup>a</sup>aPR: adjusted prevalence ratio.  
<sup>b</sup>REF: reference.  
<sup>c</sup>MSM: men who have sex with men.  
<sup>d</sup>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,
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.

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

<sup>a</sup>aPR: adjusted prevalence ratio.  
<sup>b</sup>REF: reference.  
<sup>c</sup>MSM: men who have sex with men.  
<sup>d</sup>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 portal 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


Abbreviations

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

©Cassandra Oliver Schember, Sarah E Scott, Cathy A Jenkins, Peter F Rebeiro, Megan Turner, Sally S Furukawa, Carmen Bofill, Zhou Yan, Gretchen P Jackson, April C Petit. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 25.07.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
The Impact of Telemedicine on Physicians’ After-hours Electronic Health Record “Work Outside Work” During the COVID-19 Pandemic: Retrospective Cohort Study

Katharine Lawrence1,2,3*, MPH, MD; Oded Nov1,4*, PhD; Devin Mann1,2,3*, MD; Soumik Mandal1,4*, PhD; Eduardo Iturrate2*, MD; Batia Wiesenfeld5*, 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.

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

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

<table>
<thead>
<tr>
<th>Derived metric</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scheduled days</td>
<td>For a reporting period:</td>
</tr>
<tr>
<td>Unscheduled days</td>
<td>For a reporting period:</td>
</tr>
<tr>
<td>Time outside scheduled hours per month</td>
<td>For a reporting period:</td>
</tr>
<tr>
<td>Time on unscheduled days per month</td>
<td>For a reporting period:</td>
</tr>
<tr>
<td>Clinical load</td>
<td>For a reporting period:</td>
</tr>
<tr>
<td>“Work outside work” measure</td>
<td>For a reporting period:</td>
</tr>
</tbody>
</table>

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: \( F_3(2,12822)=33.09; \ P<.001 \); WOW per appointment: \( F_3(2,12784)=42.68; \ P<.001 \)). 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).

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

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

<table>
<thead>
<tr>
<th>WOW</th>
<th>Time period</th>
<th>Median</th>
<th>Mean</th>
<th>95% CI</th>
<th>Median</th>
<th>Mean</th>
<th>95% CI</th>
<th>Median</th>
<th>Mean</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>WOW per day</td>
<td>Before pandemic</td>
<td>27.19</td>
<td>34.50</td>
<td>33.52-35.47</td>
<td>23.96</td>
<td>30.20</td>
<td>29.50-30.91</td>
<td>26.94</td>
<td>34.11</td>
<td>33.31-34.91</td>
</tr>
<tr>
<td></td>
<td>During acute pandemic</td>
<td>9.70</td>
<td>10.03</td>
<td>8.92-9.65</td>
<td>7.52</td>
<td>11.68</td>
<td>11.31-12.05</td>
<td>6.04</td>
<td>10.03</td>
<td>9.70-10.37</td>
</tr>
<tr>
<td>WOW per appointment</td>
<td>After acute pandemic</td>
<td>9.70</td>
<td>10.03</td>
<td>8.92-9.65</td>
<td>7.52</td>
<td>11.68</td>
<td>11.31-12.05</td>
<td>6.04</td>
<td>10.03</td>
<td>9.70-10.37</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Study variables</th>
<th>Normalized WOW</th>
<th>Step 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unstandardized coefficient</td>
<td>Standard error</td>
</tr>
<tr>
<td>COVID-19 time period</td>
<td>-0.27</td>
<td>0.13</td>
</tr>
<tr>
<td>Telemedicine intensity</td>
<td>6.67</td>
<td>0.32</td>
</tr>
<tr>
<td>Telemedicine intensity×time</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>period</td>
<td>a</td>
<td></td>
</tr>
</tbody>
</table>

aN/A: not applicable.
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
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 analysis of the impact of telemedicine on work burden during the acute pandemic period. 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


27. What is telework? United States Office of Personnel Management. URL: https://www.opm.gov/faq/QA.aspx?id=b48b83b440c-4f1e-a88c-3cd9d802a8c8&pid=75346675-3b92-4ace-831d-58cf5b0e86d2&result=1 [accessed 2021-01-04]


Abbreviations

- **EHR**: electronic health record
- **NYULH**: New York University Langone Health
- **PT**: pajama time
- **WOW**: work outside work
Extraction of Explicit and Implicit Cause-Effect Relationships in Patient-Reported Diabetes-Related Tweets From 2017 to 2021: Deep Learning Approach

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 multword 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**
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 A1c levels [5,6], and low medication adherence [7]. Reducing diabetes distress may improve hemoglobin A1c 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.”

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.
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].

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

<table>
<thead>
<tr>
<th>Sentences</th>
<th>Cause</th>
<th>Effect</th>
<th>Causal association</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes causes me to have mood swings</td>
<td>Diabetes</td>
<td>mood swings</td>
<td>1</td>
<td>Possible causal association</td>
</tr>
<tr>
<td>I just want to eat, I hate #diabetes</td>
<td>#diabetes</td>
<td>hate</td>
<td>1</td>
<td>Possible causal association related to diabetes distress</td>
</tr>
<tr>
<td>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.</td>
<td>___</td>
<td>___</td>
<td>0</td>
<td>Nondiabetes or diabetes distress–related relationship. “Flu” is not diabetes-related</td>
</tr>
<tr>
<td>Had two strokes and recover now and also have high blood pressure and diabetes.</td>
<td>___</td>
<td>___</td>
<td>0</td>
<td>Unclear cause-effect relationship. Not clear if “high blood pressure” or “diabetes” caused the stroke</td>
</tr>
<tr>
<td>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</td>
<td>glucose test</td>
<td>anxiety</td>
<td>1</td>
<td>Chaining cause-effect relationship (A-&gt;B-&gt;C)</td>
</tr>
<tr>
<td>My 14-year-old daughter is type 1 = malfunctioning pancreas, meaning not enough insulin being made to regulate</td>
<td>type 1</td>
<td>malfunctioning pancreas; not enough insulin</td>
<td>1</td>
<td>Negation in a cause/effect is considered being part of the cause/effect as it does not alter the meaning</td>
</tr>
<tr>
<td>It is not true to think that insulin makes you feel so bad</td>
<td>insulin</td>
<td>feel so bad</td>
<td>0</td>
<td>Negation is not part of cause/effect and alters the meaning</td>
</tr>
</tbody>
</table>

*aNot 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.
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 labeled 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, 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: Pre-trained 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-Goldfarb-Shanno 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.

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

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

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.

<table>
<thead>
<tr>
<th>Models</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT_FFL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I-C</td>
<td>0.48</td>
<td>0.46</td>
<td>0.47</td>
</tr>
<tr>
<td>I-E</td>
<td>0.20</td>
<td>0.48</td>
<td>0.29</td>
</tr>
<tr>
<td>O</td>
<td>0.91</td>
<td>0.77</td>
<td>0.83</td>
</tr>
<tr>
<td>macro</td>
<td>0.53</td>
<td>0.57</td>
<td>0.53</td>
</tr>
<tr>
<td>WE_BERT_CRF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I-C</td>
<td>0.63</td>
<td>0.61</td>
<td>0.62</td>
</tr>
<tr>
<td>I-E</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>O</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>macro</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>FastText_CRF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I-C</td>
<td>0.59</td>
<td>0.57</td>
<td>0.58</td>
</tr>
<tr>
<td>I-E</td>
<td>0.45</td>
<td>0.38</td>
<td>0.41</td>
</tr>
<tr>
<td>O</td>
<td>0.92</td>
<td>0.38</td>
<td>0.93</td>
</tr>
<tr>
<td>macro</td>
<td>0.65</td>
<td>0.63</td>
<td>0.64</td>
</tr>
</tbody>
</table>

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 (e.g., diabetes, #diabetes, diabetes mellitus) followed by “death” with 16,989 (e.g., passed away, killed, died, suicide) and “insulin” (e.g., 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.

<table>
<thead>
<tr>
<th>Parent cluster</th>
<th>Cluster</th>
<th>Value (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>diabetes</td>
<td>66,775</td>
</tr>
<tr>
<td>Death</td>
<td>death</td>
<td>16,989</td>
</tr>
<tr>
<td>Insulin</td>
<td>insulin</td>
<td>14,148</td>
</tr>
<tr>
<td>Diabetes</td>
<td>type 1 diabetes</td>
<td>11,693</td>
</tr>
<tr>
<td>Emotions</td>
<td>fear</td>
<td>10,160</td>
</tr>
<tr>
<td>Glycemic variability</td>
<td>hypoglycemia</td>
<td>9547</td>
</tr>
<tr>
<td>Symptoms</td>
<td>sick</td>
<td>6549</td>
</tr>
<tr>
<td>Nutrition</td>
<td>overweight</td>
<td>5186</td>
</tr>
<tr>
<td>Diabetes</td>
<td>type 2 diabetes</td>
<td>4909</td>
</tr>
<tr>
<td>Complications and comorbidities</td>
<td>neuropathy</td>
<td>4481</td>
</tr>
<tr>
<td>Health care system</td>
<td>medication</td>
<td>4389</td>
</tr>
<tr>
<td>Diabetes Technology</td>
<td>insulin pump</td>
<td>4307</td>
</tr>
<tr>
<td>Nutrition</td>
<td>nutrition</td>
<td>4230</td>
</tr>
<tr>
<td>Emotions</td>
<td>anger</td>
<td>4149</td>
</tr>
<tr>
<td>Health</td>
<td>oral glucose tolerance test</td>
<td>4053</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>hypertension</td>
<td>3782</td>
</tr>
<tr>
<td>Health care system</td>
<td>finance</td>
<td>3767</td>
</tr>
<tr>
<td>Nutrition</td>
<td>reduce weight</td>
<td>3589</td>
</tr>
<tr>
<td>Insulin</td>
<td>unable to afford insulin</td>
<td>3381</td>
</tr>
<tr>
<td>Nutrition</td>
<td>diet</td>
<td>3325</td>
</tr>
<tr>
<td>Emotions</td>
<td>sadness</td>
<td>3153</td>
</tr>
<tr>
<td>Glycemic variability</td>
<td>hyperglycemia</td>
<td>3144</td>
</tr>
<tr>
<td>Diabetes</td>
<td>suffer</td>
<td>3132</td>
</tr>
<tr>
<td>Diabetes Distress</td>
<td>depression</td>
<td>2810</td>
</tr>
<tr>
<td>Health care system</td>
<td>hospital</td>
<td>2721</td>
</tr>
<tr>
<td>Diabetes Distress</td>
<td>stress</td>
<td>2681</td>
</tr>
<tr>
<td>Nutrition</td>
<td>sugar</td>
<td>2369</td>
</tr>
<tr>
<td>Nutrition</td>
<td>fasting</td>
<td>2363</td>
</tr>
<tr>
<td>Insulin</td>
<td>rationing insulin</td>
<td>2244</td>
</tr>
<tr>
<td>Health</td>
<td>gestational diabetes</td>
<td>2076</td>
</tr>
</tbody>
</table>
Table 5. The most frequent cause-effect relationships excluding the cluster “diabetes” with the number of occurrences.

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


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]


Abbreviations

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

©Adrian Ahne, Vivek Kheter, Xavier Tannier, Md Imbesat Hassan Rizvi, Thomas Czernichow, Francisco Orchard, Charline Bour, Andrew Fano, Guy Fagherazzi. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.07.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Accurate Forecasting of Emergency Department Arrivals With Internet Search Index and Machine Learning Models: Model Development and Performance Evaluation

Bi Fan¹, PhD; Jiaxuan Peng², MSc; Hainan Guo¹, PhD; Haobin Gu³, MSc; Kangkang Xu⁴, PhD; Tingting Wu¹, MSc

¹College of Management, Institute of Business Analysis and Supply Chain Management, Shenzhen University, Shenzhen, China
²Faculty of Science, University of St Andrews, St Andrews, United Kingdom
³School of Management Science and Engineering, Dongbei University of Finance and Economics, Dalian, China
⁴School 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...
Introduction

Background

The emergency department (ED), which provides instant and efficient medical services for patients all day, is one of the most important parts of 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 semurplent 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.
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 2. Weekly patient arrivals to ED. ED: emergency department.

![Boxplot of patient volumes of ED visits per month. ED: emergency department.](image)

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

<table>
<thead>
<tr>
<th>Aspects</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Names</td>
<td>癌症 (cancer), 流感 (influenza), abortion, flu, h1n1 symptoms</td>
</tr>
<tr>
<td>Causes</td>
<td>天气 (weather), 病毒 (virus), pregnancy, skin problem, tobacco</td>
</tr>
<tr>
<td>Symptoms</td>
<td>喉咙痛 (sore throat), 发烧 (fever), 出汗 (sweat), infections</td>
</tr>
<tr>
<td>Treatments</td>
<td>克拉汀 (claritin), 按摩 (massage)</td>
</tr>
<tr>
<td>Others</td>
<td>蜂蜜 (honey), 醫生 (doctor), 冬季 (winter), depression</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Number</th>
<th>Index</th>
<th>Aspects</th>
<th>Lag$^a$</th>
<th>Correlation coefficient</th>
<th>$P$ value$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ginger</td>
<td>Treatment</td>
<td>1</td>
<td>-0.33</td>
<td>.02</td>
</tr>
<tr>
<td>2</td>
<td>swine flu$^c$ symptoms</td>
<td>Disease</td>
<td>1</td>
<td>0.50</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>3</td>
<td>Infect</td>
<td>Symptom</td>
<td>1</td>
<td>0.36</td>
<td>.01</td>
</tr>
<tr>
<td>4</td>
<td>衛生署 (Department of Health)</td>
<td>Others</td>
<td>1</td>
<td>0.43</td>
<td>.00</td>
</tr>
<tr>
<td>5</td>
<td>fever</td>
<td>Symptom</td>
<td>2</td>
<td>0.31</td>
<td>.04</td>
</tr>
<tr>
<td>6</td>
<td>豬流感診所 (swine flu clinic)</td>
<td>Disease</td>
<td>2</td>
<td>0.49</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>7</td>
<td>牙醫 (dentist)</td>
<td>Others</td>
<td>2</td>
<td>-0.32</td>
<td>.04</td>
</tr>
<tr>
<td>8</td>
<td>腸病毒 (enterovirus)</td>
<td>Disease</td>
<td>6</td>
<td>0.38</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>9</td>
<td>cough</td>
<td>Symptom</td>
<td>7</td>
<td>-0.42</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

$^a$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).
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 腸病毒 (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), 腸病毒 (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 $\mathbf{x}_i$ as the input and $y_i$ as 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:

$$\mathbf{H}^+ \mathbf{y} = \mathbf{Y} \quad (4)$$

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; $\mathbf{Y}$ 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 $\mathbf{y}$.

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

$$H\mathbf{y} = \mathbf{Y} \quad (4)$$

where $H$ is the hidden layer output matrix.

Through the least-squares method, the output weight $\mathbf{y}$ 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:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%$$

where $N$ is the number of observations; $y_i$ is the real value; and $\hat{y}_i$ 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:

$$D = \frac{1}{N} \sum_{i=1}^{N} \left( y_i - \hat{y}_{re,i} - (y_i - \hat{y}_{te,i}) \right)$$

where $y_i$ is the actual value; and $\hat{y}_{re,i}$ and $\hat{y}_{te,i}$ are forecasting values of the reference and test models, respectively; $N$ is the number of observations; $h(>1)$ is $h$-step-ahead forecasts; $\sigma^2$ 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.

<table>
<thead>
<tr>
<th>Models</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAPE&lt;sup&gt;a&lt;/sup&gt; (%)</td>
<td>RMSE&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Patient arrivals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARIMA&lt;sup&gt;c&lt;/sup&gt;</td>
<td>3.6</td>
<td>17.02</td>
</tr>
<tr>
<td>ANN&lt;sup&gt;d&lt;/sup&gt;</td>
<td>3.5</td>
<td>19.19</td>
</tr>
<tr>
<td>SVM&lt;sup&gt;e&lt;/sup&gt;</td>
<td>2.2</td>
<td>18.52</td>
</tr>
<tr>
<td>RF&lt;sup&gt;f&lt;/sup&gt;</td>
<td>2.5</td>
<td>19.36</td>
</tr>
<tr>
<td>LSTM&lt;sup&gt;g&lt;/sup&gt;</td>
<td>2.9</td>
<td>16.93</td>
</tr>
<tr>
<td>ELM&lt;sup&gt;h&lt;/sup&gt;</td>
<td>2.8</td>
<td>16.52</td>
</tr>
<tr>
<td>Patient arrivals + traditional variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLM&lt;sup&gt;i&lt;/sup&gt;</td>
<td>3.2</td>
<td>16.79</td>
</tr>
<tr>
<td>ARIMAX&lt;sup&gt;j&lt;/sup&gt;</td>
<td>3.5</td>
<td>17.85</td>
</tr>
<tr>
<td>ANN</td>
<td>3.4</td>
<td>16.10</td>
</tr>
<tr>
<td>SVM</td>
<td>2.8</td>
<td>16.24</td>
</tr>
<tr>
<td>RF</td>
<td>2.9</td>
<td>17.05</td>
</tr>
<tr>
<td>LSTM</td>
<td>3.2</td>
<td>16.43</td>
</tr>
<tr>
<td>ELM</td>
<td>2.7</td>
<td>13.17</td>
</tr>
<tr>
<td>Patient arrivals + traditional variables + internet search index</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLM</td>
<td>3.2</td>
<td>16.24</td>
</tr>
<tr>
<td>ARIMAX</td>
<td>3.4</td>
<td>17.84</td>
</tr>
<tr>
<td>ANN</td>
<td>3.0</td>
<td>14.51</td>
</tr>
<tr>
<td>SVM</td>
<td>2.6</td>
<td>14.84</td>
</tr>
<tr>
<td>RF</td>
<td>2.9</td>
<td>15.92</td>
</tr>
<tr>
<td>LSTM</td>
<td>3.0</td>
<td>15.15</td>
</tr>
<tr>
<td>ELM</td>
<td>2.6</td>
<td>13.10</td>
</tr>
</tbody>
</table>

<sup>a</sup>MAPE: average mean absolute percentage error.
<sup>b</sup>RMSE: root mean square error.
<sup>c</sup>ARIMA: autoregressive integrated moving average model.
<sup>d</sup>ANN: artificial neural network.
<sup>e</sup>SVM: support vector machine.
<sup>f</sup>RF: random forest.
<sup>g</sup>LSTM: long short-term memory.
<sup>h</sup>ELM: extreme learning machine.
<sup>i</sup>GLM: generalized linear model.
<sup>j</sup>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.

<table>
<thead>
<tr>
<th>Test model</th>
<th>Reference model(^b,c)</th>
<th>GLM(^d)</th>
<th>ARIMAX(^e)</th>
<th>ANN(^f)</th>
<th>SVM(^g)</th>
<th>RF(^h)</th>
<th>LSTM(^i)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient arrivals + traditional variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELM</td>
<td>2.8297 (&lt;.001)</td>
<td>3.0624 (&lt;.001)</td>
<td>2.012 (&lt;.001)</td>
<td>1.8178 (&lt;.001)</td>
<td>2.8481 (&lt;.001)</td>
<td>2.1002 (&lt;.001)</td>
<td></td>
</tr>
<tr>
<td>GLM</td>
<td>0.2935 (.31)</td>
<td>0.86595 (.11)</td>
<td>0.10707 (.06)</td>
<td>0.7643 (.09)</td>
<td>0.1663 (.54)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARIMAX</td>
<td>0.64435 (.17)</td>
<td>1.0691 (.06)</td>
<td>0.3957 (.23)</td>
<td>0.4876 (.68)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANN</td>
<td>0.13244 (.38)</td>
<td>0.2746 (.45)</td>
<td>1.9512 (.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td>0.5823 (.27)</td>
<td>0.8714 (.12)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td></td>
<td></td>
<td>1.0045 (.08)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Patient arrivals + traditional variables + internet search index</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELM</td>
<td>2.5062 (&lt;.001)</td>
<td>3.79 (&lt;.001)</td>
<td>2.0047 (&lt;.001)</td>
<td>2.0325 (&lt;.001)</td>
<td>2.0476 (&lt;.001)</td>
<td>1.6659 (.02)</td>
<td></td>
</tr>
<tr>
<td>GLM</td>
<td>0.32675 (.30)</td>
<td>1.1462 (.12)</td>
<td>1.6064 (.07)</td>
<td>1.0467 (.09)</td>
<td>0.3647 (.64)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARIMAX</td>
<td>1.7314 (.06)</td>
<td>2.2885 (.05)</td>
<td>1.5946 (.11)</td>
<td>1.2671 (.07)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANN</td>
<td>0.14419 (.40)</td>
<td>0.2104 (.49)</td>
<td>1.2304 (.08)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td>0.2593 (.36)</td>
<td>1.4391 (.04)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td></td>
<td></td>
<td>1.2992 (.06)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)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.

<table>
<thead>
<tr>
<th>Test model (with internet data)</th>
<th>Reference model (without internet data)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GLMb</td>
</tr>
<tr>
<td>GLM</td>
<td>2.4848 (.04)</td>
</tr>
<tr>
<td>ARIMAX</td>
<td>1.5701 (.04)</td>
</tr>
<tr>
<td>ANN</td>
<td>2.2546 (&lt;.001)</td>
</tr>
<tr>
<td>SVM</td>
<td>2.244 (&lt;.001)</td>
</tr>
<tr>
<td>RF</td>
<td>1.7886 (.02)</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.8556 (.01)</td>
</tr>
<tr>
<td>ELM</td>
<td>2.2546 (&lt;.001)</td>
</tr>
</tbody>
</table>

aValues are presented as the Diebold-Mariano statistic. The P value modified by false discovery rate is in brackets. The significance level is .05.
bGLM: generalized linear model.
cARIMAX: ARIMA with explanatory variables.
dANN: artificial neural network.
eSVM: support vector machine.
fRF: random forest.
gLSTM: long short-term memory.
hELM: extreme learning machine.

Table 6. Robustness analysis.

<table>
<thead>
<tr>
<th>SD</th>
<th>Forecasting model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GLMb</td>
</tr>
<tr>
<td>Patient arrivals + traditional variables</td>
<td></td>
</tr>
<tr>
<td>SD of MAPE (%)</td>
<td>2.5</td>
</tr>
<tr>
<td>SD of RMSE</td>
<td>15.638</td>
</tr>
<tr>
<td>Patient arrivals + traditional variables + internet search index</td>
<td></td>
</tr>
<tr>
<td>SD of MAPE (%)</td>
<td>2.4</td>
</tr>
<tr>
<td>SD of RMSE</td>
<td>15.212</td>
</tr>
</tbody>
</table>

aGLM: generalized linear model.
bANN: artificial neural network.
cSVM: support vector machine.
dARIMAX: ARIMA with explanatory variables.
eLSTM: long short-term memory.
fRF: random forest.
fELM: extreme learning machine.
hMAPE: average mean absolute percentage error.
iRMSE: 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 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


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
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

©Bi Fan, Jiaxuan Peng, Hainan Guo, Haobin Gu, Kangkang Xu, Tingting Wu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 20.07.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Classification of Twitter Vaping Discourse Using BERTweet: Comparative Deep Learning Study

William Baker¹, BSc; Jason B Colditz², MEd, PhD; Page D Dobbs³, PhD; Huy Mai¹, BSc; Shyam Visweswaran⁴, MD, PhD; Justin Zhan¹, PhD; Brian A Primack⁵, MD, PhD

¹Department of Computer Science and Computer Engineering, University of Arkansas, Fayetteville, AR, United States
²Division of General Internal Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States
³Health, Human Performance and Recreation Department, University of Arkansas, Fayetteville, AR, United States
⁴Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States
⁵College 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: pdbobs@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 (i.e., a word), but also entire sequences of data (i.e., 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 naive 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, Myslin 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 naive 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>
<thead>
<tr>
<th>Labels</th>
<th>Descriptions</th>
<th>Example quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant</td>
<td>• Is the tweet in English and related to the vaping topic at hand (eg, vape use or users, vaping devices, or products)?</td>
<td></td>
</tr>
<tr>
<td>Not relevant</td>
<td>• Typically, non-English tweets or tweets that referenced vaping cannabis products specifically.</td>
<td></td>
</tr>
<tr>
<td>Commercial</td>
<td>• Is the tweet an advertisement/marketing for vaping products?</td>
<td>Today only! Buy one JUUL get the second half price with our online coupon code #JUUL4LIFE</td>
</tr>
<tr>
<td>Noncommercial</td>
<td>• Includes tweets that demonstrate favorability toward a product but do not directly advocate for purchasing it.</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>• The tweet is associated with positive emotions or contexts regarding vaping.</td>
<td>Currently juuling in the bathroom at school!</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The tweeter shows positivity or neutral acceptance from others’ usage or others’ positive comments about vaping:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Just got Hannah to try vaping for the first time! She loved it.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>We need a Disney princess that rips her JUUL in the middle of a serious conversation.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The tweeter asks a question using first-person pronouns:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Where can I buy a JUUL?</td>
</tr>
<tr>
<td>Neutral</td>
<td>• The tweet is factual but not opinionated or is a question about unbiased facts/information about vaping:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>They are selling JUUL pens at my local tobacco shop for anyone interested.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is a JUUL?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Is a JUUL better than tobacco?</td>
</tr>
</tbody>
</table>
Table 2. Description of annotated training and test data sets (N=2401).a

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

aPercentages may not add up to 100% as classification was made for sentiment only if the tweet was relevant.
bSentiment-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 non-domain-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 $5 \times 10^{-5}$. 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:

\[
F1 = \frac{2 \times (Precision \times Recall)}{Precision + Recall} (1)
\]

Precision = True positive/(True positive + False negative) (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>
<thead>
<tr>
<th>Classifier/metric</th>
<th>Relevance</th>
<th>Commercial</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BERTweet</strong></td>
<td>F1</td>
<td>0.976</td>
<td>0.990</td>
</tr>
<tr>
<td></td>
<td>AUROC</td>
<td>0.945</td>
<td>0.993</td>
</tr>
<tr>
<td><strong>LSTM</strong></td>
<td>F1</td>
<td>0.924</td>
<td>0.727</td>
</tr>
<tr>
<td></td>
<td>AUROC</td>
<td>0.924</td>
<td>0.903</td>
</tr>
</tbody>
</table>

*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 media 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


https://jmir.mirax.org/2022/7/e33678

JMIR Med Inform 2022 | vol. 10 | iss. 7 | e33678 | p.100

(page number not for citation purposes)


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
RobBERTa: 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:
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

©William Baker, Jason B Colditz, Page D Dobbs, Huy Mai, Shyam Visweswaran, Justin Zhan, Brian A Primack. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 21.07.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Impact of a Clinical Text–Based Fall Prediction Model on Preventing Extended Hospital Stays for Elderly Inpatients: Model Development and Performance Evaluation

Yoshimasa Kawazoe1, MD, PhD; Kiminori Shimamoto1, MEng; Daisaku Shibata1, PhD; Emiko Shinohara1, PhD; Hideaki Kawaguchi2,3, MD, PhD; Tomotaka Yamamoto4, MD, PhD

1Artificial Intelligence in Healthcare, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
2Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
3Quantum Computing Center, Keio University, Tokyo, Japan
4Department 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@md.t.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.

**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](https://medinform.jmir.org/2022/7/e37913)
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 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 60,060 unfallen) were used as learning data; cases in which the day of hospitalization was in 2017 (228 fallen and 60,060 unfallen) were used as learning data; cases in which the day of hospitalization was in 2017 (228 fallen and 60,060 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.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Fallen cases (n=1728)</th>
<th>Unfallen cases (n=70,586)</th>
<th>Multivariate regression$^a$</th>
<th>Standardized difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Adjusted odds ratio (95% CI)</td>
<td>$P$ value$^b$</td>
</tr>
<tr>
<td>Hospital days, mean (SD)</td>
<td>30.3 (23.7)</td>
<td>10.6 (6.8)</td>
<td>N/A$^c$</td>
<td>N/A</td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>76.5 (6.8)</td>
<td>74.3 (SD 6.4)</td>
<td>1.03 (1.02-1.03)$^d$</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Sex (male 0, female 1), positive rate (%)</td>
<td>40.6</td>
<td>43.8</td>
<td>0.71 (0.63-0.80)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>ADL$^e$ Eats, positive rate (%)</td>
<td>9.2</td>
<td>2.4</td>
<td>1.08 (0.83-1.40)</td>
<td>.57</td>
</tr>
<tr>
<td>ADL Bathe, positive rate (%)</td>
<td>19.2</td>
<td>5.5</td>
<td>1.37 (1.06-1.77)</td>
<td>.02</td>
</tr>
<tr>
<td>ADL Dressing$^f$, positive rate (%)</td>
<td>15.3</td>
<td>4.4</td>
<td>0.76 (0.57-1.02)</td>
<td>.07</td>
</tr>
<tr>
<td>ADL Transferring$^g$, positive rate (%)</td>
<td>26.2</td>
<td>8.6</td>
<td>1.79 (1.48-2.18)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>ADL Continece$^h$, positive rate (%)</td>
<td>13.0</td>
<td>3.5</td>
<td>1.04 (0.80-1.37)</td>
<td>.75</td>
</tr>
<tr>
<td>Unconsciousness (JCS$^i$ 0, ≠0), positive rate (%)</td>
<td>18.1</td>
<td>5.6</td>
<td>1.70 (1.44-2.00)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Emergency transport, positive rate (%)</td>
<td>8.6</td>
<td>3.9</td>
<td>0.96 (0.78-1.17)</td>
<td>.68</td>
</tr>
<tr>
<td>Cognitive disorder, positive rate (%)</td>
<td>11.1</td>
<td>4.0</td>
<td>1.10 (0.92-1.32)</td>
<td>.28</td>
</tr>
<tr>
<td>Chemotherapy admission, positive rate (%)</td>
<td>11.7</td>
<td>11.4</td>
<td>1.08 (0.91-1.27)</td>
<td>.39</td>
</tr>
<tr>
<td>Past fallen, positive rate (%)</td>
<td>8.1</td>
<td>3.5</td>
<td>1.37 (1.13-1.65)</td>
<td>.001</td>
</tr>
<tr>
<td>Disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Certain infectious and parasitic diseases (A00-B99), positive rate (%)</td>
<td>8.6</td>
<td>6.8</td>
<td>0.98 (0.82-1.17)</td>
<td>.78</td>
</tr>
<tr>
<td>Neoplasms (C00-D48), positive rate (%)</td>
<td>40.8</td>
<td>41.1</td>
<td>1.10 (0.97-1.25)</td>
<td>.12</td>
</tr>
<tr>
<td>Diseases of the blood and blood-forming organs (D50-D89), positive rate (%)</td>
<td>8.3</td>
<td>6.3</td>
<td>1.28 (1.07-1.53)</td>
<td>.01</td>
</tr>
<tr>
<td>Endocrine, nutritional, and metabolic diseases (E00-E90), positive rate (%)</td>
<td>23.8</td>
<td>18.5</td>
<td>1.09 (0.96-1.24)</td>
<td>.19</td>
</tr>
<tr>
<td>Mental and behavioral disorders (F00-F99), positive rate (%)</td>
<td>4.6</td>
<td>1.1</td>
<td>2.09 (1.61-2.71)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Diseases of the nervous system (G00-G99), positive rate (%)</td>
<td>8.4</td>
<td>4.7</td>
<td>1.14 (0.95-1.38)</td>
<td>.16</td>
</tr>
<tr>
<td>Diseases of the eye and adnexa (H00-H59), positive rate (%)</td>
<td>3.8</td>
<td>13.3</td>
<td>0.47 (0.36-0.61)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Diseases of the ear and mastoid process (H60-H95), positive rate (%)</td>
<td>0.3</td>
<td>0.8</td>
<td>0.47 (0.19-1.14)</td>
<td>.10</td>
</tr>
<tr>
<td>Variables</td>
<td>Fallen cases (n=1728)</td>
<td>Unfallen cases (n=70,586)</td>
<td>Multivariate regression(^a)</td>
<td>Standardized difference</td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>-----------------------</td>
<td>---------------------------</td>
<td>-------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Adjusted odds ratio (95% CI)</td>
<td>(P) value(^b) (\times 10^2)</td>
</tr>
<tr>
<td>Diseases of the circulatory system (I00-I99), positive rate (%)</td>
<td>33.9</td>
<td>26.1</td>
<td>1.15 (1.02-1.29)</td>
<td>.02</td>
</tr>
<tr>
<td>Diseases of the respiratory system (J00-J99), positive rate (%)</td>
<td>9.5</td>
<td>6.2</td>
<td>1.01 (0.85-1.20)</td>
<td>.91</td>
</tr>
<tr>
<td>Diseases of the digestive system (K00-K93), positive rate (%)</td>
<td>17.8</td>
<td>16.6</td>
<td>0.77 (0.67-0.87)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Diseases of the skin and subcutaneous tissue (L00-L99), positive rate (%)</td>
<td>3.0</td>
<td>1.6</td>
<td>1.46 (1.09-1.95)</td>
<td>.01</td>
</tr>
<tr>
<td>Diseases of the musculoskeletal system and connective tissue (M00-M99), positive rate (%)</td>
<td>11.9</td>
<td>8.4</td>
<td>1.22 (1.04-1.43)</td>
<td>.02</td>
</tr>
<tr>
<td>Diseases of the genitourinary system (N00-N99), positive rate (%)</td>
<td>10.0</td>
<td>7.3</td>
<td>0.94 (0.79-1.12)</td>
<td>.50</td>
</tr>
<tr>
<td>Pregnancy, perinatal period, congenital malformations (O00-Q99), positive rate (%)</td>
<td>0.3</td>
<td>0.4</td>
<td>1.03 (0.42-2.52)</td>
<td>.94</td>
</tr>
<tr>
<td>Symptoms, signs, and abnormal clinical and laboratory findings (R00-R99), positive rate (%)</td>
<td>5.8</td>
<td>3.3</td>
<td>1.03 (0.83-1.28)</td>
<td>.80</td>
</tr>
<tr>
<td>Injury, poisoning and certain other consequences of external causes (S00-T98), positive rate (%)</td>
<td>5.1</td>
<td>3.1</td>
<td>1.11 (0.88-1.40)</td>
<td>.38</td>
</tr>
<tr>
<td><strong>Blood tests</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low hemoglobin (3.9% missing data), positive rate (%)</td>
<td>71.8</td>
<td>57.5</td>
<td>1.34 (1.19-1.53)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Low total protein or albumin (5.0% missing data), positive rate (%)</td>
<td>48.7</td>
<td>33.8</td>
<td>1.20 (1.08-1.34)</td>
<td>.001</td>
</tr>
<tr>
<td>High blood urea nitrogen (4.4% missing data), positive rate (%)</td>
<td>3.4</td>
<td>1.6</td>
<td>1.20 (0.90-1.61)</td>
<td>.22</td>
</tr>
<tr>
<td>High liver enzymes (AST(^i), ALT(^k); 4.0% missing data), positive rate (%)</td>
<td>6.0</td>
<td>3.6</td>
<td>1.22 (0.98-1.52)</td>
<td>.07</td>
</tr>
<tr>
<td>Low plasma glucose (19.7% missing data), positive rate (%)</td>
<td>2.5</td>
<td>1.7</td>
<td>1.14 (0.80-1.62)</td>
<td>.48</td>
</tr>
<tr>
<td>Abnormal electrolytes (Na, K, Cl; 12.1% missing data), positive rate (%)</td>
<td>35.1</td>
<td>21.6</td>
<td>1.40 (1.26-1.57)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>High C-reactive protein (6.8% missing data), positive rate (%)</td>
<td>10.9</td>
<td>5.0</td>
<td>1.12 (0.94-1.34)</td>
<td>.21</td>
</tr>
<tr>
<td><strong>Prescription</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypnotics and sedatives, anxiolytics</td>
<td>37.4</td>
<td>30.7</td>
<td>1.09 (0.97-1.22)</td>
<td>.13</td>
</tr>
<tr>
<td>Antiepileptic</td>
<td>4.4</td>
<td>1.8</td>
<td>1.30 (1.00-1.69)</td>
<td>.05</td>
</tr>
<tr>
<td>NSAIDs(^l)</td>
<td>43.5</td>
<td>32.6</td>
<td>1.21 (1.08-1.36)</td>
<td>.001</td>
</tr>
<tr>
<td>Variables</td>
<td>Fallen cases (n=1728)</td>
<td>Unfallen cases (n=70,586)</td>
<td>Multivariate regression</td>
<td>Standardized difference</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-----------------------</td>
<td>----------------------------</td>
<td>-------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Adjusted odds ratio</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(95% CI)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P value</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Before matching</td>
<td>After matching</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(n=1728 fallen cases, n=70,586 unfallen cases)</td>
<td>(n=1728 fallen cases, n=1728 unfallen cases)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antiparkinsonism</td>
<td>3.2</td>
<td>1.0</td>
<td>1.61 (1.18-2.21)</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antipsychotic</td>
<td>21.4</td>
<td>9.6</td>
<td>1.44 (1.25-1.66)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other neuroactive agents</td>
<td>13.8</td>
<td>6.6</td>
<td>1.19 (1.01-1.39)</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muscle relaxant</td>
<td>0.3</td>
<td>0.1</td>
<td>1.70 (0.60-4.85)</td>
<td>.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diuretic</td>
<td>23.4</td>
<td>13.7</td>
<td>1.33 (1.16-1.53)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antihypertensive</td>
<td>31.4</td>
<td>25.9</td>
<td>0.88 (0.77-1.00)</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes treatment</td>
<td>15.9</td>
<td>12.7</td>
<td>1.08 (0.92-1.26)</td>
<td>.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Narcotic analgesic</td>
<td>3.3</td>
<td>1.4</td>
<td>1.11 (0.81-1.51)</td>
<td>.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purgative medicine</td>
<td>38.3</td>
<td>32.6</td>
<td>1.09 (0.97-1.22)</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polypharmacy (&gt;10 drugs)</td>
<td>48.7</td>
<td>35.8</td>
<td>1.02 (0.89-1.17)</td>
<td>.77</td>
</tr>
</tbody>
</table>

\(^a\)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.

![Box-and-whisker plots](image)

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

<table>
<thead>
<tr>
<th>$\Gamma$</th>
<th>Maximum $P$ value</th>
<th>Minimum Hodges–Lehmann point estimate (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>&lt;.001</td>
<td>14.1</td>
</tr>
<tr>
<td>2.0</td>
<td>&lt;.001</td>
<td>8.6</td>
</tr>
<tr>
<td>3.0</td>
<td>&lt;.001</td>
<td>6.0</td>
</tr>
<tr>
<td>4.0</td>
<td>&lt;.001</td>
<td>4.1</td>
</tr>
<tr>
<td>5.0</td>
<td>&lt;.001</td>
<td>2.9</td>
</tr>
<tr>
<td>6.0</td>
<td>&lt;.001</td>
<td>2.0</td>
</tr>
<tr>
<td>7.0</td>
<td>.01</td>
<td>1.1</td>
</tr>
<tr>
<td>7.5</td>
<td>.05</td>
<td>0.8</td>
</tr>
<tr>
<td>8.0</td>
<td>.16</td>
<td>0.5</td>
</tr>
</tbody>
</table>

$^a$ $\Gamma$: 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Input data</th>
<th>Evaluation accuracy</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AUC</td>
<td>Sensitivity</td>
<td>Specificity</td>
<td>Precision</td>
<td></td>
</tr>
<tr>
<td>Model 1: MLP</td>
<td>✓</td>
<td>0.735</td>
<td>0.662</td>
<td>0.708</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td>Model 2: BERT+Bi-LSTM</td>
<td>✓</td>
<td>0.851</td>
<td>0.737</td>
<td>0.839</td>
<td>0.093</td>
<td></td>
</tr>
<tr>
<td>Model 3: BERT+Bi-LSTM</td>
<td>✓</td>
<td>0.850</td>
<td>0.794</td>
<td>0.776</td>
<td>0.076</td>
<td></td>
</tr>
</tbody>
</table>

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

AUC: area under the receiver operating characteristic curve.

F1 is the harmonic mean of precision and recall.

MLP: multilayer perceptron.

BERT: bidirectional encoder representations from transformers.

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 (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.

<table>
<thead>
<tr>
<th>Predictions</th>
<th>Fallen cases</th>
<th>Unfallen cases</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted fallen</td>
<td>168</td>
<td>1638</td>
<td>1806</td>
</tr>
<tr>
<td>Predicted unfallen</td>
<td>60</td>
<td>8520</td>
<td>8580</td>
</tr>
<tr>
<td>Sum</td>
<td>228</td>
<td>10,158</td>
<td>10,386</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Scenario</th>
<th>ATETa (days)</th>
<th>Fall prevention rate (%)</th>
<th>Reduced length of hospital stay (number of days per day of interventions)</th>
<th>Hospital stays that could not be reduced (number of days per day of interventions)</th>
<th>Net reduced length of hospital stay (number of days per day of interventions)</th>
<th>Net reduced daily medical costs (Yen per day)b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>17.8</td>
<td>100</td>
<td>0.154</td>
<td>0.055</td>
<td>0.099</td>
<td>3950</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>17.8</td>
<td>25</td>
<td>0.035</td>
<td>0.012</td>
<td>0.022</td>
<td>886</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>8.6</td>
<td>100</td>
<td>0.069</td>
<td>0.025</td>
<td>0.044</td>
<td>1769</td>
</tr>
<tr>
<td>Scenario 4</td>
<td>8.6</td>
<td>25</td>
<td>0.017</td>
<td>0.006</td>
<td>0.011</td>
<td>420</td>
</tr>
</tbody>
</table>

ATET: average treatment effect on treatment.

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 $\Gamma=7$. 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 $\Gamma$, of the 49 variables shown in Table 1, if there were unobserved covariates with $\Gamma=2.0$ 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

References


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

©Yoshimasa Kawazoe, Kiminori Shimamoto, Daisaku Shibata, Emiko Shinohara, Hideaki Kawaguchi, Tomotaka Yamamoto. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 27.07.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

https://medinform.jmir.org/2022/7/e37913

JMIR Med Inform 2022 | vol. 10 | iss. 7 | e37913 | p.117
(page number not for citation purposes)